A Flexible and Customizable Method for Assessing Cognitive Abilities

Andrea Civelli*
University of Arkansas

Cary Deck
University of Alabama
Economic Science Institute, Chapman University

Abstract

This paper describes the properties of a set of puzzles that are behaviorally similar to those of the common Raven Progressive Matrix test. Our puzzles consist of a three-by-three grid of images with the lower right element omitted. Each image is characterized by six characteristics that can vary along several patterns. Lab experiments demonstrate that the puzzles become more challenging as the number of characteristics that change increases. Further, the ability to correctly solve our puzzles is shown to be correlated with scores on the Raven Progressive Matrix test and with performance in a beauty contest game. Due to the manner in which our puzzles are constructed, there are a large number of unique puzzles that can be generated for use in economics experiments using software described in the paper. Thus our puzzles are well suited for use as an alternative method to assess the cognitive ability of respondents and for use as a real effort task with multiple levels of cognitive difficulty.

Keywords: Cognitive Abilities Tests, Raven's Matrices, Experimental Economics Tools
JEL Classification: C9, C90, C91.

*Corresponding author: University of Arkansas, Business Building 402, Fayetteville, AR 72701. Email: andrea.civelli@gmail.com. URL: http://comp.uark.edu/∼acivelli/.
We thank Justin LeBlanc and Diego Calderon-Rivera for the excellent research assistance they provided to us in developing the experiment and the interface pf the software application.

1 Introduction

The Raven's Progressive Matrices test (RPM) is a measurement strategy for analogical reasoning and deduction abilities of an individual, which are related to her analytical intelligence (see Raven, Court, and Raven, 1998; Carpenter, Just, and Shell, 1990). For this reason, the RPM is being widely applied in Psychology, Behavioral Economics, and Neuroscience research when an assessment of the subjects' cognitive ability is desired. For example, Carpenter, Graham, and Wolf (2013) show that people who perform well on the RPM are strategically more sophisticated in that they give better responses in a beauty contest game and deviate less from the best response to their stated beliefs in such games. Benito-Ostolaza, Hernandez, and Sanchis-Llopis (2016) also show that those who score highly on the RPM behave more strategically while Burks, Carpenter, Goette, and Rustichini (2009) report that RPM performance is correlated with taking calculated risks, social awareness, and job perseverance. Al-Ubaydli, Jones, and Weel (2016) report that pairs with higher average RPM scores are better able to sustain cooperation. Duttle (2015) finds that high a RPM score is associated with less overconfidence from better calibrated interval forecasts. In experimental asset markets, Cueva and Rustichini (2015) finds that higher ability individuals earn greater profits and that higher ability groups exhibit less market volatility.

The RPM consists of visual problems where the respondent must identify what image completes a given pattern. The full RPM involves 60 tasks varying from very simple patterns to highly complex ones. One of the advantages to this type of procedure is that it is relatively easy to explain and free of context, language and culture, making it an ideal tool for use in economics experiments. However, the relatively small number of matrices of a given difficulty level limits its usefulness in some important ways. First, given the popularity of the procedure, some respondents are likely to have previously seen the specific puzzles being used.[1] Second, the RPM cannot finely partition people of similar, but distinct, abilities the way it could if it involved many many puzzles calibrated to respondent ability. Additionally, there are an insufficient number of matrices of a given difficulty level to use them as a real effort task despite the matrices being well suited to such use.

The shortage of matrices in the RPM led Matzen, Benz, Dixon, Posey, Kroger, and Speed (2010) to develop algorithms to generate tasks that are similar to the RPM tasks.[2] This paper reports an effort similar to Matzen, Benz, Dixon, Posey, Kroger, and Speed (2010) in that it generates puzzles that have behavioral properties like those of the RPM and makes those puzzles freely available to researchers. To be clear, we do not see our work as proposing an alternative test of cognitive ability. By contrast, Condon and Revelle (2014) develop a public-domain alternative approach to measuring cognitive ability that uses a limited number of RPM like tasks along with several other techniques such as three dimensional rotation and face detection.[3]

Specifically, our approach:

[1] Familiarity with the specific tasks that are used in the RPM may partially explain why scores on such tests are increasing over time. Dubbed the Flynn Effect, Flynn (1987) has documented increasing IQ scores over time across several countries.

[2] In personal correspondence, Matzen, Benz, Dixon, Posey, Kroger, and Speed (2010) indicated the software they developed is no longer available for distribution.

[3] See https://icar-project.com/ for details.

1. Allows researchers to measure cognitive ability without relying on a small set of specific puzzles.

2. Allows researchers to finely partition respondents by incorporating more puzzles of a given difficulty level calibrated to the average ability in the group.

3. Provides researchers a real effort task that can be calibrated to each specific subject.

4. Provides researchers with a set of uniform real effort tasks for use in situations where subjects can select among difficulty levels.

The paper proceeds in three steps. First, we describe the characteristics of the proposed puzzles and we link the degree of difficulty of the puzzles to combinations of these characteristics. We follow Matzen, Benz, Dixon, Posey, Kroger, and Speed (2010), who analyze the types of underlying relations that appear in original Raven's matrices, to specify and develop the structure of our puzzles. Second, we use a series of lab experiments conducted at the University of Arkansas and at Chapman University to provide an accurate calibration of subject performance as a function of the characteristics of the matrix and the solution set. Third, we verify that the proposed puzzles rely upon the same cognitive characteristics as the RPM. This is achieved both by a within-subject comparison of performance on the proposed matrices and the RPM and by a replication of Carpenter, Graham, and Wolf (2013) with the proposed matrices substituted for the RPM.

2 The Matrix Puzzles

The puzzles are 3×3 graphical matrices in which each entry contains an image with specific attributes. The bottom right image in the matrix is hidden by a question mark. Subjects must understand the patterns followed by the attributes of the images and identify the correct image to complete the puzzle from a pool of given options. An example puzzle is given in panel (a) of Figure 1. Panel (b) of the same Figure illustrates an example of a pool of solution options.

