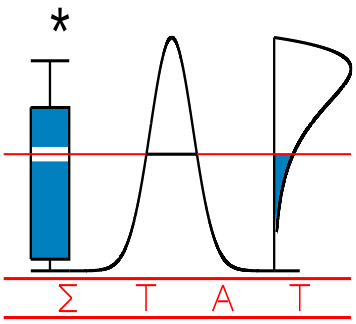
T E C H N I C A L

R E P O R T

0652

PERSON FIT FOR TEST SPEEDEDNESS : NORMAL CURVATURE AND LIKELIHOOD RATIO BASED TESTS

GOEGEBEUR Y., DE BOECK P., and G. MOLENBERGHS



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Person fit for test speededness: normal curvature and likelihood ratio based tests

Yuri Goegebeur *
Paul De Boeck ‡
Geert Molenberghs §

January 26, 2007

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). The performance of both methods is illustrated on the Chilean SIMCE mathematics test data.

*University of Southern Denmark, Department of Statistics, J.B. Winsløws Vej 9B, DK-5000 Odense C, 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

1 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 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, and hence can be seen as the IRT analogues of the global influence diagnostics, 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 illustrate the possibilities the local influence diagnostics, introduced by Cook (1986) as general measures to assess the impact of minor model perturbations, offer for detecting test speededness, and compare in this respect their performance with the optimal person fit statistic proposed by Levine and Drasgow (1988). 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. On the other hand, the item difficulty parameters of items administered late in the test tend to be overestimated (Douglas *et al.*, 1998 and Oshima, 1994). Item response models accommodating test speededness were proposed by Yamamoto and Everson (1997), Bolt *et al.* (2002), Wollack and Cohen (2004) and Goegebeur *et al.* (2005a,b). The analysis described in this paper will be based on the model Goegebeur *et al.* (2005a) 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. As this model builds upon classical IRT models, it is instructive to review some of these.

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 random intercept 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 guessing parameter c_i represents the probability of a correct answer under random guessing.

The remainder of this paper is organized as follows. In Section 2 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 s/he encounters an item. In Section 3 we illustrate how the optimal person fit test of Levine and Drasgow (1988) and the local influence diagnostics of Cook (1986) can be used to highlight examinees affected by test speededness. In Section 4 we illustrate both methods with the Chilean SIMCE mathematics placement test data.

2 A model for test speededness and omitted items

In this section we introduce 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 proposed model is taken from Goegebeur *et al.* (2005a), where it proved useful for modeling test speededness and non-response.

The model can be motivated as follows. When subject p encounters item i s/he is either knowledgeable or ignorant. If knowledgeable the probability of a correct answer, denoted $P_i(\theta_p)$, is given by (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

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

The random effect ξ_{0p} can be seen as the initial propensity of examinee p to omit items, while ξ_{1p} reflects the examinee-specific effect of test speededness, where speeding increases the probability of an omitted response. Speededness is assumed to be a function of the item number, which explains the covariate i/I. 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.

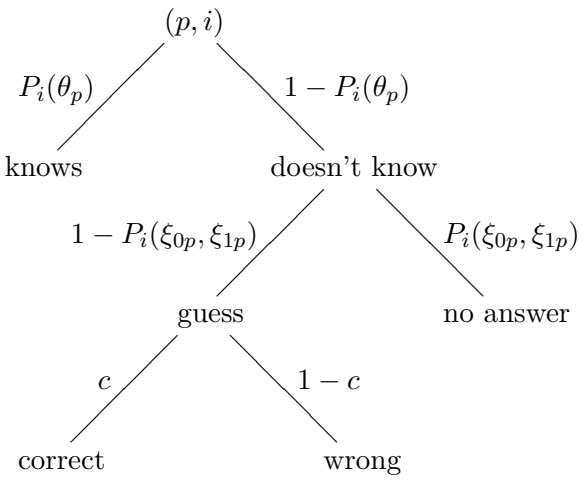


Figure 1: Decision tree representation of the test speededness model.

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:

$$\pi_{pi0} = [1 - P_i(\theta_p)]P_i(\xi_{0p}, \xi_{1p}), \tag{4}$$

$$\pi_{pi1} = [1 - P_i(\theta_p)][1 - P_i(\xi_{0p}, \xi_{1p})](1 - c), \tag{5}$$

$$\pi_{pi2} = [1 - P_i(\xi_{0p}, \xi_{1p})]c + \{1 - [1 - P_i(\xi_{0p}, \xi_{1p})]c\}P_i(\theta_p). \tag{6}$$

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

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

with $\boldsymbol{\mu}' = (0, \mu_{\xi_0}, \mu_{\xi_1})$ and Ω a positive definite covariance matrix. 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.

