Tilburg University

On the usefulness of a multilevel logistic regression approach to person-fit analysis

Conijn, J.M.; Emons, W.H.M.; van Assen, M.A.L.M.; Sijtsma, K.

Published in:
Multivariate Behavioral Research

DOI:
[10.1080/00273171.2010.546733](https://doi.org/10.1080/00273171.2010.546733)

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
Conijn, J. M., Emons, W. H. M., van Assen, M. A. L. M., & Sijtsma, K. (2011). On the usefulness of a multilevel logistic regression approach to person-fit analysis. *Multivariate Behavioral Research*, 46(2), 365-388.
<https://doi.org/10.1080/00273171.2010.546733>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

This article was downloaded by: [Conijn, Judith M.]

On: 20 April 2011

Access details: Access Details: [subscription number 936641722]

Publisher Psychology Press

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Multivariate Behavioral Research

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t775653673>

On the Usefulness of a Multilevel Logistic Regression Approach to Person-Fit Analysis

Judith M. Conijn^a; Wilco H. M. Emons^a; Marcel A. L. M. van Assen^a; Klaas Sijtsma^a

^a Tilburg University,

Online publication date: 19 April 2011

To cite this Article Conijn, Judith M. , Emons, Wilco H. M. , van Assen, Marcel A. L. M. and Sijtsma, Klaas(2011) 'On the Usefulness of a Multilevel Logistic Regression Approach to Person-Fit Analysis', Multivariate Behavioral Research, 46: 2, 365 — 388

To link to this Article: DOI: 10.1080/00273171.2010.546733

URL: <http://dx.doi.org/10.1080/00273171.2010.546733>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

On the Usefulness of a Multilevel Logistic Regression Approach to Person-Fit Analysis

Judith M. Conijn, Wilco H. M. Emons,
Marcel A. L. M. van Assen, and Klaas Sijtsma
Tilburg University

The logistic person response function (PRF) models the probability of a correct response as a function of the item locations. Reise (2000) proposed to use the slope parameter of the logistic PRF as a person-fit measure. He reformulated the logistic PRF model as a multilevel logistic regression model and estimated the PRF parameters from this multilevel framework. An advantage of the multilevel framework is that it allows relating person fit to explanatory variables for person misfit/fit. We critically discuss Reise's approach. First, we argue that often the interpretation of the PRF slope as an indicator of person misfit is incorrect. Second, we show that the multilevel logistic regression model and the logistic PRF model are incompatible, resulting in a multilevel person-fit framework, which grossly violates the bivariate normality assumption for residuals in the multilevel model. Third, we use a Monte Carlo study to show that in the multilevel logistic regression framework estimates of distribution parameters of PRF intercepts and slopes are biased. Finally, we discuss the implications of these results and suggest an alternative multilevel regression approach to explanatory person-fit analysis. We illustrate the alternative approach using empirical data on repeated anxiety measurements of cardiac arrhythmia patients who had a cardioverter-defibrillator implanted.

Reise (2000) proposed a multilevel logistic regression (MLR) approach to the assessment of person fit in the context of the 1- and 2-parameter logistic item

Correspondence concerning this article should be addressed to Judith M. Conijn, Department of Methodology and Statistics FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: j.conijn@uvt.nl

response theory (IRT) models for dichotomous item scores. Henceforth, we call this approach multilevel person-fit analysis (PFA). Whereas traditional methods for PFA (Karabatsos, 2003; Meijer & Sijtsma, 1995, 2001) provide little more than a yes/no decision rule for whether test performance is aberrant, Reise's proposal offers great potential for explaining person misfit by including explanatory variables in the statistical analysis. Several studies provide real-data examples of this potential (Wang, Reise, Pan, & Austin, 2004; Woods, 2008). For example, multilevel PFA was used to study faking on personality scales (LaHuis & Copeland, 2009) and to explain aberrant responding of military recruits to personality scales (Woods, Oltmanns, & Turkheimer, 2008).

What none of these studies have questioned is whether the combination of MLR and a logistic IRT model for the person-response probability as a function of item location, here denoted person response function (PRF; Sijtsma & Meijer, 2001), is compatible and produces correct statistical information for PFA. Our study demonstrates that the combination is incompatible, assesses the degree of bias the inconsistency causes in the multilevel-model parameter estimates used for person-fit assessment, and discusses the consequences for the viability of MLR for PFA.

PFA studies the fit of IRT models to individual examinees' item-score vectors of 0s (e.g., for incorrect answers) and 1s (for correct answers) on the J items from the test of interest. The 1- and 2-parameter logistic models (1PLM, 2PLM; Hambleton & Swaminathan, 1985) assume that one underlying ability or trait affects an examinee's responses to the items. However, for some examinees unwanted attributes may affect the responses. For example, in ability testing test anxiety, incorrect learning strategy, answer copying, and guessing may affect responses in addition to an examinee's ability level. In personality assessment response styles, faking, and untraitedness (Reise & Waller, 1993; Tellegen, 1988) may produce item scores different from what was expected from the trait level alone. Aberrant responding produces item-score vectors that are inconsistent with the IRT model and likely results in invalid latent-variable estimates (Meijer & Nering, 1997). Identification of such item-score vectors is imperative so as to prevent drawing the wrong conclusions about examinees.

PFA based on the 1PLM or the 2PLM identifies item-score vectors, which are either consistent or inconsistent with these models. Inconsistent vectors contain unusually many 0s where the IRT model predicts more 1s and 1s where more 0s are expected. A limitation of traditional PFA is that it only identifies fitting and misfitting item-score vectors but leaves the researcher speculating about the causes of the misfit. Multilevel PFA attempts to move PFA from only signaling person misfit to also understanding its causes by introducing an explanatory model of the misfit. It uses the PRF for this purpose (Emons, Sijtsma, & Meijer, 2004, 2005; Lumsden, 1977, 1978; Nering & Meijer, 1998; Sijtsma & Meijer,

2001; Trabin & Weiss, 1983). For dichotomously scored items, the PRF provides the relationship between an examinee's probability of having a 1 score on an item as a function of the item's location. Lumsden (1978), Ferrando (2004, 2007), and Emons et al. (2005) noticed that the PRF based on the 1PLM decreases. Emons et al. (2005) argued that a PRF that increases locally indicates misfit to the 1PLM and that the location of the increase in the PRF on the latent scale and also the shape of the PRF provide diagnostic information about misfit. For example, for average-ability examinees low probabilities of correct responses on the first and easiest items might signal test anxiety, and for low-ability examinees high probabilities of correct responses on the most difficult items might signal cheating.

Reise's (2000) multilevel PFA is based on logistic PRFs to assess person fit in the context of the 1PLM and the 2PLM. Multilevel PFA focuses on the PRF slope, which is taken as a person-fit measure quantifying the degree to which examinees are sensitive to differences in item locations. The MLR framework allows modeling variation in PRF slopes using explanatory variables such as verbal skills, motivation, anxiety, and gender. This renders multilevel PFA useful for explaining person misfit and investigating group differences in person fit.

