Response to reviewers 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

Formatting note

Formatting note: in the document below, reviewers’ comments are indicated in bold typeface. Quoted text from the revised manuscript is presented in quotation blocks; within these excerpts, text that is new is indicated in italicized typeface. All changes to text in the revised manuscript are detailed below. Furthermore, changes to text in the manuscript have been highlighted (note: highlighted section headers indicate all text in that section is new or has been revised).

Reviewer #1

This paper proposes the validation of a matrix item bank to assess cognitive ability and provides three parallel forms with some evidence of reliability and convergent validity. For this purpose, it starts from a public matrix item bank of items with 4 response options, described in Chierchia (2019). The main contribution is that here a larger sample is used (1501 people vs. 659 in the original study), as well as a better application design (since in Chierchia’s study, the test was applied under time-limited conditions which led to a high presence of missing values in many of the items). The tested bank is composed of 64 tested item templates and 384 derived clones.

In general, I believe that this study can be useful for researchers by making available a calibrated tool for the evaluation of abstract reasoning. In addition, the results found concerning the prediction of difficulty in this type of item are in line

with previous studies. On the other hand, the paper is well written and the analyses, quite sophisticated, have been carried out correctly. It is also of particular value that the authors provide the code and scripts. The modeling of multilevel item structure, the application of item response theory, and the use of response times to improve the estimation of item parameters also add value.

However, I have some concerns regarding the tests generated, which I describe below.

First, several of the item response options have been developed using the "minimum difference" strategy. The use of this strategy poses a great potential threat to the validity of the test, since the correct option can be identified by looking at the answer options only (the correct answer is the one that is minimally different from the rest of the options). Even though this problem is not detected in this empirical sample, it is potentially an important problem. On the other hand, since the article evaluates the performance of different strategies in the construction of response options, it should devote more space to explaining these strategies (rather than referring to another paper). It is also striking that the final tests are composed almost exclusively of PD items, which would indicate that the MD items are far from optimal.

We agree with the reviewer that the validity of items with MD-type distractors would be compromised if participants solved them just by looking at the response options, and that this is an important issue to consider when evaluating these items. However, our results suggest that this is not a major issue in practice given that, in this dataset, those items are on average more difficult than items with PD-type distractors (as the reviewer themselves notes). Moreover, we suspect there is a simpler reason why our new test forms are primarily composed of items with PD-type distractors: when selecting items to maximize test information, all else equal, more difficult items contribute less information than easier items. This is because participants' responses to difficult items are more likely to be contaminated by guessing, and are thus less informative about participants' latent ability [1]. As a consequence, an optimal test assembly procedure maximizing test information (i.e. test reliability) is more likely to select the easier, PD-type item clones.

Regardless, we agree with the reviewer that it would be useful to include additional text describing these issues in the manuscript rather than referring readers to other articles. As such, we have expanded the methods to include the following text (page 9, first full paragraph):

> Importantly we also counterbalanced the assignment of item clones across participants, such that we had an approximately uniform number of responses available for each clone by shape set (1, 2, or 3) and distractor type (minimal difference, MD, or paired difference, PD). *Distractor type refers to the two strategies used to generate the distractor response options in the MaRs-IB. The MD strategy produces*

> *distractors that are variations of the target response. The MD strategy has the advantage of preventing pop-out effects (where the correct response "pops out" among the possible responses), but has the disadvantage of theoretically allowing participants to solve an item by looking at the response options only. In contrast, the PD strategy produces distractors that are have at least one component in common with the target response. The PD strategy prevents participants from solving items by looking at the response options alone, but can potentially induce pop-out effects. We note that, for complex items with many elements, it is not possible to prevent both at the same time.*

Second, the small sample sizes (considering the complexity of the models applied) are striking. Although the total sample is 1501 persons, the 64 item templates have been applied to subsamples of about 375 persons (a somewhat small size) and the item clones to samples of about 62 subjects (see page 9 of 54). Although auxiliary information is used in the calibration of item parameters (response times, item characteristics, etc.), this sample does not seem to be sufficient, and may certainly lead to low power for some contrasts.

