## Supplement to "Person-Fit Statistics for Joint Models for Accuracy and Speed"

A Bayesian residual analysis for the joint model is proposed. The residual statistics can be computed as by-products of an MCMC algorithm. In correspondence to a Bayesian joint modeling approach, Bayesian residual analysis are considered to evaluate the fit of the joint model. Following the procedure of Albert and Chib (1993) and Fox (2010), expressions are given to estimate residuals related to response accuracy (RA) patterns and response time (RT) patterns. Furthermore, the extremeness of each realized residual is evaluated by computing the posterior probability that it is greater than a certain threshold value.

The normality assumption of each item's RT residuals is evaluated using the Kolmogorov-Smirnov test. The empirical distribution is compared to the normal distribution and a Bayesian p-value is computed, which quantifies the extremeness of the violation of the normality assumption. The tools for model evaluation are also applied to the joint model with a three-parameter model for ability to account for guessing behavior. The fit statistics and computation of residuals are also applicable to this joint model. MCMC will be used for parameter estimation, which closely connects to the MCMC method of Klein Entink, Fox, et al. (2009) and Fox (2010), which are extended to handle guessing behavior.

### Residual Analysis

Albert and Chib (1995), Johnson and Albert (1999), and Fox (2010) considered Bayesian residuals to evaluate the fit of an IRT model. Their methods can be applied to the joint model (i.e, ). Latent residuals, $e_{ik}$, are considered, which represent the difference between the latent continuous response and the mean. These residuals are computed to identify extreme outliers and the total percentage of extreme outliers per item and per test taker. Following the procedure of Albert and Chib (1995), the expressions of the conditional expected latent residuals is given by

$$E\left(e_{ik}\left|Y_{ik}=0,a_k,b_k,\theta_i\right.\right)=\frac{-\varphi\left(b_k-a_k\theta_i\right)}{\Phi\left(b_k-a_k\theta_i\right)} \tag{S1}$$

and

$$E\left(e_{ik}\left|Y_{ik}=1,a_k,b_k,\theta_i\right.\right)=\frac{\varphi\left(b_k-a_k\theta_i\right)}{\Phi\left(a_k\theta_i-b_k\right)}. \tag{S2}$$

The extremeness of a latent residual is computed by the posterior probability that the latent residual is greater than a specific threshold *r.* It follows that

$$P\left(\left|e_{ik}\right|>r\left|Y_{ik}=0,a_k,b_k,\theta_i\right.\right)=\frac{\Phi\left(-r\right)}{1-\Phi\left(a_k\theta_i-b_k\right)} \tag{S3}$$

and

$$P\left(\left|e_{ik}\right|>r\left|Y_{ik}=1,a_k,b_k,\theta_i\right.\right)=\frac{\Phi\left(-r\right)}{\Phi\left(a_k\theta_i-b_k\right)}. \tag{S4}$$

Next, the log-RT residuals, $\varepsilon_{ik}$, are considered, which represent the difference between the RT and the mean. The extremeness of the RT residual can be expressed as the posterior probability that the residual is greater than a threshold *q*. In the same way, it follows that,

$$P\left(\left|\varepsilon_{ik}\right| > q \middle| \zeta_i, \lambda_k, \phi_k, rt_{ik}^*\right) = \Phi\left(-q - \frac{\varepsilon_{ik}}{\sigma_k}\right) + 1 - \Phi\left(q - \frac{\varepsilon_{ik}}{\sigma_k}\right), \tag{S5}$$

where $\varepsilon_{ik} = rt_{ik}^* - \left(\lambda_k - \phi_k \zeta_i\right)$. The extremeness of the latent residuals and RT residuals are computed as by-products of the MCMC algorithm for the joint model, thereby accounting for relationships between the joint model parameters.

**Evaluating Distributional Assumptions**

The fit of the distribution of the RT residuals can be evaluated using the Kolmogorov-Smirnov test (KS test). This is a non-parametric test that is used to compare the empirical distribution of the residuals for each item to the assumed normal distribution. The distance between the distribution of the realized residuals and the normal cumulative distribution function is evaluated in each MCMC iteration. Under the null hypothesis, the RT residuals are assumed to be normally distributed, and a significant statistic value shows evidence for non-normally distributed RT residuals. In theory, the KS test could also be applied to evaluate the normality assumption of the latent residuals. However, this application of the KS test does not have power, since the latent residuals already depend on simulated normally distributed responses.

The KS test is used as a Bayesian goodness-of-fit-test by computing the marginal posterior probability that each vector of item residuals is non-normally distributed. For the computed residuals of item *k* of *n* test-takers, the empirical distribution is given by

$$F_n\left(\varepsilon\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(\varepsilon_{ik} < \varepsilon\right)\left(\varepsilon_{ik}\right) \tag{S6}$$

where $\varepsilon_{ik} = rt_{ik}^* - \left(\lambda_k - \phi_k \zeta_i\right)$ and $I\left(.\right)$, the indicator function, equals one when $\varepsilon_{ik} < \varepsilon$ and zero otherwise. The KS test statistic is given by,

$$D_n = \sup_{\varepsilon} \left|F_n\left(\varepsilon\right) - \Phi\left(\varepsilon\right)\right|. \tag{S7}$$

The distribution of $D_n$ is the Kolmogorov distribution and the $D_n$ converges to zero when the residuals are normally distributed. Subsequently, the posterior significance probability is computed in each MCMC iteration,

$$p_{ks}\left(\lambda_k, \phi_k, \boldsymbol{\zeta}\right) = P\left(D_n > c \middle| \mathbf{rt}_k^*, \lambda_k, \phi_k, \boldsymbol{\zeta}\right), \tag{S8}$$

where the average significance probability over MCMC iterations is used as an estimate of the marginal posterior probability that the residuals of item *k* are non-normally distributed.