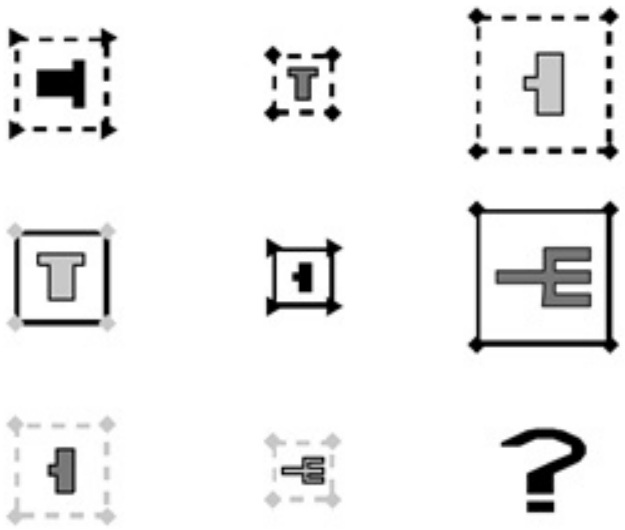
The characteristics of a matrix are defined by the number of varying attributes of the images and the patterns along which these attributes change. There are six attributes of an image: (1) shape, (2) size, (3) shade of the filling, (4) orientation, (5) border style, and (6) corner marker style. There are six schemes of patterns as well. These can be divided in groups of two each: (a) orthogonals - along rows and columns, (b) diagonals - along main or counter-diagonal, and (c) corners - from NW to SE and from SW to NE. These patterns are illustrated in panel (c) of Figure 1.

The difficulty level of a matrix is defined by the number of attributes allowed to change; hence we have up to 6 levels, although this could be extended. Each attribute can take one of four possible values, except for the shape which has 15 values. As shown by Figure 1 panel (c), each scheme re-arranges the attributes in three sets of the same type over the nine elements of a matrix. The example shown in Figure 1 is a level 6 puzzle and the correct solution is f in panel (b) as the shape varies from the NW corner to SE corner, the size varies across the columns, the shading varies along the main-diagonal, the orientation varies

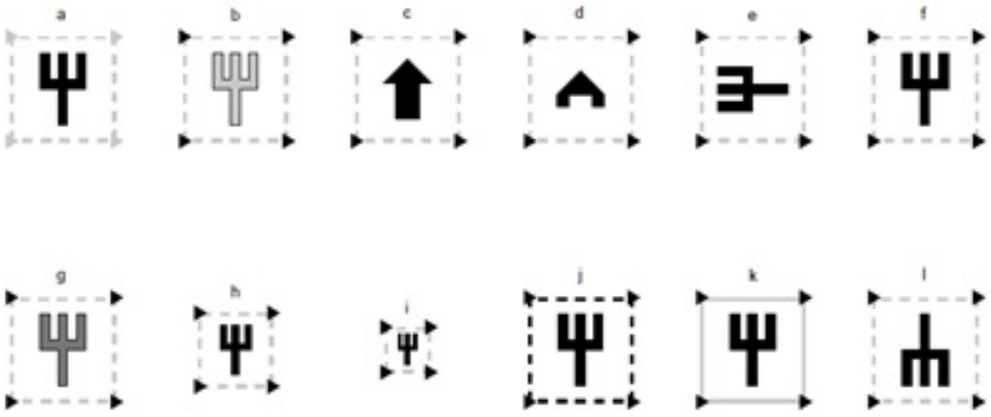
Figure 1: The Matrix Puzzles

An example of matrix puzzle and a possible solution set.

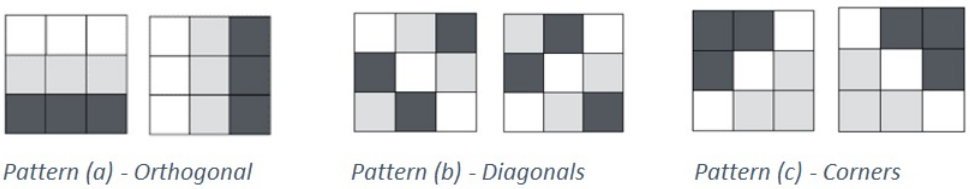
(a) Puzzle



(b) Solution set



(c) Patterns schemes



along the counter-diagonal, the border style varies by the row, and the border corners vary from the NE corner to the SW corner.

The number of multiple choice options and the construction of the incorrect responses can also contribute to a puzzle's difficulty. We vary the number of multiple choice options from 4 to 12, but ultimately this aspect of the problem has only marginal impact on performance. We also vary how the incorrect responses are generated, which is found to have a significant effect on subject performance. In the remainder of the paper we use the notation x_y to indicate a matrix with x varying attributes and y options in the solution set. The individual images shown in Figure 1 can be represented by a 6-element vector, one for each attribute. Wrong responses are created by starting with the vector of the correct answer and randomly modifying some of its attributes. We consider two versions of this procedure. In the first version, which we refer to as basic, only one attribute of the correct answer is randomly modified. In the second version, a mutation of the correct solution in which two attributes are randomly modified is first generated and included in the options. Then, the remaining options of the solution set are created by randomly selecting either the correct solution or the mutation and then modifying two randomly selected attributes. We refer to this version as mutation. As shown in Section 4, problems with the second version of the solution set have a surprisingly higher rate of correct responses.

The matrices are generated via a simple and intuitive software tool, which allows the user to quickly generate a very large number of matrices with specified characteristics. The user can select the level of difficulty of the matrix, the attributes to vary, and which groups of schemes they will follow. The software then randomly picks the three values that each varying attribute will exhibit and then randomly matches attributes to the patterns. Because the software is extremely flexible and matrices are highly customizable, researchers can generate distinct sets of matrices to be employed in repeated tasks during an experiment (e.g. to use in a real effort task with varying levels of effort cost), in addition to basic assessment of subjects' abilities. Further, the difficulty of the tasks can be precisely measured and easily modulated with respect to the subject's level of sophistication. The matrix generation software is briefly discussed in Appendix A.[4]

3 Experimental Procedures

Three main separate experiments, numbered chronologically, were conducted. In all three experiments, undergraduate subjects were recruited from the given lab's standing subject pool for a 90 minute session. In each case, subjects received a participation payment of $7 plus salient earnings. For our puzzles, subjects were paid $0.50 per correct answer with the ordering of difficulty level, number of options, and specific images characteristics randomized. We also discuss some results from additional data collected in two additional experiments, which were designed for independent purposes but employed our matrices to measure the cognitive ability of the subjects. Table 1 summarizes the characteristics of these experiments.

Experiment 1 was conducted at the University of Arkansas. These 17 subjects read instructions as shown in Appendix B and then completed 50 tasks of varying difficulty and

[4]The software is freely available subject to the license agreement from the authors' web-page where additional details are provided.

Table 1: Experiments Used in the Paper

Main features and differences between the five lab experiments from which the empirical analysis of the paper is obtained.