Some remarks apply. First, the probability of a missing value depends on unobserved information (the random effects) 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. 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 (2) and to the 1PL extended with guessing (1PLc) if $P_i(\theta_p)$ is given by (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 follows

immediately from a simple rearrangement of terms. Under the proposed model the probability of a correct answer is given by

$$\pi_{pi2} = P_i(\theta_p) + [1 - P_i(\xi_{0p}, \xi_{1p})]c[1 - P_i(\theta_p)],$$

while under the 3PL the success probability is 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

$$P_i(\xi_{0p}, \xi_{1p}|\omega_p) = \frac{\exp(\xi_{0p} + \omega_p \xi_{1p}\, i/I)}{1 + \exp(\xi_{0p} + \omega_p \xi_{1p}\, i/I)}. \tag{7}$$

Under this parametrization, the reduced model is obtained for $\omega_p = 0$, $p = 1, \ldots, P$, whereas the test speededness model results from setting $\omega_p = 1$, $p = 1, \ldots, P$, see also Goegebeur *et al.* (2005a).

3 Person fit for test speededness

3.1 Optimal person fit test

Drasgow and Levine (1986) and Levine and Drasgow (1988) used the Neyman-Pearson lemma 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 in Section 2 and denoting $\boldsymbol{Y}_p = (\boldsymbol{Y}_{p1}, \ldots, \boldsymbol{Y}_{pI})'$, the decision about the nature of test taking behavior will be based on the ratio

$$\Lambda_p = \frac{P(\boldsymbol{Y}_p = \boldsymbol{y}_p \mid \text{aberrant })}{P(\boldsymbol{Y}_p = \boldsymbol{y}_p \mid \text{normal })} \tag{8}$$

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

$$\begin{aligned} A_p(\omega_p) &= P(\boldsymbol{Y}_p = \boldsymbol{y}_p | \theta_p, \xi_{0p}, \xi_{1p}, \omega_p) \\ &= \prod_{i=1}^{I} P(\boldsymbol{Y}_{pi} = \boldsymbol{y}_{pi} | \theta_p, \xi_{0p}, \xi_{1p}, \omega_p) \\ &= \prod_{i=1}^{I} [\pi_{pi0}(\omega_p)]^{y_{pi0}} [\pi_{pi1}(\omega_p)]^{y_{pi1}} [\pi_{pi2}(\omega_p)]^{1-y_{pi0}-y_{pi1}}, \end{aligned} \tag{9}$$

where f denotes the joint density function of the random effects. In (9), $\pi_{pi0}(\omega_p)$, $\pi_{pi1}(\omega_p)$ and $\pi_{pi2}(\omega_p)$ are given by (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 $\log \Lambda_p$.

Note that the likelihood ratio statistic depends on unknown model parameters that hence need to be estimated. Moreover, application of (8) requires that both the reduced and the test speededness model are fitted to the available data. This can be done by using for instance the SAS NLMIXED procedure; example SAS code is given in the appendix. 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).

3.2 Local influence diagnostics

Global influence diagnostics are based on a case-deletion approach (Chatterjee and 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. One of the main problems with the method applied to more general settings is (1) that 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. For instance, Beckman *et al.* (1987) used local influence to assess the impact of perturbing the error variances, the random effect variances and the response vector in the linear mixed model. In the same context, Lesaffre and Verbeke (1998) illustrated that the local influence approach is also useful for the detection of influential subjects in a longitudinal analysis.

In the present context, we use local influence diagnostics to assess the impact introducing a random test speededness effect has on the key model parameter estimates. This can be done by considering (7) 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.

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

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

in which $\ell_p(\boldsymbol{\theta}|\omega_p)$ denotes the log-likelihood contribution of examinee p, i.e.

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

with

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

$A_p(\omega_p)$ is given by (9), $\boldsymbol{\omega}' = (\omega_1, \ldots, \omega_P)$, $\boldsymbol{\theta}' = (\beta_1, \ldots, \beta_I, \sigma^2_\theta, \mu_{\xi_0}, \sigma^2_{\xi_0}, \sigma_{12}, c)$ and where $\sigma_{12} = \text{Cov}(\theta, \xi_0)$. 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 $\boldsymbol{\omega}_0 \in \tilde{\Omega}$, $\ell(\boldsymbol{\theta}|\boldsymbol{\omega_0})$ corresponds to a model without test speededness effects, and this for all values of $\boldsymbol{\theta}$.

Let $\widehat{\boldsymbol{\theta}}$ be the maximum likelihood estimator for $\boldsymbol{\theta}$, obtained by maximizing $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}_0)$, and let $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ denote the maximum likelihood estimator for $\boldsymbol{\theta}$ under $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$. The local influence approach compares $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$. 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 $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ by the likelihood displacement, defined by

$$LD(\boldsymbol{\omega}) = 2[\ell(\widehat{\boldsymbol{\theta}}|\boldsymbol{\omega}_0) - \ell(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}|\boldsymbol{\omega}_0)]. \tag{10}$$

