Person Fit for Test Speededness Normal Curvatures, Likelihood Ratio Tests and Empirical Bayes Estimates

Peer-reviewed author version

Goegebeur, Yuri; De Boeck, Paul & MOLENBERGHS, Geert (2010) Person Fit for Test Speededness Normal Curvatures, Likelihood Ratio Tests and Empirical Bayes Estimates. In: METHODOLOGY-EUROPEAN JOURNAL OF RESEARCH METHODS FOR THE BEHAVIORAL AND SOCIAL SCIENCES, 6(1). p. 3-16.

DOI: 10.1027/1614-2241/a000002
Handle: http://hdl.handle.net/1942/11051

Person fit for test speededness: normal curvatures, likelihood ratio tests and empirical Bayes estimates

Yuri Goegebeur *
Paul De Boeck ‡
Geert Molenberghs §

Abstract

The local influence diagnostics, proposed by Cook (1986), provide a flexible way to assess the impact of minor model perturbations on key model parameters' estimates. In this paper, we apply the local influence idea to the detection of test speededness in a model describing non-response in test data, and compare this local influence approach to the optimal person fit index proposed by Drasgow and Levine (1986), and the empirical Bayes estimate of the test speededness random effect. The performance of the methods is illustrated on the Chilean SIMCE mathematics test data. The data example indicates that the three statistics are promising when it comes to the detection of special profiles, and besides overlap to a considerable extent. Given that the statistics were developed for different purposes, they

*University of Southern Denmark, Department Mathematics and Computer Science, Campusvej 55, DK-5230 Odense M, Denmark, and K.U.Leuven, Department of Psychology, Higher Cognition and Individual Differences, Tiensestraat 102, B-3000 Leuven, Belgium. Email: yuri.goegebeur@stat.sdu.dk

‡K.U.Leuven, Department of Psychology, Higher Cognition and Individual Differences, Tiensestraat 102, B-3000 Leuven, Belgium. Email: paul.deboeck@psy.kuleuven.be

§Hasselt University, Center for Statistics, Agoralaan - Building D, B-3590 Diepenbeek, Belgium. Email: geert.molenberghs@uhasselt.be

react of course differentially to the various characteristics of the response profiles, and hence also exhibit some specificity.

Keywords: missing data, sensitivity analysis, local influence, empirical Bayes, likelihood ratio test.

Introduction

Person fit or appropriateness measurement refers to a collection of statistical techniques for evaluating the misfit of individual test performances to an item response theory (IRT) model or to other item-score patterns in a sample of persons. Generally, these methods do not allow for the recovery of the mechanism that created the deviant item-score patterns, that is, they do not give the user information on why a profile is deviant, and hence can be seen as the IRT analogues of the global influence diagnostics in the field of statistics, see, for instance, Cook and Weisberg (1982), and Chatterjee and Hadi (1988). However, some recent contributions explicitly test against specific violations of a test model assumption or particular types of deviant item-score patterns. For an up to date overview of the available person fit methodology we refer to Meijer and Sijtsma (2001).

In the present paper we introduce and evaluate three indices for identifying response profiles affected by test speededness effects. Test speededness refers to testing situations in which some examinees do not have ample time to answer all questions. Speededness effects are often detrimental to the intended functioning of the test in the sense that the speed with which one responds is usually not an important part of the construct of interest, yet examinees affected by test speededness hurry through, randomly guess on or even fail to complete items, usually at the end of the test, and hence receive ability estimates that underestimate their capacities. In this respect it may be interesting to supplement test scores or response profiles with an index that reflects the examinee's sensitivity to test speededness. Besides this underestimation of the ability parameters due to speededness, the item difficulty parameters of items administered late in the test tend to be overestimated (Douglas, Kim, Habing, & Gao, 1998 and Oshima, 1994). Item response models accommodating test speededness were proposed by Bolt, Cohen, and Wollack (2002); Goegebeur, De Boeck, Wollack, and Cohen (2008); Wollack and Cohen (2005) and Yamamoto and Everson (1997). Although these models provide improved parameter estimates, they do not explicitly allow for omissions. However, omissions occur in testing situations, especially when tests are administered under rather stringent time constraints, and provide information about unobservable quantities such as the examinee ability, propensity to

omit and test speededness, which implies that they cannot be ignored. The analysis described in this paper is based on the model Goegebeur, De Boeck, Molenberghs, and del Pino (2006) developed for explaining non-response in test data. Under this model, non-response emerges from a general tendency to omit in case one does not know the answer and a test speededness effect, both taken to be examinee specific. The present paper extends the analysis described in Goegebeur et al. (2006) in that the normal curvatures for test speededness described in the latter are supplemented with and compared to two new indices that can be used to identify test speededness: a likelihood ratio test statistic and the empirical Bayes estimate of the test speededness parameter.

Given that the model under consideration builds upon classical IRT models, and furthermore fits in the missing data framework established by Rubin (1976) and Little and Rubin (2002), it is instructive to review some of these concepts. Let Y_{pi} denote the binary response (correct/incorrect, coded $Y_{pi} = 1$ and $Y_{pi} = 0$, respectively) of examinee p, $p = 1, \ldots, P$, to item i, $i = 1, \ldots, I$. In the classical one-parameter Rasch model (1PL) (Rasch, 1960), Y_{pi} depends on the examinee's ability θ_p and item difficulty β_i in the following way

$$Y_{pi}|\theta_p \quad \sim \quad \text{Bern}(P_i(\theta_p)),$$

with

$$P_i(\theta_p) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}, \qquad \beta_i, \theta_p \in \mathbb{R} \tag{1}$$

and $\theta_p \sim N(0, \sigma_\theta^2)$. Moreover, conditional on θ_p, all responses of subject p are assumed independent, the so-called *local item independence condition*. The Rasch model has been extended in several ways. In the two-parameter logistic model (2PL) (Birnbaum, 1968) the ability parameter θ_p is weighted by an item parameter α_i:

$$P_i(\theta_p) = \frac{\exp[\alpha_i(\theta_p - \beta_i)]}{1 + \exp[\alpha_i(\theta_p - \beta_i)]}, \qquad \alpha_i > 0, \beta_i, \theta_p \in \mathbb{R} \tag{2}$$

so that the influence of the examinee's ability on outcome depends on the item. The three-parameter logistic model (3PL) (Birnbaum, 1968) extends the 2PL with an item-specific guessing parameter c_i:

$$P_i(\theta_p) = c_i + (1 - c_i)\frac{\exp[\alpha_i(\theta_p - \beta_i)]}{1 + \exp[\alpha_i(\theta_p - \beta_i)]}, \qquad \alpha_i > 0, \beta_i, \theta_p \in \mathbb{R}, c_i \in [0, 1).$$

The parameter c_i is the horizontal asymptote of the item characteristic curve (the graph of $P_i(\theta_p)$ as a function of θ_p) for $\theta_p \to -\infty$, and reflects that even individuals with a very low ability have a positive probability of producing a correct answer to the item as they may simply guess the correct answer. We refer to San Martín, del Pino, and De Boeck (2006) for a deeper discussion and some extensions of the 3PL model.

Rubin (1976) and Little and Rubin (2002, chap. 6) established a framework to distinguish between different missing values processes. A missing value process is said to be missing completely at random (MCAR) if missingness is independent of both observed and unobserved data and missing at random (MAR) if, conditional on the observed data, missingness does not depend on the unobserved data; otherwise the missingness process is termed non-random (MNAR). If the missingness process is random and the parameters of the observation process are functionally independent of the parameters describing the missingness process, then a valid statistical analysis can be obtained through a likelihood based analysis (or a Bayesian analysis) that ignores the missigness mechanism. This situation is termed *ignorable* by Rubin (1976) and Little and Rubin (2002).

While historically most methods were framed within the MCAR category, for computational and other simplicity reasons, more work has been done in the MAR and more recently in the MNAR category (see for instance Hogan & Laird, 1997; Little, 1995; Little & Rubin, 2002; Molenberghs & Kenward, 2007; Molenberghs & Verbeke, 2005; Schafer, 1997; Verbeke & Molenberghs, 2000, and the references therein). In many testing situations, including our context, missingness often depends on latent data such as examinee ability and sensitivity to test speededness. This would point to MNAR, which is nonignorable, regardless the inferential mode chosen. Many authors have warned for too firm a belief in a single (MNAR) model since, due to the very nature of incompleteness, such a model cannot be verified from observed data only. This implies great sensitivity to model assumptions (Molenberghs, Beunckens, Sotto, & Kenward, 2008; Molenberghs & Kenward, 2007; Molenberghs & Verbeke, 2005; Verbeke & Molenberghs, 2000). These issues are compounded when, in addition to incomplete data, the models feature latent structure, (unobserved) random effects, etc. We are in need of a model that combines all of these.

Apart from random guessing, random subject effects, and test speededness, incompleteness occurs and there are likely interrelationships between these entities.

The remainder of this paper is organized as follows. First, we introduce a model for omitted responses and test speededness. This model is derived from a decision tree that describes the student's possible states and actions when he/she encounters an item. Second, we discuss how the optimal person fit test of Levine and Drasgow (1988), the empirical Bayes estimate for the test speededness effect and the local influence diagnostics of Cook (1986) can be used to highlight examinees affected by test speededness. Finally, we illustrate the three methods with the Chilean SIMCE mathematics placement test data.

A Model for Test Speededness and Omitted Items

In this section we describe a model that provides a possible explanation for non-response in test data. Under the postulated model, non-response arises from a tendency to omit in case one does not know the answer and a test speededness effect, both taken to be examinee specific. The model is discussed in full detail in Goegebeur et al. (2006), where it proved useful for modeling test speededness and non-response.