Experiment	Location	Subjects	Our Puzzles	Other Tasks
Experiment 1	University of Arkansas	17	50	None
Experiment 2	Chapman University	40	40	Carpenter, Graham, and Wolf (2013): Beauty contest
Experiment 3	University of Arkansas	36	30	Bilker, Hansen, Brensinger, Richard, Gur, and Gur (2012): 9 question version of RPM
Lee, Nayga, Deck, and Drichoutis (2018)	University of Arkansas	120	21	Full 60 question RPM and second price auction
Deck, Jahedi, and Sheremeta (2017)	Chapman University	120	5	Risk, allocation, and math problems under cognitive load

number of options, with options generated using the basic method described previously.

Experiment 2 was conducted at Chapman University. These 40 subjects were first given a paper handout that contained a version of the beauty contest game and a survey. The beauty contest game closely follows that of Carpenter, Graham, and Wolf (2013). The subjects were placed in groups of 10 and asked to guess a number between 0 and 20, with the person guessing closest to half of the average receiving a $10 payment. Subjects could also earn a $10 payment by having the most accurate prediction of the other nine guesses as well. The survey consisted of two (unpaid) questions from Carpenter, Graham, and Wolf (2013) for the Hit15 game and basic demographic questions. A copy of this handout is also provided in Appendix B. Next, subjects read the same directions as in Experiment 1 except that subjects were informed they would answer 40 matrix problems. For this experiment, wrong answers were generated using the mutation method discussed above.

Experiment 3 was conducted at the University of Arkansas. These 36 subjects completed both the Bilker, Hansen, Brensinger, Richard, Gur, and Gur (2012) 9 question version of the RPM and 30 of our puzzles, again using the same directions as in Experiment 1 updated for the number of tasks. As is common, the RPM score did not impact a subject's payment. Half of our puzzles used the basic method for generating options and half used the mutation method. Finally, the order of the RPM test and our puzzles was varied.

Additional data is obtained from two other experiments. One was conducted at the University of Arkansas. This experiment is the basis of a separate paper examining the relationship between cognitive ability and bidding behavior (see Lee, Nayga, Deck, and Drichoutis, 2018). These 120 subjects completed the full 60 question RPM and 21 of our puzzles in which the basic solution method is used. The RPM score did not impact a

subject's payment, and our puzzles were always the last activity the subjects experienced in this study.[5] The other additional source of data is an experiment conducted at Chapman University involving a separate group of 120 subjects who faced our puzzles along with various other tasks including allocation games and decisions under uncertainty while subject to different techniques of inducing cognitive load, the results of which are reported in Deck, Jahedi, and Sheremeta (2017). In this experiment puzzle performance was incentivized, but only easy and moderately difficult puzzles were used.[6]

Our procedures for administering our puzzles differ from the standard application of the RPM in several ways. For one, the RPM is typically not incentivized when used by psychologists as a measure of cognitive ability whereas following the traditional approach of experimental economics, the subjects in our main experiments were incentivized. Second, the RPM is typically administered in stages with increasing degrees of difficulty, whereas our subjects completed one stage with the difficulty level randomized over the full set of tasks.[7] Finally, the RPM is often administered with a time limit, whereas our subjects did not face a time limit. These changes in part reflect the way in which we expect that our puzzles will be useful to experimental economists. Further, that the results, discussed below, indicate that the two sets of tasks yield similar patterns despite these procedural changes is strong evidence that our puzzles are an expansion of the set of tasks in the RPM.

4 Puzzle Performance

We focus on two main aspects of performance on the puzzles: how the performance of the subjects varies as a function of the degree of difficulty of the matrices and the differences between the two methods of generating options.

Panel (a) of Figure 2 reports the average percentage of correct answers by level of complexity of the puzzles in Experiment 1; the average time to solve a puzzle is shown in panel (b) of the same Figure. In this experiment we also compare puzzles with the same number of varying attributes, but different numbers of options in the solution set. We find the puzzles are gradually more difficult for the average subject as the degree of complexity of the matrix increases. Similarly, it takes the subjects longer to solve a more complicated problem. This is a desirable property of the task in that it helps identify cutoffs for the subjects' types. We also observe that the number of options has a less pronounced impact on the performance of the subjects, and it is usually less decisive for the difficulty of the puzzles than an increase in number of varying attributes.

[5]The first session of this experiment did not include our puzzles. Our puzzles were simply tacked onto this other study already employing the RPM, which is why the order is fixed.

[6]The puzzle directions used in Deck, Jahedi, and Sheremeta (2017) differed from the other experiments. Their directions simply stated "The fourth type of question is a pattern recognition task. For each task, you will be shown a 3x3 table and asked to identify the missing element that completes the pattern of shapes. There are 8 multiple choice answers to choose from." Thus the directions do not need to be extensive, although we do not test the effects of different directions or the degree to which the puzzles can be understood by less literate populations. However, we suspect that our puzzles and those of the RPM are equally easy to explain to a given population since the same set of directions could be used for either task.

[7]As pointed out by a reviewer, it is possible that the detailed instructions helped subjects deal with the random ordering of puzzle difficulty in our experiments.

Similar results are found in Figure 3 for Experiment 2, which uses the mutation methodology to construct the solution sets. The percentage of correct solutions falls as the matrix complexity increases (panel a) and response time increases (panel b). However, the deterioration is not as pronounced as in Experiment 1. Because the first two experiments vary both the subject pool and the method for generating options, we rely upon the results of Experiment 3 to disentangle these possible causes. In fact, this separation was one of the main motivations for conducting Experiment 3.[8]

We quantify the effects of attributes and options more formally with the regression estimates reported in Table 2, where the left panel refers to Experiment 1 and the right one to Experiment 2 (basic and mutation options respectively). In both cases, the effect of the number of varying attributes on the accuracy of the solutions is negative and statistically strongly significant. On the contrary, the effect of the number of options is negative, but very small and not significant. For the logit model, we report the estimates of the odds ratios for which a value smaller than 1 indicates a negative effect. For instance, the .469 coefficient estimated for "Attributes" in the first logit model in Table 2 tells us that increasing the matrix difficulty level by 1, while holding the number of options constant, would cut the odds of a correct response roughly in half (53%). The odds only drop by 34% in Experiment 2, an effect smaller by a third than in Experiment 1.[9]

The effects for the basic solution methodology are about 2 times bigger than those for the mutation options. The same analysis holds for the response time, except for the coefficient on the number of options in Experiment 1, but not Experiment 2, which is now significant. Despite affecting response time, however, the number of options does not affect the degree of difficulty of the puzzles. This result is also consistent with the results of analyzing learning, which we now discuss.