Note that the log-likelihood function of the postulated model is evaluated in both $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ and hence $LD(\boldsymbol{\omega}) \geq 0$. Note also that the likelihood displacement takes the variability of $\widehat{\boldsymbol{\theta}}$ into account. Indeed, $LD(\boldsymbol{\omega})$ will be large if $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}_0)$ is strongly curved at $\widehat{\boldsymbol{\theta}}$, which means that $\boldsymbol{\theta}$ 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 $\tilde{\Omega}$. 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 $\boldsymbol{l}$, with $\boldsymbol{l}$ a P dimensional vector of unit length. These normal curvatures can be easily calculated as

$$C_{\boldsymbol{l}} = 2|\boldsymbol{l}'\Delta'\ddot{L}^{-1}\Delta\boldsymbol{l}|, \tag{11}$$

with

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

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

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

The normal curvature (11) can be used in several ways to study the influence graph $\alpha(\boldsymbol{\omega})$, each one corresponding to a particular direction $\boldsymbol{l}$ in $\tilde{\Omega}$. One evident choice is the vector $\boldsymbol{l}_p$ 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 (11) reduces to

$$C_p = 2|\Delta_p'\ddot{L}^{-1}\Delta_p|. \tag{12}$$

Other important directions are the directions of minimal and maximal curvature, denoted $\boldsymbol{l}_{\min}$ and $\boldsymbol{l}_{\max}$, respectively, obtained as solutions to the minimization and maximization, respectively, of $C_{\boldsymbol{l}}$ over the space of all vectors of unit length. It can be shown that $C_{\boldsymbol{l}_{\min}}$ and $C_{\boldsymbol{l}_{\max}}$ correspond to the smallest and largest eigenvalues of $-2\Delta'\ddot{L}^{-1}\Delta$ and $\boldsymbol{l}_{\min}$ and $\boldsymbol{l}_{\max}$ 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 $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}_0)$, 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.* (2005a) and will not be repeated here. The authors developed a

Fortran program to compute the elements of Δ, the normal curvatures $C_{\boldsymbol{l}}$, and the direction of maximal curvature $\boldsymbol{l}_{\max}$. 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).

Note that the perturbation scheme as defined above involves, besides $\boldsymbol{\theta}$, also the parameters μ_{ξ_1}, $\sigma^2_{\xi_1}$, σ_{13} and σ_{23}, where $\sigma_{13} = \mathrm{Cov}(\theta, \xi_1)$ and $\sigma_{23} = \mathrm{Cov}(\xi_0, \xi_1)$. These additional parameters have to be fixed by the user since a null model fit only produces estimates for the components of $\boldsymbol{\theta}$. However, this more general parametrization allows to assess the impact of perturbing the null model with an extra random effect, in particular a random test speededness effect, that may be correlated with the random effects in the null model. If one is only interested in the effect of perturbing the model with a fixed, i.e. non-random, test speededness effect, one simply fixes σ_{13} and σ_{23} at zero. Doing so, the mean of ξ_{1p} appears as a common scale factor in the expressions for the elements of Δ, and hence can be safely ignored, see Goegebeur *et al.* (2005a).

So far, the discussion of local influence diagnostics was focused on the complete $\boldsymbol{\theta}$ vector. Similar principles can be applied to obtain the local influence of perturbations on subsets of $\boldsymbol{\theta}$, see Cook (1986), Verbeke *et al.* (2001) and Goegebeur *et al.* (2005a). This will not be pursued in the current paper.

4 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 normal curvatures and the likelihood ratio statistic we will use a sample of 3,000 examinees randomly drawn from this database. In Figure 2, 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. 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. The proportions of wrong answers are in the range $[0.1600, 0.7889]$ with mean 0.4745 and standard deviation 0.1589. Finally, the proportions of correct answers are between 0.1913 and 0.8380 with

mean 0.5079 and standard deviation 0.1660.

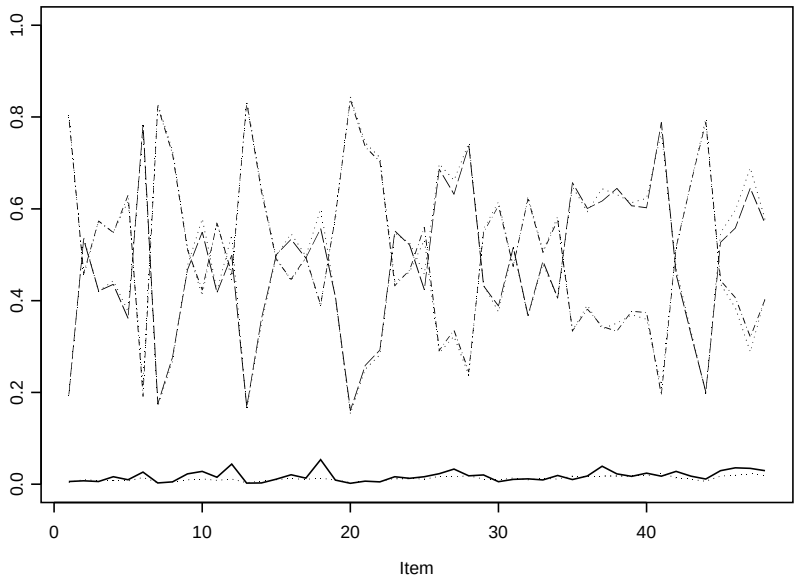


Figure 2: Proportion of missing data (solid line), wrong answers (dashed line) and correct answers (dashed-dotted line) together with the estimated theoretical proportion under the test speededness model (dotted line).

