1 Supplemental materials

1.1 Identification

We discuss two issues in identification here. First, the relation of the score parameters \( \lambda^1 \) and \( \lambda^2 \) of booklet \( x \) and \( y \) to the complete set of score parameters \( \lambda^* \). Second, the relation of the item parameters \( b \) and \( c \) in the just created booklets to the item parameters of the complete set \( b^* \).

1.1.1 Identification of the score parameters. The relation between score parameters \( \lambda^1 \) and \( \lambda^* \) will be described below, the relation between \( \lambda^2 \) and \( \lambda^* \) follows easily. The fitted model to subset \( x \) concerning only the score parameters of this subset \( \lambda^1 \) is:

\[
P(x|b, \lambda^1) = \frac{\prod_{i=1}^{m} b_i x_i}{\sum_{s=0}^{m} \gamma_s(b) \lambda^1_s}
\]

while the full model regarding the complete set of score parameters \( \lambda^* \) and difficulty parameters \( c \) of booklet \( y \) is:

\[
P(x|b, c, \lambda^*) = \frac{\prod_{i=1}^{m} b_i x_i \sum_{t=0}^{n} \gamma_t(c) \lambda^*_{s+t}}{\sum_{s=0}^{m} \gamma_s(b) \sum_{t=0}^{n} \gamma_t(c) \lambda^*_{s+t}}.
\]

We can substitute \( \sum_{t=0}^{n} \gamma_t(c) \lambda^*_{s+t} \) in the numerator and denominator of Equation 2 with \( \lambda^1 \) from Equation 1 if for all \( s \):

\[
\lambda^1_s = \sum_{t=0}^{n} \gamma_t(c) \lambda^*_s + t
\]

and likewise concerning \( \lambda^2 \) for all \( t \):

\[
\lambda^2_t = \sum_{t=0}^{m} \gamma_t(b) \lambda^*_s + t
\]

The vector of score parameters of the complete set \( \lambda^* \) is identified if all its elements occur in one of the linear transformations \( T^1 \) and \( T^2 \):

\[
\lambda^1 = T^1 \lambda^*
\]

\[
\lambda^2 = T^2 \lambda^*
\]

which equals that the total matrix of transformations in Equation 5 is of full column rank. The identification of the item parameters is discussed next.

1.1.2 Identification of item difficulty parameters. The identification of the difficulty parameters \( b \) in booklet \( x \) and \( c \) in booklet \( y \) in relation to the full set of item parameters \( b^* \) is straightforward given the identification of the score parameters \( \lambda^1 \) and \( \lambda^2 \). The relation can be directly observed in Equation 3 and Equation 4 where \( b \) and \( c \) are related through the score parameters \( \lambda^* \).
1.2 Simulating response patterns for data augmentation

Response patterns need to be simulated to perform data augmentation in the proposed application, where a sum score for the missing part is provided by the observed part, see Equation 11 in the main text.

To impute scores on the unobserved items, we make use of the sufficiency of the sum score $y_+$ in the RM:

$$ P(y|\theta) = \prod_{j=1}^{n} \frac{e^{y_j(\theta - \log(c_j))}}{1 + e^{\theta - \log(c_j)}} $$

$$ \downarrow $$

$$ P(y|y_+ = t) = \frac{\prod_{j=0}^{n} e^{-y_j \log(c_j)}}{\gamma_t(e^{-\log(c_j)})} \quad \forall \theta $$

Note that we introduce a regular RM notation here involving $\theta$ for efficient sampling. We utilize a method shown by Marsman, Maris, Bechger, and Glas (2017) to efficiently sample from conditional distributions. The main idea is, that if we keep sampling from the distribution in Equation 7 until we by chance simulate the sum score we need, then the simulated realization is distributed according to Equation 8, regardless of which value of $\theta$ was used for the simulation. Since the probability with which the needed sum score is generated depends on the choice of $\theta$, we choose it to be the maximum likelihood (ML) estimate corresponding to the imputed sum score.

We provide a simulation to demonstrate the performance of sampling response patterns. Response patterns with a score $t = 70$ are generated using an implementation based on the following pseudo-code:

$$ \hat{\theta} : \mathcal{E}(Y_+|\hat{\theta}) = y_+ $$

repeat
  $$ y \leftarrow P(y|\hat{\theta}) $$
until $$ y_+ = t $$

A sequence of 100 items, with item difficulties ranging between -2 and 2, is used, and $\hat{\theta}$ is estimated with ML estimation for $t = 70$.

The distribution of waiting time in iterations is illustrated in Figure 1 using 10,000 replications. The graph illustrates that the number of iterations required is limited due to the estimation of $\hat{\theta}$. In addition, iterations are computationally light, i.e., the average waiting time is about 15 ms on a mainstream laptop from 2010 using a script written in R (R Core Team, 2015).

References

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2017). Turning simulation into estimation: generalized exchange algorithms for exponential family models. PLOS ONE, 12(1), e0169787. doi:10.1371/journal.pone.0169787

R Core Team. (2015). R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from [http://www.R-project.org/]
Figure 1. Distribution of numbers of iterations required for obtaining a response pattern with sum score $t = 70$, using 100 items.