Figure 4 provides insight into subject learning over the course of the experiment. In the two panels of this figure we compute the average percentage of correct answers and the average response time by attempt, pooling together across subjects and matrix types all the observations for the n*th* repetition of each matrix type. The figure suggests that people do not become more accurate, but do respond more quickly as they gain experience. As a practical matter, this pattern suggests that subjects should be given some practice problems if a researcher wants to use the matrices as a repeated task (e.g. a real effort task with varying difficulty). We formally test this result with the econometric specification in Table 3, where the left panel refers to Experiment 1 and the right panel refers to Experiment 2. Success in solving the puzzles and response time are regressed on a set of dummies for the matrix difficulty level (the number of Attributes that vary) and interaction terms between these dummies and dummies for the number of times a subject has already attempted a difficulty level (Attempts). For the sake of clarity and convenience, in the table we only report the coefficients of the interaction terms for the easiest and hardest matrix levels although all interactions terms are included in the specification.[10] In both experiments, the effect on accuracy of the number of varying attributes is negative, and statistically significant. On

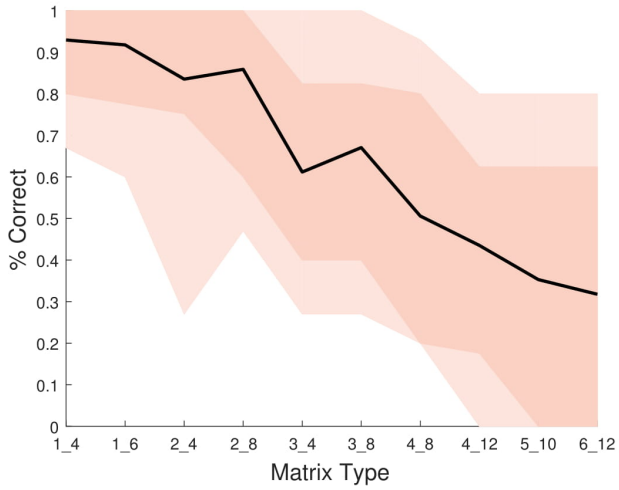
[8]The other motivation was to make direct comparison to the RPM task.

[9]In experiment 1, for instance, the odds of a correct answer going from difficulty level 1 to level 2 (with 4 options) drops from about .46 to .21. The same increase in difficulty reduces the odds in Experiment 2 from .56 to .37.

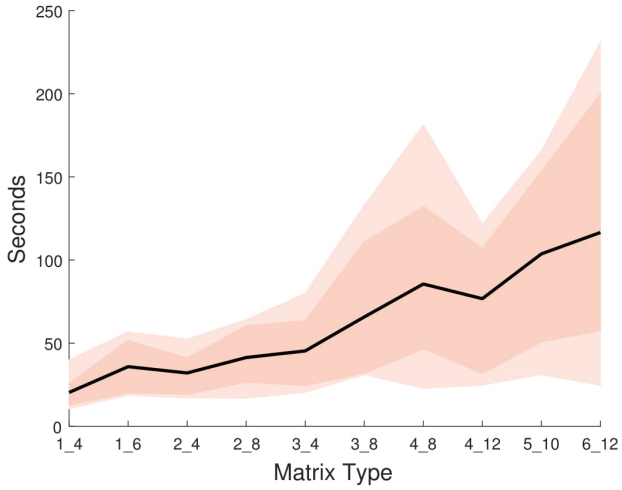
[10]The full econometric results are available from the authors upon request.

Figure 2: Subjects' Performance in Experiment 1

Average responses by matrix type. Matrix type x_y indicates x varying attributes and y options in the solution set. Options generated by basic methodology. Study conducted at the University of Arkansas. Light and dark shading correspond to $14/86th$ and $5/95th$ percentiles of the distribution across subjects, respectively.



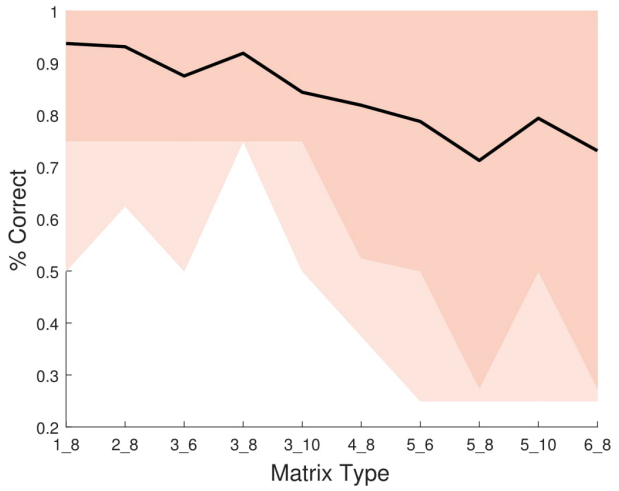
(a) Correct answers.



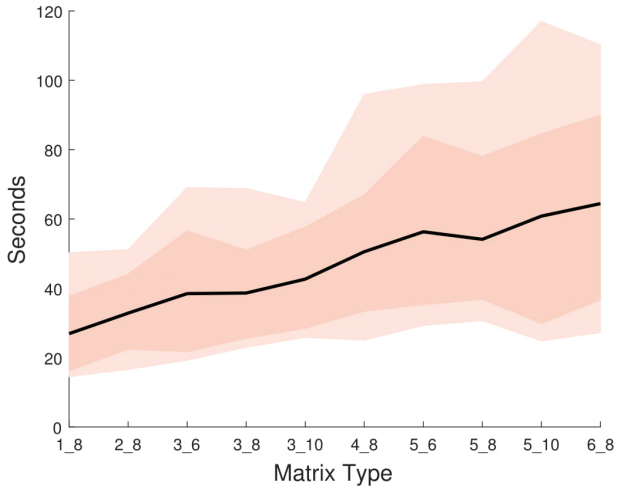
(b) Time.

Figure 3: Subjects' Performance in Experiment 2

Average responses by matrix type. Matrix type x_y indicates x varying attributes and y options in the solution set. Options generated by mutation methodology. Study conducted at Chapman University. Light and dark shading correspond to $14/86th$ and $5/95th$ percentiles of the distribution across subjects, respectively.



(a) Correct answers.



(b) Time.