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 analyses are performed under the assumption of independent random effects. 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, there is overwhelming 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. 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. Figure 3 contains for both models the estimates of the item difficulty parameters.

To obtain an indication about the fit of the test speededness model to the SIMCE mathematics data, we show in Figure 2 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

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

and correct answers (dotted lines), given by

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

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 2, the empirical and estimated theoretical proportions agree quite well, indicating a good fit of the test speededness model.

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 (8), with unknown parameters replaced by their maximum likelihood estimates, and the normal curvature (12), for $p = 1, \ldots, 3000$. Since interest is in the extreme cases, i.e. the most significant likelihood ratio test and largest normal curvatures, we examine the 20 largest values of both. In Figure 4 we show the response profiles in the intersection of the sets of examinees with the 20 largest observations for C_p and Λ_p. This intersection contains 10 response profiles, of which all clearly

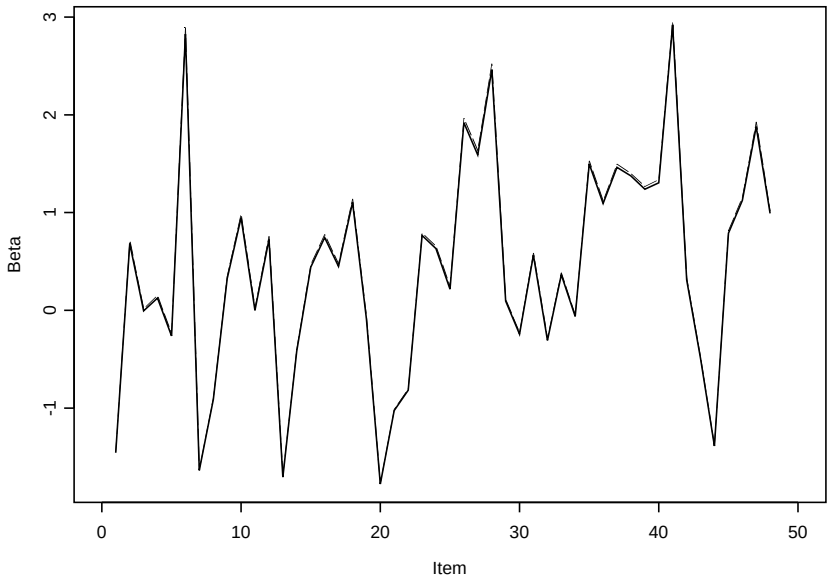


Figure 3: Estimated item difficulties under the reduced (solid line) and the test speededness model (dashed line).

contain a large number of omissions near the end of the test. Moreover, the transition from responses (correct or wrong) to omissions is quite abrupt. With the exception of one profile (examinee 1536), the same holds for the 10 non-overlapping cases of Λ_p, see Figure 5. Based on Figure 6, the remaining non-overlapping cases identified by C_p still contain a lot of omissions, especially near the end of the test, but the transition from responses to omissions is no longer always clear cut. 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 $\boldsymbol{\theta}$, i.e. have a log-likelihood contribution that is more strongly curved at $\widehat{\boldsymbol{\theta}}$, than those with less variable profiles, and hence will be more likely to be included in the set of extreme C_p measures.

To assess the correspondence between the sets of extreme cases, identified by the two procedures, we computed the proportion of overlap in the highlighted examinees for the largest k values of C_p and Λ_p, with $k = 5, \ldots, 300$, see Figure 7. As is clear from this figure, the overlap varies between 40% and 60%, indicating that the methods agree quite well in the identification of examinees affected by test speededness.