## Simulation Study

The simulated RTs under the joint model were manipulated to simulate positively (right-) skewed RT errors. Therefore, Gamma distributed noise (with different shape and rate parameters) was added to the simulated log-normally distributed RTs for items 1 to 5 of a 20-item test. Three conditions were considered for the Gamma distributed noise. For each condition, a shape parameter of *2* was specified, but the rate parameter was different and equaled *1* (distributed noise with mean *2* and variance *2*), 1.25 (distributed noise with mean *1.6* and variance *1.28*), or 1.5 (distributed noise with mean *1.33* and variance *.89*). The positively distributed errors led to positively skewed RT distributions.

A total of *50%* Gamma distributed noise was added to the RTs of item 1, *33%* to item 2, *25%* to item 3, *20%* to item 4, and *17%* to item 5 for the *1,000* persons. The posterior probability of non-lognormally distributed residuals was computed for each item and each specification of the Gamma distribution of added noise. The joint model with time discriminations equal to one was used to simulate the RA and RT data. This joint model was also estimated given the RA and RT data with positively skewed distributed noise.

The detection rates of non-normally distributed item residuals were computed for *100* replicated data sets using a significance level of *.05* using Equation (S8). The estimated detection rates for the *5* items and the three noise conditions are given in Table S1.

The detection rates were around *1* for all items when the added noise was distributed with a rate parameter of *1*. The added Gamma distributed noise was larger than the normally distributed errors under the model leading to a violation of normality. When increasing the rate parameter of the Gamma distributed noise the detection rates are decreasing. When the percentage of added noise is smaller than *33%*, the detection rates decreased for decreasing percentages of noise added. The item residuals of the remaining items 6 to 20 did not show significant violations of normality, since for these items the logarithm of RTs were generated from normal distributions.

TABLE S1

*Performance of the KS test in identifying non-normally distributed item residuals. The reported detection rates are based on one hundred replications using a significance level of .05.*

| Item | % | Gamma ($\alpha, \beta$) | | |
|---|---|---|---|---|
| | | $\alpha = 2, \beta = 1$ | $\alpha = 2, \beta = 1.25$ | $\alpha = 2, \beta = 1.5$ |
| 1 | 50 | 1 | 0.91 | 0.68 |
| 2 | 33.33 | 1 | 0.89 | 0.68 |
| 3 | 25 | 0.98 | 0.88 | 0.58 |
| 4 | 20 | 0.99 | 0.76 | 0.49 |
| 5 | 16.67 | 0.94 | 0.64 | 0.45 |

## Real Data Analysis

The joint model was fitted to a real data set of Cizek and Wollack (2016), referred to as the credentialing data, concerning *1,636* test takers who applied for licensure. The KS test was used to check the normality assumption of the RT item residuals. It followed that for *18* of the *170* items (around *11%*) the normality assumption was rejected with a significance level of *.05*, where the p-values ranged from *.000* to *.042*. The extremeness of the residuals was investigated, using Equation (S5), with *q* equal to *2*. Around *12.3%* of the standardized residuals were considered extreme, which partly explained the non-normality residual distribution of the *18* items. For most of the identified extreme residuals, the observed RT was larger than *200*s and ranging to *612*s, where *128*s was the highest population-average response time. It was concluded that the log-normal distribution failed to include this relatively high percentage of extreme RTs since the tails of the log-normal distribution were too flat.

The extremeness of the latent residuals was also investigated, using Equation (S3) and Equation (S4), where *r=2*. Around *1.9%* of the latent response residuals were marked as extreme. These residuals corresponded to incorrect responses of high-ability test takers (with an ability level around *0.5*) to easy items (with a difficulty level below *-1*).

It can be concluded that the residual statistics concerning both patterns can be used to evaluate the fit of the joint model. The residuals and residual statistics are easily computed as by-products of the MCMC algorithm. The KS test, based on the realized residuals, showed good performance to identify non-normally distributed RT item residuals. The KS test was used as a Bayesian significance test, where the posterior probability of an extreme KS test statistic was computed. Other residual statistics can be considered to further develop Bayesian significance testing for the joint model.

## References

Albert, J., & Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, *82*(4), 747-769. doi: 10.1093/biomet/82.4.747.

Cizek, G. J., Wollack, J. A. (2016). Exploring cheating on tests: The context, the concern, and the challenge. In G. Cizek & J. A. Wollack (Eds.), *Handbook of Detecting Cheating on Tests.* New York, NY: Routeledge.

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and applications*. New York: Springer. doi: 10.1007/978-1-4419-0742-4

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

Johnson, V. and Albert, J. (1999). *Ordinal Data Modeling*. New York: Springer.