Table 2: Difficulty of Puzzles & Accuracy of Solutions in Experiments 1 and 2

Effects of number of attributes varrying and number of options on accuracy and response time of the subjects in Experiment 1 and 2. Experiment 1 uses the basic solution methodology; Experiment 2 the mutation approach. Standard errors reported in parenthesis; (R) if robust correction; significance at 1%, 5%, and 10% level is respectively indicated by *, **, and ***. Panel F.E. models estimation follows Cameron and Miller (2015), with robust S.E. for linear model and no-adjustment required for logit. Odds ratios, p-values, and pseudo R^2 are reported for logit.

	Experiment 1			Experiment 2		
	Accuracy		Time	Accuracy		Time
	lin.	*logit*	*lin.*	*lin.*	*logit*	*lin.*
Attributes	-0.134	0.469	14.752	-0.048	0.664	7.858
	(0.012)***	(0.000)***	(2.420)***	(0.007)***	(0.000)***	(0.845)***
Options	-0.002	0.996	2.598	-0.005	0.960	0.905
	(0.005)	(0.925)	(0.536)***	(0.008)	(0.482)	(0.733)
F.E.	Y	Y	Y	Y	Y	Y
S.E. Adjust.	R		R	R		R
Obs.	850	850	850	1600	1600	1600
N. Subjects	17	17	17	40	40	40
R^2	0.23	0.21	0.24	0.04	0.06	0.14

the contrary, the number of attempts has a small and insignificant effect across the different levels of matrix difficulty. However attempts matter for response time. More experience leads to substantial decreases in response times.

Table 4 reports the results of several joint hypotheses tests based on the analysis in Table 3. Each test investigates if the set of coefficients for the interaction terms associated with a given matrix difficulty level are all zero. These tests allow us to assess the cumulative effect of multiple attempts for each difficulty level. For response accuracy, we fail to reject the null of no learning in 11 of 12 cases, with the exception being difficulty level 3 in Experiment 2. However, for response time more practice has a significant negative effect in most cases, with the exceptions being the two easiest difficulty levels in Experiment 2, where accuracy is high and responses are fast initially.

Overall, the evidence from Experiments 1 and 2 suggests that the most effective criterion to classify our matrices by difficulty level is to use the number of varying attributes. This is the approach we follow in Experiment 3 where we hold the number of options fixed and the recommendation we would give to other users of these matrices. In Experiment 3 we directly compare the two methodologies for generating options. Figure 5 reports the same output as in Figures 2 and 3, but for Experiment 3. The plots clearly indicate that the basic methodology leads to lower performance than the mutation methodology. Table 5 reveals that this conclusion is also supported econometrically. The Table replicates the regressions estimated in Table 2 with the options variable removed since the number of options is fixed and a dummy added to indicate the mutation solution methodology. In addition to this

Figure 4: Attempts and Accuracy of the Solutions

Average accuracy and response time by attempt sequence for the two solution methodologies. In Experiment 1 each matrix difficulty is encountered five times, while we have four repetitions in Experiment 2.

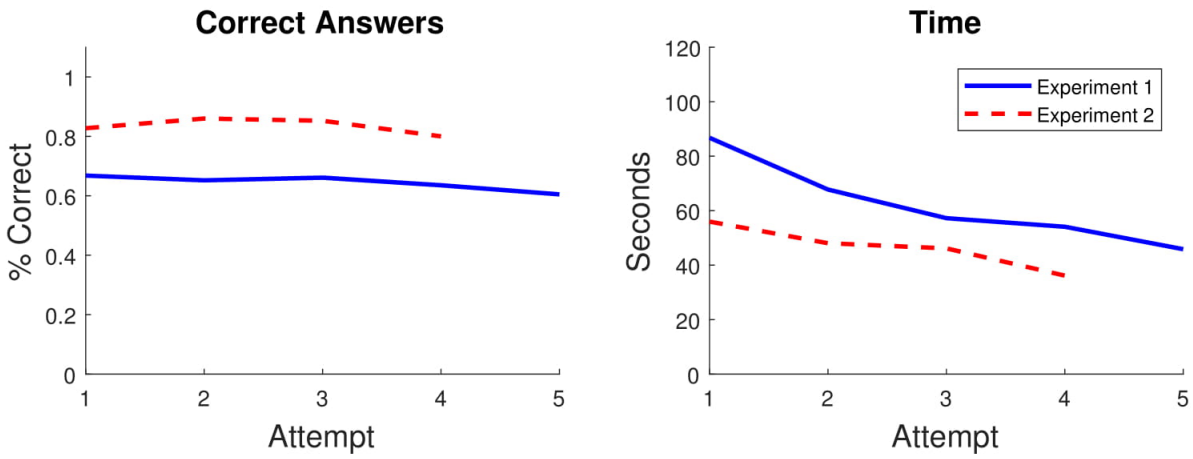
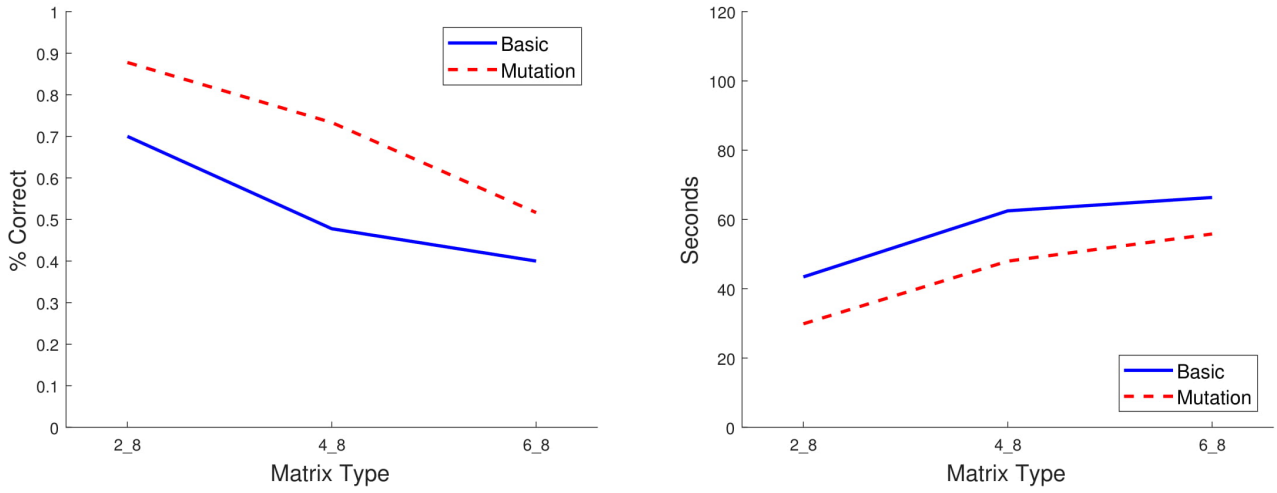


Figure 5: Subjects' Performance in Experiment 3

Average responses by matrix type. Matrix type x_y indicates x varying attributes and y options in the solution set. Comparison between the two methodologies for generating options. Study conducted at the University of Arkansas.