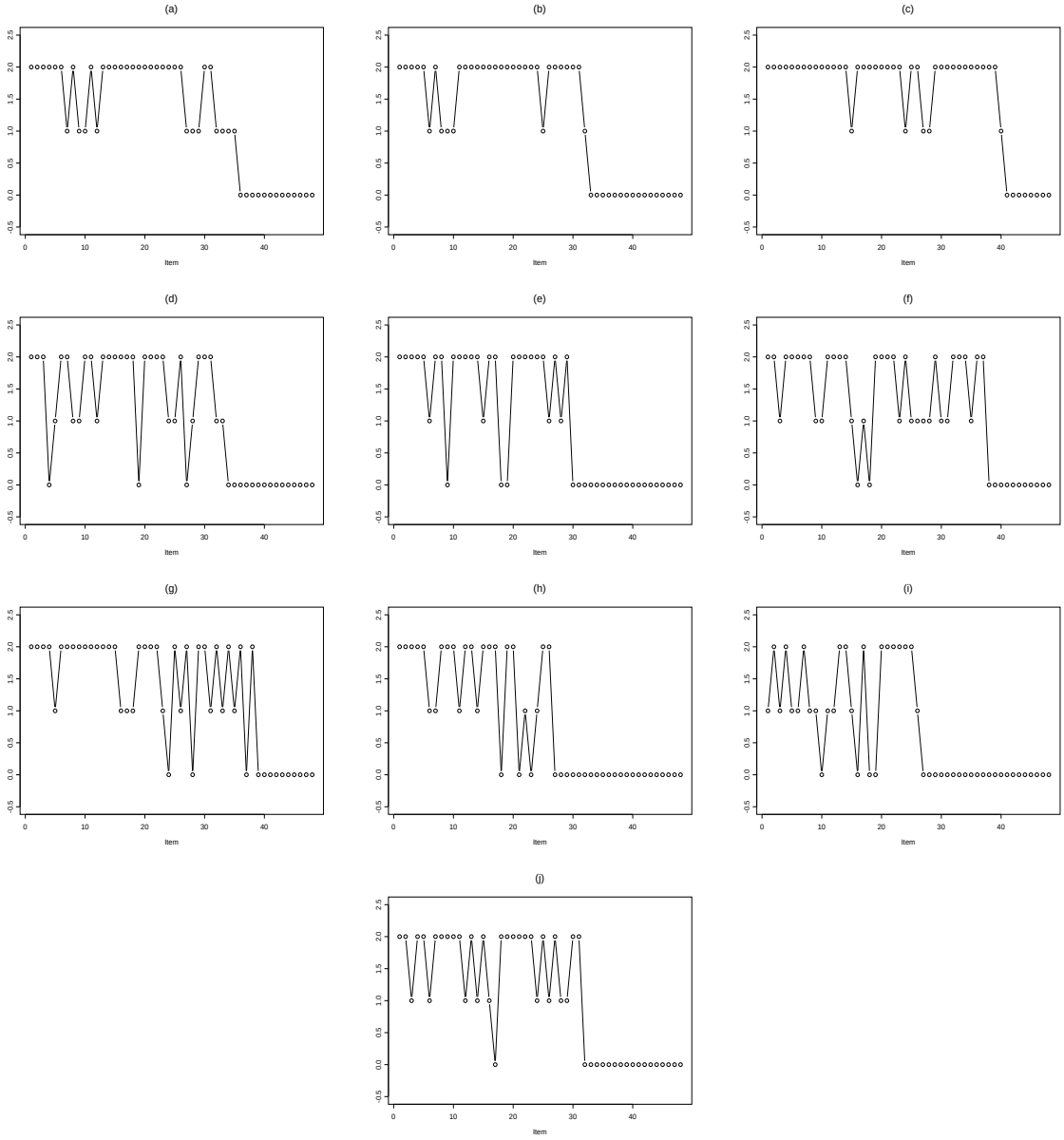


Figure 4: Intersection of the sets of examinees with the 20 largest values for C_p and Λ_p: (a) examinee 99, (b) examinee 192, (c) examinee 497, (d) examinee 827, (e) examinee 846, (f) examinee 866, (g) examinee 1637, (h) examinee 2013, (i) examinee 2377 and (j) examinee 2769.

5 Discussion and conclusion

In this paper we compared the performance of the optimal appropriateness statistic proposed by Drasgow and Levine (1986) and the local influence approach of Cook (1986) with respect to

