Supplemental materials for "An item response theory analysis of the Matrix Reasoning Item Bank (MaRs-IB)"

Samuel Zorowitz[1], Gabriele Chierchia[3], Sarah-Jayne Blakemore[3], Nathaniel D. Daw[1,2]

[1]Princeton Neuroscience Institute, Princeton University, USA

[2]Department of Psychology, Princeton University, USA

[3]Department of Psychology, University of Cambridge, Downing Street, Cambridge, UK

Supplemental materials for "An item response theory analysis of the Matrix Reasoning Item Bank (MaRs-IB)"

Speed-accuracy trade-offs in Chierchia et al. (2019)

To investigate the possibility of speed-accuracy trade-offs in the MaRs-IB response data collected by Chierchia et al. (2019), we looked at the proportion of correct responses to the easiest items (dimension 1 & 2 items) as a function of the number of participants having reached that item. The logic is that, if items that appeared later in the fixed-order test were disproportionately reached by participants sacrificing accuracy for speed, then we should observe a positive correlation between the total number of available responses and proportion correct amongst the easiest items.

We detected strong positive correlations between proportion correct and number reached (dimension 1 items: $\rho = 0.514$, p = 0.050; dimension 2 items: $\rho = 0.767$, p $<$ 0.001; combined: $\rho = 0.695$, p $<$ 0.001). This result supports the hypothesis that participants that did reach items later in the test did so by prioritizing speed at the expense of accuracy. This suggests that the summary statistics of later items released as part of Chierchia et al. (2019) are likely biased indicators of item difficulty.

Defining a threshold for rapid guessing

In online testing environments, it is inevitable that some participants will not engage meaningfully with an experiment and instead in engage in careless or insufficient effort responding. On matrix reasoning tasks, one such low-effort strategy is rapid guessing wherein participants response in such a short time that there is no way they could have meaningfully considered an item (Wise, 2017). In sufficient quantities, the presence of rapid guesses in data can systematically bias estimates of item parameters. Thus, if possible, rapid guess responses should ideally be identified and removed.

There are a number of approaches for identifying rapid guess responses (for a review, see Wise (2017)). We opted for a threshold approach, in which responses taking less than a particular time would be denoted as rapid guesses and participants exhibiting too many rapid guessing responses would be excluded from the data. To

define this threshold, we fit an extended version of the effort-moderated item response theory (EM-IRT) model (Wise & DeMars, 2006) to a small dataset of responses collected during piloting. In the EM-IRT, the probability of correct responding for participant i to item j is defined as the following mixture:

$$p(y_{i,j} = 1) = (1 - w_{ij}) \cdot \gamma + w_{ij} \cdot \text{logit}^{-1}(\alpha_j \cdot \theta_i - \beta_j)$$

where θ_i is the latent ability for person i, and β_j, α_j, and γ_j are the difficulty, discrimination, and guessing parameters for item j. As in the main text, here we fixed the guessing parameter for every item clone to the nominal guessing rate ($\gamma_j = 0.25$). Crucially, w_{ij} is a weight parameter, bounded between zero and one, that controls whether a participant is responding effortfully in accordance with their ability ($w_{ij} \to 0$) or engaging in rapid guessing responding ($w_{ij} \to 1$).

Here we defined the rapid guessing weight as a function of participant's response time on that trial:

$$w_{ij} = \text{logit}^{-1}\left(\zeta_0 + z_{ij} \cdot \zeta_1 + z_{ij}^2 \cdot \zeta_2\right)$$

where z_{ij} is the (log-transformed) response time for participant i for item j, and ζ_n are regressing coefficients mapping responses times to rapid guessing weights. Thus, this form of the EM-IRT model learns a function in a data-driven fashion to classify responses as having originated from effortful or rapid guessing response strategies.