(a) Correct answers.

(b) Time.

Table 3: Learning in Experiments 1 and 2

Effects of number of attributes varying and number of options on accuracy and response time in Experiment 1 and 2. Standard errors reported in parenthesis; (R) if robust correction; significance at 1%, 5%, and 10% level is respectively indicated by *, **, and ***. Panel F.E. models estimation follows Cameron and Miller (2015), with robust S.E. for linear model and no-adjustment required for logit. Odds ratios, p-values, and pseudo R^2 are reported for logit. Interaction terms for Attr_2 to Attr_5 are not reported, but are available from the authors upon request.

	Experiment 1			Experiment 2		
	Accuracy		Time	Accuracy		Time
	lin.	*logit*	*lin.*	*lin.*	*logit*	*lin.*
Attributes_2	-0.088	0.418	5.092	0.025	1.566	5.096
	(0.078)	(0.268)	(5.165)	(0.057)	(0.64)	(3.39)
Attributes_3	-0.147	0.278	38.135	-0.033	0.653	16.688
	(0.095)	(0.093)*	(12.857)***	(0.051)	(0.533)	(3.727)***
Attributes_4	-0.441	0.063	59.394	-0.05	0.549	33.893
	(0.074)***	(0.000)***	(8.237)***	(0.072)	(0.447)	(7.053)***
Attributes_5	-0.676	0.02	107.404	-0.208	0.178	38.686
	(0.123)***	(0.000)***	(17.641)***	(0.063)***	(0.008)***	(5.368)***
Attributes_6	-0.441	0.063	111.151	-0.225	0.162	56.919
	(0.144)***	(0.001)***	(33.092)***	(0.067)***	(0.011)**	(6.600)***
$\text{Attr}_1 \times \text{Atte}_2$	0	1	-12.841	0.025	1.566	-0.098
	(0.076)	(1)	(4.710)**	(0.044)	(0.64)	(3.638)
$\text{Attr}_1 \times \text{Atte}_3$	0	1	-21.34	0.025	1.566	-5.168
	(0.062)	(1)	(4.288)***	(0.044)	(0.64)	(3.874)
$\text{Attr}_1 \times \text{Atte}_4$	0.059	3.372	-20.429	0	1	-6.18
	(0.06)	(0.313)	(5.116)***	(0.062)	(1)	(3.442)*
$\text{Attr}_1 \times \text{Atte}_5$	0	1	-28.201			
	(0.076)	(1)	(5.402)***			
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\text{Attr}_6 \times \text{Atte}_2$	-0.294	0.214	-39.138	0.125	2.201	-24.614
	(0.169)	(0.064)*	(36.653)	(0.082)	(0.168)	(7.239)***
$\text{Attr}_6 \times \text{Atte}_3$	-0.059	0.766	-47.434	-0.025	0.875	-27.424
	(0.106)	0.715	(30.283)	(0.076)	(0.796)	(7.577)***
$\text{Attr}_6 \times \text{Atte}_4$	-0.176	0.434	-29.575	0.025	1.149	-37.037
	(0.13)	(0.27)	(49.65)	(0.099)	(0.792)	(8.333)***
$\text{Attr}_6 \times \text{Atte}_5$	-0.235	0.314	-80.116			
	(0.139)	(0.139)	(34.157)**			
F.E.	Y	Y	Y	Y	Y	Y
S.E. Adjust.	R		R	R		R
Obs.	850	850	850	1600	1600	1600
N. Subjects	17	17	17	40	40	40
R^2	0.25	0.24	0.32	0.05	0.08	0.21

Table 4: Learning in Experiments 1 and 2: Tests on Interacted Terms

Hypothesis tests for the null hypothesis that the coefficients of the interacted terms from the models in Table 3 are all jointly zero by matrix difficulty level. Experiment 1 uses the basic solution methodology; Experiment 2 the mutation approach. P-values reported in parenthesis.

	Experiment 1			Experiment 2		
	Accuracy		Time	Accuracy		Time
Null Hypothesis:	*lin.*	*logit*	*lin.*	*lin.*	*logit*	*lin.*
$\text{Attr}_1 \times \text{Atte}_i = 0 \; \forall i$	F(4,16)=1.04 (0.418)	$\chi^2(4)$=1.28 (0.865)	F(4,16)=11.18 (0.000)	F(3,39)=0.18 (0.911)	$\chi^2(3)$=0.44 (0.932)	F(3,39)=1.41 (0.253)
$\text{Attr}_2 \times \text{Atte}_i = 0 \; \forall i$	F(4,16)=1.98 (0.146)	$\chi^2(4)$=4.31 (0.365)	F(4,16)=5.55 (0.005)	F(3,39)=0.27 (0.848)	$\chi^2(3)$=1.09 (0.780)	F(3,39)=1.74 (0.175)
$\text{Attr}_3 \times \text{Atte}_i = 0 \; \forall i$	F(4,16)=0.83 (0.525)	$\chi^2(4)$=3.55 (0.471)	F(4,16)=5.60 (0.005)	F(3,39)=3.46 (0.025)	$\chi^2(3)$=12.31 (0.006)	F(3,39)=7.41 (0.000)
$\text{Attr}_4 \times \text{Atte}_i = 0 \; \forall i$	F(4,16)=0.66 (0.631)	$\chi^2(4)$=1.32 (0.858)	F(4,16)=2.74 (0.065)	F(3,39)=0.61 (0.612)	$\chi^2(3)$=1.23 (0.746)	F(3,39)=7.35 (0.000)
$\text{Attr}_5 \times \text{Atte}_i = 0 \; \forall i$	F(4,16)=0.43 (0.787)	$\chi^2(4)$=1.65 (0.799)	F(4,16)=6.39 (0.003)	F(3,39)=0.54 (0.657)	$\chi^2(3)$=2.83 (0.419)	F(3,39)=12.92 (0.000)
$\text{Attr}_6 \times \text{Atte}_i = 0 \; \forall i$	F(4,16)=1.22 (0.342)	$\chi^2(4)$=4.85 (0.303)	F(4,16)=3.74 (0.025)	F(3,39)=1.03 (0.391)	$\chi^2(3)$=2.92 (0.404)	F(3,39)=6.84 (0.001)

baseline specification, we also consider a model with no fixed-effects, but clustered standard errors, which allows us to include a dummy that takes the value 1 when the Bilker, Hansen, Brensinger, Richard, Gur, and Gur (2012) 9-task RPM is executed by the subjects before our matrices.[11]