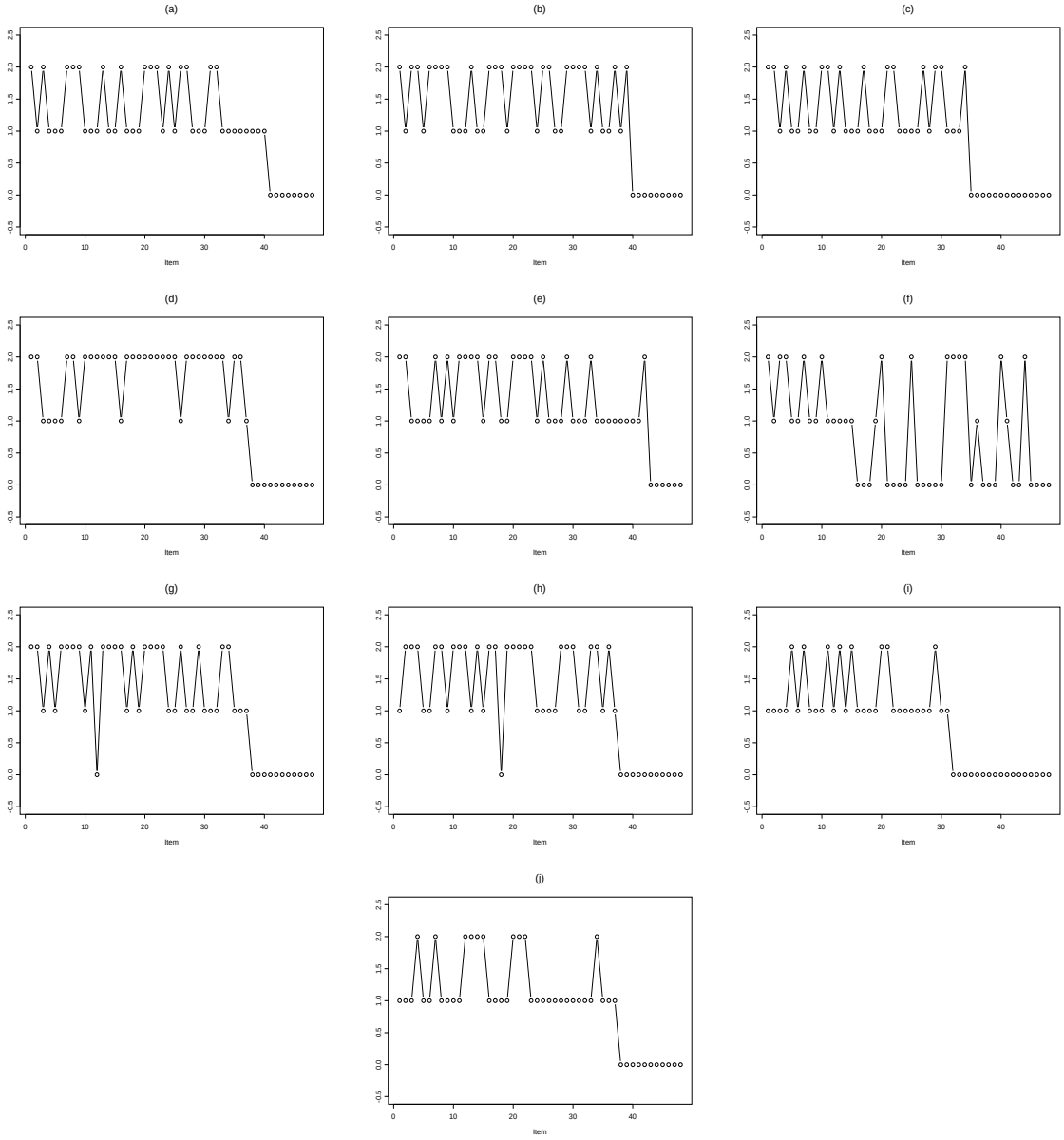


Figure 5: Likelihood ratio - non-overlapping cases: (a) examinee 48, (b) examinee 826, (c) examinee 1027, (d) examinee 1181, (e) examinee 1267, (f) examinee 1536, (g) examinee 1821, (h) examinee 1945, (i) examinee 2216 and (j) examinee 2946.

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.* (2005a). 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