We fit this model to data collected from a total of N=180 participants recruited from the Prolific Academic platform as part of a pilot experiment (independent of the experiments presented in the main text). Each participant completed one of two sets of eight items from the MaRs-IB. The EM-IRT model estimated within a Bayesian framework using Hamiltonian Monte Carlo as implemented in Stan (v2.22) (Carpenter et al., 2017). Four separate chains with randomised start values each took 3,000 samples from the posterior. The first 2,000 samples from each chain were discarded. As such, 4,000 post-warmup samples from the joint posterior were retained. The $\hat{R}$ values for all parameters were equal to or less than 1.01, indicating acceptable convergence

between chains, and there were no divergent transitions in any chain.

The estimated weights (w_{ij}) across all subjects and items are plotted as a function of their corresponding response time in Figure S3. As can be observed, the weights quickly approach 1 for response times faster than 5 seconds. Interestingly, weights begin to rise again for responses after 20 seconds suggesting that participants have an internal awareness of the amount of time that has elapsed.

To define a threshold for rapid guessing, we found the corresponding response time for which $w = 0.5$. This was at approximately 3 seconds. Thus, in the main experiments we defined rapid guessing as responses taking fewer than 3 seconds.

Organization of geometric stimuli in the MaRs-IB

Every item template in the MaRs-IB has three unique versions that differ only in the geometric shapes populating its cells. There are 45 unique geometric shapes in total, organized into nine stimulus sets, where each item clone draws a subset of shapes from one of the nine sets (Figure S2). As such, shape set — indicating whether a clone is the 1st, 2nd, or 3rd version of an item template — is not a meaningful nominal variable. Furthermore, item clones drawing from the same shape set are not always populated by the same geometric shapes, making complicated the possibility of modeling the nine stimulus sets instead. The most rigorous way of modeling the presence or absence of a given geometric shape on item functioning would be to include each as a binary attribute predicting item difficulty and discrimination. In order to keep our models simple, we elected to leave shape set unmodeled and accounted for by the residual variability terms instead.

References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1), 1–32.

Chierchia, G., Fuhrmann, D., Knoll, L. J., Pi-Sunyer, B. P., Sakhardande, A. L., & Blakemore, S.-J. (2019). The matrix reasoning item bank (MaRs-IB): novel, open-access abstract reasoning items for adolescents and adults. *Royal Society open science*, *6*(10), 190232.

Chiesi, F., Morsanyi, K., Donati, M. A., & Primi, C. (2018). Applying item response theory to develop a shortened version of the need for cognition scale. *Advances in cognitive psychology*, *14*(3), 75.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: development of the subjective numeracy scale. *Medical Decision Making*, *27*(5), 672–680.

Iverson, G. L., Marsh, J. M., Connors, E. J., & Terry, D. P. (2021). Normative reference values, reliability, and item-level symptom endorsement for the PROMIS® v2. 0 cognitive function-short forms 4a, 6a and 8a. *Archives of Clinical Neuropsychology*.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated irt model. *Journal of Educational Measurement*, *43*(1), 19–38.

Model	psis-loco	Δ psis-loco (se)
1r	28376.7	635.3 (29.7)
1	27741.4	-
2r	27627.6	665.6 (32.2)
2	26962.1	-
3r	27218.7	633.1 (33.9)
3	26585.5	-

Table S1

Comparison of the first wave of item response models to their equivalents without a guessing parameter ($\gamma = 0$). LOO-CV values are presented in deviance scale (i.e. smaller values indicate better fit). Abbreviations: PSIS = Pareto-smoothed importance sampling; LOCO = leave-one-cluster-out.

			Spearman rank correlation			
Measure	Mean (SD)	IQR	NFC-10	PCF-8a	SNS	MaRs-SF
NFC-10	25.03 (8.27)	20.00 – 31.00	-			
PCF-8a	22.17 (6.29)	18.75 – 26.25	0.27**	-		
SNS	29.07 (7.48)	25.00 – 35.00	0.46**	0.29**	-	
MaRs-SF	8.00 (2.53)	6.00 – 10.00	-0.04	0.04	0.14*	-

Table S2

Correlations between performance on the MaRs-IB short forms and self-report measures. Abbreviations: IQR = interquartile range; NFC-10 = need for cognition (10-item) scale (Chiesi, Morsanyi, Donati, & Primi, 2018); PCF = PROMIS cognitive functioning scale (8a; Iverson, Marsh, Connors, & Terry 2021); SNS = subjective numeracy scale (Fagerlin et al., 2007). ** $p < 0.001$, * $p < 0.05$ (not corrected for multiple comparisons)