While the number of attributes again affects the matrix difficulty level, we find that on average the mutation solution is positively and significantly associated with a higher accuracy across the board. Similarly, the mutation solution significantly reduces the average response time, reflecting the lower difficulty of this alternative solution set. We speculate that the puzzles using the mutations may be easier for people to solve because wrong answers are more likely to differ from correct responses along more dimensions.[12] On the contrary, the dummy for the 9-task RPM executed first is not significant in any specification either for accuracy or for response time. The results for this second dummy are relevant for the discussion of the broader validity of our matrices in Section 5. Whether the (unpaid) 9-task RPM is carried out before or after our matrices does not affect the performance of the subjects in solving our puzzles (the paid task), but the order does have a dramatic effect on performance in the unpaid RPM task itself, as we explain in the next section.

[11] Clustering standard errors and using fixed effects when there are only a few clusters, as in our case, can be problematic due to overfitting and unrealistically small confidence intervals. As discussed by Cameron and Miller (2015), the most appropriate way to estimate this type of model is by using panel fixed-effects with a robust correction for standard errors. In order to also include the Raven-9 dummy in Table 5, which drops out in the panel estimation, it is necessary to turn to a specification with no fixed-effects. The estimates of the model are very robust to this change.

[12] If a person knows the answer must have certain characteristics, then a wrong response is more likely to be identified if it differs from the correct answer in more dimensions.

Table 5: Difficulty of Puzzles & Accuracy of Solutions in Experiment 3

Effects of number of attributes varrying and number of options on accuracy and response time of the subjects in Experiment 3. Clustered by subject (CL) or robust (R) standard errors reported in parenthesis; significance at 1%, 5%, and 10% level is respectively indicated by *, **, and ***. Panel F.E. models estimation follows Cameron and Miller (2015), with robust S.E. for linear model and no-adjustment required for logit. Odds ratios, p-values, and pseudo R^2 are reported for logit.

	Experiment 3					
	Accuracy				Time	
	lin.	*lin.*	*logit*	*logit*	*lin.*	*lin.*
Attributes	-0.083	-0.083	0.683	0.630	6.109	6.109
	(0.009)***	(0.009)***	(0.000)***	(0.000)***	(0.891)***	(0.891)***
Mutation	0.183	0.183	2.823	2.360	-12.865	-12.865
	(0.022)***	(0.022)***	(0.000)***	(0.000)***	(2.213)***	(2.214)***
Raven-9		-0		1		10.698
		(0.068)		(0.999)		(6.889)
F.E.	Y	N	Y	N	Y	N
S.E. Adjust.	R	CL		CL	R	CL
Obs.	1080	1080	1080	1080	1080	1080
N. Subjects	36	36	36	36	36	36
R^2	0.14	0.11	0.14	0.09	0.10	0.09

5 Validity of the Matrices as a Substitute for the RPM

In this section we consider how well our puzzles capture similar information to that captured by the RPM and thus the degree to which they may be viewed as substitutes. One of the primary uses of the RPM has been to classify respondents. Figure 6 plots the percentage of correct answers for the easiest matrices in Experiment 3, type 2_8, versus the percentage of correct answers for the most difficult matrices, type 6_8. Separate plots are given for the two methods for generating solutions since they lead to different performance levels. The size of the markers in the figure reflects the distribution of subjects across performance levels, while the color of the dots is used to group subjects by performance.

The first important take away from Figure 6 is that individual subjects consistently do better on the easier matrices (i.e. those that have fewer dimensional changes). This is apparent from the lack of observations in the top left portion of the figure in both panels. Observations in that region would indicate increasing the number of dimensions changed did not make the puzzles harder.[13]

The patterns in Figure 6 also suggest a tradeoff between the two methodologies for generating options. As shown in the previous section, response times are faster with the mutation

[13]We observe only a single instance of such behavior. Shown as the darkest dot in the left panel of the Figure 6, one subject answered 60% percent of the 5 difficult questions correctly and only 40% percent of the 5 easy questions correctly using the basic methodology.

methodology and thus it may be better suited for experiments in which the puzzles are nested in a larger decision problem, but these puzzles do not generate as much separation between subjects. By contrast, the basic methodology provides a finer assessment of subject abilities and also reduces the clumping of upper end performance. With the basic methodology we can identify three types of subjects: those that are low ability who do not perform well even on the easy puzzles (lower-left region); those that perform well on the easy puzzles but not the hard ones (lower-right region); and those that perform well on both the easy and the hard puzzles (upper-right region). In contrast, the mutation methodology decisively shifts the distribution of outcomes outwards for both matrix levels, leaving only two clearly distinguishable subject types.

Subjects in Experiment 3 also complete the Bilker, Hansen, Brensinger, Richard, Gur, and Gur (2012) 9 task RPM. Overall, the correlation between the number of our puzzles a subject answered correctly and her score on the shortened RPM was $\rho = 0.43$ (with $p\text{-}value = 0.009$). However, there was a considerable difference in RPM performance when this unpaid procedure occurred before our paid puzzles ($mean = 7.47$) and after our paid puzzles ($mean = 6.53$, with a two-sample equal-mean t-test $p\text{-}value = 0.069$). If we look only at those who complete the RPM prior to our paid puzzles, the correlation in scores increases to $\rho = 0.57$ ($p\text{-}value = 0.010$).[14] Separating puzzle performance by how the options in the solutions are generated, the correlation of RPM and puzzles with basic options is 0.47 ($p\text{-}value = 0.041$) and it is 0.61 ($p\text{-}value = 0.005$) between RPM and puzzles with mutation options.

The 120 subjects in the additional experiment involving bidding in auctions (Lee, Nayga, Deck, and Drichoutis, 2018) completed the full 60 question RPM along with 21 of our puzzles, 7 for each of the three difficulty levels: 2_6, 4_6, and 6_6. For these subjects, the correlation between the two tests of cognitive ability was $\rho = 0.51$ ($p\text{-}value < 0.001$).[15] Figure 7 graphically illustrates the strong correlation between our measure of cognitive ability and the RPM test. The figure shows the scatter plot of the total score obtained by the subjects of this experiment in the two tests, with our measures on the vertical axis, along with the corresponding regression line. The slope estimate is $\beta = .45$ (with $p\text{-}value < 0.001$).