Multilevel PFA is valuable and original but also evokes the question whether the multilevel model and the logistic PRF model are compatible. Hence, we submitted multilevel PFA to a thorough logical analysis and a Monte Carlo simulation study. First, we discuss the PRF definition used in multilevel PFA. Second, we explain multilevel PFA. Third, unlike Reise (2000) and Woods (2008), we argue that the interpretation of the PRF slope as a person-fit measure is only valid for the 1PLM but invalid for the 2PLM. Fourth, we show that the PRF model under the 1PLM is not compatible with the MLR framework from which the PRF parameters are estimated. Fifth, the results of a Monte Carlo study show the effect of the model mismatch on the bias in the estimates of distribution parameters of PRF intercepts and slopes. Sixth, we discuss our findings and their consequences for multilevel PFA. Seventh, we suggest an alternative multilevel approach to explanatory PFA. We illustrate the alternative approach using empirical data on repeated anxiety measurements of cardiac arrhythmia patients who had a cardioverter-defibrillator implanted. Finally, we discuss the viability of multilevel PFA and our proposed alternative approach to explanatory PFA.

THEORY OF MULTILEVEL PERSON-FIT ANALYSIS

Person Response Function

Let θ denote the latent variable and $P_j(\theta)$ the conditional probability of a 1

FIGURE 1 Two item response functions under the 1-parameter logistic model (solid curves) and 2-parameter logistic model (dashed curves).

Note. $\delta_1 = -1, \alpha_1 = 1; \delta_2 = 0, \alpha_2 = 1; \delta_3 = 1, \alpha_3 = 0.5; \delta_4 = 1.5, \alpha_4 = 1.5$.

score on item j ($j = 1, \dots, J$; we also use k as item index), also known as the item response function (IRF). Let δ_j be the location or difficulty parameter of item j and α_j the slope or discrimination parameter. The IRF of the 2PLM for item j is defined as

$$P_j(\theta) = \frac{\exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]}. \quad (1)$$

The 1PLM is obtained by setting $\alpha_j = \alpha = 1$. Figure 1 shows two IRFs for the 1PLM (solid curves) and two IRFs for the 2PLM (dashed curves).

The PRF reverses the roles of examinees and items. For examinee v (we also use u and w as examinee indices), the PRF provides the relationship between the probability of a 1 score and the item location, δ . Reise (2000) and Ferrando (2004, 2007) defined a logistic PRF, which introduces a person parameter α_v in addition to θ_v . Parameter α_v quantifies the slope of the PRF for examinee v . Latent variable value θ_v is the location of the PRF of examinee v for which $P_v(\delta) = .5$. This PRF is defined as follows (Ferrando, 2004, 2007; Reise, 2000, p. 55):

$$P_v(\delta) = \frac{\exp[\alpha_v(\delta - \theta_v)]}{1 + \exp[\alpha_v(\delta - \theta_v)]}. \quad (2)$$

FIGURE 2 Two person response functions (PRFs).
Note. Dashed PRF: $\alpha_v = -2$, $\theta_v = 0$; solid PRF: $\alpha_w = -0.2$, $\theta_w = 1$.

Figure 2 shows a steep decreasing PRF for examinee v (dashed curve) of which the large negative slope parameter ($\alpha_v = -2$) indicates a strong relation between item location and correct-response probability. Figure 2 also shows a nearly flat PRF for examinee w (solid curve) of which the small negative slope parameter ($\alpha_w = -0.2$) indicates a weak relation. Large negative slopes indicate high person reliability (Lumsden, 1977, 1978), low individual trait variability (Ferrando, 2004, 2007), and good person fit (Reise, 2000).

Multilevel PFA rephrases Equation 2 as a 2-level logistic regression model and estimates the PRF parameters from the latter model. This is innovative relative to existing methods. For example, Ferrando (2004, 2007) developed a PRF model based on Lumsden's Thurstonian model (1977), and Strandmark and Linn (1987) formulated the PRF as a generalized logistic response model. In the context of nonparametric IRT, Sijtsma and Meijer (2001) and Emons et al. (2004, 2005) estimated PRFs using nonparametric regression methods such as binning and kernel smoothing, and for parametric IRT, Trabin and Weiss (1983) and Nering and Meijer (1998) used binning to estimate the PRF.

Multilevel Approach to Person-Fit Analysis

This section discusses multilevel PFA as proposed and explained by Reise (2000). In the 2-level logistic regression model, the item scores are the Level 1 units, which are nested in the examinees, who are the Level 2 units. Following Reise, we rewrite Equation 2 as a logit and then reparameterize the logit by means of $b_{0v} = -\alpha_v\theta_v$ and $b_{1v} = \alpha_v$, so that Level 1 of the multilevel

PFA model equals

$$\text{logit}[P_v(\delta)] = \log \left[\frac{P_v(\delta)}{1 - P_v(\delta)} \right] = -\alpha_v \theta_v + \alpha_v \delta = b_{0v} + b_{1v} \delta. \quad (3)$$

Intercept b_{0v} and slope b_{1v} are random effects across examinees and are modeled at the second level. Reise treats intercept b_{0v} as an analog to θ_v . After having accounted for variation in θ_v , remaining variation in intercepts b_{0v} is a sign of multidimensionality in the item scores. Reise interprets slope b_{1v} as a person-fit measure. Hence, variation in slopes indicates differences in person fit.

Reise (2000, pp. 558–562) distinguishes three steps in multilevel PFA. These steps are preceded by the estimation of the item locations δ_j and the latent variable values θ_v from either the 2PLM or the 1PLM.

Step 1 estimates the PRF in Equation 3. For this purpose, the item location estimates, $\hat{\delta}_j$, are used. In the Level 2 model, the Level 1 intercept b_{0v} is split into an average intercept γ_{00} and a random intercept effect u_{0v} and the slope b_{1v} into an average slope γ_{10} and a random slope effect u_{1v} , so that

$$\begin{aligned} b_{0v} &= \gamma_{00} + u_{0v}, \\ b_{1v} &= \gamma_{10} + u_{1v}. \end{aligned} \quad (4)$$

Step 2 explains the variance of the estimated intercept b_{0v} , which is denoted $\tau_{00} = \text{Var}(b_0) = \text{Var}(u_0)$. For this purpose, the estimated latent variable, $\hat{\theta}$, is used as an explanatory variable of intercept b_0 , so that the Level 2 model equals

$$\begin{aligned} b_{0v} &= \gamma_{00} + \gamma_{01} \hat{\theta}_v + u_{0v}, \\ b_{1v} &= \gamma_{10} + u_{1v}. \end{aligned} \quad (5)$$

Reise (2000) claims that under a fitting IRT model, variation in $\hat{\theta}$ explains all intercept variance, so that $\hat{\tau}_{00}$ is not significantly greater than 0.

Step 3 estimates the variance in the slopes, denoted $\tau_{11} = \text{Var}(b_1) = \text{Var}(u_1)$. For this purpose, the Level 1 intercepts are fixed given $\hat{\theta}_v$, meaning that $\tau_{00} = 0$, and the Level 1 slopes, b_{1v} , are assumed random, so that

$$\begin{aligned} b_{0v} &= \gamma_{00} + \gamma_{01} \hat{\theta}_v, \\ b_{1v} &= \gamma_{10} + u_{1v}. \end{aligned} \quad (6)$$

Significant slope variance, $\hat{\tau}_{11}$, indicates systematic differences in person fit, and the Empirical Bayes (EB) estimates, $\hat{b}_{1v}$, are used as individual person-fit

measures. Larger negative values of $\hat{b}_{1v}$ reflect greater sensitivity to item location and are interpreted as a sign of person fit, whereas smaller negative values and positive values of $\hat{b}_{1v}$ are interpreted as signs of person misfit. One may include explanatory variables in the Level 2 model for the slope to explain variation in person fit. Reise (2000) discussed the multilevel PFA approach only for the 1PLM but also claimed applicability to the 2PLM.