We thank the reviewer for this important point and agree that we should have included further justification of our sample size and clear discussion about the implications of the sample size for the study's various goals and conclusions. We have therefore conducted a parameter recovery and power analysis, which we now report in the Supplementary Materials (see "Parameter recovery & power analysis" section beginning on page S6). The complete details of the methods and results of the analysis can be found there. In general, we find that the sample was adequately powered for the analyses underlying our positive conclusions (about associations between item characteristics and item difficulty) and, importantly, for guiding the assembly of reliable test forms; but less well-powered for testing associations between item characteristics and item discrimination (where we indeed had not reported positive results).

In slightly more detail, based on our study design, the new analyses found that our statistical models were capable of accurately estimating item difficulty parameters with high precision, but were able to accurately estimate item discrimination parameters with only adequate precision. Accordingly, we found that we were well-powered to detect moderate-to-large associations between item attributes and item difficulty, but only adequately powered to detect large associations (i.e. explaining 14% or more variance) between item attributes and item discrimination. The latter results have minimal impact on the main conclusions of our study. We were more than sufficiently powered to detect associations between item attributes and item difficulty; therefore, we can be confident in our conclusion that items clones differ systematically in difficulty by distractor type and are not therefore exchangeable. Furthermore, as we now demonstrate in the Supplementary Materials, the somewhat noisier estimation of item discrimination parameters leads only to a negligible loss in the reliability of test forms generated from our optimal test assembly procedure. Therefore, we believe our sample size and experiment design was sufficient for the primary objectives of the item calibration study.

We now point out the implications of these power considerations at appropriate points in the manuscript, and in particular stress the caveat that asking more detailed questions about item discrimination would require a larger followup sample. Concretely, in the *Goodness-of-fit* section of the methods for the Calibration Study, we have included the following paragraph (page 16):

> *As a final validation of the best-fitting model, we also performed a parameter recovery analysis. In this analysis, we generated 100 artificial datasets with sampling and statistical properties matched to what we observed empirically. That is, we generated datasets with 1500 participants and 384 item clones (nested in 64 item templates), where the item parameters were randomly generated and matched to the observed distribution of item parameters. We then fit the best-fitting model to each artificial dataset, and quantified the consistency between the ground-truth and recovered model parameters. The complete details of this procedure can be found in the supplementary materials.*

In the *Goodness-of-fit* section of the results for the Calibration Study, we have included the following paragraph (page 22):

> *Finally, we inspected the results of the parameter recovery analysis (the complete results are reported in the supplementary materials). Briefly, we found we were able to recover item difficulty parameters with excellent precision. Conversely, we observed only adequate recovery of the item discrimination parameters. The results may in turn explain why we detected only one credible association between item attributes and item discrimination. Although we were adequately-powered to detect large associations between item attributes and discrimination (i.e. explaining 14% or more variance in the latter), we are poorly powered to detect smaller associations. This was not true for item difficulty, where we were well-powered to detect associations of the magnitude reported above. Importantly, we found through follow-up analyses that the less-than-perfect recovery of item discrimination parameters yielded only negligible effects on the reliability of test forms produced using optimal test assembly procedures. Therefore, our analyses were sufficiently powered for the primary objectives of this study.*

Finally, the discussion of the Calibration Study, we have included the following sentences (page 23):

> We also investigated how item complexity shapes item functioning. We found that element number and rule number were both positively associated with item