The model can be motivated as follows. Let ξ_{0p} denote an examinee specific initial propensity to omit items and ξ_{1p} an examinee specific effect of test speededness. When examinee p encounters item i he/she is either knowledgeable or ignorant. If knowledgeable, the probability of a correct answer, denoted $P_i(\theta_p)$, is given by Equation 1 or 2. If ignorant, the examinee omits the item with probability $P_i(\xi_{0p}, \xi_{1p})$ and guesses at random with probability $1 - P_i(\xi_{0p}, \xi_{1p})$, where we assume

$$P_i(\xi_{0p}, \xi_{1p}) = \frac{\exp(\xi_{0p} + \xi_{1p}\, i/I)}{1 + \exp(\xi_{0p} + \xi_{1p}\, i/I)}; \qquad \xi_{0p} \in \mathbb{R},\ \xi_{1p} > 0. \tag{3}$$

Note that speededness is assumed to be a function of the item number, which explains the covariate i/I. Moreover speededness increases the probability of an omitted response. In case the examinee guesses at random, the answer is correct with probability c. In Figure 1 the process described above is visually represented by a decision tree.

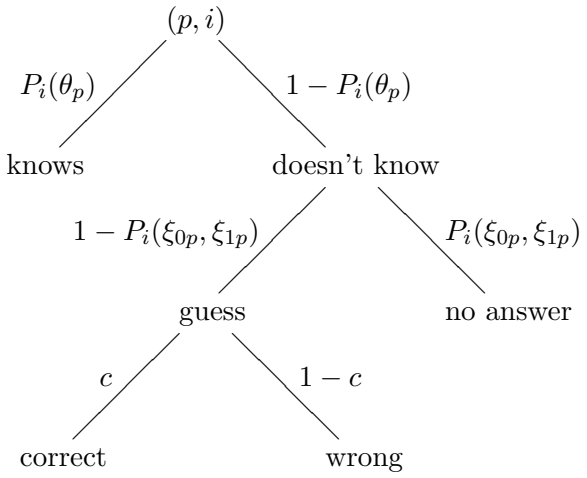

Figure 1: Decision tree representation of the test speededness model.

Clearly, this decision tree involves a categorical response variable with 3 possible levels: no answer, wrong answer and correct answer, coded $\boldsymbol{Y}'_{pi} := (Y_{pi0}, Y_{pi1}) = (1, 0)$, $\boldsymbol{Y}'_{pi} = (0, 1)$, and $\boldsymbol{Y}'_{pi} = (0, 0)$, respectively. The corresponding conditional probabilities will be denoted by π_{pi0}, π_{pi1}, and π_{pi2}, and have expressions that follow immediately from Figure 1:

$$\begin{aligned}
\pi_{pi0} &= [1 - P_i(\theta_p)]P_i(\xi_{0p}, \xi_{1p}), && (4)\\
\pi_{pi1} &= [1 - P_i(\theta_p)][1 - P_i(\xi_{0p}, \xi_{1p})](1 - c), && (5)\\
\pi_{pi2} &= [1 - P_i(\xi_{0p}, \xi_{1p})]c + \{1 - [1 - P_i(\xi_{0p}, \xi_{1p})]c\}P_i(\theta_p)\\
&= P_i(\theta_p) + [1 - P_i(\xi_{0p}, \xi_{1p})]c[1 - P_i(\theta_p)] && (6)
\end{aligned}$$

The random effects θ_p, ξ_{0p}, and $\log \xi_{1p}$ are assumed to follow a multivariate normal distribution:

$$\begin{pmatrix} \theta_p \\ \xi_{0p} \\ \log \xi_{1p} \end{pmatrix} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Omega}), \qquad (7)$$

with $\boldsymbol{\mu}' = (0, \mu_{\xi_0}, \mu_{\xi_1})$ and $\boldsymbol{\Omega}$ a positive definite covariance matrix. Given that $\xi_{1p} > 0$, we assumed that normality holds for $\log \xi_{1p}$. Otherwise stated, ξ_{1p} is log-normally distributed. Conditional on the random effects θ_p, ξ_{0p} and ξ_{1p}, the responses of examinee p to the I items are

assumed to be independent. Under the model proposed, the probability of an omission, given by Equation 4, increases with the difficulty of the item, the initial propensity to omit answers and the sensitivity to test speededness effects, but it decreases with the examinee's ability.

Some remarks apply. First, the probability of a missing value depends on unobserved information (the random effects that underlie the data) and hence missingness is allowed to be missing not at random (MNAR). Second, the dropout and measurement processes are allowed to have some parameters in common, turning it into a shared-parameter model (Molenberghs & Verbeke, 2005). As is clear from the model statement, the probabilities related to the measurement process (π_{pi1} and π_{pi2}), and the probability related to the missingness process (π_{pi0}) share the random effects θ_p, ξ_{0p} and ξ_{1p}. This implies that apart from the correct/wrong answers, also missingness contains information about item difficulty and person ability. Third, if $P_i(\xi_{0p}, \xi_{1p}) = 0$ then the proposed model reduces to the 3PL in case $P_i(\theta_p)$ is given by Equation 2 and to the 1PL extended with guessing (1PLc) if $P_i(\theta_p)$ is given by Equation 1. Fourth, if $P_i(\xi_{0p}, \xi_{1p}) > 0$, π_{pi2} is smaller than the probability of a correct answer under the 3PL or the 1PLc. This becomes immediately clear from a comparison of the success probability under the proposed model, as given in Equation 6, with the 3PL success probability, given by

$$P_i(\theta_p) + c[1 - P_i(\theta_p)].$$

As a direct consequence, the lower asymptote (for $\theta_p \to -\infty$) of the proposed model, given by $[1 - P_i(\xi_{0p}, \xi_{1p})]c$, is smaller than the lower asymptote of the 3PL or the 1PLc (which is c).

Since the purpose of the paper is to identify examinees with response profiles affected by test speededness effects, we will need to compare two models: a model without test speededness (the reduced model, also referred to as the null model) and a test speededness model. To facilitate the comparison and to introduce a generic formulation, we extend the model by including weight parameters ω_p, $p = 1, \ldots, P$, in the probability of an omitted item in the following way

$$P_i(\xi_{0p}, \xi_{1p}|\omega_p) = \frac{\exp(\xi_{0p} + \omega_p \xi_{1p}\, i/I)}{1 + \exp(\xi_{0p} + \omega_p \xi_{1p}\, i/I)}. \tag{8}$$

Under this parametrization, the reduced model is obtained for $\omega_p = 0$, $p = 1, \ldots, P$, whereas the test speededness model results from setting $\omega_p = 1$, $p = 1, \ldots, P$.

Optimal Person Fit Test

Drasgow and Levine (1986) and Levine and Drasgow (1988) used the Neyman-Pearson lemma (see e.g., Lehmann & Romano, 2005, p 59) to construct optimal person fit indices. In this, 'optimal' means that for a given level of significance no other procedure can attain a higher probability of detecting aberrant response patterns. The basic idea is to compute the probability of a response vector $\boldsymbol{Y}_p$ under two competing models, describing normal and aberrant test taking behavior, respectively, followed by a decision on the basis of their ratio. In their work, Drasgow and Levine (1986), and Levine and Drasgow (1988), concentrated mainly on the detection of spuriously low (e.g., due to alignment errors, atypical education) and high (copying answers, cheating) response patterns, but of course the procedure can be equally well applied to detect other forms of aberrant behavior. In the current paper, normal test taking behavior refers to non-speeded examinees whereas aberrant test taking behavior refers to examinees affected by test speededness effects. In this respect, for the model proposed above and denoting $\boldsymbol{Y}_p = (\boldsymbol{Y}_{p1}, \ldots, \boldsymbol{Y}_{pI})'$, the decision about the nature of the test taking behavior of examinee p will be based on the ratio

$$\Lambda_p = \frac{P(\boldsymbol{Y}_p = \boldsymbol{y}_p \mid \text{aberrant })}{P(\boldsymbol{Y}_p = \boldsymbol{y}_p \mid \text{normal })} \tag{9}$$

with

$$\begin{aligned} P(\boldsymbol{Y}_p = \boldsymbol{y}_p \mid \text{aberrant }) &= \int_{\mathbb{R}^2} \int_0^\infty A_p(1) f(\theta_p, \xi_{0p}, \xi_{1p}) d\xi_{1p} d\xi_{0p} d\theta_p, \\ P(\boldsymbol{Y}_p = \boldsymbol{y}_p \mid \text{normal }) &= \int_{\mathbb{R}^2} \int_0^\infty A_p(0) f(\theta_p, \xi_{0p}, \xi_{1p}) d\xi_{1p} d\xi_{0p} d\theta_p, \\ &= \int_{\mathbb{R}^2} A_p(0) f(\theta_p, \xi_{0p}) d\xi_{0p} d\theta_p, \end{aligned}$$

and

$$\begin{aligned} A_p(\omega_p) &= P(\boldsymbol{Y}_p = \boldsymbol{y}_p|\theta_p, \xi_{0p}, \xi_{1p}, \omega_p) \\ &= \prod_{i=1}^{I} P(\boldsymbol{Y}_{pi} = \boldsymbol{y}_{pi}|\theta_p, \xi_{0p}, \xi_{1p}, \omega_p) \\ &= \prod_{i=1}^{I} [\pi_{pi0}(\omega_p)]^{y_{pi0}} [\pi_{pi1}(\omega_p)]^{y_{pi1}} [\pi_{pi2}(\omega_p)]^{1-y_{pi0}-y_{pi1}}, \end{aligned} \tag{10}$$

where f denotes the joint density function of the random effects. In Equation 10, $\pi_{pi0}(\omega_p)$, $\pi_{pi1}(\omega_p)$ and $\pi_{pi2}(\omega_p)$ are given by Equation 4, 5 and 6, respectively, with $P_i(\xi_{0p}, \xi_{1p})$ replaced by $P_i(\xi_{0p}, \xi_{1p}|\omega_p)$. The hypothesis of normal test behavior of examinee p is rejected at level α in favor of aberrant test behavior, in casu speeded test behavior, if Λ_p is too large, or formally, if $\log \Lambda_p > c_\alpha$, where c_α is quantile $1 - \alpha$ of the null distribution of $\log \Lambda_p$.

