Data for the Matrix Reasoning Item Bank fits the dichotomous Rasch model. This may be useful for (1) further internet based research aiming to include measures of matrix reasoning, and (2) further developing psychometric tools for use in internet based research.

Preliminary Rasch Analysis of the Matrix Reasoning Item Bank

S.B. Galvin [1]
@S_B_Galvin
shane.galvin@ucc.ie
R. Murphy [1]
[1] *School of Applied Psychology, University College Cork*

Introduction

Abstract reasoning is the ability to abstract relationships between observable stimuli into inernalised concepts which are generalised to understand relationships between new stimuli. Extensive research has shown abstract reasoning (often measured via matrix reasoning tasks) to be a highly g-loading construct, strongly related to a number of cognitive processes (Kane et al., 2004).

The ubiquity of internet based research, and prevalence of open science research methods signifies a need for an open use, psychometrically calibrated matrix reasoning test with (1) clear usage procedures, and (2) well-understood measurement properties which may help to allay frequently observed measurement issues in psychology (see Flake & Fried (2020)). In particular, scale shortening can be supported using computerised adaptive testing (CAT) administration, e.g. Harrison et al. (2017). As an alternative to a large CAT item bank or parallel test forms, Explanatory Item Response Models (EIRM) (De Boeck & Wilson, 2004) allow for the automatic writing of Rasch compliant items (Gierl & Haladyna, 2012). This would assist with deployment of procedures in internet based testing with less concern for test security and scale validity threats. However, extension into advanced modeling and test administration requires the fitting of a measurement model prior to explanatory modeling.

The present study is a step within a larger project that aims to detail the measurement properties and task structure of the Matrix Reasoning Item Bank (MaRs-IB) (Chierchia et al., 2019); a free-to-use, automatically generated bank of matrix reasoning items. This step aims to assess the psychometric properties of the MaRs-IB via Rasch analysis (Andrich, 1988; Rasch, 1960) to enable further validation research using EIRMs and extension into applied domains.

Method

Participants: 485 participants completed an online procedure containing several demographic variables, a digits span forwards test, and a selection of 45 MaRs-IB items. Of the 485 participants, 443 participants remain; with 42 participants removed due to inappropriate response patterns, person misfit, or 0% / 100% score rate.

Ethics: Ethical approval was sought from and approved by the UCC school of Applied Psychology Research Ethics Committee. To allow for participants retaining agency over their data they were permitted to end the procedure at any time, and their data would not be stored. Participants were recruited using convenience and snowball sampling.

Design: The first 5 MaRs items were administered as practice items, with the remaining items presented in a pseudo-randomised order. Items were presented alongside 4 multiple choice response options, where the graphical position of each response options were also randomised. Items were delivered with a 30 second item-wise time limit. Of the 40 MaRs items scheduled for Rasch analysis (excluding practice items) 19 items remain.

Rasch Analysis was conducted using Conditional Maximum Likelihood (CML) estimation as this conditions out the effect of the person ability distribution from item parameter estimation, and most closely represents the idea that the raw score should be a sufficient statistic for scaling items (Fischer & Molenaar, 2012).

Three comparison variables; device type (mobile vs desktop), biological sex (Female vs Male), and age group (age ≥ 44, age < 44); were used for conducting goodness of fit via Andersen's likelihood Ratio Test (Andersen, 1973) (LRT), and differential item functioning (DIF) using a Wald Z test.

Analysis of the relative fit of the dRM against a 2 parameter Item Response Theory (2pl IRT) model was also conducted. Bayesian Information Criterion (BIC) and Akaike's Information Criterion (AIC) where used to identify which model possessed a better fit to the data, where lower relative values indicate better fit.

Software: Stimuli were presented using jsPsych (De Leeuw, 2015) (v 6.1.0). The procedure was administered to participants over the internet, who then used their own device and web browser. Statistical analysis was conducted using R (R Core Team, 2021), and Rasch Analysis via CML was conducted using the eRm package (Mair & Hatzinger, 2007) (v 1.0-2). Rasch and 2pl IRT models were estimated via Marginal MAximum Likelihood (MML) estimation using the TAM package (Robitzsch et al., 2021) (v 3.6-45).

Results

Descriptive Statistics: Participant scores were approximately normally distributed with a mean of 22.21, and sd 7.03. A summary visualisation of demographic variables are presented in figure 1.