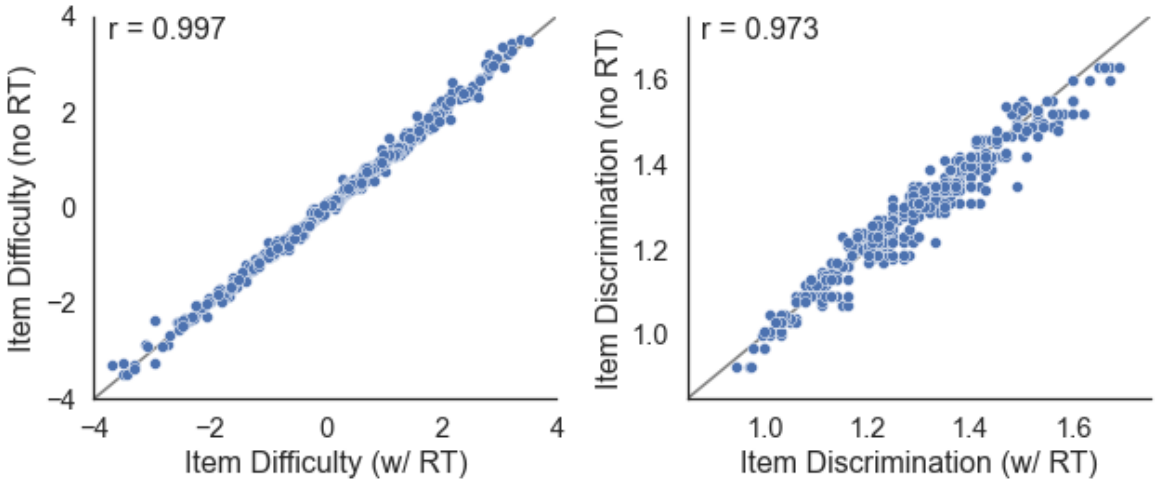

Figure S1. Item parameter estimates for the best-fitting model (Model 5) with and without including mean response time as a clone-level attribute.