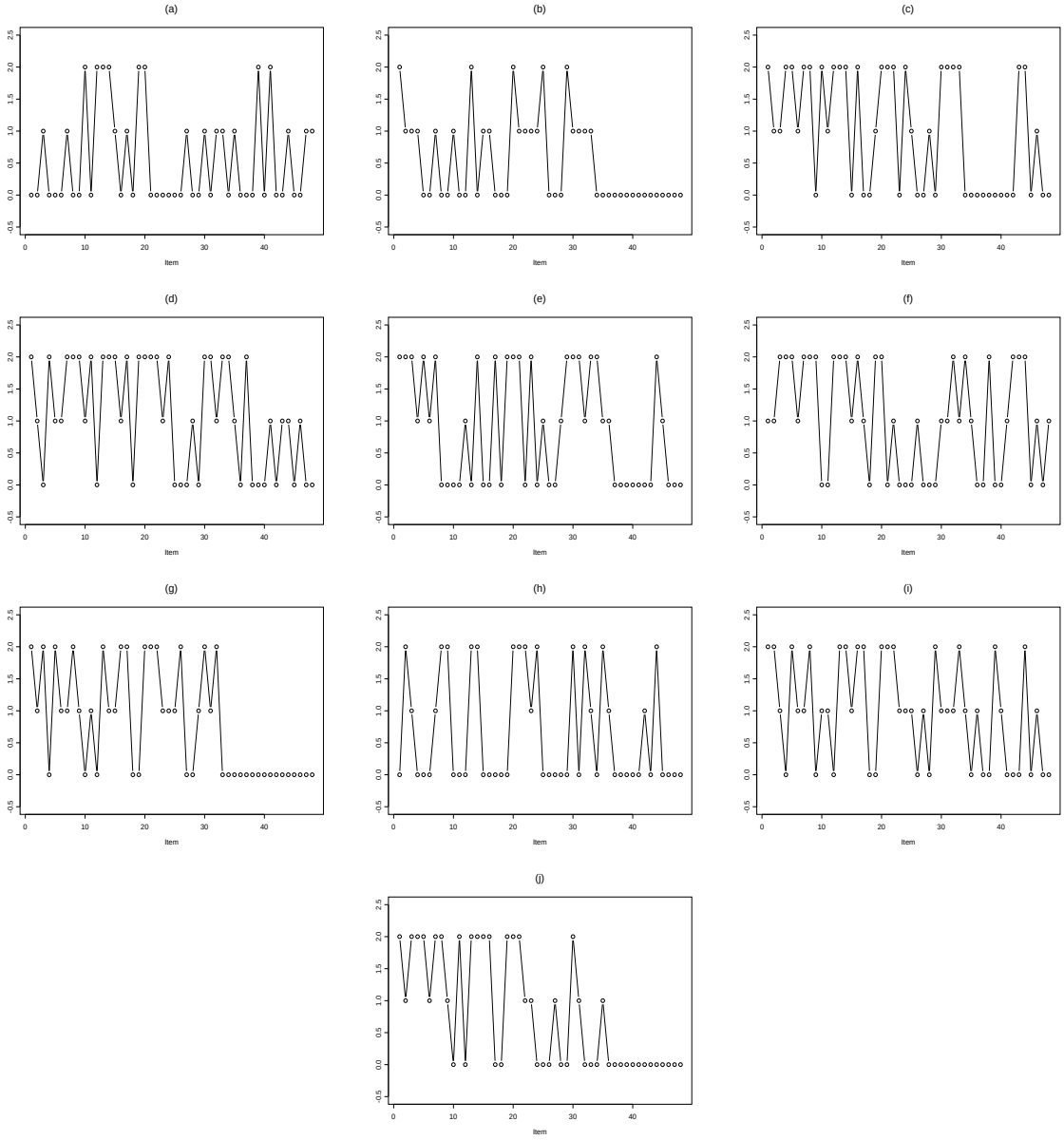


Figure 6: Normal curvatures - non-overlapping cases: (a) examinee 188, (b) examinee 193, (c) examinee 215, (d) examinee 554, (e) examinee 1133, (f) examinee 1753, (g) examinee 1767, (h) examinee 2322, (i) examinee 2330 and (j) examinee 2489.

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

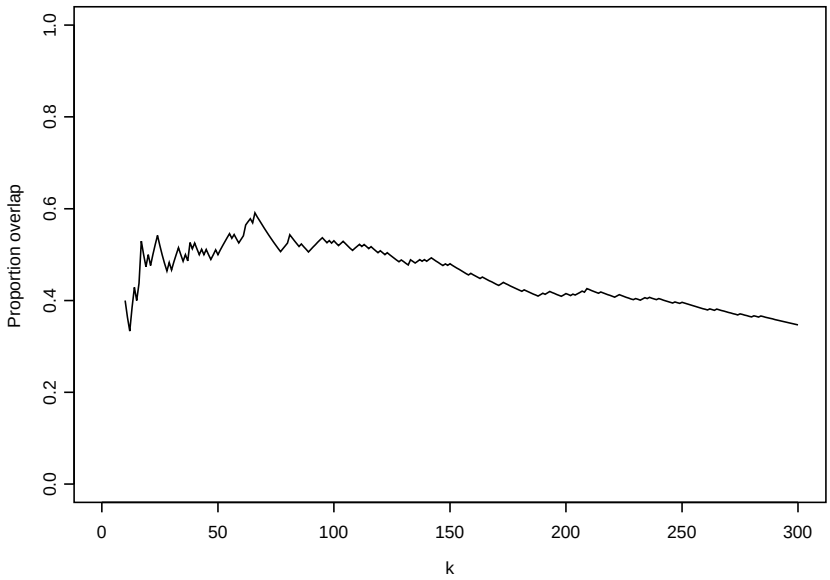


Figure 7: Proportion of overlap in the largest k values of C_p and Λ_p.

higher detection rate. On the other hand, 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 have on the parameter estimates. 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, the results obtained on the SIMCE test data showed that both 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.

The local influence approach offers a very general and flexible framework for assessing the local impact of model perturbations and the possibility to define these perturbations on an examinee basis entails virtually unlimited capacities for tackling person fit problems. In fact, every aspect of the fit of a postulated model can be scrutinized by introducing perturbation parameters on the appropriate places in this model. Doing so on an examinee basis allows to identify examinees that have a considerable impact on the key model parameter estimates when the model under consideration is slightly altered in the direction of an alternative description of test-taking behavior.

References

[1] Beckman, R.J., Nachtsheim, C.J., and Cook, R.D., 1987. Diagnostics for mixed-model analysis of variance. *Technometrics*, **29**, 413-426.

[2] Birnbaum, A., 1968. Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. and Novick, M.R. (Eds), *Statistical Theories of Mental Test Scores*, pp. 394–479. Addison-Wesly.

[3] Bolt, D.M., Cohen, A.S. and 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.

[4] Chatterjee, S. and Hadi, A.S., 1988. *Sensitivity Analysis in Linear Regression*. Wiley.