We return to Step 2 and notice that significant intercept variance provides evidence of multidimensionality in the form of either violation of local independence (or unidimensionality) or differential test functioning (Reise, 2000, pp. 560–561). Following Reise, LaHuis and Copeland (2009) suggest including exploratory variables in the intercept model to study causes of this model misfit.

EVALUATION OF MULTILEVEL PERSON-FIT ANALYSIS

We identify two problems with respect to multilevel PFA. First, the interpretation of the PRF slopes α_v in Equation 2 and b_{1v} in Equation 3 as person-fit measures is only valid under restrictive assumptions for the items. Second, the PRF model (Equation 2) and the multilevel PFA models (Equations 3 through 6) used to estimate the PRF are incompatible. Next, we discuss these problems and their implications for multilevel PFA.

Problem 1: Interpretation of the Variance in PRF Slope Parameters in PFA

Multilevel PFA posits that when either the 1PLM or the 2PLM is the true model, all examinees have the same negative PRF slope parameter (Reise, 2000, pp. 560, 563, spoke of nonsignificant variation in person slopes). However, Sijtsma and Meijer (2001; see also Emons et al., 2005) showed that PRFs are only monotone nonincreasing if the IRFs of the items in the test do not intersect anywhere along the θ scale. In the 2PLM, IRFs intersect by definition if item discrimination varies over items, and PRFs are not decreasing functions but show many local increases. Hence, PRF slope parameters do not have a clear-cut definition, and we therefore ask whether Reise's position concerning variation in PRF slopes is correct. First we discuss this question for the 1PLM and then for the 2PLM.

Based on the IRF defined in Equation 1, we write the difference of the logits for examinee v and arbitrary items j and k as

$$\text{logit}[P_k(\theta_v)] - \text{logit}[P_j(\theta_v)] = \theta_v(\alpha_k - \alpha_j) - \alpha_k\delta_k + \alpha_j\delta_j. \quad (7)$$

For the 1PLM, by definition $\alpha_j = \alpha_k = \alpha$, so that Equation 7 reduces to $\alpha(\delta_j - \delta_k)$. Hence, the difference depends on item parameters α , δ_j , and δ_k

but not on θ_v . Furthermore, for arbitrary item locations such that $\delta_j < \delta_k$ the difference is negative; hence the PRF decreases. Thus, under the 1PLM the PRF slope parameters are equal and negative. Figure 3a shows two 1PLM IRFs ($\alpha = 1$) and the response probabilities for examinees u , v , and w expressed as probabilities, and Figure 3b shows the logits. Figure 3c shows the corresponding parallel decreasing PRFs for examinees u , v , and w expressed as logits (PRF-slope parameters are $\alpha_u = \alpha_v = \alpha_w = -1$).

If a sample also includes examinees for whom the 1PLM is the incorrect model, observed variance in PRF slope parameters by definition means variation in person fit, and nonnegative PRF slope parameters definitely indicate person misfit. This interpretation of variance in PRF slopes α_v is identical to the interpretation under multilevel PFA. This means that under the 1PLM observed variance in PRF slopes can be validly interpreted as variation in person fit across examinees.

Under the 2PLM, Equation 7 clarifies that, if $\alpha_j \neq \alpha_k$, the difference in logits for two items also depends on an examinee's θ_v value; hence, differences in θ cause differences in PRF slopes. Moreover, the difference in logits is not always negative for $\delta_j < \delta_k$. For instance, if $\theta_v = 0$ then the difference is positive for those items j and k for which $\frac{\alpha_j}{\alpha_k} \delta_j > \delta_k$; hence, for examinee v the PRF slope does not decrease everywhere.

Figure 3d shows two 2PLM IRFs and the response probabilities for examinees u , v , and w expressed as probabilities, and Figure 3e shows the logits. Figure 3f shows the corresponding PRFs for examinees u , v , and w expressed as logits. For IRF slopes $\alpha_j = 2$ and $\alpha_k = 0.5$, the two IRFs intersect. Consequently, the resulting PRFs have different slopes, and the PRF for examinee u even increases. This result illustrates that under the 2PLM, PRF slopes vary and PRFs do not necessarily decrease monotonically and may even increase monotonically. In Figure 3f, the large variation in PRF slopes is due to the large difference between IRF slopes α_j and α_k given the difference between IRF locations δ_j and δ_k (Figure 3d and 3e) but smaller IRF-slope differences also lead to variation in PRF slopes. Sijtsma and Meijer (2001) and Emons et al. (2005) discuss similar results. Thus, under the 2PLM, the PRF slopes are expected to show variation also in the absence of person misfit.

To conclude, under the multilevel PFA model variation in person slopes provides valid information about person fit only if the items vary in difficulty but not in discrimination power (i.e., the items satisfy the 1PLM). If items also vary in their discrimination power (i.e., items satisfy the 2PLM), PRF slopes will vary even in the absence of person misfit. Hence, in real data, for which the 1PLM is often too restrictive and more flexible IRT models such as the 2PLM are appropriate, relating person fit to PRF slopes may lead to overestimation of individual differences in person fit and increases the risk of incorrectly identifying an examinee as misfitting or fitting.

FIGURE 3 Item response functions (IRFs) and corresponding person response functions (PRFs) under the 1-parameter logistic model (upper panels) and the 2-parameter logistic model (lower panels).
Note. $\delta_j = -1$, $\delta_k = 1$; $\theta_u = -2.5$, $\theta_v = 0$, $\theta_w = 2.5$. Upper panels: item slopes $\alpha_j = \alpha_k = 1$, PRF slopes equal to -1 . Lower panels: item slopes $\alpha_j = 2$, $\alpha_k = 0.5$. PRF slopes equal to 0.6 , -1.3 , and -3.1 , for examinees u , v , and w , respectively.

Problem 2: Incompatibility Between the PRF Model and the Multilevel PFA Model

We assume that the 1PLM holds (i.e., items only differ in difficulty) in the population of interest but that the fit of individual examinees varies randomly, which is reflected by positive PRF-slope variance. Under this assumption, slope variance only reflects random variation in person fit and does not result from differences in item discrimination. For multilevel PFA (Equations 3 through 6), we discuss whether under these conditions the MLR formulation of the logistic PRF model leads to correct estimates of the means and the variances of the slopes and the intercepts in the PRF model. If estimates are biased, analyzing PRF slope variance based on multilevel PFA would be misleading with respect to the true variation in person fit.

The MLR Level 1 intercept and slope parameters (Equation 3) and the PRF examinee parameters (Equation 2) are related by $b_{0v} = -\alpha_v\theta_v$ and $b_{1v} = \alpha_v$. For the multilevel PFA model, in the intercept $b_{0v} = \gamma_{00} + \gamma_{01}\theta_v + u_{0v}$ (Equation 5) the effect γ_{01} of θ_v is fixed across examinees. For the PRF model, in the intercept $b_{0v} = -\alpha_v\theta_v$ (Equation 3) the effect $-\alpha_v$ of θ_v is variable. Hence, the models do not match. This mismatch has the following consequences:

In multilevel models, the Level 2 random effects, u_{0v} and u_{1v} , are assumed to be bivariate normal (Raudenbush & Bryk, 2002, p. 255; Snijders & Bosker, 1999, p. 121). It may be noted that, from $b_{0v} = -\alpha_v\theta_v$ and $b_{1v} = \alpha_v$, it follows that $b_{0v} = -b_{1v}\theta_v$. Thus, intercept b_{0v} depends on slope b_{1v} , and in subgroups having the same slope value (i.e., $b_{1v} = b_1$) intercept variance across examinees is smaller the closer the slope value is to 0 (from $\sigma_{b_0|b_1}^2 = b_1^2\sigma_\theta^2$). This dependence implies a violation of bivariate normality of u_{0v} and u_{1v} . The next example illustrates this violation.