It is important to keep in mind that the likelihood ratio test statistic in Equation 9 will only be optimal if the two probabilities are correct. In a study involving real data, the likelihood ratio test will be accurate to the extent that (i) there is little misspecification of the two models and (ii) the item parameters have been precisely estimated. From a computational point of view, application of Equation 9 requires that both the reduced and the test speededness model are fitted to the available data. This can be done by marginal maximum likelihood estimation, for instance, using the SAS NLMIXED procedure (example SAS code can be obtained upon request). For the actual computation of Λ_p the authors developed a Fortran program. In this program the numerical integrations are performed by the NAG library subroutines D01BBF and D01FBF (NAG, 1993). For the numerical integration related to the speededness effect ξ_{1p}, the quadrature points were taken from the standard normal distribution and transformed to the log-normal scale by $\exp(\mu_{\xi_1} + \sigma_{\xi_1} z)$, where z denotes a quadrature point for the standard normal distribution.

Empirical Bayes Estimates

Although the model estimation implies an estimate of the parameters of the marginal distribution of $\boldsymbol{Y}$, it is common practice in psychometrics to also calculate the estimations of the person

parameters. These are in the case of the test speededness model given by Equations 4, 5 and 6, the ability parameter θ_p, the initial propensity to omit ξ_{0p}, and the test speededness parameter ξ_{1p}. These random effects estimates give an idea about the between-subject variability, and hence provide information that is helpful for detecting special profiles, say outlying individuals, or groups of individuals evolving differently in time, in our context individuals affected by test speededness effects. To obtain estmates for the random effects, we need their conditional posterior distribution. Let $\boldsymbol{\psi}_1$ denote the parameter vector of the test-speededness model, with $P_i(\theta_p)$ modeled by a 1PL, that is, $\boldsymbol{\psi}_1' = (\beta_1, \ldots, \beta_I, c, \mu_{\xi_0}, \mu_{\xi_1}, \sigma^2_\theta, \sigma^2_{\xi_0}, \sigma^2_{\xi_1}, \sigma_{12}, \sigma_{13}, \sigma_{23})$, where σ^2_θ, $\sigma^2_{\xi_0}$ and $\sigma^2_{\xi_1}$ denote the variance of the examinee ability, the initial propensity to omit, and the log-transformed test speededness random effect, respectively, and $\sigma_{12} = \text{Cov}(\theta, \xi_0)$, $\sigma_{13} = \text{Cov}(\theta, \log \xi_1)$ and $\sigma_{23} = \text{Cov}(\xi_0, \log \xi_1)$ (these variances and covariances are the elements of the covariance matrix $\boldsymbol{\Omega}$ in Equation 7). For notational convenience we split $\boldsymbol{\psi}_1$ into subvectors $\boldsymbol{\psi}_{11}$ and $\boldsymbol{\psi}_{12}$, with $\boldsymbol{\psi}_{11}' = (\beta_1, \ldots, \beta_I, c)$ and $\boldsymbol{\psi}_{12}' = (\mu_{\xi_0}, \mu_{\xi_1}, \sigma^2_\theta, \sigma^2_{\xi_0}, \sigma^2_{\xi_1}, \sigma_{12}, \sigma_{13}, \sigma_{23})$. Using Bayes' rule we have

$$p(\theta_p, \xi_{0p}, \xi_{1p} \,|\boldsymbol{y}_p, \boldsymbol{\psi}_1) = \delta_p p(\boldsymbol{y}_p \,|\theta_p, \xi_{0p}, \xi_{1p}, \boldsymbol{\psi}_{11}) p(\theta_p, \xi_{0p}, \xi_{1p} \,|\boldsymbol{\psi}_{12}), \tag{11}$$

where δ_p is the normalizing constant, that is, $\delta_p = 1/p(\boldsymbol{y}_p \,|\boldsymbol{\psi}_1)$, $p(\boldsymbol{y}_p \,|\theta_p, \xi_{0p}, \xi_{1p}, \boldsymbol{\psi}_{11})$ is given by $A_p(1)$, see Equation 10, and

$$p(\theta_p, \xi_{0p}, \xi_{1p} \,|\boldsymbol{\psi}_{12}) = \frac{1}{(2\pi)^{3/2} |\boldsymbol{\Omega}|^{1/2} \xi_{1p}} \, e^{-\boldsymbol{\gamma}' \boldsymbol{\Omega}^{-1} \boldsymbol{\gamma}/2}, \tag{12}$$

with $\boldsymbol{\gamma}' = (\theta_p, \xi_{0p} - \mu_{\xi_0}, \ln \xi_{1p} - \mu_{\xi_1})$. The mode of the conditional posterior density given in Equation 11 is used as point estimate for θ_p, ξ_{0p} and ξ_{1p}. More specifically, the empirical Bayes estimate $(\hat{\theta}_p, \hat{\xi}_{0p}, \hat{\xi}_{1p})$ is the value for $(\theta_p, \xi_{0p}, \xi_{1p})$ that maximizes $p(\theta_p, \xi_{0p}, \xi_{1p} \,|\boldsymbol{y}_p, \boldsymbol{\psi}_1)$, in which the unknown parameters in $\boldsymbol{\psi}_1$ have been replaced by their estimates obtained from the marginal maximum likelihood estimation. Computation of the empirical Bayes estimates requires the estimation of the model for omissions with test speededness, followed by an optimization of Equation 11 for each of the examinees. The computation of the empirical Bayes estimates for the random effects is clearly the most direct approach to the identification of examinees, whose performance is vulnerable to test speededness, but this approach does not take

the fit of a particular model to a response profile into account.

Local Influence Diagnostics

Global influence diagnostics are based on a case-deletion approach (Chatterjee & Hadi, 1988). Broadly, all or part of a subject's measurements are deleted and key aspects of the model refitted, such as the likelihood value, parameter estimates, etc. When the distance between the overall and the refitted measure is large in a precisely defined sense, a case is considered influential. Global influence or case-deletion diagnostics have been well developed, for example, for linear regression and explicit forms derived. The main problems with the method applied to more general settings are that (1) the application of the method can be computer-intensive since no closed form expressions exist and (2) it may be difficult to gain further insight as to why a certain subject, observation, or set of observations is influential.

To overcome these limitations, local influence methods have been suggested, see Cook (1986). The principle of these is to investigate how the results of an analysis change under infinitesimal perturbations of the model. In the present context, we use local influence diagnostics to assess the impact of introducing a random test speededness effect on the key model parameter estimates. This can be done by considering Equation 8 as the mechanism describing non-response in case one does not know the answer to a particular item. Indeed, the case $\omega_p = 0$, $p = 1, \ldots, P$, corresponds to a model without a test speededness effect. If a small perturbation of a particular ω_p leads to large differences in the parameter estimates, then examinee p exerts an unusually large impact on the model. We will now sketch the basic principles of local influence analysis and apply these to our test speededness problem. In this we assume $P_i(\theta_p)$ is modeled by a 1PL.

We denote by $\boldsymbol{\omega}$ the P dimensional vector of perturbation parameters, that is, $\boldsymbol{\omega}' = (\omega_1, \ldots, \omega_P)$, and by $\boldsymbol{\psi}$ the $(I + 5)$ dimensional vector of parameters associated with the postulated model, that is, $\boldsymbol{\psi}' = (\beta_1, \ldots, \beta_I, c, \mu_{\xi_0}, \sigma^2_\theta, \sigma^2_{\xi_0}, \sigma_{12})$. Note that the perturbation scheme as defined in Equation 8, with infinitesimal small changes in the direction of test speededness, also involves the parameters μ_{ξ_1}, $\sigma^2_{\xi_1}$, σ_{13} and σ_{23}. These additional parameters must be fixed by the user,

since the local influence approach only considers the impact of perturbations on the parameters of the null model. However, this more general parameterization allows us to assess the effect of perturbing the postulated model with an extra random effect, in particular a random test speededness effect, that may be correlated with the random effects in the model postulated. In case one is interested only in the effect of perturbing the model with a fixed, that is, non-random, test speededness effect, one simply fixes σ_{13} and σ_{23} at 0. Doing so the mean of ξ_1 appears as a common scale factor in the expressions for the normal curvatures, and hence can be safely ignored. For the technical details of this we refer to Goegebeur et al. (2006).