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

[6] Cook, R.D. and Weisberg, S., 1982. *Residuals and Influence in Regression*. Chapman and Hall.

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

[8] Drasgow, F. and Levine, M.V., 1986. Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement*, **10**, 59–67.

[9] Goegebeur, Y., De Boeck, P., Molenberghs, G. and del Pino, G., 2005a. A local influence based diagnostic approach to a speeded IRT model. Technical report.

[10] Goegebeur, Y., De Boeck, P., Wollack, J.A. and Cohen, A.S., 2005b. A speeded item response model with gradual process change. Technical report.

[11] Lesaffre, E. and Verbeke, G., 1998. Local influence in linear mixed models. *Biometrics*, **54**, 570–582.

[12] Levine, M.V. and Drasgow, F., 1988. Optimal appropriateness measurement. *Psychometrika*, **53**, 161-176.

[13] Meijer, R.R. and Sijtsma, K., 2001. Methodology review: evaluating person fit. *Applied Psychological Measurement*, **25**, 107-135.

[14] NAG, 1993. *NAG Fortran Library Manual - Mark 19*. The Numerical Algorithms Group Limited.

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

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

[17] Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. and Kenward, M.G., 2001. Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics*, **57**, 7-14.

[18] Wollack, J.A. and Cohen, A.S., 2005. A model for simulating speeded test data. Technical report.

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

Appendix: Example SAS code

Model without test speededness effect

```
data geg;
infile 'e:\simce\simce1.txt';
input y i person x1-x48;
run;

ods output CovMatParmEst=covm ParameterEstimates=estimates
Hessian=hessian ;

proc nlmixed data=geg method=gauss technique=newrap noad
maxiter=5000 maxfu=500000 qpoints=5 cov hess;

parms b1-b48=-1 s2=1 mxi1=0 vxi1=1 c=.25;

bounds 0 < c < .5;

b = b1*x1+b2*x2+b3*x3+b4*x4+b5*x5+b6*x6+b7*x7+b8*x8+b9*x9+b10*x10+
 b11*x11+b12*x12+b13*x13+b14*x14+b15*x15+b16*x16+b17*x17+b18*x18+b19*x19+b20*x20+
 b21*x21+b22*x22+b23*x23+b24*x24+b25*x25+b26*x26+b27*x27+b28*x28+b29*x29+b30*x30+
 b31*x31+b32*x32+b33*x33+b34*x34+b35*x35+b36*x36+b37*x37+b38*x38+b39*x39+b40*x40+
 b41*x41+b42*x42+b43*x43+b44*x44+b45*x45+b46*x46+b47*x47+b48*x48;

pi = exp(xi1)/(1+exp(xi1));

p = exp(theta-b)/(1+exp(theta-b));

if (y=0) then prob=(1-p)*pi; else if (y=1) then
prob=(1-p)*(1-pi)*(1-c); else if (y=2) then
prob=(1-pi)*c+(1-(1-pi)*c)*p;

ll=log(prob);

model y ~ general(ll);

random theta xi1 ~ normal([0,mxi1],[s2,0,vxi1]) subject=person;

run;
```

Test speededness model

```
data geg;
infile 'e:\simce\simce1.txt';
input y i person x1-x48;
run;

ods output CovMatParmEst=covm ParameterEstimates=estimates
Hessian=hessian ;

proc nlmixed data=geg method=gauss technique=newrap noad
maxiter=5000 maxfu=500000 qpoints=5 cov hess;

parms b1-b48=-1 s2=1 mxi1=0 vxi1=1 mxi2=0 vxi2=1 c=.25;

bounds 0 < c < .5;

b = b1*x1+b2*x2+b3*x3+b4*x4+b5*x5+b6*x6+b7*x7+b8*x8+b9*x9+b10*x10+
 b11*x11+b12*x12+b13*x13+b14*x14+b15*x15+b16*x16+b17*x17+b18*x18+b19*x19+b20*x20+
 b21*x21+b22*x22+b23*x23+b24*x24+b25*x25+b26*x26+b27*x27+b28*x28+b29*x29+b30*x30+
 b31*x31+b32*x32+b33*x33+b34*x34+b35*x35+b36*x36+b37*x37+b38*x38+b39*x39+b40*x40+
 b41*x41+b42*x42+b43*x43+b44*x44+b45*x45+b46*x46+b47*x47+b48*x48;

xi2s=exp(xi2);
pi = exp(xi1+xi2s*i/48)/(1+exp(xi1+xi2s*i/48));

p = exp(theta-b)/(1+exp(theta-b));

if (y=0) then prob=(1-p)*pi; else if (y=1) then
prob=(1-p)*(1-pi)*(1-c); else if (y=2) then
prob=(1-pi)*c+(1-(1-pi)*c)*p;

ll=log(prob);

model y ~ general(ll);

random theta xi1 xi2 ~ normal([0,mxi1,mxi2],[s2,0,vxi1,0,0,vxi2])
subject=person;

run;
```