We consider that a PRF model in which $\alpha \sim N(-2, 1)$ and $\theta \sim N(0, 1)$ generated the data. Figure 4a shows the resulting bivariate distribution of u_{0v} and u_{1v} for the Level 2 model without θ_v (Equation 4). We computed u_{0v} based on $b_{0v} = -\alpha_v\theta_v$ and u_{1v} based on $b_{1v} = \alpha_v$ (the note below Figure 4 provides computational details). Random effect u_{0v} is the person-specific intercept deviation from the mean b_{0v} (i.e., the mean of $-\alpha_v\theta_v$, which equals γ_{00} ; see Equation 4), and u_{1v} is the person-specific slope deviation from the mean b_{1v} (i.e., the mean of α_v , which equals γ_{10} ; see Equation 4). It follows that the u_{0v} values on the ordinate in Figure 4a equal the corresponding b_{0v} values (because $\gamma_{00} = 0$ if $\mu_\theta = 0$). The u_{1v} values on the abscissa correspond to b_{1v} values between -6 and 2 (because $\gamma_{10} = \mu_\alpha = -2$).

Figure 4a shows that bivariate normality is violated in the multilevel PFA model defined by Equations 3 and 4. The figure shows smaller variation in u_{0v} for large positive u_{1v} (corresponding to near-0 b_{1v}) than for large negative u_{1v} (corresponding to large negative b_{1v}). Thus, poorly fitting examinees who

FIGURE 4 Bivariate distribution of random slope effect (u_{1v}) and random intercept effect (u_{0v}) for multilevel person-fit analysis model excluding θ_v (Panel a) and including θ_v (Panel b). *Note.* $\theta \sim N(0, 1)$ and $\alpha \sim N(-2, 1)$; $u_{1v} = \alpha_v - \text{MEAN}(\alpha_v)$. In Panel a, u_{0v} is computed for Equation 4: $u_{0v} = -\alpha_v \times \theta_v - \text{MEAN}(-\alpha_v \times \theta_v)$, and in Panel b, u_{0v} is computed for Equation 5: $u_{0v} = -\alpha_v \times \theta_v - [\text{MEAN}(-\alpha_v) \times \theta_v]$.

have near-0 PRF slopes (i.e., large positive random slope effects) have smaller intercept variation than well-fitting examinees who have steep negative PRF slopes (i.e., large negative random slope effects). The explanation is that differences in θ are ineffective when examinees respond randomly (reflected by flat PRFs) but effective when examinees respond according to the 1PLM (reflected by decreasing PRFs) because then differences in θ determine differences in response probabilities. Figure 4b shows that when θ is included in the multilevel

PFA model to explain intercept variance (Equation 5), the joint distribution of u_{0v} and u_{1v} again is not bivariate normal. The examples in Figure 4 show that one consequence of using the MLR framework for estimating the distribution of PRF parameters is that estimates are based on assumptions that are unreasonable when data satisfy the logistic PRF model (Equation 3).

The mismatch of the multilevel PFA model and the PRF model also affects the usefulness of Reise's (2000) three-step procedure. In Step 2, residual intercept variance is taken as a sign of multidimensionality. However, because the effect $-\alpha_v$ of θ_v on the intercept b_{0v} (i.e., $b_{0v} = -\alpha_v\theta_v$) is perfectly negatively related to the PRF slope (α_v), this effect differs across examinees when there is variation in PRF slopes. As a result, if the PRF slope varies θ_v cannot be expected to explain all variation in the intercepts and, therefore, residual intercept variance in the multilevel PFA model does not necessarily represent multidimensionality. This is illustrated by Figure 4b in which the ordinate values show variability in u_{0v} after having accounted for differences in θ_v . If u_{1v} equals 0, the standard deviation of u_{0v} equals 0. The standard deviation appears to increase linearly in $|u_{1v}|$. This shows that if PRF slopes vary, residual intercept variance is larger than 0. This result has consequences for the usefulness of Step 3 in multilevel PFA. In Step 3, PRF slope variation is studied restricting the residual intercept variance to 0. However, residual intercept variance is only 0 if slope variance is 0 (i.e., all u_{1v} s equal 0), rendering Step 3 useless. Thus, only Step 1 and Step 2 are meaningful.

To conclude, the multilevel PFA model is incompatible with the PRF model even if the items satisfy the 1PLM. The mismatch refutes the interpretation of positive intercept variance as an unambiguous sign of multidimensionality because in multilevel PFA slope variance necessarily implies intercept variance. Apart from whether multilevel PFA model parameters can be interpreted meaningfully in each situation, the mismatch also questions the validity of the parameter estimates under the multilevel PFA model. We showed that the multilevel model does not adequately capture the bivariate distribution of residuals (u_{0v} and u_{1v}) to be expected if data comply with the PRF model. So the more problematic consequence of the mismatch is that the multilevel model may produce biased estimates of means and variances of PRF slopes and intercepts, as we demonstrate next.

MONTE CARLO STUDY: BIAS DUE TO MODEL MISMATCH

We conducted a Monte Carlo study to examine whether estimates of multilevel PFA model parameters γ_{00} , γ_{01} , γ_{10} , and τ_{11} (Equation 5; Step 2 in Reise's [2000] three-step procedure) are biased due to the mismatch between the multilevel PFA

model and the PRF model and the resulting violation of bivariate normality of Level 2 random effects. We focused primarily on slope variance τ_{11} , which is most relevant for explaining and detecting person misfit.

We compared bias in the absence of model mismatch with bias in the presence of mismatch. Mismatch of the multilevel PFA model with the PRF model is absent if in the latter the effect of θ_v is equal across examinees. We call this version of the PRF model the Compatible PRF (C-PRF) model. Let μ_α denote the fixed effect of θ_v . The C-PRF model is defined as

$$P_v(\delta) = \frac{\exp(\alpha_v \delta - \mu_\alpha \theta_v)}{1 + \exp(\alpha_v \delta - \mu_\alpha \theta_v)}. \quad (8)$$

If the C-PRF model underlies the data and we find bias in the multilevel PFA model estimates, this bias is inherent in MLR. However, if the PRF model generated the data, bias is caused by both MLR and model mismatch. Thus, if model mismatch also causes bias, we expect bias to be larger under the PRF model than the C-PRF model.

Method

We simulated data consistent with the C-PRF model (Equation 8) and the PRF model (Equation 2). Item and person parameters were estimated under the 1PLM. Bias in multilevel PFA was studied under four conditions. In conditions C-PRF true and PRF true, we used the parameter values of δ and θ to estimate the multilevel PFA model. In conditions C-PRF est and PRF est, we used the parameter estimates $\hat{\delta}$ and $\hat{\theta}$ to examine the bias found in practical data analysis where the true parameter values are unknown and substituted by their sample estimates.