The log-likelihood function of the perturbed model is given by

$$\ell(\boldsymbol{\psi}|\boldsymbol{\omega}) \quad = \quad \sum_{p=1}^{P} \ell_p(\boldsymbol{\psi}|\omega_p)$$

in which $\ell_p(\boldsymbol{\psi}|\omega_p)$ denotes the log-likelihood contribution of examinee p, that is,

$$\ell_p(\boldsymbol{\psi}|\omega_p) = \ln P(\boldsymbol{Y}_p = \boldsymbol{y}_p|\omega_p, \boldsymbol{\psi}),$$

with

$$P(\boldsymbol{Y}_p = \boldsymbol{y}_p|\omega_p, \boldsymbol{\psi}) \quad = \quad \int_{\mathbb{R}^2} \int_0^\infty A_p(\omega_p) f(\theta_p, \xi_{0p}, \xi_{1p}) d\xi_{1p} d\xi_{0p} d\theta_p,$$

$A_p(\omega_p)$ is given by Equation 10. It is assumed that $\boldsymbol{\omega}$ belongs to an open subset $\tilde{\Omega}$ of $\mathbb{R}^P$. For $\boldsymbol{\omega}$ equal to $\boldsymbol{\omega}_0 = (0, \ldots, 0)'$, with $\boldsymbol{\omega}_0 \in \tilde{\Omega}$, $\ell(\boldsymbol{\psi}|\boldsymbol{\omega_0})$ corresponds to a model without test speededness effects, and this for all values of $\boldsymbol{\psi}$.

Let $\widehat{\boldsymbol{\psi}}$ be the maximum likelihood estimator for $\boldsymbol{\psi}$, obtained by maximizing $\ell(\boldsymbol{\psi}|\boldsymbol{\omega}_0)$, and let $\widehat{\boldsymbol{\psi}}_{\boldsymbol{\omega}}$ denote the maximum likelihood estimator for $\boldsymbol{\psi}$ under $\ell(\boldsymbol{\psi}|\boldsymbol{\omega})$. The local influence approach compares $\widehat{\boldsymbol{\psi}}$ and $\widehat{\boldsymbol{\psi}}_{\boldsymbol{\omega}}$. Similar estimates indicate that the parameter estimates are stable with respect to the proposed perturbations of the postulated model. Strongly different estimates indicate that the estimation procedure is highly sensitive with respect to perturbations. Cook (1986) proposed to measure the distance between $\widehat{\boldsymbol{\psi}}$ and $\widehat{\boldsymbol{\psi}}_{\boldsymbol{\omega}}$ by the so-called likelihood displacement, defined by

$$LD(\boldsymbol{\omega}) = 2[\ell(\widehat{\boldsymbol{\psi}}|\boldsymbol{\omega}_0) - \ell(\widehat{\boldsymbol{\psi}}_{\boldsymbol{\omega}}|\boldsymbol{\omega}_0)]. \tag{13}$$

Note that the log-likelihood function of the postulated model is evaluated in both $\widehat{\boldsymbol{\psi}}$ and $\widehat{\boldsymbol{\psi}}_{\boldsymbol{\omega}}$ and hence $LD(\boldsymbol{\omega}) \geq 0$. Note also that the likelihood displacement takes the variability of $\widehat{\boldsymbol{\psi}}$ into account. Indeed, $LD(\boldsymbol{\omega})$ will be large if $\ell(\boldsymbol{\psi}|\boldsymbol{\omega}_0)$ is strongly curved at $\widehat{\boldsymbol{\psi}}$, which means that $\boldsymbol{\psi}$ is estimated with high precision. From this perspective, a graph of $LD(\boldsymbol{\omega})$ versus $\boldsymbol{\omega}$ contains essential information on the influence of the perturbation scheme of interest. It is useful to view this graph as the geometric surface formed by the $P+1$ dimensional vector

$$\alpha(\boldsymbol{\omega}) = \begin{pmatrix} \boldsymbol{\omega} \\ LD(\boldsymbol{\omega}) \end{pmatrix}$$

as $\boldsymbol{\omega}$ varies throughout $\tilde{\Omega}$, see Figure 2 for in illustration in case $P = 2$.

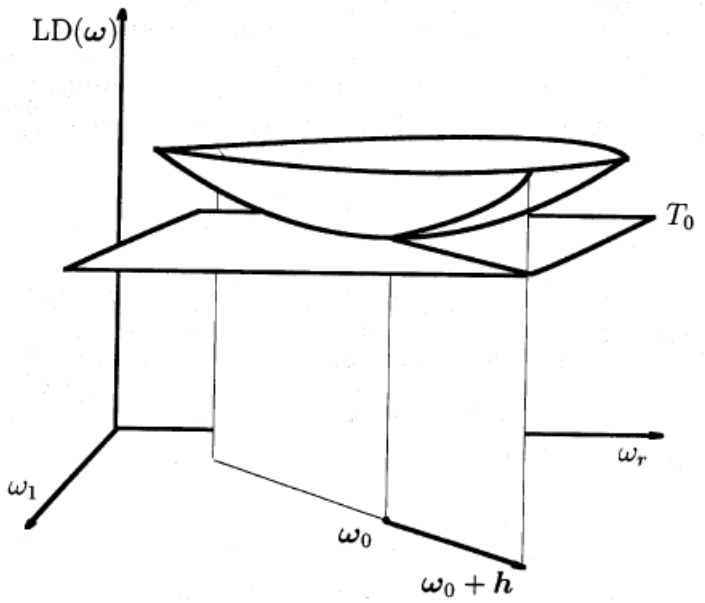

Figure 2: Illustration of the likelihood displacement.

Since this surface, the so-called influence graph, can only be depicted when $P \leq 2$, Cook (1986) proposed to look at normal curvatures of $\alpha(\boldsymbol{\omega})$ in $\boldsymbol{\omega}_0$ in a direction $\boldsymbol{h}$, with $\boldsymbol{h}$ a P dimensional vector of unit length. These normal curvatures can be easily calculated as

$$C_{\boldsymbol{h}} = 2|\boldsymbol{h}'\Delta'\ddot{L}^{-1}\Delta\boldsymbol{h}|, \tag{14}$$

with

$$\ddot{L} = \left. \frac{\partial^2 \ell(\boldsymbol{\psi}|\boldsymbol{\omega}_0)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} \right|_{\boldsymbol{\psi}=\widehat{\boldsymbol{\psi}}}$$

and Δ a $(I+5) \times P$ matrix of which the $p-$th column Δ_p is given by

$$\Delta_p = \left. \frac{\partial^2 \ell_p(\boldsymbol{\psi}|\omega_p)}{\partial \boldsymbol{\psi} \partial \omega_p} \right|_{\boldsymbol{\psi}=\widehat{\boldsymbol{\psi}},\omega_p=0}.$$

Figure 2 illustrates graphically the basic idea behind the normal curvature computed in $\boldsymbol{\omega}_0$ in the direction $\boldsymbol{h}$. The normal curvature, Equation 14, can be used in several ways to study the influence graph $\alpha(\boldsymbol{\omega})$, each one corresponding to a particular direction $\boldsymbol{h}$ in $\tilde{\Omega}$. One evident choice is the vector $\boldsymbol{h}_p$ which has a one on position p and zeros elsewhere, corresponding to a perturbation of the postulated model by weight ω_p only. In this case Equation 14 reduces to

$$C_p = 2|\Delta_p' \ddot{L}^{-1} \Delta_p|. \tag{15}$$

Other important directions are the directions of minimal and maximal curvature, denoted $\boldsymbol{h}_{\min}$ and $\boldsymbol{h}_{\max}$, respectively, obtained as solutions to the minimization and maximization, respectively, of $C_{\boldsymbol{h}}$ over the space of all vectors of unit length. It can be shown that $C_{\boldsymbol{h}_{\min}}$ and $C_{\boldsymbol{h}_{\max}}$ correspond to the smallest and largest eigenvalues of $-2\Delta' \ddot{L}^{-1} \Delta$ and $\boldsymbol{h}_{\min}$ and $\boldsymbol{h}_{\max}$ are the corresponding eigenvectors. Note that, compared to Λ_p, the computation of C_p requires only a null model fit, yielding significant gains in computation time, especially on large data sets.

The calculation of the local influence measures can be carried out as soon as expressions for $\ddot{L}$ and Δ have been obtained. The elements of $\ddot{L}$ are not computed analytically as these can be easily obtained from the maximization of $\ell(\boldsymbol{\psi}|\boldsymbol{\omega}_0)$, for instance by using the SAS NLMIXED procedure. The elements of the columns Δ_p of Δ and some theoretical properties thereof are given in Goegebeur et al. (2006) and will not be repeated here. The authors developed a Fortran program to compute the elements of Δ, the normal curvatures $C_{\boldsymbol{h}}$, and the direction of maximal curvature $\boldsymbol{h}_{\max}$. In this program, the numerical integrations are performed by the NAG library subroutines D01BBF and D01FBF, and the direction of maximal curvature is computed using subroutine F02FCF (NAG, 1993).

So far, the discussion of local influence diagnostics was focused on the complete $\boldsymbol{\psi}$ vector. Similar principles can be applied to obtain the local influence of perturbations on subsets of $\boldsymbol{\psi}$, see Cook (1986); Verbeke, Molenberghs, Thijs, Lesaffre, and Kenward (2001) and Goegebeur et al. (2006). This will not be pursued in the current paper.