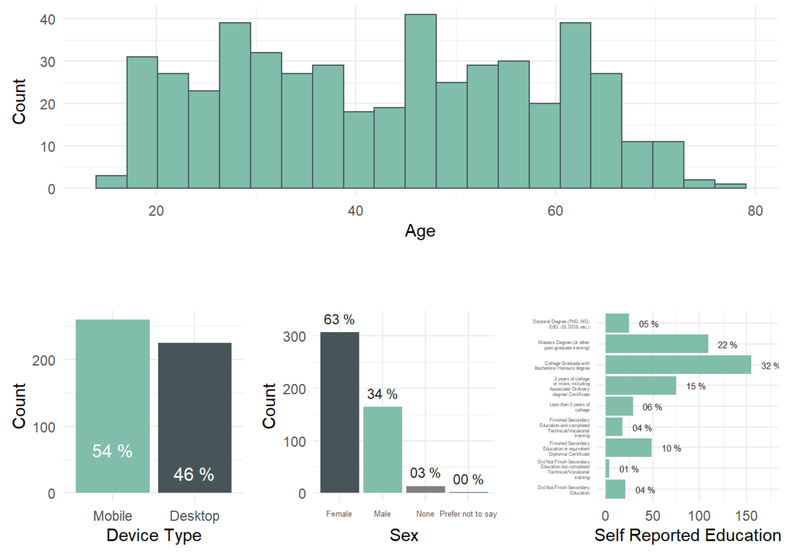

Figure 1: Sample descriptive statistics. Including Age (median = 44), Device type which appears to be evenly distributed between mobile and desktop devices, biological sex where Female participates make up the majority of the sample, and self-reported level of education where ~32% of the sample posses a bachelors degree.

Goodness of fit for the dichotomous Rasch model (dRM) was assessed via Andersen's LRT, no invariance comparisons exhibited high χ^2 values (table 1), indicating that the items fit to the model well.

Table 1: Goodness of Fit - Andersen LR test

Comparison Group	χ^2	df	p
Median Score Split	20.287	18	0.317
Device Type (mobile and desktop)	13.120	18	0.784
Sex (male and female)	16.636	18	0.548
Age (over and under 44 yrs)	9.143	18	0.956

Item difficulty ranged from -2.008 for the easiest item, and 1.977 for the most difficult item (β column, table 2). The test difficulty was 0.006 on the logit scale (Test information = 4.172, Test target ability = 0.029) (fig. 2), resulting in an adequate separation reliability (SepRel = 0.7), and internal consistency (Cronbach's coefficient α = 0.728).

Figure 2: Wright map for the study data. The histogram represents the estimated distribution of person ability, with the average and 95% SE shown underneath. The Points in the bottom half of the plot represent CML estimated item difficulty along the θ scale.

Items were examined for DIF, and no instance of DIF was observed across the comparison groups in the final batch of items (fig. 3) .

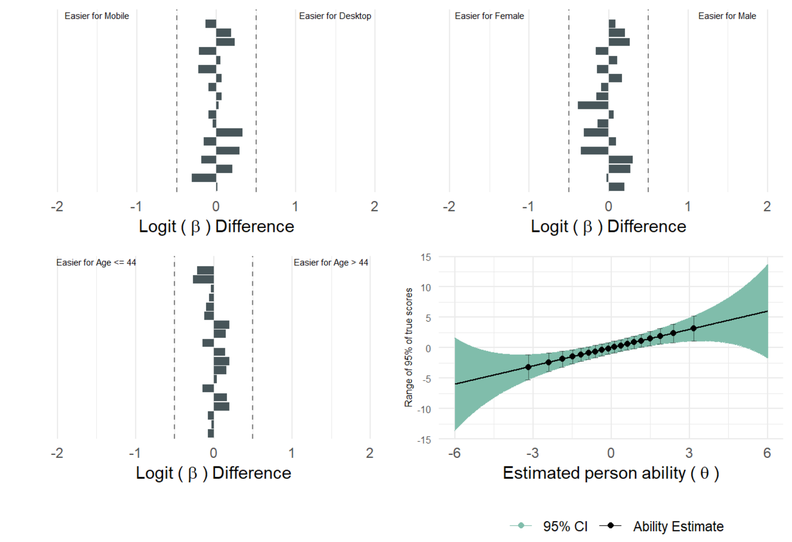

Figure 3: Differential Item Functioning with the Wald Z test. No item in the comparison groups appeared to be biased in favour of either subgroup and no item exhibits a difficulty difference ≥ 0.5 logits. This is shown alongside the precision of the ability estimates (95% SE) for the sample (black features) and across the hypothetical θ range (green error band 95% SE).