> difficulty, the effects of which were of approximately equal magnitude (i.e. a one-unit change in either was independently associated with a roughly 10% reduction in performance). This finding is interesting in light of previous investigations of nonverbal reasoning tasks, which have found a greater influence on item difficulty from either the number of elements (Bethell-Fox, Lohman, & Snow, 1984) or number of rules (Mulholland, Pellegrino, & Glaser, 1980). One possible reason for this discrepancy is that these two attributes are largely uncorrelated across items in the MaRs-IB, which allows for unconfounded estimates of their effects. Together element number and rule number explained 67.6% of the variance in difficulty across templates (and 38.6% of the variance across item clones), which is in line with previous investigations of matrix reasoning tasks (P. A. Carpenter, Just, & Shell, 1990; Matzen, Van der Molen, & Dudink, 1994). These findings further validate the design of the MaRs-IB insofar that item complexity is a primary determinant of item difficulty. *Finally, we found that rule number, but not element number, was associated with item discrimination. However, this finding must be interpreted with caution as our parameter recovery analysis revealed our sample size was adequately powered to detect only larger effects (attributes explaining 14% or more variance in item discrimination). Future studies with larger samples will be required to more thoroughly investigate the relationship between item complexity and discrimination.*

Third, some norms should be provided to be useful for researchers. For this, the test should be applied to more representative samples. I see the author's analysis as a first step.

We wholeheartedly agree with the reviewer that populations norms for MaRs-IB test scores would be helpful in order for researchers to compare the outcomes of specific individuals or groups against the performance of the general population. However, we believe that providing population norms is outside the scope of the aims of the current article. Here, we are aiming to address a growing need for individual differences researchers in this current era of online convenience samples: a paucity of unique and reusable matrix-reasoning items for measuring cognitive ability. To address this need, we have provided an in-depth demonstration that the MaRs-IB is both psychometrically sound and ideal for designing parallel test forms for online convenience samples.

Regardless, we agree that representative normative samples would also be useful for many other reasons. Therefore we mention in the discussion that this is a limitation we should address in future work (page 31):

> The current investigation of the MaRs-IB is not without its own limitations. One notable limitation is our sample. Here we analyzed response data collected from an online adult sample which was relatively young and well-educated. We cannot guarantee that the psychometric properties of the MaRs-IB reported here will gen-

> eralize to other populations or testing contexts. Future researchers should consider replicating the current study in other populations of interest (e.g. children, clinical samples). By using item response models in this study, however, we make possible the opportunity for future IRT "linking" studies. IRT linking describes a set of methods to establish the comparability of item and ability parameters estimated from response data collected from two or more groups that differ in ability (Lee & Lee, 2018). Future studies might exploit these methods to provide new insights, not only in how the functioning of the MaRs-IB may differ in across populations, but also in how matrix reasoning ability changes across populations. *Future studies involving larger samples could also provide populations norms for MaRs-IB test scores, which would enable researchers to compare the outcomes of specific individuals or groups (e.g. clinical groups) against the performance of the general population (of a particular country or region).*

Fourth, in the validation study, three forms of 12 items are proposed. This seems a very small number of items if adequate psychometric properties are to be obtained. The authors report that they have used the ordinal alpha coefficient, which may be an overestimate of reliability as it does not estimate the reliability of the observed score (instead, provides the reliability of the sum of the continuous variables underlying the items; see Chalmers, 2008; Flora, 2020). Authors should report the Green and Yang coefficient (also provided by the semTools package). For example, the values of these coefficients (omega2 or omega3) are below 0.75 in several of the tests (in the first test, omega3 = 0.65). It would also be important to report the traditional measures of model fit for these three tests (RMSEA, CFI, etc.). The CFI for the third test does not seem adequate (i.e., it is lesser than 0.9). It seems a limitation that the proposed tests do not have fully satisfactory psychometric properties.

We thank the reviewer for this crucial feedback. After reading Flora (2020) [2], we agree that the ordinal alpha coefficient is not a measure of score reliability and is therefore not useful to report. We are not enthusiastic however, about instead calculating score reliability using the Green and Yang coefficient. The GY coefficient is calculated based on pairwise item tetrachoric correlations using a probit response function. This neglects that our data are contaminated by guessing (i.e. item response functions are characterized by a nonzero lower asymptote). In our view, it is more appropriate to calculate and report the IRT test reliability coefficient [3, 4], which is based on the empirical item response functions. Based on the IRT test reliability coefficient, we find that the three short-form measures possess only adequate score reliabilities ($\rho > 0.7$). Thus, the reviewer was right to suspect that we overestimated the reliability of the MaRs-IB short-form measures. The main text has been updated to reflect this.