SIMCE Mathematics Test Data

The SIMCE (Sistema de Medición de la Calidad de la Educación) project in Chile has developed mandatory language and mathematics tests to assess on a regular basis the educational progress in three levels: 4th, 8th and 10th graders. All students in the grade level in the country (public, private and mixed support schools) are expected to take the tests when they are scheduled (every 3 or 4 years). In this paper we will consider the data from the 2001 administration of the SIMCE mathematics test to the 10th graders in public schools. The mathematics test contains 48 items, each having 4 response alternatives, and covers topics such as problem formulation, functions, simple algebra, geometry and probability. For instance, simplifying $\frac{4}{x^2}/\frac{2}{x}$, or computing 30% of USD 2,000 in the context of an applied problem. The test is administered under a fixed time limit of 90 minutes. The database under consideration contains response profiles of 36,118 examinees. To illustrate the use of the likelihood ratio statistic, the empirical Bayes estimates and the normal curvatures we will use a sample of 3,000 examinees randomly drawn from this database. In Figure 3, the sample is summarized by plotting the proportions of omitted answers (solid line), wrong answers (dashed line) and correct answers (dashed-dotted line) as a function of the item number. The proportions of omitted answers vary between 0.0020 and 0.0537 with mean 0.0176 and standard deviation 0.0117. Out of the 3,000 examinees, 626 (20.87%) have a response profile with at least one omitted answer, so a complete case analysis would, besides being inappropriate given the type of missingness, also entail a substantial loss of information. Note also that the proportion of omitted items slightly increases with the item number, an effect that may be due to the fixed time limit administration of the test.

In Table 1 the reduced model and the test speededness model are compared on the basis of -2ℓ, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). All the

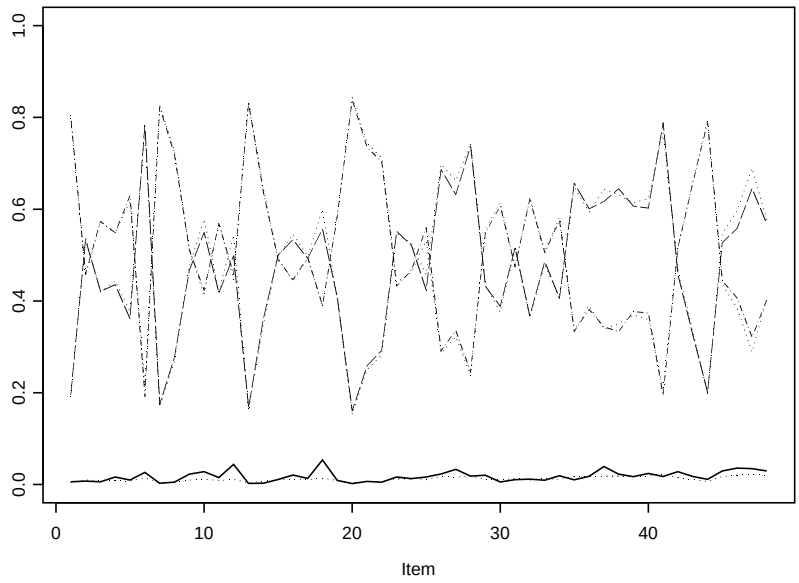

Figure 3: Proportion of missing data (solid line), wrong answers (dashed line) and correct answers (dashed-dotted line) together with the estimated theoretical proportion under the test speededness model (dotted line).

analyses are performed under the assumption of independent random effects. This does however not imply that the missingness mechanism is ignorable, given that we are dealing with a shared parameter model. Note that the reduced model is nested in the test speededness model and hence will always have a larger -2ℓ value. The difference of the -2ℓ values can be used to construct a likelihood ratio test for the null hypothesis of the reduced model. Given a difference of 530 for only two model parameters, there is strong evidence in favor of the test speededness model. Also the AIC and BIC indicate the test speededness model as the most appropriate one to describe the SIMCE mathematics test data.

To obtain an indication about the fit of the test speededness model to the SIMCE mathematics data, we show in Figure 3 also the estimated theoretical proportions of omissions, wrong answers

Table 1: Goodness-of-fit statistics for the reduced and the test speededness model.

	Reduced model	Speeded model
-2ℓ	184,526	183,996
AIC	184,630	184,104
BIC	184,943	184,428

and correct answers (dotted lines), given by

$$
\begin{aligned}
P(Y_{pi0}=1,Y_{pi1}=0) &= \int_{\mathbb{R}^2}\int_0^\infty [1-P_i(\theta_p)]P_i(\xi_{0p},\xi_{1p})dF_3(\xi_{1p})dF_2(\xi_{0p})dF_1(\theta_p),\\
P(Y_{pi0}=0,Y_{pi1}=1) &= (1-c)\int_{\mathbb{R}^2}\int_0^\infty [1-P_i(\theta_p)][1-P_i(\xi_{0p},\xi_{1p})]dF_3(\xi_{1p})dF_2(\xi_{0p})dF_1(\theta_p),\\
P(Y_{pi0}=0,Y_{pi1}=0) &= c\int_{\mathbb{R}}\int_0^\infty [1-P_i(\xi_{0p},\xi_{1p})]dF_3(\xi_{1p})dF_2(\xi_{0p}) +\\
&\quad \int_{\mathbb{R}^2}\int_0^\infty \{1-[1-P_i(\xi_{0p},\xi_{1p})]c\}P_i(\theta_p)dF_3(\xi_{1p})dF_2(\xi_{0p})dF_1(\theta_p),
\end{aligned}
$$

respectively, with F_1, F_2 and F_3 denoting the distribution functions of examinee ability, initial propensity to omit and examinee-specific effect of test speededness, respectively, and with the unknown parameters replaced by their respective maximum likelihood estimate, as a function of item number. As is clear from Figure 3, the empirical and estimated theoretical proportions agree quite well, indicating a good fit of the test speededness model. Note that this comparison involves only marginal probabilities and hence gives only a partial picture of the model fit. For a more elaborate goodness-of-fit evaluation, involving also the fit of the model to the conditional response distributions, we refer to Goegebeur et al. (2006). The results presented there indicate that the assumption of a common guessing parameter c is not too restrictive.

Table 2 shows the estimates of the parameters related to the random effects and the random guessing parameter c, under both the reduced model and the test speededness model. Focusing on the test speededness model, the magnitudes of the estimates for the variances of the random effects indicate that the examinees clearly differ from each other with respect to their ability, their initial propensity to omit answers, and their speededness parameter. In this respect it is worthwhile to mention that the model without the test speededness random effect (the reduced

model in Table 2) gives a fit to the univariate marginal distributions that is nearly indistinguishable from the test speededness model (cf Figure 3). However, according to Table 2, the test speededness effect is important, and as a consequence the simpler model without a test speededness effect will give a worse description of the joint marginal distribution - and hence the dependence structure - of the item responses.

Table 2: Parameter estimates under the reduced model and the test speededness model.

	Reduced model		Speeded model	
Parameter	estimate	standard error	estimate	standard error
σ^2_θ	0.9928	0.0324	1.0155	0.0330
μ_{ξ_0}	-5.3783	0.0750	-5.7481	0.0965
$\sigma^2_{\xi_0}$	3.4854	0.1083	3.4794	0.1372
μ_{ξ_1}	-	-	-1.7657	0.2845
$\sigma^2_{\xi_1}$	-	-	1.7933	0.2558
c	0.1472	0.0050	0.1524	0.0049

We now try to identify the examinees with response profiles affected by test speededness effects. This is performed by computing the likelihood ratio test statistic, Equation 9, with unknown parameters replaced by their maximum likelihood estimates, the empirical Bayes estimate $\hat{\xi}_{1p}$, and the normal curvature, Equation 15, for $p = 1, \ldots, 3000$. Since interest is in the extreme cases, that is, the most significant likelihood ratio test, and the largest empirical Bayes estimates and normal curvatures, we examine the 20 largest values of each statistic. In Figures 4-6 we show the response profiles of the 20 examinees having the largest value for Λ_p, $\hat{\xi}_{1p}$ and C_p, respectively, sorted in ascending order. The response profiles show the item responses, where the correct answers are coded as 2, the wrong answers as 1, and the omissions as 0, as a function of the item number. Clearly, all highlighted profiles contain a lot of omissions, especially near the end of the test. Further, the likelihood ratio test and the empirical Bayes estimates identify almost exclusively response profiles with a quite abrupt transition from responses (correct or wrong), to omissions. The C_p criterion on the other hand also identifies cases with lots of

omissions, but where the transition from responses to omissions is less clear cut.

To assess the correspondence between the sets of extreme cases, identified by the three procedures, we computed the proportion of overlap in the highlighted examinees for the largest k values of Λ_p, $\hat{\xi}_{1p}$ and C_p, with $k = 5, \ldots, 500$. The results of this are presented in Figure 7. As is clear from this figure, for k values up to 100, Λ_p and $\hat{\xi}_{1p}$ show an overlap of about 90% in the highlighted cases (dashed line), whereas Λ_p and C_p (dotted line), and $\hat{\xi}_{1p}$ and C_p (solid line), show an overlap between 50 and 60%. The methods agree quite well in their identification of examinees with special response profiles although Λ_p and $\hat{\xi}_{1p}$ seem to show a closer correspondence to each other compared to C_p. An alternative way to evaluate the overlap and specificity of the three measures under consideration consists in examining their pairwise correlations. Given the extreme skewness of the distributions of Λ_p and $\hat{\xi}_{1p}$, the correlations involving the latter were computed after having taken the log-transform of these. As is clear from Table 3, and ignoring the row labeled $z_{\tilde{\xi}_{1p}}$ for the time being, $\ln \Lambda_p$ and $\ln \hat{\xi}_{1p}$ show a stronger linear dependency with each other than with C_p, a result that is in line with the earlier findings based on the overlap in the identified extreme cases.