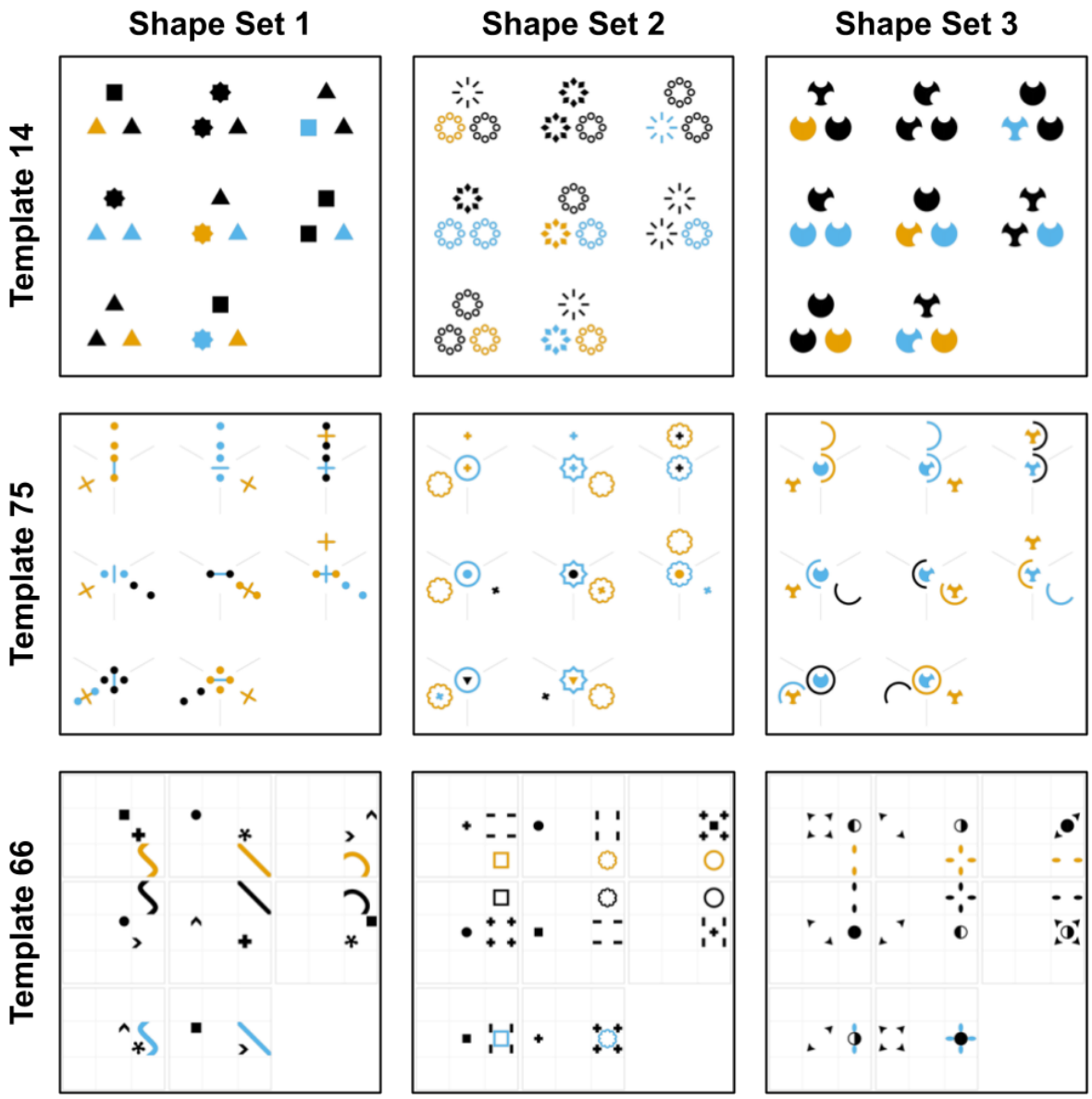

Figure S2. An example item clone from each of the nine stimulus sets.

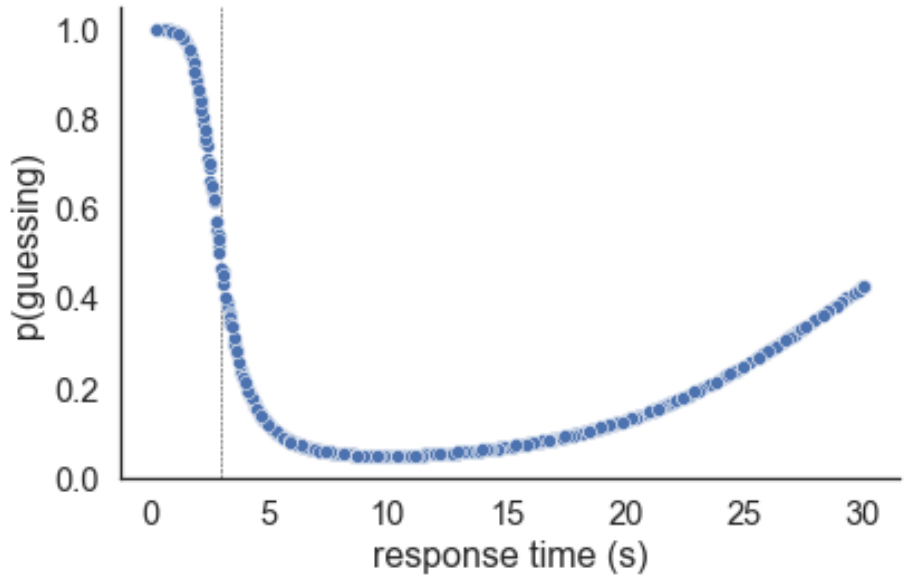

Figure S3. Estimates of the rapid guessing weights (w_{ij}) and their corresponding responses times from the effort-moderated item response theory (EM-IRT) model fit to response data from N=180 pilot participants. The dashed line indicates the chosen rapid guessing threshold, i.e. where $w = 0.5$.