Parameters used to generate the data were distributed as $\alpha \sim N(\mu_\alpha, \sigma_\alpha^2)$ and $\theta \sim N(\mu_\theta, \sigma_\theta^2)$ and, following Reise (2000), the item location was an equidistant sequence from $\delta \sim U(-2, 2)$, with increments of 0.08. In the “true” conditions we assessed bias of estimates of the C-PRF model and the PRF model using $2 \times 4 \times 2 \times 2$ combinations of μ_α (valued 1, 2), σ_α^2 (0, 0.1, 0.5, 1), μ_θ (0, 1), and σ_θ^2 (0.2, 1). The C-PRF model and the PRF model coincide in the eight combinations with $\sigma_\alpha^2 = 0$; that is, for both models the effect of θ_v equals μ_α for all testees. The values for μ_α and σ_α^2 are based on empirical multilevel PFA results by Woods (2008) and Woods et al. (2008), who used multilevel PFA to analyze empirical data. The conditions with the largest σ_α^2 , which are $\alpha \sim N(-1, 1)$ and $\alpha \sim N(-2, 1)$, resulted in 16% and 2% increasing PRFs ($\alpha_v > 0$), respectively, and 14% and 4% nearly flat PRFs ($-0.5 < \alpha_v < 0$).

For the “est” conditions, we studied fewer combinations because this study focused more on bias due to model mismatch than on bias due to estimates

$\hat{\delta}$ and $\hat{\theta}$. In the “est” conditions, we assessed bias of the C-PRF and the PRF models in 2×2 combinations of μ_α (1, 2) and σ_α^2 (0.1, 1) using $\theta \sim N(0, 1)$ throughout. In all conditions, $\gamma_{00} = 0$ (because it is the adjusted mean outcome; see Raudenbush & Bryk, 2002, pp. 112–113), $\gamma_{01} = -\mu_\alpha$, $\gamma_{10} = \mu_\alpha$, and $\tau_{11} = \sigma_\alpha^2$.

We generated 1,000 data sets for each combination of parameter values. Because Moineddin, Matheson, and Glazier (2007) showed that a Level 1 sample size of at least 50 is required to obtain unbiased MLR parameter estimates, we chose a test length of 50 items. For several C-PRF conditions, we tried different Level 2 sample sizes and concluded that a Level 2 sample size of 500 examinees throughout resulted in sufficient precision. The Appendix provides information on the software used in this study.

Results

Condition C-PRF true. Table 1 shows that bias in $\hat{\tau}_{11}$ ranged from -0.57 to 0.01 , meaning that $\hat{\tau}_{11}$ was underestimated. Bias in other estimates was small: parameter γ_{01} was estimated without bias, γ_{00} was slightly underestimated, and estimate $\hat{\gamma}_{10}$ was pulled a little toward 0 (results not tabulated). Bias for $\hat{\tau}_{11}$ was small for $\theta \sim N(0, 1)$ (bias ranged from -0.02 to 0.01) and particularly high when $\alpha \sim N(-2, 0.5)$ and $\theta \sim N(1, 0.2)$ (relative bias, i.e., bias/τ_{11} , equaled $0.27/0.5 = 0.54$), and $\alpha \sim N(-2, 1)$ and $\theta \sim N(1, 0.2)$ (relative bias equaled $0.57/1 = 0.57$).

Condition PRF true. Similar to the C-PRF true conditions, $\hat{\gamma}_{10}$ and $\hat{\tau}_{11}$ were pulled toward 0 but in contrast to the C-PRF true conditions, γ_{00} was overestimated and γ_{01} underestimated (results only tabulated for $\hat{\tau}_{11}$).

Mean bias difference between conditions. Table 2 shows the mean bias difference between the C-PRF true and the PRF true conditions (i.e., mean bias PRF true – mean bias C-PRF true) and its range as a function of μ_α , σ_α^2 , μ_θ , and σ_θ^2 . Compared with the C-PRF true conditions, the bias in the PRF true conditions was larger for $\hat{\gamma}_{10}$ and $\hat{\gamma}_{01}$. For γ_{10} this means that estimates were pulled more toward 0. The bias in $\hat{\gamma}_{00}$ was also larger in the PRF true than in the C-PRF true condition, but the sign was opposite. With the exception of $\hat{\tau}_{11}$ for $\alpha \sim N(-2, 0.5)$ and $\theta \sim N(1, 0.2)$, and $\alpha \sim N(-2, 1)$ and $\theta \sim N(1, 0.2)$, bias in $\hat{\tau}_{11}$ was larger (pulled more toward 0) in the PRF true conditions (Table 1 and last column of Table 2).

Table 2 shows that the mean bias difference in $\hat{\gamma}_{00}$ (second column) was larger for larger negative μ_α , increased in σ_α^2 and μ_θ , and decreased in σ_θ^2 . The bias differences in $\hat{\gamma}_{01}$, $\hat{\gamma}_{10}$, and $\hat{\tau}_{11}$ (third to fifth column) were larger for

TABLE 1
Mean Bias (SD in Parentheses) in Estimated Slope Variance τ_{11}

α Distribution	Model	θ Distribution			
		$N(0, 1)$	$N(1, 1)$	$N(0, 0.2)$	$N(1, 0.2)$
$N(-1, 0)$	C-PRF true	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$N(-2, 0)$	C-PRF true	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
$N(-1, 0.1)$	C-PRF true	0.00 (0.01)	-0.01 (0.01)	0.00 (0.01)	-0.02 (0.01)
	PRF true	-0.03 (0.01)	-0.02 (0.01)	-0.01 (0.01)	-0.01 (0.02)
	C-PRF est	-0.03 (0.01)	—	—	—
$N(-2, 0.1)$	PRF est	-0.03 (0.01)	—	—	—
	C-PRF true	0.00 (0.02)	-0.01 (0.02)	0.00 (0.02)	-0.04 (0.01)
	PRF true	-0.07 (0.01)	-0.07 (0.02)	-0.03 (0.02)	-0.02 (0.03)
$N(-1, 0.5)$	C-PRF est	-0.10 (0.01)	—	—	—
	PRF est	-0.10 (0.01)	—	—	—
	C-PRF true	-0.01 (0.03)	-0.04 (0.04)	-0.01 (0.04)	-0.05 (0.03)
$N(-2, 0.5)$	PRF true	-0.10 (0.03)	-0.08 (0.03)	-0.07 (0.03)	-0.05 (0.03)
	C-PRF true	-0.01 (0.04)	-0.12 (0.04)	-0.01 (0.05)	-0.27 (0.03)
	PRF true	-0.25 (0.04)	-0.22 (0.08)	-0.13 (0.04)	-0.10 (0.06)
$N(-1, 1)$	C-PRF true	-0.01 (0.05)	-0.08 (0.05)	-0.01 (0.10)	-0.09 (0.05)
	PRF true	-0.20 (0.05)	-0.18 (0.05)	-0.12 (0.05)	-0.10 (0.05)
	C-PRF est	0.03 (0.10)	—	—	—
$N(-2, 1)$	PRF est	0.17 (0.15)	—	—	—
	C-PRF true	-0.02 (0.06)	-0.20 (0.09)	-0.02 (0.06)	-0.57 (0.05)
	PRF true	-0.39 (0.11)	-0.27 (0.06)	-0.27 (0.06)	-0.18 (0.07)
	C-PRF est	-0.51 (0.06)	—	—	—
	PRF est	-0.49 (0.05)	—	—	—

Note. C-PRF = compatible person response function; PRF = person response function. “est” and “true” refer to whether or not θ and δ were estimated, respectively. “—” indicates that for this condition no simulations were done.

larger negative μ_α , increased in σ_α^2 and σ_θ^2 , and decreased in μ_θ . In sum, model mismatch and violation of bivariate normality caused biased estimates.