Given that each of the parallel short-form measures possess less-than-desirable score reliability lower, we have designed two 24-item long-form measures. Based on the IRT test reliability coefficient, these long-form measures exhibit acceptable score reliability ($\rho > 0.8$). We then subjected these two measures to the same empirical test as the short-form measures. That is, we recruited an additional N=300 participants to complete the two long-form measures and the abbreviated Raven's progressive matrices measure. The methods and results of this additional study have been incorporated into the Validation Studies section (starting on page 24).

The reviewer also suggests reporting model fit indices for all the MaRs-IB test form measures. We agree with the reviewer that such information is important to include in the manuscript. In doing so, however, we look for an alternative to the suggestion of reporting traditional (SEM-based) measures of model fit (e.g. RMSEA, CFI, TLI). This is because a number of studies have demonstrated the limitations of these indices for categorical data [5], especially for evaluating models involving dichotomous data [6], small numbers of participants, and few degrees of freedom [7]. Accordingly, we instead rely on an established posterior predictive model check: the χ^2_{NC} discrepancy measure, which compares the observed and model-predicted proportion of participants at each total score level [8]. Using this discrepancy measure, we find that the distribution of observed test scores for each test form is not credibly different from model predictions using the item parameter estimates from the calibration study (all posterior predictive p-values > 0.05).

In sum, based on the reviewer's excellent feedback, we have improved the psychometric evaluation of our test form measures. We have replaced estimates of reliability using the ordinal alpha coefficient with the more appropriate IRT test reliabiltiy coefficient. In response to the lower-than-anticipated reliability of the 12-item short form measure, we have designed and in an additional study validated two 24-item long-form measures with good reliability ($\rho > 0.8$). Thus, we provide researchers with two possible choices of MaRs-IB test forms that balance reliability and administration time trade-offs. Finally, we demonstrate that the observed score distributions of each test form conforms to expectations based on the results of the item calibration study.

Minor issues:

1) The prior distributions used should be reported.

We thank the author for this point. We have now described the priors for the additive multilevel item structure (AMIS) models in the supplement (beginning on page S4). Moreover, we have added a sentence to the methods to alert readers to this section (page 15).

2) In the test assembly, performed by linear programming, the constraint is that the alpha coefficient should be greater than or equal to 0.8, but the reported values are slightly less than 0.8.

We thank the reviewer for this point. We have updated the text to reflect the revised reliability estimates (page 25):

> *Each short form was required to contain 12 items. This was chosen to minimize the administration time of a given form (2-4 minutes on average) while achieving a score reliability* ≥ 0.7.

Reviewer #2

I think the paper is of high quality. However, I found it difficult to understand for readers. I suggest Authors to simplify some parts. Thanks.

We thank the reviewer for the positive review. We have attempted to clean up and simplify the manuscript in parts. For example, we have simplified and shortened our description of the additive multilevel item structure models. We have also rewritten and streamlined our reporting of the second (i.e. validation) study.

References

1. Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement* **28,** 989–1020 (1968).
2. Flora, D. B. Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science* **3,** 484–501 (2020).
3. Kim, S. & Feldt, L. S. The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review* **11,** 179–188 (2010).
4. Nicewander, W. A. Conditional reliability coefficients for test scores. *Psychological Methods* **23,** 351 (2018).
5. Reise, S. P., Cook, K. F. & Moore, T. M. in *Handbook of item response theory modeling* 31–58 (Routledge, 2014).
6. Clark, D. A. & Bowles, R. P. Model fit and item factor analysis: Overfactoring, underfactoring, and a program to guide interpretation. *Multivariate behavioral research* **53,** 544–558 (2018).
7. Kenny, D. A., Kaniskan, B. & McCoach, D. B. The performance of RMSEA in models with small degrees of freedom. *Sociological methods & research* **44,** 486–507 (2015).
8. Sinharay, S., Johnson, M. S. & Stern, H. S. Posterior predictive assessment of item response theory models. *Applied Psychological Measurement* **30,** 298–321 (2006).