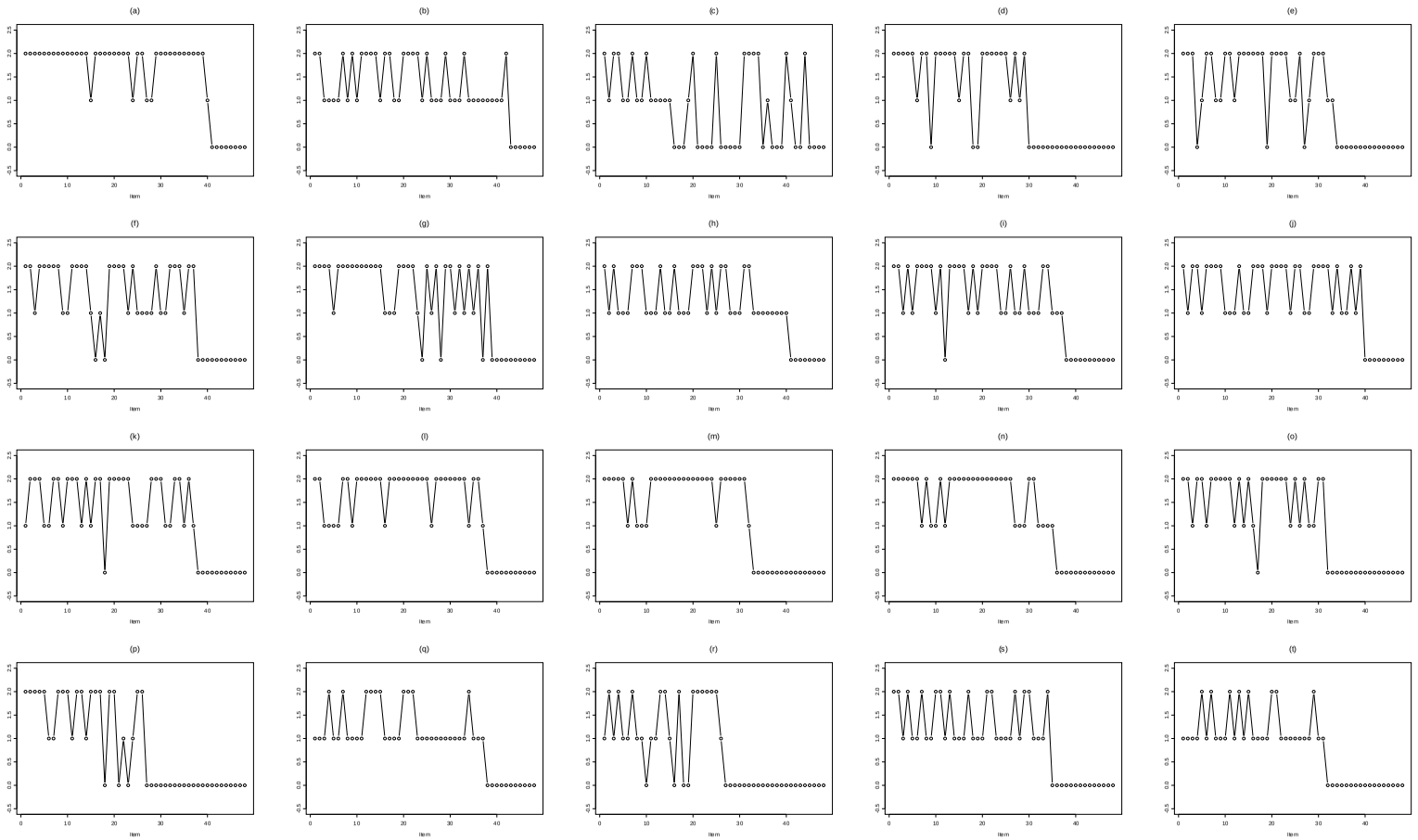

Figure 4: Profiles of the examinees with the 20 largest values for Λ_p, sorted in ascending order, with the response values 2, 1 and 0 representing the correct, wrong and omitted answers, respectively. Examinees (a) 497, (b) 1267, (c) 1536, (d) 846, (e) 827, (f) 866, (g) 1637, (h) 48, (i) 1821, (j) 826, (k) 1945, (l) 1181, (m) 192, (n) 99, (o) 2769, (p) 2013, (q) 2946, (r) 2377, (s) 1027 and (t) 2216.

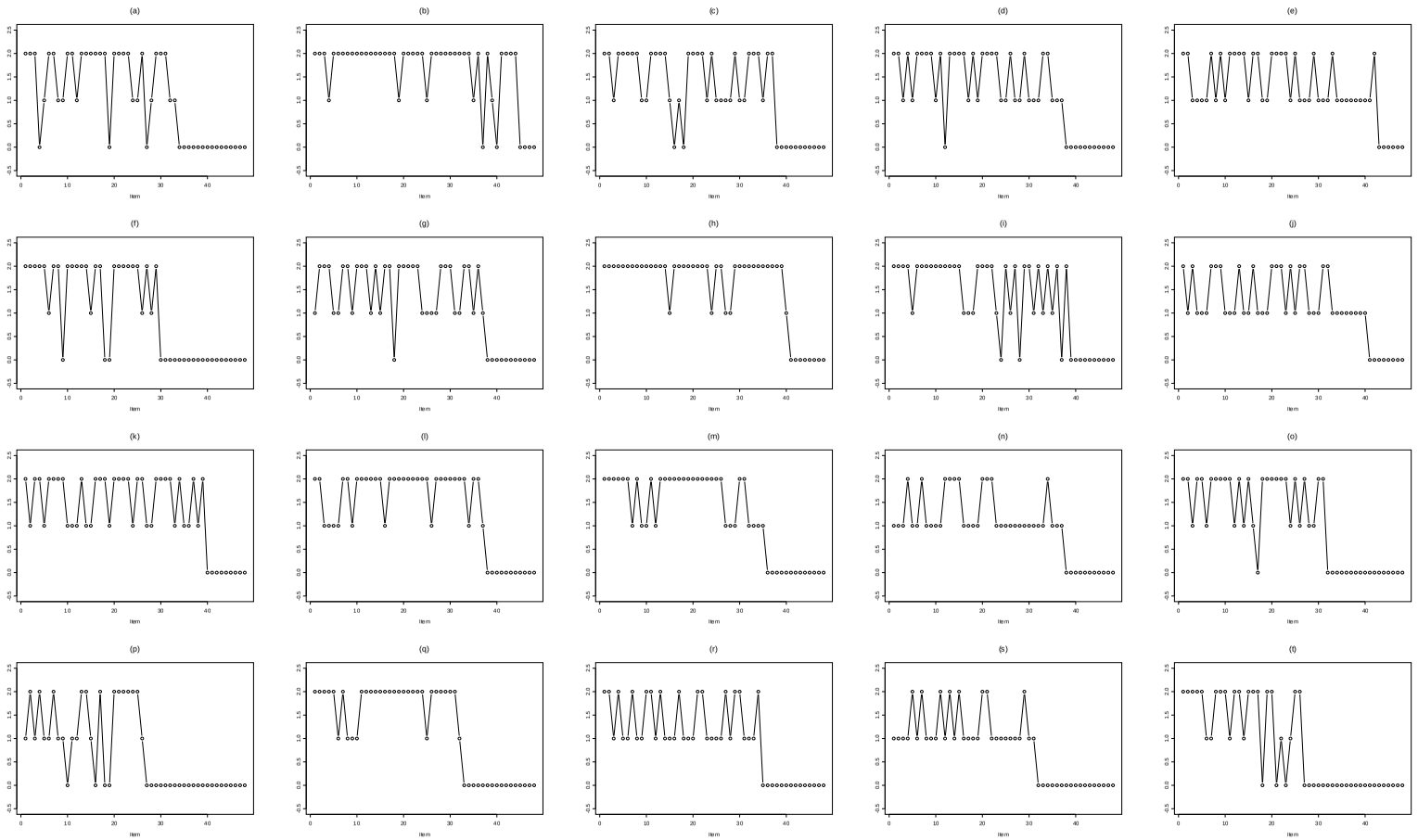

Figure 5: Profiles of the examinees with the 20 largest values for $\hat{\xi}_{1p}$, sorted in ascending order, with the response values 2, 1 and 0 representing the correct, wrong and omitted answers, respectively. Examinees (a) 827, (b) 1275, (c) 866, (d) 1821, (e) 1267, (f) 846, (g) 1945, (h) 497, (i) 1637, (j) 48, (k) 826, (l) 1181, (m) 99, (n) 2946, (o) 2769, (p) 2377, (q) 192, (r) 1027, (s) 2216 and (t) 2013.

