Appendix

1. Scoring stage: Technical details

Consider a test made up of \(j=1,2,\ldots,n\) items, and let \(x_i\) be the full vector of responses given by individual \(i\). We shall use the generic expression \(P(X_j|\theta_i,\sigma_i^2)\) to denote the conditional probability (discrete case) or conditional density (continuous case) assigned to a specific item score for fixed \(\theta_i\) and \(\sigma_i^2\). For the M-DTCRM the conditional density is normal, with mean and variance given in equation (17). For the M-DTGRM, the conditional probability is that given in equation (34).

The likelihood of \(x_i\) can then be written generically as

\[
L(x_i | \theta, \sigma^2) = \prod_{j=1}^{n} P(X_j | \theta, \sigma^2) \tag{A1}
\]

For a person parameter \(\xi (\xi = \theta_1, \ldots, \theta_m\) or \(\sigma_i^2\)), the EAP point estimate is the mean of the posterior distribution of \(\xi\) given the respondent’s item response pattern

\[
EAP = \hat{\xi} = E(\xi | x_i) = \frac{\int_{\theta} \int_{\sigma^2} \xi L(x_i | \theta, \sigma^2) g(\theta) h(\sigma^2) d\sigma^2 d\theta}{\int_{\theta} \int_{\sigma^2} L(x_i | \theta, \sigma^2) g(\theta) h(\sigma^2) d\sigma^2 d\theta} \tag{A2}
\]

The joint distribution of \(\theta_1, \ldots, \theta_m\) and \(\sigma_i^2\) is approximated by the product \(g(\theta_1)\cdots g(\theta_m) h(\sigma_i^2)\). By default, \(g(\theta)\) is taken as standard normal, while \(\sigma_i^2\) is scaled inverse \(\chi^2\) with \(d\) degrees of freedom and scaling parameter \(t\) (Novick & Jackson, 1974). Ferrando (2019) provides a complete discussion about the choices for this prior. From a practical point of view, however, the main choice entails balancing two opposite aims: (a) provide plausible estimates for all the respondents, and (b) avoid excessive tightness and regression toward the prior mean (which would imply that all the respondents receive a very similar PDD.
point estimate). Ample experience with the unidimensional modeling suggests that \( d = 5 \) and \( t = 3 \) is a reasonable choice.

As in the unidimensional case, the multiple integral in (A2) is approximated by using rectangular numerical quadrature. We found, that beyond 40 equally-spaced points, increases in accuracy were virtually inexistent, and this is in our experience the best choice. However, for solutions with more than 2 dimensions, particularly for the M-DTGRM to use 40 can be very time consuming.

2. Simulation Study

Random samples of \( N = 200 \), \( N = 500 \), and \( N = 1000 \) simulated responses were generated according to the M-DTCRM and the M-DTGRM for (a) three item-to-factor ratios: \( n = 14/f = 2 \), \( n = 21/f = 3 \) and \( n = 36/f = 3 \); and (b) two levels of IDDs: low (average \( \alpha \) value of 0.55) and high (average \( \alpha \) value of 0.70) by using MATLAB programs written by the authors. The factor solutions were oblique and independent-cluster, and the inter-factor correlations were all 0.40. In all cases, 50 replicas per condition were used, the distribution of the \( \theta \)'s was standard normal, and the distribution of \( \sigma^2 \) was inverse chi-square with \( t = 3 \) and \( d = 5 \). The \( \beta \) item locations were uniformly distributed between -1.5 and 1.5, and, for the M-DTGRM items were discretized in 5 response categories using Muthén & Kaplan’s (1985) thresholds for obtaining centred distributions. It should be noted, however, that the combination of the chosen \( \beta \) values with the standard thresholds gave rise in some cases to quite skewed item distributions.

First part: item calibration

The simulated responses were calibrated by fitting a multiple FA model with the correct number of factors (2 or 3) to the product-moment (M-DTCRM) or the polychoric
(M-DTGRM) inter-item correlation matrices by using ULS estimation. The direct canonical solution was next rotated using Browne’s (1972) semi-specified oblique target rotation. The recovery of the generating parameters was assessed with the Burt-Tucker congruence coefficient between the ‘true’ and estimated factor patterns. Goodness of fit was assessed with two statistics: the root mean squared residual (RMSR) and the GFI goodness-of-fit index (see McDonald, 1999). The results are in table 5 below.

The recovery of the ‘true’ solution in table 5 is consistently good, both in terms of congruence and in terms of goodness of fit. So, the possible differential trends are not very pronounced. In general, results are slightly better for the M-DTCRM, as expected (loss of information due to categorization). For both models, results systematically improve when item discrimination is high and as the sample size increases, so that for the largest samples and the high-discrimination conditions, the differences between the M-DTCRM and the M-DTGRM virtually vanish. These are also expected results. The size of the model, however, does not appear to have a systematic effect here.

Second part: individual scoring

For each simulee, EAP estimates of both $\theta$'s and $\sigma^2$ were obtained by using rectangular quadrature over 40 equally spaced points and the default prior distribution discussed above. The individual estimates were based on the ‘true’ structural solutions, and so, the scoring mimics the usual scenario in which the calibration results are obtained based on a large and representative sample and can be taken as fixed and known. The measures of accuracy in this case were the product-moment correlations between the individual estimates and the corresponding true values. With regards to the trait estimates, the single reported correlation is the average over factors. Results are in table 6 below.
Results in table 6 generally behave according to the theoretical expectations. For both \( \theta \)'s and \( \sigma^2 \) the accuracy increases with model size, item discriminating power, and, to a lesser extent, sample size. The same as in the previous study, accuracy is better for the M-DTCRM, as expected. However, in this case the differences between both models are much more pronounced.

