RAVING ABOUT RAVENS: MODELLING SPEED-ACCURACY IN INTELLIGENCE TESTS

Diana Eugenie Kornbrot

Psychology Department, University of Hertfordshire d.e.kornbrot@herts.ac.uk.

Abstract

The effect of time pressure on performance on intelligence tests is a long standing problem. In this study a computerised version of the Ravens Advanced Progressive Matrices was administered using 3 different forms of instructions: control, speed pressure, and accuracy pressure. Analyses used Rasch measures of participant ability and item difficulty, and the time each participant took to solve each problem. Raw scores were, surprisingly, more useful than Rasch measures. The time pressure group were faster but scored less well than the other two groups. Raw score had a small but significant correlation with total test time. Brighter participants took less time for easy items, but more time for hard items, which were both slower and more variable than easier items. Mean and SD were more consistent for total time than for either correct or error time. Effective models will need to incorporate these diverse results.

The Raven's Progressive Matrices is one of the most successful culture fair tests of cognitive functioning (Raven, 1956; Raven, Court, & Raven, 1988; Salthouse, 2005). Like all tests and examinations, performance is potentially subject to time pressures. The first, and most practical aim, of this study is to make use of computerised administration to investigate those time pressures at the item level. A more theoretical aim is to produce a model of performance for this complex task that parallels models of simple perceptual decision making. The aim is to generate for each individual one parameters that corresponds to rate of accumulation of useful information, and one parameter that corresponds to speed bias (like the separation of barriers in a random walk). There have been many studies comparing performance on elementary tasks such as simple and choice reaction time with performance on some version of the Raven's. Typically, such studies use mean reaction time and d' or Luce's choice, ln(η) for the elementary processes, but only raw score for the Raven's (Beh, Roberts, & Prichard Levy, 1994; Fink & Neubauer, 2001; Salthouse, 2005). Even total time, easily obtained with a stopwatch, is rarely used as a peformance measure for the Raven's. Studies looking at time for individual items, requiring computerised administration are even rarer. There are however, several studies that analyse Ravens results using the Rasch model (Alderton & Larson, 1990; Forbes, 1964; Gallini, 1983; Green & Kluever, 1992; Pitariu, 1986). The two or three parameter Rasch models give participant ability and item difficulty measures that depend on logits of probabilities and are hence very similar to the bias and sensitivity measure of choice model.

The present study measures time per item for each person for each item. It explores four measures of individual person performance: raw score, and one, two and three parameter ability. The relation between time per item for the 36 items and individual performance is explored using exploratory principal components analysis. The relation between item difficulty (proportion of people giving correct response or one, two or three parameter item difficulty) to time per item was also explored.

Method

Participants & Design

60 female and male participants, aged from 18 to 72 were recruited from the university population. Each was randomly allocated to one of 3 groups: control, time pressure or accuracy pressure.

Apparatus and Materials

The apparatus was a computerised version of the Ravens© Advanced Progressive Matrices, comprising 12 practice and 36 test items, presented on a MAC computer. The instructions preceding the item administration included the following text presented on the computer screen.

> This is a test of observation and clear thinking. On each of the screens that follow you will see a pattern with a piece cut out of it. Look at the pattern. Think what piece is needed to complete it correctly both along and down must look like. Then find the right piece out of the eight bits shown below. When you think you have found the right piece, click on the piece with the mouse. Your selected answer will be displayed alongside the problem number on the right-hand side of the screen. If you make a mistake or want to change your answer, click on the appropriate piece and your answer will be updated. Select each problem from the list displayed on the right-hand side of the screen by clicking on the appropriate number. You can complete the problems in any order you like. However, the problems are simple at the beginning and get harder as you go along. There is no catch. If you pay attention to the way the answers to the easy problems are found you will find the later ones less difficult.
>
> PRACTICE PHASE
>
> There are 12 practice problems. Try each in turn. The first problem will be demonstrated for you. Then complete the 11 remaining practice problems. When you have completed all 12 problems click the 'Finished' button and await further instructions from the experimenter.
>
> TEST PHASE
>
> *Group specific instructions* preceding the following text to all participants
>
> You will be presented with 36 problems. Do not miss any out. If you are unsure of your answer, guess as guesses are sometimes right. If you get stuck, move onto the next problem and then come back to the one you had difficulty with. When you have completed all 36 problems click the 'Finished' button and inform the experimenter.
>
> GROUP SPECIFIC INSTRUCTIONS
>
> Control Group,1: *None*
>
> Time Pressure Group 2: You are required to perform a general intelligence test. We are interested in how successfully you can complete the test. Success will not only be measured by the number of correct answers obtained, you will also receive credit for the speed with which you complete the task. The participant with the highest overall score based on the number of correct responses <u>and</u> the time to complete the task will receive a prize.
>
> Accuracy Pressure Group, 3: You are required to perform a general intelligence test. We are interested in how successfully you can complete the test. Success will be measured by the number of correct answers obtained. The participant with the highest overall score based on the number of correct responses will receive a prize

