National Institute of Mental Health

Grant # R01MH118514

Item Response Theory Analyses of Matrix Reasoning: Towards a New Short Form or Adaptive Test?

Reise, S.R.[1,] Widaman, K.[2,] Bauer, R.M.[3], Draine, D.L.[4,5], Loring, D.[4], Umfleet, L.[6], Wahlstrom, D.[7], Enriquez, K.[1,8], Wong, E.F.[1], Hubbard, A.S.[1], & **Bilder, R.M.**[1,8]

[1]UCLA Department of Psychology, [2]UC Riverside Department of Psychology, [3]University of Florida, [4]Emory University, [5]Emory University School of Medicine, [6]Medical College of Wisconsin, [7]Pearson Clinical Assessment, [8]Semel Institute for Neuroscience and Human Behavior at UCLA

UCLA

Abstract

An Item Response Theory (IRT) 2-Parameter Logistic Model (2PL) was applied to N = 550 responses to the WAIS-IV Matrix Reasoning Test (MR) drawn from a clinical sample. Our primary goal was to explore the effects of shortening MR using a simulated computerized adaptive testing (CAT) strategy. We found that using an algorithm that administered at most 10 items (adaptively) and stopping the test when the standard error was below .4, CAT and full-length trait estimates correlated .97, suggesting examinee relative ordering remains the same, and 70% of the sample required only 5 items to reach the stopping criterion. The average standard error was .36 and .28 for CAT and full-length, respectively.

Introduction

- Wechsler Adult Intelligence Scale (WAIS): one of the most widely used batteries for assessing general intelligence and often included in neuropsychological assessment.
- The Matrix Reasoning (MR) subtest is one of the core subtests in the Perceptual Reasoning Index, a major supplementary dimension of the WAIS
- The present study aims examine and explore:
  1) MR from an item response theory (IRT) perspective
  2) potential for item reduction using computer adaptive testing (CAT)
- The present study leveraged nationwide data collected by the National Neuropsychology Network (NNN).
- NNN aims to transform clinical practice to digitalizing tests and coordinating with numerous clinics across the U.S. to contribute data to the NIMH Data Archive (NDA).

Methods

- Analyzed data collected by the NNN, N = 550 (complete cases).
- We used the WAIS-IV. The MR subscale consists of 26 items.
- All analyses conducted in R (v. 3.6.3) and excluded first 5 items.
- IRT analyses: "mirt" (v. 1.32.1); estimated 2 parameter logistic (2Pl) model.
- Computer Adaptive Testing (CAT) simulations: "Firestar" (v. 1.9.2); used real data CAT simulation, set minimum items administered = 5 and SE <= 0.40 stopping rule. The most informative item was administered first, the next item selected was based on maximum information, and expected a posteriori trait scoring (EAP) was used.

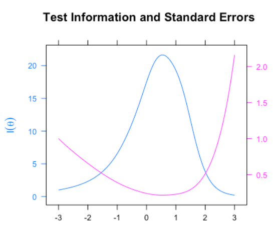
Classical Test Theory and IRT Statistics

Table 1 Proportion correct (first row), item-test correlations (second row), IRT-estimated slopes (third row), and IRT-estimated locations (fourth row).

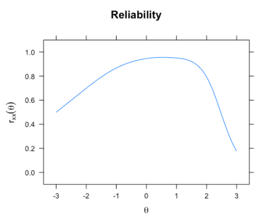
Item	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
%	0.92	0.90	0.76	0.79	0.70	0.73	0.64	0.65	0.59	0.50	0.49	0.37	0.49	0.34	0.36	0.33	0.36	0.20	0.17	0.11	0.08
r	0.34	0.35	0.53	0.56	0.67	0.57	0.64	0.73	0.65	0.66	0.72	0.65	0.75	0.62	0.65	0.66	0.66	0.54	0.52	0.44	0.38
Slope	1.17	1.07	1.34	1.64	2.19	1.53	1.83	2.81	1.94	2.15	2.83	2.49	3.45	2.5	2.78	3.4	3.03	3.12	3.36	3.65	3.29
Location	-2.55	-2.44	-1.1	-1.19	-0.59	-0.91	-0.43	-0.37	-0.22	0.11	0.15	0.49	0.19	0.58	0.52	0.59	0.51	0.97	1.05	1.25	1.46

IRT

A



B



C

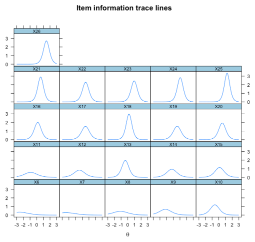


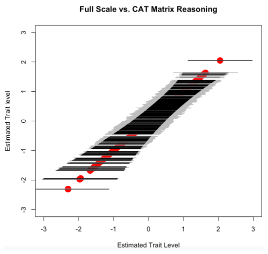
Fig A This figure shows the "psychometric information" (i.e. discrimination ability) and standard errors as function of the latent trait (theta; standardized metric) based on all items.

Fig B This figure shows the approximate "reliability" of trait level estimates as a function of the latent trait (θ; standardized metric). This figures helps to translate the IRT information results to a more commonly understood metric. The reliability is above .80 from around θ = -1 to +2.

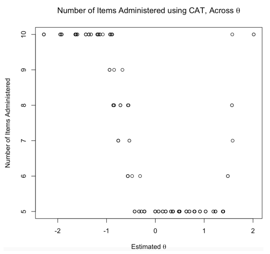
Fig C This is a graph of the item information curves for each item; More information mean the items is more discriminating. The first couple of items are not very discriminating, and it appears that the items become more discriminating as they get more difficult.

Computer Adaptive Testing (CAT)

D



E



F

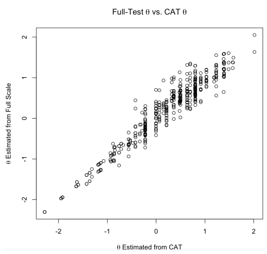


Fig D The black confidence bands for each trait level estimate based on scoring all items using the IRT model. Confidence bands are narrower for trait ranges where there is more information. The gray confidence bands are based on simulated computer adaptive testing algorithm (EAP scoring) item administration; these are, of course, wider, but not by much. The average standard error was .36 and .28 for CAT and full-length, respectively.

Max items administered = 10
- 70% required only 5 items, 20% required all 10 items.
- CAT: 13% had SE's ≥ .45

Fig E This is a graph of estimated trait level and number of items administered in the simulated CAT to reach the criterion. For people with trait level estimates between -.5 and 1.2 (roughly) only 5 items were needed (because that is where the test is most informative). As trait levels get higher or lower, more items are needed to achieve the criterion. When trait levels are very low, the criterion is not reached even after 10 items.

Fig F A scatterplot of theta estimated using full test vs. CAT (*r* = 0.976).

Discussion

- When fit to an IRT model, MR items are ordered from "easy" to more "difficult" in a way consistent with the design of the test (as judged by item proportion correct or location parameters).
- Under an IRT framework the MR test has highly "peaked" information indicating better measurement in the middle of the trait range.
- Simulated CAT results based on patient response patterns suggest that either a short form test, or a CAT, can achieve similar results to those achieved through standard administration practice.