In the conditions used in the study, the accuracy of the \( \theta \) estimates is acceptable in all cases, and, for the high item discrimination conditions and the two largest models can be considered good for both models. The accuracy of the \( \sigma^2 \) estimates, however, is clearly lower, as expected. Even so, the present results suggest that individual estimates of \( \sigma^2 \), which are accurate enough for practical purposes might be obtained from tests made up of 20 items with reasonably low IDDs (continuous model), or of 35 good items for the graded model.

New references used in the appendix

Browne, M. (1972). Oblique rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology*, 25, 207-212.

McDonald, R.P. (1999). *Test theory: A unified approach*. Mahwah: LEA.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, 38, 171-189.
TABLE 5

Results of the simulation study. Item calibration.

M-DTCRM

| Model size | 14/2 | 21/3 | 36/3 |
|------------|------|------|------|
| Sample size | 200 | 500 | 1000 | 200 | 500 | 1000 | 200 | 500 | 1000 |
| Item Disc | Low | 0.95 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.97 (0.01) | 0.99 (0.01) | 0.98 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.96 (0.01) | 0.97 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) |
| GFI | High | 0.98 (0.01) | 0.99 (0.01) | 0.99 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.97 (0.01) | 0.99 (0.01) | 0.98 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.96 (0.01) | 0.97 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) |
| z rmsr | Low | 0.04 (0.01) | 0.03 (0.01) | 0.02 (0.01) | 0.02 (0.01) | 0.04 (0.01) | 0.04 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.04 (0.01) | 0.02 (0.01) | 0.04 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.02 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.02 (0.01) |
| BT Congruence | High | 0.97 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.97 (0.01) | 0.98 (0.01) | 0.98 (0.01) | 0.95 (0.01) | 0.97 (0.01) | 0.97 (0.01) | 0.98 (0.01) | 0.95 (0.01) | 0.97 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) |

M-DTGRM

| Model size | 14/2 | 21/3 | 36/3 |
|------------|------|------|------|
| Sample size | 200 | 500 | 1000 | 200 | 500 | 1000 | 200 | 500 | 1000 |
| Item Disc | Low | 0.92 (0.01) | 0.96 (0.01) | 0.97 (0.01) | 0.99 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.97 (0.01) | 0.99 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.96 (0.01) | 0.97 (0.01) | 0.99 (0.01) | 0.96 (0.01) | 0.97 (0.01) | 0.99 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) |
| GFI | High | 0.96 (0.01) | 0.97 (0.01) | 0.99 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.97 (0.01) | 0.99 (0.01) | 0.98 (0.01) | 0.96 (0.01) | 0.97 (0.01) | 0.97 (0.01) | 0.98 (0.01) | 0.95 (0.01) | 0.97 (0.01) | 0.99 (0.01) | 0.96 (0.01) | 0.97 (0.01) | 0.99 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) |
| z rmsr | Low | 0.05 (0.01) | 0.03 (0.01) | 0.02 (0.01) | 0.02 (0.01) | 0.05 (0.01) | 0.05 (0.01) | 0.04 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.05 (0.01) | 0.05 (0.01) | 0.04 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.02 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.03 (0.01) | 0.02 (0.01) |
| BT Congruence | High | 0.95 (0.01) | 0.97 (0.01) | 0.98 (0.01) | 0.98 (0.01) | 0.92 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.97 (0.01) | 0.98 (0.01) | 0.98 (0.01) | 0.93 (0.01) | 0.96 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.97 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) | 0.95 (0.01) | 0.96 (0.01) | 0.98 (0.01) | 0.99 (0.01) |
TABLE 6

Results of the simulation study. Individual scoring.

**M-DTCRM**

| Model size | 14/2 | 21/3 | 36/3 |
|------------|------|------|------|
| Sample size | 200  | 500  | 1000 | 200  | 500  | 1000 | 200  | 500  | 1000 |
| Item Disc | Low | High | Low | High | Low | High | Low | High | Low | High | Low | High |
| \( r(\hat{\theta}, \theta) \) | 0.86 (0.01) | 0.88 (0.01) | 0.88 (0.01) | 0.92 (0.01) | 0.86 (0.01) | 0.90 (0.01) | 0.86 (0.01) | 0.91 (0.01) | 0.87 (0.01) | 0.91 (0.01) | 0.90 (0.01) | 0.94 (0.01) |
| \( r(\sigma^2, \sigma^2) \) | 0.60 (0.06) | 0.73 (0.04) | 0.62 (0.04) | 0.74 (0.02) | 0.66 (0.03) | 0.76 (0.02) | 0.68 (0.04) | 0.80 (0.02) | 0.71 (0.03) | 0.80 (0.02) | 0.72 (0.03) | 0.83 (0.02) |

**M-DTGRM**

| Model size | 14/2 | 21/3 | 36/3 |
|------------|------|------|------|
| Sample size | 200  | 500  | 1000 | 200  | 500  | 1000 | 200  | 500  | 1000 |
| Item Disc | Low | High | Low | High | Low | High | Low | High | Low | High | Low | High |
| \( r(\hat{\theta}, \theta) \) | 0.82 (0.01) | 0.84 (0.01) | 0.84 (0.01) | 0.89 (0.01) | 0.84 (0.01) | 0.90 (0.01) | 0.84 (0.01) | 0.90 (0.01) | 0.86 (0.01) | 0.90 (0.01) | 0.87 (0.01) | 0.92 (0.01) |
| \( r(\sigma^2, \sigma^2) \) | 0.49 (0.06) | 0.63 (0.06) | 0.50 (0.06) | 0.65 (0.05) | 0.55 (0.05) | 0.69 (0.04) | 0.56 (0.04) | 0.71 (0.03) | 0.57 (0.03) | 0.71 (0.03) | 0.62 (0.03) | 0.73 (0.02) |