Procedure

Each participant signed the consent form and was then seated in front of the computer and informed of the task, as describe in the reference manual. This was followed by computerised presentation of instructions followed the practice and test administration. A conspicuous clock was present on the screen throughout the time pressure condition. There was a pause for participants to relax and ask questions between the practice and test phases of the experiment.

Results

One, two and three parameter Rasch analyses were performed on the frequency of a correct response by each participant to each item. Each analysis gave an item difficulty score or each item and an ability score for each participant. The performance of these measures can then be compared with raw score for each participant and proportion correct for each item. All statistical tests were conducted at the 95% confidence level, with confidence levels in parentheses, as appropriate.

Group and Participant Performance

Table 1 shows means and standard deviations for test score for all 3 groups. ANOVA gave a significant effect of condition for both score, F(57,2) = 8.1, p = .001, effect size partial η^2 = .221 and total time, F(57,2) =3.3, p = .043, partial η^2 = .10. Since there was a predicted order for both score and time such that time pressure < control < accuracy pressure, post hoc tests were conducted without multiple comparison correction. With this proviso, the time pressure group was significantly less accurate than the other two groups, and also took significantly less time,. There were no significant differences between the control and accuracy pressure groups.

Table 1
Mean & standard deviation for raw Raven's score and total time to complete the test for all group

	control N=20		time pressure N=21		accuracy pressure N=19	
	mean	sd	mean	sd	mean	sd
test score/36	23.5	4.3	20.0	5.8	26.2	4.4
total time, mins	52.6	16.0	42.1	13.2	54.6	20.0

Alternative performance measures gave similar effect sizes to those obtained for the raw score. Partial eta squared = .224 for logit(probability correct), .209 for probability correct times logit(probability correct); and .198 for information favouring correct response. Group differences did not show up at all using ability measures from any of the Rasch models.

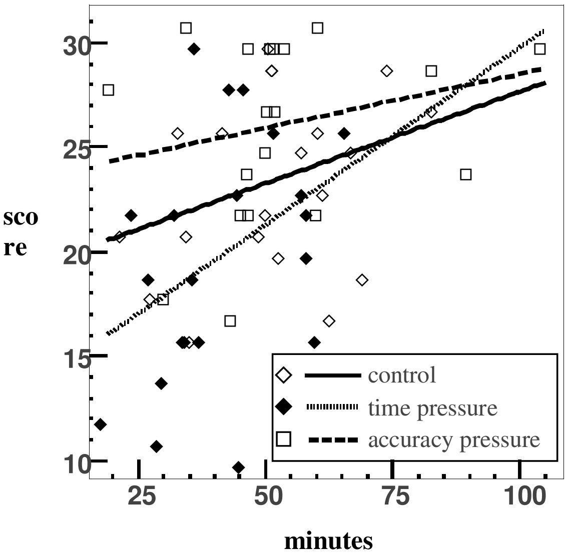

Figure 1. Raven's score as a function time to complete test for 3 groups.

Figure 1 shows score as a function of time to complete the test for all 3 groups. An ANCOVA with score as response variable and time, condition and time*conditions as explanatory variables gave a main effects of condition, F(2,54) = 8.6, p = .001, partial η^2 = .24; and time, F(1,54) = 5.6, p =.022, partial η^2 = .09; but no significant interaction term. The model with only main effects gave a slope of .090 confidence limits (.014, .166) and intercepts for control = 18.8 confidence limits(16.7, 20.9); time pressure = 16.2 with confidence limits(14.1, 18.3); accuracy pressure = 21.3 confidence limits(19.3, 23.4). So participants scored .9 points for each extra 10 minutes spent on the problem. Regression of the Rasch ability measures on total time in minutes were also conducted. Results were similar to those for raw scores but accounted for slightly *less* variance, so they are not shown

The relation between measures of correctness and time spent on each separate question was investigated by performing exploratory principal components analysis with time spent on each of the 36 problems and some measure of ability. Four separate analyses were conducted using the 4 possible measures of ability, total score, ability1, ability2 and ability3, where the digit following ability indicates the number of parameters in the generating Rasch model. The results were essentially identical for all ability measures. Table 2 shows the varimax rotated loadings for the solution using total score as the ability measure. Two components accounted for 41% of the variance. The first loaded *positive* on score and more than .45 on time spent on the more *difficult* items 19 onwards. The second loaded *negatively* on score and more than .45 on time spent on the *easier* items 1-17. Exceptions were problems 8 & 18, loading nearly equally on both components. It appears that abler participants spend less time on easy items but more time on difficult items.