Table 2: dRM - Item Fit Statistics

						Outfit		Infit	
Item	β	SE	χ^2	df	p	MSQ	z	MSQ	z
06	-0.001	0.100	469.041	440	0.155	1.066	1.223	1.013	0.360
09	-2.008	0.132	346.049	440	1.000	0.786	-1.524	0.933	-0.811
12	1.977	0.141	323.353	440	1.000	0.735	-1.809	0.922	-0.835
17	0.414	0.102	441.201	440	0.462	1.003	0.065	0.982	-0.414
19	-0.125	0.100	463.556	440	0.202	1.054	1.002	1.030	0.803
22	-0.859	0.104	403.576	440	0.886	0.917	-1.152	0.914	-2.012
23	-0.406	0.100	413.049	440	0.808	0.939	-1.077	0.966	-0.879
25	-0.997	0.106	433.606	440	0.564	0.985	-0.159	1.033	0.729
27	0.370	0.102	448.431	440	0.367	1.019	0.338	1.030	0.712
28	-0.115	0.100	432.331	440	0.581	0.983	-0.310	1.000	0.005
29	0.262	0.101	472.902	440	0.127	1.075	1.294	1.017	0.440
30	0.370	0.102	424.407	440	0.683	0.965	-0.578	0.975	-0.582
35	0.305	0.101	415.123	440	0.788	0.943	-0.970	0.966	-0.833
39	0.083	0.100	463.169	440	0.205	1.053	0.972	1.069	1.763
45	0.177	0.100	392.882	440	0.944	0.893	-1.982	0.936	-1.632
69	-0.053	0.100	405.085	440	0.876	0.921	-1.509	0.937	-1.698
70	-1.044	0.106	448.563	440	0.366	1.019	0.265	0.990	-0.199
75	1.080	0.112	439.823	440	0.480	1.000	0.026	1.010	0.191
80	0.569	0.104	453.598	440	0.305	1.031	0.481	1.027	0.616

The differences between β parameter estimates for the estimated models were minimal, and the correlation between all estimates was high ($\rho(17) > .98$) (table 3).

Table 3: Pearson correlation matrix and p-values between estimation method parameter estimates

	Correlation		P value	
	CML	MML	CML	MML
2pl	0.9887	0.9892	p < 0.0001	p < 0.0001
CML		0.9999		p < 0.0001

The MML dRM and 2pl IRT model were compared using analysis of variance; where lower goodness of fit index values (AIC and BIC) indicate a better fit for the Rasch model (table 4). At this point it is reasonable to conclude that the data adequately fit to the CML Rasch model.

Table 4: Comparison of MML Rasch and 2pl IRT models

Model	Log Likelihood	Deviance	Npar	AIC	BIC	χ^2	df	p
dRM	-5128.278	10256.56	20	10296.56	10378.43	18.39	18	0.43
2pl IRT	-5119.083	10238.17	38	10314.17	10469.72			

Discussion

The fit of MaRs-IB items to the dRM confer several advantages over standard approaches to test validation. While the number of scaled items seems low; it is a good starting point for further research to test the design structure of MaRs-IB items via EIRMs; thereby providing an account of how item design features relate to item difficulty. As the current batch of MaRs-IB items stems from a set of item-model designs, producing many parallel test forms, comparison test forms should be tested for measurement invariance to further validate the parent item models. If measurement invariance can be demonstrated across the items in MaRs-IB test forms, then the existing MaRs-IB could be used with minimal concern for test security.

Limitations: The range of participant ability was a limiting factor in the analysis of the remaining items; as a result most of the fitted items are clustered closely together in terms of difficulty, leading to a limitation of abilty estimation at the tails of the scale distribution (Estimated person ability plot, fig. 3) If one is to plan research in populations with extreme scores; new items should be constructed that may fit the Rasch model. Alternatively, a sampling approach to collect data on extreme scoring participants with the existing items may be beneficial.

Further Research: The next phase of this project aims to fit an EIRM (e.g. the linear logistic test model (Fischer, 1973)) to the project data based on the matrix reasoning rule taxonomy developed by Carpenter, Just, and Shell (1990).

References

Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, *38*(1), 123–140.
Andrich, D. (1988). *Rasch models for measurement* (Vol. 68). Sage.
Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the raven progressive matrices test. *Psychological Review*, *97*(3), 404.
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.
De Boeck, P., & Wilson, M. (2004). *Explanatory item response models*.
De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.
Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374.
Fischer, G. H., & Molenaar, I. W. (2012). *Rasch models: Foundations, recent developments, and applications*. Springer Science & Business Media.
Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465.
Gierl, M. J., & Haladyna, T. M. (2012). *Automatic item generation: Theory and practice*. Routledge.
Harrison, P. M., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports*, *7*(1), 1–18.
Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189.
Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*. http://www.jstatsoft.org/v20/i09
R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.*
Robitzsch, A., Kiefer, T., & Wu, M. (2021). *TAM: Test analysis modules*. https://CRAN.R-project.org/package=TAM

UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh | School of Applied Psychology
IRISH RESEARCH COUNCIL An Chomhairle um Thaighde in Éirinn