Conditions C-PRF est and PRF est. Table 1 (third column) shows the bias in $\hat{\tau}_{11}$ in the “est” conditions when $\theta \sim N(0, 1)$. Parameter τ_{11} was overestimated in the conditions in which $\alpha \sim N(-1, 1)$ but underestimated in all other α conditions. Bias also differed from the “true” conditions; except for $\alpha \sim N(-1, 1)$, bias in $\hat{\tau}_{11}$ was larger and bias in the C-PRF est and PRF est conditions was equal. Interestingly, mean $\hat{\tau}_{11}$ was 0 if $\alpha \sim N(-2, 0.1)$ in both the C-PRF est and PRF est conditions. Thus, person misfit was not detected in the “est” conditions when misfit was modest but it was detected in the “true” conditions. Estimate $\hat{\gamma}_{00}$ was unbiased but $\hat{\gamma}_{01}$ and $\hat{\gamma}_{10}$ were substantially biased in most of the “est” conditions. Thus, multilevel PFA also yields biased

TABLE 2
Mean and Range (Between Brackets) of Mean Bias Difference Between C-PRF true Conditions and PRF true Conditions in Which $\sigma_{\alpha}^2 > 0$ as Function of PRF Properties

Distribution Values	$\hat{\gamma}_{00}$	$\hat{\gamma}_{01}$	$\hat{\gamma}_{10}$	$\hat{\tau}_{11}$
Slope mean μ_{α}				
-1	0.04 [0.00, 0.12]	-0.11 [-0.19, -0.04]	0.03 [-0.01, 0.08]	-0.05 [-0.19, 0.01]
-2	0.06 [0.00, 0.27]	-0.20 [-0.40, 0.04]	0.07 [-0.07, 0.22]	-0.06 [-0.37, 0.39]
Slope variance σ_{α}^2				
0.1	0.01 [0.00, 0.03]	-0.05 [-0.06, -0.04]	0.02 [0.00, 0.04]	-0.02 [-0.07, 0.02]
0.5	0.05 [0.00, 0.15]	-0.17 [-0.23, -0.10]	0.06 [0.00, 0.15]	-0.06 [-0.24, 0.16]
1	0.08 [0.00, 0.27]	-0.24 [-0.40, -0.12]	0.07 [-0.07, 0.22]	-0.09 [-0.37, 0.39]
Latent variable mean μ_{θ}				
0	0.00 [0.00, 0.00]	-0.18 [-0.40, -0.05]	0.07 [0.01, 0.22]	-0.13 [-0.37, -0.01]
1	0.09 [0.01, 0.27]	-0.13 [-0.27, -0.04]	0.03 [-0.07, 0.15]	0.02 [-0.10, 0.39]
Latent variable variance σ_{θ}^2				
0.2	0.06 [0.00, 0.27]	-0.15 [-0.40, -0.04]	0.02 [-0.07, 0.12]	0.00 [-0.25, 0.39]
1	0.03 [0.00, 0.14]	-0.16 [-0.38, -0.05]	0.08 [0.03, 0.22]	-0.11 [-0.37, -0.02]

Note. C-PRF = compatible person response function; PRF = person response function. $\hat{\gamma}_{00}$ = estimated average intercept; $\hat{\gamma}_{01}$ = estimated effect of θ ; $\hat{\gamma}_{10}$ = estimated average slope; $\hat{\tau}_{11}$ = estimated slope variance.

estimates when using $\hat{\delta}$ and $\hat{\theta}$, and the results suggest that multilevel PFA does not detect person misfit in some conditions when the variance in PRF slopes is small.

Intercept variance. Results for $\hat{\tau}_{00}$ were troublesome. Agreeing with our theoretical analysis, if $\sigma_{\alpha}^2 > 0$, in the “true” conditions $\hat{\tau}_{00} > 0$ but in the “est” conditions surprisingly we found $\hat{\tau}_{00} \approx 0$. This result suggests that true intercept variance may be concealed when estimated item and person parameters are used in multilevel PFA. Indeed, additional simulations showed that also when multidimensionality holds one may find $\hat{\tau}_{00} = 0$ in the “est” conditions. Thus, finding $\hat{\tau}_{00} = 0$ does not imply unidimensionality because including $\hat{\theta}$ in the multilevel PFA model may render multidimensionality undetectable.

Summary of Monte Carlo Study

The Monte Carlo study showed that due to the mismatch between MLR and the PRF model MLR yields biased estimates of the distributions of the person intercepts and slopes from the PRF model. The variance of the PRF slopes, which is of primary interest in PFA, tended to be underestimated in most cases. The other parameters were also biased, but no clear trends in the direction of the bias were found. Bias became even more serious when estimated person and item parameters were used.

CONCLUSIONS ON MULTILEVEL PERSON-FIT ANALYSIS

Multilevel PFA has serious limitations. First, multilevel PFA takes the slope of the PRF as a valid person-fit measure, which is only correct under the 1PLM but contrary to Reise's (2000) suggestion not under the 2PLM. Second, MLR is incompatible with the PRF model even if items satisfy the 1PLM. As a result, the assumption of bivariate normality of random effects is violated when PRF slopes are different. Third, the mismatch between MLR and the PRF model leads to biased estimates of multilevel PFA model parameters. Most important, PRF-slope variance is underestimated or not even detected.

Part of the problem revolves around the interpretation of PRF slope variation. Reise's (2000) methodology argues that variation in PRF slopes indicates variation in person fit but does not recognize that under the 2PLM, in which items have different discrimination parameters, PRF slopes vary *by definition* because the PRF slope depends on the examinee's latent variable value. This also means that, as a person-fit measure, the PRF slope is inherently contaminated by the latent variable value. Obviously, this is an undesirable property for person-fit statistics. Using PRF slopes for assessing person fit is even more problematic because near-0 or positive PRF slopes, which Reise qualifies as indicators of uninterpretable item-score patterns, can be fully consistent with the 2PLM. Thus, person-fit assessment based on the PRF slopes is inappropriate under the 2PLM. On the other hand, under the 1PLM, PRF slope variance is 0 by definition and deviant PRF slopes found in a sample may flag person misfit.

The other part of the problem involves using the MLR framework for estimating the PRF model and appears fundamental. In the PRF model, both the location and slope vary over examinees and need to be estimated as random effects. The multilevel approach assumes bivariate normality for the Level 2 random effects. We showed that the PRF slope restricts the variation in the intercept and, as a result, the Level 2 random effects do not follow a bivariate normal distribution.

Our simulation study using item and person parameters showed that multilevel PFA produces biased estimates of the systematic differences in person fit. Studies in other research areas also found that nonnormally distributed random effects in MLR lead to bias in variance and fixed effects estimates (Heagerty & Kurland, 2001; Litière, Alonso, & Molenberghs, 2007, 2008). The PRF-slope variance was underestimated; hence, differences in person fit came out too small. The underestimation of PRF-slope variance became greater when item and person parameter estimates were used, which is what researchers do, thus showing that the problem is greater in real-data analysis. Ironically, multilevel PFA only provides correct estimates when PRF slopes are equal but then person misfit is absent. In real data it is unknown whether there is variation in person fit or no

misfit at all; this is exactly what multilevel PFA was designed to find out. Finally, we found that multilevel PFA sometimes does not pick up multidimensionality (Step 2).