Table 2. *Rotated principal components loadings for score and time to complete each problem*

Variable	score	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18
Component 1	.64			.34		.36			.48										.50
Component 2	-.48	.51	.54	.62	.64	.60	.72	.69	.43	.62	.70	.66	.38	.40	.60	.53	.48	.62	.49

Variable	T19	T20	T21	T22	T23	T24	T25	T26	T27	T28	T29	T30	T31	T32	T33	T34	T35	T36
Component 1	.50	.63	.33	.43	.71	.40	.67	.68	.70	.73	.61	.78	.66	.79	.57	.54	.67	.65
Component 2				.36														

Item Performance

The item difficulty level could be measured in 5 ways: by problem number (the original validation has problems in order of difficulty), by proportion of participants getting correct response, Pcor, and by the 3 different Rasch difficulties. Figure 2 shows mean of all times and sd off all times as function of number correct, together with sd as a function of mean. Each graph has 36 points generated by the 36 different items. Adjusted r^2 were as follows: regression of mean on number correct, r^2 = .820; for regression of SD on number correct, r^2 = .822; for SD on mean, r^2 = .937.

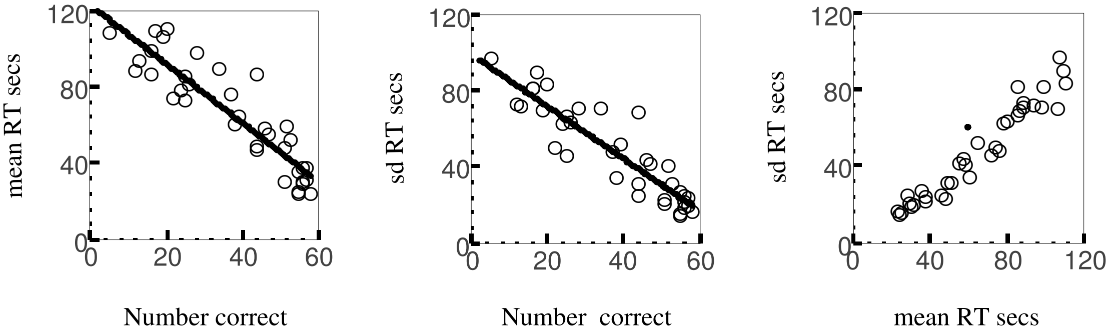

Figure 2. Mean RT as function number correct (left panel); and SD RT as function number correct (middle panel); SD RT as a function of mean RT (right panel).

The mean and s.d of total time, time for correct response and time for error response for each item was regressed separately on each of the five measures of item difficulty. Higher linear correlation coefficients were found for total time than for either correct or error response time. Furthermore there was no significantly difference between mean error and mean correct times. The highest correlations were obtained for regression on number correct for both mean and standard deviation. Means and standard deviations were highly correlated.

Summary

The general level of performance is much a would be anticipated for a university community (Bors & Stokes, 1998; Bors & Vigneau, 2001; Paul, 1985). There was a quite substantial effect of time pressure on both performance and total time taken. However, there was no difference between the standard administration control group and the accuracy pressure group. This is encouraging suggesting that left to themselves people do indeed opt of the accuracy strategy. Nevertheless, in the time pressure group only 1 person took more than an hour whereas in the other two groups combined 11/39, i.e. more than 25% took more than 1 hour. Standard administration is often limited to 1 hour. Clearly this can cause underestimation of performance.

There was a small positive correlation between time and performance, the same for all groups. Participants' performance was better by nearly 1 whole point for each extra 10 minutes spent. However, the direction of causality is unknown. Brighter participants may choose to take longer.

The results of the principal components analysis are new and interesting. Factor analysis of Raven's item themselves shows only a single factor. Nevertheless, abler participants are faster on easy items, but slower on difficult items. This may explain some of the relation between total time and raw score. The less able participants may guess on the difficult items, which is less time consuming than problem solving.

The relation between item difficulty, as measured by participant success rate, to time per item is also interesting. People spend longer on the more difficult items, but performance is also more variable as shown in Figure 2.