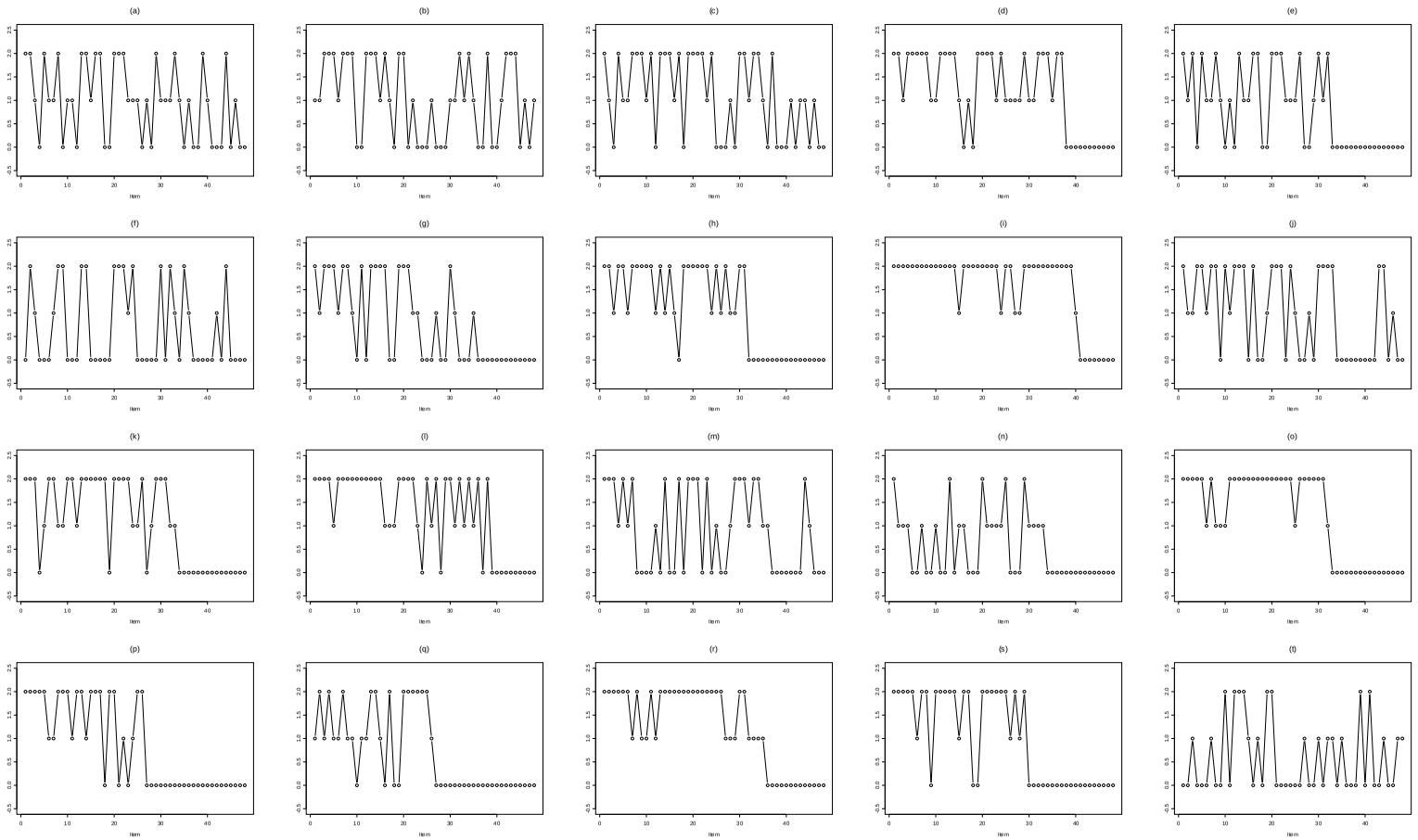

Figure 6: Profiles of the examinees with the 20 largest values for C_p, sorted in ascending order, with the response values 2, 1 and 0 representing the correct, wrong and omitted answers, respectively. Examinees (a) 2330, (b) 1753, (c) 554, (d) 866, (e) 1767, (f) 2322, (g) 2489, (h) 2769, (i) 497, (j) 215, (k) 827, (l) 1637, (m) 1133, (n) 193, (o) 192, (p) 2013, (q) 2377, (r) 99, (s) 846 and (t) 188.

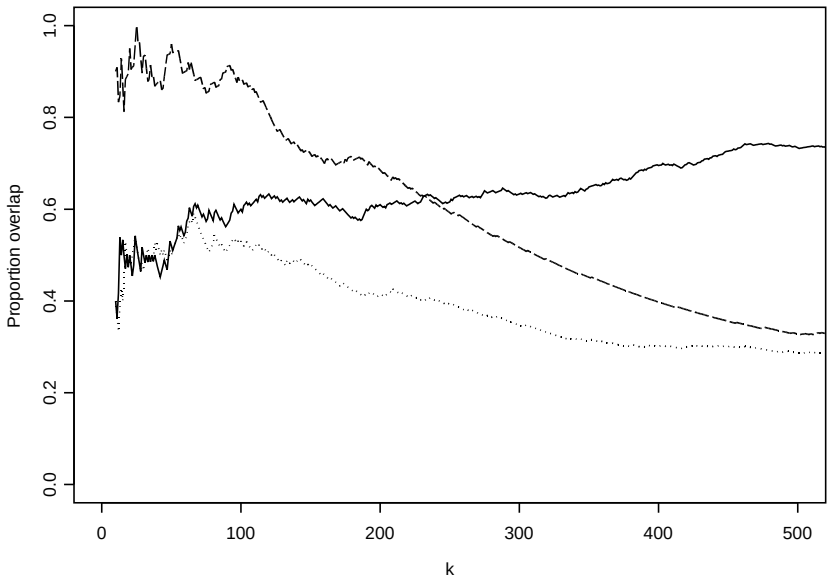

Figure 7: Proportion of overlap in the largest k values of (a) $\hat{\xi}_{1p}$ and C_p (solid line), (b) Λ_p and $\hat{\xi}_{1p}$ (dashed line), and (c) Λ_p and C_p (dotted line).

Table 3: Sample correlation between $\ln \Lambda_p$, $\ln \hat{\xi}_{1p}$, C_p, and $z_{\tilde{\xi}_{1p}}$.

	$\ln \Lambda_p$	$\ln \hat{\xi}_{1p}$	C_p
$\ln \hat{\xi}_{1p}$	0.803		
C_p	0.512	0.626	
$z_{\tilde{\xi}_{1p}}$	0.893	0.925	0.669

To gain insight in the determinants of the statistics under consideration we computed their sample correlations with the following characteristics of the response profiles: the proportions of answers omitted, the proportions of answers omitted in the first and the second half of the test, the 'empirical slope', being the difference between the proportions of answers omitted in the second and the first half of the test, and the variance of the response variable (taking the

values 0,1 and 2, cf Figures 4-6). The results of this analysis are presented in Table 4. The statistics $\ln \Lambda_p$ and $\ln \hat{\xi}_{1p}$ show a quite similar behavior in the sense that they correlate rather weakly with the proportions of answers omitted in the first half of the test, and rather strongly with the proportions of answers omitted in the second half of the test, as well as with the empirical slope. The latter is an important observation, as this 'empirical slope' can be seen as a possible proxy for the way test speededness was defined according to our model, namely the degradation in response quality – in casu more frequent omissions – as the test progresses. The normal curvatures show a somewhat deviant behavior in the sense that they correlate high with the total proportion of omissions, both the proportions of omissions in the first and the second half of the test, and with the variance of the response profile, a behavior that is consistent with the earlier visual impressions obtained from Figure 6. A possible explanation for this phenomenon could be that, given the relatively high variability of the item difficulties, examinees with quite variable response profiles contain more information about $\boldsymbol{\psi}$, that is, have a log-likelihood contribution that is more strongly curved at $\widehat{\boldsymbol{\psi}}$, than those with less variable profiles.

Table 4: Sample correlation between the person fit indices and the proportion of omissions (p), the proportion of omissions in the first half of the test (p_1), the proportion of omissions in the second half of the test (p_2), the difference between the proportions of omissions in the second and first half of the test, and the variance of the response.

Index	p	p_1	p_2	$p_2 - p_1$	$\mathrm{Var}(Y)$
$\ln \Lambda_p$	0.428	0.018	0.606	0.766	0.351
$\ln \hat{\xi}_{1p}$	0.570	0.195	0.709	0.766	0.492
C_p	0.894	0.668	0.899	0.654	0.807
$z_{\tilde{\xi}_{1p}}$	0.611	0.193	0.768	0.843	0.545

The computation of the empirical Bayes estimate is clearly the most direct approach to the identification of examinees affected by test speededness. The likelihood ratio statistic, originating from statistical test theory, takes a different point of view in the sense that the likelihood of

a response profile is evaluated under two competing models, and the examinee is assigned to the speeded class if the likelihood for the latter is considerably larger than the likelihood for the nonspeeded class. The likelihood ratio statistic evaluates two models with respect to each other, making it a relative criterion, and hence it gives no guarantee that the model for the group to which the examinee is assigned provides absolutely a good fit, that is, an examinee may be assigned to the speeded class without actually being speeded. Finally, the normal curvature is an influence diagnostic, measuring the impact of small model perturbations - here small perturbations in the direction of test speededness - on the estimates for the key model parameters. As such, the local influence diagnostics allow one to identify the set of observations that drive the conclusion of a statistical analysis in the direction of a particular model, when two models are under consideration. The latter consideration formed the motivation for the analyses performed in Goegebeur et al. (2006), see also Molenberghs et al. (2001) and Thijs et al. (2000). This may imply that examinees which do not fit the postulated model (the model without speededness) get highlighted as being locally influential because they do not fit the null model, but for another reason than being speeded, that is, the extra flexibility offered by the perturbation will be used by the respective observations as a way out of the pinching model, without test speededness being the cause of the misfit to the postulated model. A possible explanation for why the empirical Bayes estimate and the likelihood ratio test statistic show a closer correspondence to each other than to the normal curvature is that both statistics are based on the full model, compared to the normal curvature which only considers a perturbation of the reduced model. The above considerations seem to be confirmed by the scatter plots of the statistics versus the proportion of omissions in the second half of the test, given in Figure 8. The empirical Bayes estimates identify an clearly outlying, that is, speeded, group op observations, plotted by diamonds. The normal curvatures do not perform well in separating this speeded group from the non-speeded group, whereas the likelihood ratio statistic assumes a somewhat intermediate position.