The key advantage of multilevel PFA over traditional person-fit methods is to detect systematic individual differences in person fit and explain these differences by including explanatory variables in the model. The multilevel PFA model parameter estimates were expected to provide information about person-fit variation and explanatory variables included to explain this variation. However, we showed that multilevel parameters are biased and that under the 2PLM the PRF slope is confounded with the latent variable distribution. These results suggest that multilevel PFA has limited value as an explanatory tool in person-fit research. Contrary to Reise's (2000) suggestions we also found that multilevel PFA is inappropriate for studying multidimensionality.

Furthermore, Reise (2000) proposed to use the EB slopes from the multilevel PFA model for identifying respondents having aberrant item-score patterns. Woods (2008) studied the Type I error and the power of the EB slope in multilevel PFA and concluded that in most conditions its performance was adequate. However, Woods also found occasionally increased Type I error rates for the EB slopes and showed that it is difficult to specify the cutoff criteria for EB slopes needed to operationalize misfit. Thus, even though these results suggest that EB slopes have potential for identifying person misfit, their usefulness requires additional research. However, given the theoretical limitations of interpreting EB slopes as a measure of person fit, and also the bias in EB slope estimates caused by biased slope variance estimates of the multilevel model (e.g., Collett, 2003, pp. 274–275), we consider further study on the usefulness of the EB slopes not a fruitful contribution to person-fit assessment.

AN ALTERNATIVE EXPLANATORY MULTILEVEL PERSON-FIT APPROACH: REAL-DATA EXAMPLE

An alternative multilevel PFA approach that we have started pursuing in our research has similarities to Reise's (2000) approach and aims but avoids the problems we identified. We tentatively advocate this approach using what we believe is an interesting data example concerning cardiac patients who had a cardioverter-defibrillator implanted, inducing anxiety in many patients due to anticipation of a sudden, painful electrical shock responding to cardiac arrhythmia. A sample of cardiac patients and their partners ($N = 868$) completed the state-anxiety scale from the State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983) in a longitudinal study comprising five measurement occasions. Here, the repeated measurements constitute the multilevel nature of the data. Using multilevel modeling, we assessed whether

person fit is a reliable individual-difference variable that may be explained by demographic, personality, medical, psychological distress, and mood variables.

At each occasion, we used the widely accepted and much used l_z person-fit statistic (Drasgow, Levine, & McLaughlin, 1987; Drasgow, Levine, & Williams, 1985; Li & Olejnik, 1997) for assessing person fit on the anxiety-state scale of the STAI. Given the 4-point rating-scale data collected by means of the STAI, we used statistic l_z to assess person fit relative to the graded response model (GRM; Samejima, 1997). We assessed goodness of fit of the GRM to the data for each measurement occasion and found satisfying results (Conijn, Emons, van Assen, Pedersen, & Sijtsma, 2011). Several authors noticed that, in particular for small numbers of dichotomous items, the sampling distribution of statistic l_z depends on latent-variable level (Nering, 1995; Snijders, 2001; Van Krimpen-Stoop & Meijer, 1999). We implemented a parametric bootstrap procedure developed by De la Torre and Deng (2008) to make sure that the l_z statistic is standard normally distributed under the GRM at all values of the latent variable.

The l_z statistic was modeled as a dependent variable in a 2-level model. As independent variables we used measures of mood state and psychological distress, which are time dependent, and demographic characteristics, personality traits, and medical conditions, known to be stable across time. The Level 1 model describes within-individual variation in person fit across repeated measures, and the Level 2 model describes variation across individual respondents. An unconditional random intercept model estimated within-person and between-person variance in statistic l_z . The intraclass correlation (ICC; Snijders & Bosker, 1999, pp. 16–18) provides evidence for or against substantial systematic between-person differences in the data and indicates whether a multilevel approach is useful. If significant between-person variance is found, respondents differ systematically in person fit, and given this result, this variation may be explained using the independent variables at Level 1 and Level 2. Explanatory variables specific to measurement occasions at Level 1 may be added to explain within-person variation in statistic l_z .

The results are as follows: The ICC equaled 0.31, suggesting that multilevel analysis was appropriate and that of the total variation in l_z 31% was attributable to differences between persons and 69% to differences within persons. The unconditional random intercept model revealed significant between-person variance in l_z . We were able to explain 8% of the between-person differences and 4% of the within-person differences in person fit. Respondents having more psychological problems, higher trait anger, and lower education level showed more person misfit. When respondents had higher anxiety level at the measurement occasion than usual they also showed more misfit than usual. Thus, respondents showing poor fit at previous measurements, having low education level, and experiencing psychological problems are at risk of producing invalid test results. Also, assessment shortly before implantable cardioverter-defibrillator

implantation likely produces person misfit due to higher state anxiety. Our results show that multilevel modeling can be highly useful in gaining a better understanding of the person and situational characteristics that may produce person misfit and, consequently, distort valid test performance.

One final remark is that in other studies researchers may not have access to repeated measures but multilevel modeling of person misfit may well be possible, thus facilitating the explanatory analysis so badly needed in person-fit research. For example, for data based on one measurement occasion the multilevel aspect may be the person-fit statistic obtained on scales measuring different attributes or even on subsets of items coming from the same scale.

DISCUSSION

We showed that Reise's (2000) multilevel PFA approach suffers from serious theoretical and statistical problems, rendering the method questionable as an explanatory tool in PFA. Exactly because the idea of constructing such an explanatory tool was so strong, and because multilevel analysis is a powerful approach that produces explanations at different levels in the data, we suggested a simple alternative that avoids the technical problems of Reise's approach and maintains the explanatory ambitions so badly needed in PFA.

A reviewer suggested finding a solution for the problem of nonnormally distributed random effects in the multilevel PFA model by estimating the bivariate distribution of the random effects from the data. Thus far, for generalized linear models only methods have been developed for estimating the univariate distribution of random effects (Chen, Zhang, & Davidian, 2002; Litière et al., 2008). Maybe these methods could be extended to the bivariate case, but if they could, implementation of these extensions would only possibly repair the 1PLM version but not the much more flexible and for practitioners more interesting 2PLM version of the multilevel PFA model. Moreover, for researchers advocating the 1PLM our alternative approach may be used because statistic l_z is also adequate for 1PLM data (and Snijders, 2001, solved the distributional problems due to dependence on the latent-variable level). As an aside, one may note that our approach does not hinge on statistic l_z . For example, when the 1PLM is consistent with the data one may use a statistic proposed by Molenaar and Hoijtink (1990) as the dependent variable, and if parametric IRT models are inconsistent but a nonparametric model does fit, the normed count of Guttman errors (Emons, 2008) may be used. Most important is the awareness that our approach uses the multilevel model in a regular context without the technical problems induced by Reise's (2000) multilevel PFA model and that the choice of the most appropriate dependent variable for person fit is up to the researcher.