In this study raw scores outperformed Rasch measures of ability on all fronts. Raw person ability scores showed larger group difference effect sizes, stronger correlations with total reaction time and more variance accounted for in principal component analysis. Raw item difficulty measures showed stronger correlations with time per item than Rasch difficulty measures. This is good news for simplicity of analysis. There is a very simple message. Stick with raw scores.

However, the poor performance of the Rasch measures is disappointing, and in my view surprising, from the perspective of modelling performance. I had anticipated that Rasch measures, similar to parameters of choice model would be more closely related to time measures. From this perspective it is interesting that total time is a more reliable measure than either correct or error time, as would be predicted by relative judgement theory. Unfortunately this might be an artefact of the limitation that some error or correct times were based on small numbers of observations, while total time was always based on all 36 items. In addition, information measure of performance were just as bad as the (related) Rasch measures in terms of variance accounted for in the various analyses.

In summary, Raven's matrices remain an excellent tool for assessing cognitive ability. Raw scores seem better than any transformation. Time pressure, i.e. a fixed time limit is not a good idea. Modelling of processes involved in solving the Raven's has a very long way to go. Such modelling will need to take into account the new finding that more able participants are only faster on easy items. For harder items, they appear to strategically increase the time spent to accommodate the fact that more information is needed to solve the problems.

Acknowledgements

Thanks to Campbell Thomson & McLaughlin Ltd, who were extremely helpful and permitted the production of the computerised version of the test. Raven's Advanced Progressive Matrices Test. Raven Progressive Matrices are stored digitally for this study with the permission of J.C. Raven Ltd. Advanced Progressive Matrices © John C. Raven 1943, 1947, 1962. © J.C. Raven Ltd 1976. All rights in Raven Progressive Matrices are reserved, and they may not be reproduced copied down-loaded stored in a data base or published in whole or in part in any form except with the express permission of the copyright owners. Enquiries about copyright in the RPM may be made to Campbell Thomson & McLaughlin Ltd, 1 King's Mews, London WC1N 2JA, tel: 020-7242-0958, fax: 020-7242-2408. Thanks also to Jed Everitt who programmed the computerised version of the test and to David Wellsted who tested the main bulk of participants.

References

Alderton, D. L., & Larson, G. E. (1990). Dimensionality of Raven's Advanced Progressive Matrices items. Educational and Psychological Measurement, 50, 887–900.

Beh, H. C., Roberts, R. D., & Prichard Levy, A. (1994). The relationship between intelligence and choice reaction time within the framework of an extended model of Hick's Law: A preliminary report. Personality and Individual Differences, 16(6), 891-897 Record 812 of 813 in PsycINFO 1992-1994.

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. Educational and Psychological Measurement, 58(3), 382-398 Record 388 of 313 in PsycINFO 1995-1997.

Bors, D. A., & Vigneau, F. (2001). The effect of practice on Raven's Advanced Progressive Matrices. Learning and Individual Differences, 13(4), 291-312.

Fink, A., & Neubauer, A. C. (2001). Speed of information processing, psychometric intelligence: And time estimation as an index of cognitive load. Personality and Individual Differences, 30(6), 1009-1021 Record 1006 of 1013 in PsycINFO 1998-1999.

Forbes, A. R. (1964). An item analysis of the Advanced Matrices. British Journal of Educational Psychology, 34, 223-236.

Gallini, J. K. (1983). A Rasch analysis of Raven item data. Journal of Experimental Education, 52(1), 27-32.

Green, K. E., & Kluever, R. C. (1992). Components of item difficulty of Raven's matrices. Journal of General Psychology, 119(2), 189-199.

Paul, S. M. (1985). The advanced Raven's Progressive Matrices: normative data for an American university population and an examination of the relationship with Spearman's g. Journal of Experimental Education, 54, 95–100.

Pitariu, H. (1986). Item analysis and standardization of advanced progressive matrices (APM) / Analiza de itemi si standardizarea matricilor progresive avansate (MPA). Revista de Psihologie, 32(1), 33-43.

Raven, J. C. (1956). Progressive Matrices, sets A, B, C, D and E. London.

Raven, J. C., Court, J. H., & Raven, J. (1988). Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4: Advanced Progressive Matrices, sets I and II. (1988 ed ed.). Oxford: Oxford Psychologists Press.

Salthouse, T. A. (2005). Relations Between Cognitive Abilities and Measures of Executive Functioning. Neuropsychology, 19(4), 532-545 Record 532 of 513 in PsycINFO 2004 Part B.