Following a suggestion made by one of the referees, we also considered a standardized empirical Bayes estimate for ξ_{1p}, and hence took the variability of the estimate for ξ_{1p} into account. Given the skewness of the conditional posterior density of the random effects, we worked with the log-

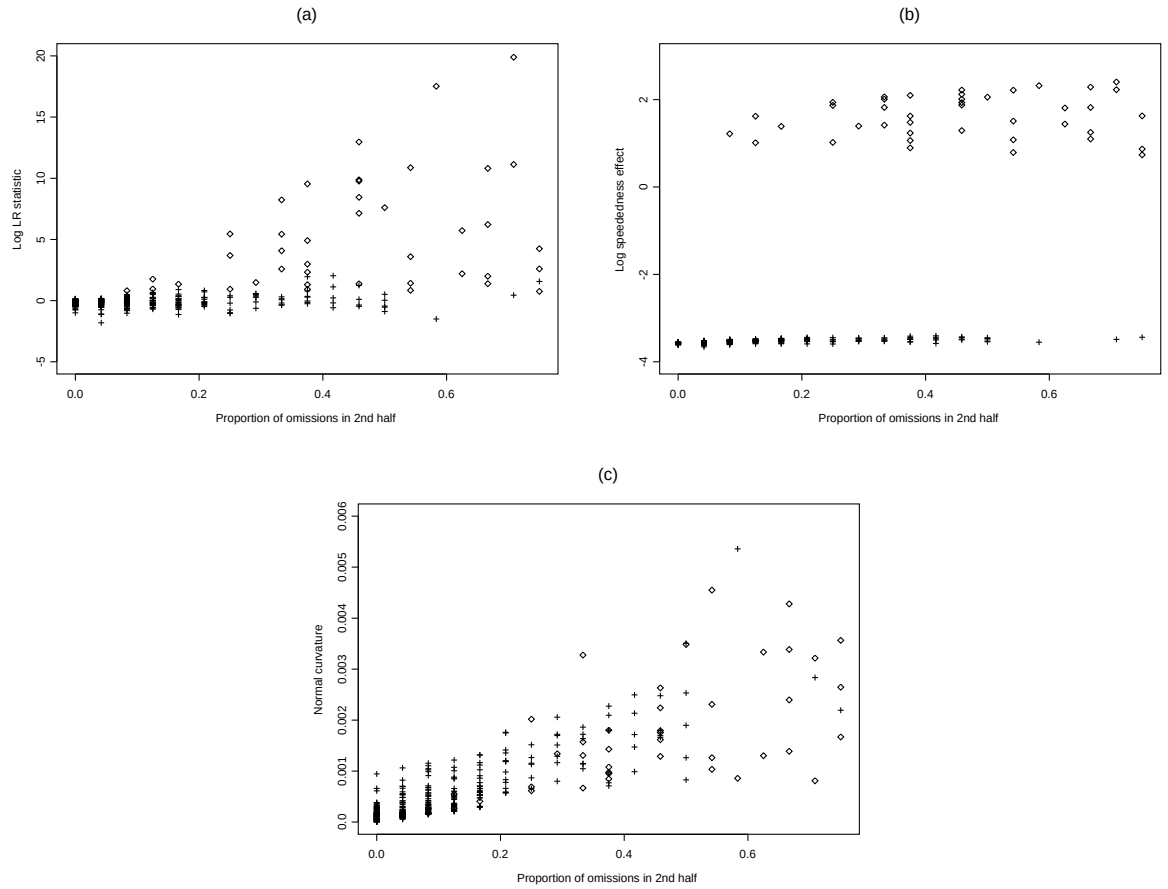

Figure 8: Scatter plots of (a) $\ln \Lambda_p$, (b) $\ln \hat{\xi}_{1p}$ and (c) C_p versus the proportion of omissions in the second half of the test. The observations plotted by a diamond represent the outlying group identified by $\hat{\xi}_{1p}$.

transformed test speededness parameter, denoted $\tilde{\xi}_{1p}$, and used the variance of the multivariate normal approximation to the joint density of $(\theta_p, \xi_{0p}, \tilde{\xi}_{1p})$ as an indicator of precision. The correlations of this standardized empirical Bayes estimate, denoted $z_{\tilde{\xi}_{1p}}$, with the other statistics and the properties of the response profiles are also given in Table 3 and Table 4, respectively. Compared to the 'raw' empirical Bayes estimate, this statistic correlates - as expected - better with C_p although the gain is rather small. Also the percentages of overlap in the extreme cases differed only in a minor way from the previously obtained ones.

Discussion and Conclusion

In this paper we compared the performance of the optimal appropriateness statistic proposed by Drasgow and Levine (1986), the empirical Bayes estimate for the test speededness random effect and the local influence approach of Cook (1986) with respect to the detection of test scores affected by test speededness effects. The framework for this person fit analysis was the model for omitted responses in test data recently proposed by Goegebeur et al. (2006). Under this model, non-response emerges from a general tendency to omit answers in case one does not know the answer, and a test speededness effect, both taken to be examinee specific. Under the optimal appropriateness approach, two models are compared, a model with and one without test speededness, and the decision about the nature of an examinee's test taking behavior is based on the ratio of the response profile probabilities under both models. This approach is optimal in the sense that no other procedure with the same size can yield a higher detection rate. The local influence approach starts from a postulated model, here a model without a test speededness effect, and looks at the impact minor model perturbations in the direction of test speededness have on the parameter estimates. Finally, according to the empirical Bayes approach one obtains for each examinee an estimate for the test speededness random effect, and hence this approach can be considered as the most direct one when interest is in detecting examinees affected by test speededness. Although the statistics considered are developed for quite different purposes, hypothesis testing in case of the optimal person fit test versus assessment of local influence in case of the normal curvatures and estimation in case of empirical Bayes, and hence will exhibit specificity, the results obtained on the SIMCE test data indicated that there is also overlap, and that all offer promising perspectives with respect to detecting test speededness. To get a better understanding of the true virtues of these methods in this respect, a more thorough examination is needed, for instance on the basis of an extensive simulation study. Work on this is in progress.

Persons identified as being speeded can be removed from the data set for purposes of estimation quality of the other parameters, such as the item parameters, and in order to avoid the consequences of a misspecified model when speededness is not incorporated in the model for reasons of simplicity. The local influence diagnostic is a direct indication of how large the impact is of

a given person on the key model parameters, and it is therefore a highly interesting indicator of cases to be removed. Interestingly, the proportions of persons identified as being speeded tells us for which proportion the model complexity is required in order to obtain a good fit. Perhaps these persons should be tested in a different way in order to obtain a more valid ability estimate. When keeping the ability estimates of such persons based on the test with speededness effects, these estimates should be treated with more caution. Identifying persons with a speededness profile is like identifying persons with a poor person fit.

Acknowledgements

Yuri Goegebeur's research was supported by a grant of the Danish Natural Science Research Council. Paul De Boeck and Geert Molenberghs gratefully acknowledge support from IAP Research Network P6/03 of the Belgian Government (Belgian Science Policy).

References

Beckman, R.J., Nachtsheim, C.J., & Cook, R.D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics, 29*, 413–426.

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. 394–479). Reading, MA: Addison-Wesley.

Bolt, D.M., Cohen, A.S., & Wollack, J.A. (2002). Item parameter estimation under conditions of test speededness: application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331–348.

Chatterjee, S., & Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression.* New York: John Wiley & Sons.

Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B, 48*, 133–169.

Cook, R.D., & Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman & Hall.

Douglas, J., Kim, H.R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics, 23*, 129–151.

Drasgow, F., & Levine, M.V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10*, 59–67.

Goegebeur, Y., De Boeck, P., Molenberghs, G., & del Pino, G. (2006). A local influence based diagnostic approach to a speeded IRT model. *Journal of the Royal Statistical Society, Series C, 55*, 647–676.

Goegebeur, Y., De Boeck, P., Wollack, J.A., & Cohen, A.S. (2008). A speeded item response model with gradual process change. *Psychometrika, 73*, 65–87.

Hogan, J.W., & Laird, N.M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine, 16*, 239–258.

Lehmann, E.L., & Romano, J.P. (2005). *Testing Statistical Hypothesis*. New York: Springer.

Lesaffre, E., & Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics, 54*, 570–582.

Levine, M.V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53*, 161–176.

Little, R.J.A. (1995). Modeling the drop-out mechanism in repeated measures studies. *Journal of the American Statistical Association, 90*, 1112–1121.

Little, R.J.A., & Rubin, D.B. (2002). *Statistical Analysis with Missing Data (2nd ed.)*. New York: John Wiley and Sons.

Meijer, R.R., & Sijtsma, K. (2001). Methodology review: evaluating person fit. *Applied Psy-*

chological Measurement, 25, 107–135.

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M.G. (2008). Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B, 70*, 371–388.

Molenberghs, G., & Kenward, G. (2007). *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons.

Molenberghs, G., & Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., & Kenward, M.G. (2001). Influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis, 37*, 93–113.

NAG (1993). *NAG Fortran Library Manual - Mark 19*. The Numerical Algorithms Group Limited.

Oshima, T.C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*, 200–219.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Danish Institute for Educational Research, Copenhagen, Denmark.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.

San Martín, E., del Pino, G., & De Boeck, P. (2006). IRT models for ability based guessing. *Applied Psychological Measurement, 30*, 183–203.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Thijs, H., Molenberghs, G., & Verbeke, G. (2000). The milk protein trial: influence analysis of the dropout process. *Biometrical Journal, 42*, 617–646.

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., & Kenward, M.G. (2001). Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics, 57*, 7–14.

Wollack, J.A., & Cohen, A.S. (2005). A model for simulating speeded test data. Technical report.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the hybrid model. In J. Rost, & R. Langeheine (Eds.), *Applications of Latent Trait and Latent Class Models in the Social Sciences* (pp. 89–99). New York: Waxmann.