Another reviewer suggested that PFA in general has been rarely applied to real-data problems, which questions the usefulness of PFA. Although some promising examples are available (e.g., Conrad et al., 2010; Engelhard, 2009; Meijer, Egberink, Emons, & Sijtsma, 2008; Tatsuoaka, 1996), we agree that more applications are needed. Conijn et al. (2010) further elaborated the example using the sample of cardiac patients and their partners. More generally, PFA suffers from low power because the number of items in the test is the “sample size” that determines the power of a person-fit statistic (e.g., Emons et al., 2005; Meijer & Sijtsma, 2001), and this is a problem that is not easily solved. Nevertheless, the assessment of individual test performance is highly important, and highly invalid item-score vectors can be identified, even if the power for finding moderate violations is low and some invalid vectors may be missed.

Approaches focusing on PRFs and multilevel models have in common that they try to incorporate PFA in an explanatory framework, thus strengthening the methods and lending them more practical relevance. We believe that in spite of the problems such attempts must be further pursued so as to improve the assessment of individual test performance.

REFERENCES

- Bates, D., Maechler, M., & Dai, B. (2008). lme4: Linear mixed effects models using S4 classes [computer software]. Retrieved from <http://cran.r-project.org/web/packages/lme4/index.html>
- Chen, J., Zhang, D., & Davidian, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random-effects distribution. *Biostatistics*, 3, 347–360.
- Collett, D. (2003). *Modelling binary data* (2nd ed.). London, UK: Chapman & Hall/CRC.
- Conijn, J. M., Emons, W. H. M., van Assen, M. A. L. M., Pedersen, S. S., & Sijtsma, K. (2011). *Response consistency on the State-Trait Anxiety Inventory in cardiac patients*. Manuscript submitted for publication.
- Conrad, K. J., Bezruczko, N., Chan, Y. F., Riley, B., Diamond, G., & Dennis, M. L. (2010). Screening for atypical suicide risk with person fit statistics among people presenting to alcohol and other drug treatment. *Drug and Alcohol Dependence*, 106, 92–100.
- De la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, 45, 159–177.
- Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11, 59–79.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with poly-chotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67–86.
- Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, 32, 224–247.
- Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person response function in person-fit analysis. *Multivariate Behavioral Research*, 39, 1–35.
- Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person response functions. *Psychological Methods*, 10, 101–119.

- Engelhard, G. (2009). Using item response theory and model data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, 69, 585–602.
- Ferrando, P. J. (2004). Person reliability in personality measurement: An item response theory analysis. *Applied Psychological Measurement*, 28, 126–140.
- Ferrando, P. J. (2007). A person-type-VII item response model for assessing person fluctuation. *Psychometrika*, 72, 25–41.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer Academic.
- Heagerty, P. J., & Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88, 973–985.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16, 277–298.
- LaHuis, D. M., & Copeland, D. (2009). Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organizational Research Methods*, 12, 296–319.
- Li, M. F., & Olejnik, S. (1997). The power of Rasch person fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, 21, 215–231.
- Litière, S., Alonso, A., & Molenberghs, G. (2007). Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics*, 63, 1038–1044.
- Litière, S., Alonso, A., & Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on maximum likelihood estimation in generalized linear mixed models. *Statistics in Medicine*, 27, 3125–3144.
- Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement*, 1, 477–482.
- Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, 31, 19–26.
- Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unsalable item score patterns using item response theory: An illustration with Harter's Self-Perception Profile for Children. *Journal of Personality Assessment*, 90, 227–238.
- Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement*, 21, 321–336.
- Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new developments. *Applied Measurement in Education*, 8, 261–272.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person-fit. *Applied Psychological Measurement*, 25, 107–135.
- Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, 7, 34–43.
- Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person-fit indices. *Psychometrika*, 55, 75–106.
- Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, 19, 121–129.
- Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the I_2 person-fit statistic. *Applied Psychological Measurement*, 22, 53–69.
- Pan, T. (2010). *Comparison of six IRT computer programs in estimating the Rasch model*. Unpublished manuscript.
- Partchev, I. (2008). irtoys: Simple interface to the estimation and plotting of IRT models [computer software]. Retrieved from <http://cran.r-project.org/web/packages/irtoys/index.html>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
- Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2008). *HLM: Hierarchical linear and nonlinear modeling (Version 6.06)* [computer software]. Lincolnwood, IL: Scientific Software International.

- Raudenbush, S. W., Yang, M. L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, 9, 141–157.
- Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research*, 35, 543–568.
- Reise, S. P., & Waller, N. G. (1993). Traitendness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, 65, 143–151.
- Rizopoulos, D. (2009). ltm: An R package for latent variable modeling and item response analysis [computer software]. Retrieved from <http://cran.r-project.org/web/packages/ltm/index.html>
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer.
- Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, 66, 191–207.
- Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, 66, 331–342.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.
- Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory (Form Y)*. Palo Alto, CA: Consulting Psychologists Press.
- Strandmark, N. L., & Linn, R. L. (1987). A generalized logistic item response model parameterizing test score inappropriateness. *Applied Psychological Measurement*, 11, 355–370.
- Tatsuoka, K. K. (1996). Use of generalized person-fit indexes, zetas for statistical pattern classification. *Applied Measurement in Education*, 9, 65–75.
- Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality*, 56, 621–663.
- Thissen, D., Chen W. H., & Bock R. D. (2003). MULTILOG for Windows (Version 7) [computer software]. Lincolnwood, IL: Scientific Software International.
- Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item theory models. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 83–108). New York, NY: Academic Press.
- Van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, 23, 327–345.
- Wang, L., Reise, S. P., Pan, W., & Austin, J. T. (2004, April). *Multilevel modeling approach to detection of differential person functioning in latent trait models*. Paper presented at the meeting of the American Educational Research Association, San Diego, CA.
- Woods, C. M. (2008). Monte Carlo evaluation of two-level logistic regression for assessing person-fit. *Multivariate Behavioral Research*, 43, 50–76.
- Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2008). Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological Assessment*, 20, 159–168.

APPENDIX: SOFTWARE

We used the ltm R-package (Rizopoulos, 2009) to obtain the marginal maximal likelihood estimates of δ under the 1-parameter logistic model (1PLM). We used the irtoys R-package (Partchev, 2008) to obtain the expected a posteriori estimates of θ_v given the δ estimates from the ltm R-package. Pan (2010) found

that the ltm R-package provided parameter estimates at least as accurate as the estimates item response theory (IRT) programs such as MULTILOG (Thissen, Chen, & Bock, 2003) provide.

We used HLM 6.06 (Raudenbush, Bryk, & Congdon, 2008) to estimate the multilevel person-fit analysis (PFA) model. Parameter estimation was done with the Laplace6 (Raudenbush, Yang, & Yosef, 2000) procedure in HLM 6.06. Laplace6 uses a sixth order approximation to the likelihood based on a Laplace transform, using the EM algorithm. The maximum number of iterations was set at 20,000. If convergence was not achieved, the parameter estimates were not included in computing summary statistics on the bias. Simulation of data sets was continued until the number of converged models was 1,000 in each condition.

Raudenbush et al. (2000) found that Laplace6 provided more accurate parameter estimates than penalized quasi-likelihood and was at least as accurate as Gauss-Hermite quadrature using 10 to 40 quadrature points and adaptive Gauss-Hermite quadrature using 7 quadrature points. Furthermore, Laplace6 was faster in terms of processing time than (adaptive) Gauss-Hermite quadrature. An additional reason to use Laplace6 instead of adaptive Gauss-Hermite quadrature was that the latter method converged slowly in the person response function (PRF) conditions when the lme4 package (Bates, Maechler, & Dai, 2008) was used in R. Laplace6 did not provide any serious convergence problems.