Subjects in Experiment 2 competed in a paid beauty contest game similar to that in Carpenter, Graham, and Wolf (2013). In Table 7 of that paper, the authors report that individuals who score higher on the RPM make better predictions of the guesses of others, provide guesses that are closer to the best responses to one's own stated beliefs, and have guesses that are closer to the wining guess. In Table 6 we report similar evidence for our subjects using our puzzles, although we do note that our subjects were compensated for puzzle performance whereas subjects in Carpenter, Graham, and Wolf (2013) were not compensated for RPM performance. For consistency with Carpenter, Graham, and Wolf (2013) we report the coefficients for gender, but suppress those for ethnicity and class standing.[16] In each case the coefficient on puzzle score has the anticipated sign and is significant in 2 of

[14] On the contrary, the correlation falls to 0.3 for those that complete the RPM second.

[15] We also find a decreasing percentage of correct answers and an increasing time of response perfectly consistent with those in Experiment 1-3. These results, not reported here for sake of brevity, are available from the authors.

[16] Demographic information was not collected in Experiments 1 and 3. It was captured in Experiment 2 to enable the direct comparison to Carpenter, Graham, and Wolf (2013).

Figure 6: Difficulty of Puzzles & Classification of Subjects

Percentage of correct answers for the easiest matricies in Experiment 3, type 2_8, versus the percentage of correct answers for the hardest matrices, type 6_8. The size of the dots is proportional to the distribution of subjects across performance levels. The colors identify groups of subjects with similar abilities as defined by correctly answering more than or fewer than 50% of the matricies of a given diffculty level.

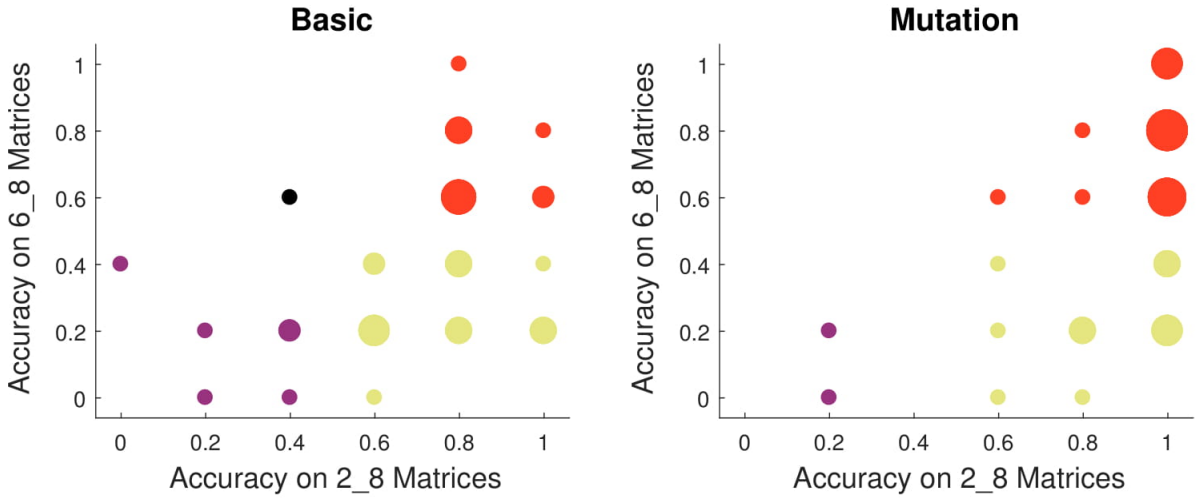


Figure 7: Comparative Performance of Subjects in Lee et al.

Scatter plot of the RPM test total score (maximum of 60 points) versus the total correct answers for our puzzles (maximum score 21 points). Regression line in red ($\beta = .45$ and $p\text{-}value < 0.001$).

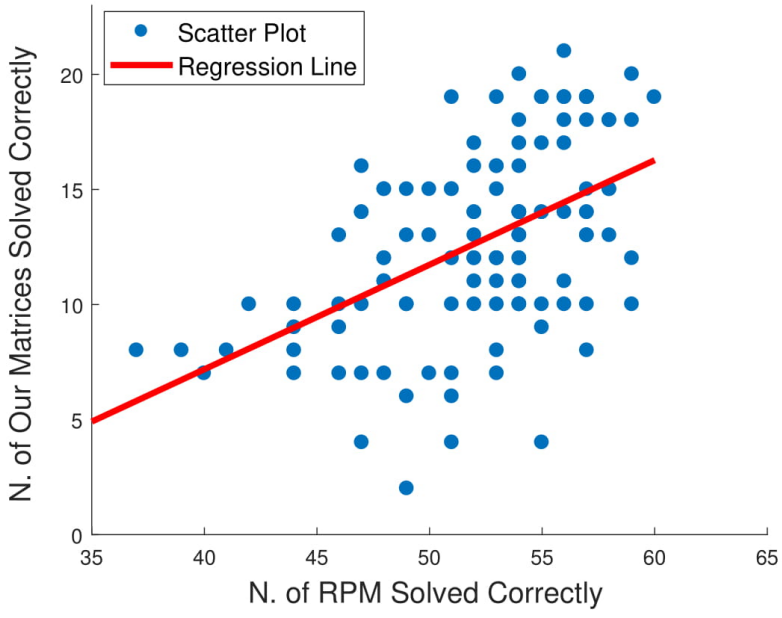


Table 6: Beauty Contest Game & Puzzle Score in Experiment 2

Effects of subject performance on our puzzle and other demographic characteristics on behavior in the beauty contest game. The p-values for the Constant are two-sided, but the p-values for Puzzle Score are one-sided based on the results presented by Carpenter et al. (2013). Controls included class standing and ethnicity, as in Carpenter et al. (2013). Standard errors in parenthesis. Significance at 1%, 5%, and 10% level is respectively indicated by *, **, and ***.

	Own Guess	*Mean of Others' Guesses*	*Deviation from Best Response*	*Deviation from Winning Guess*
Puzzle Score	-0.32	-0.10	-0.34	-0.27
	(0.16)**	(0.11)	(0.13)***	(0.14)**
Female	-0.68	0.33	-0.87	-0.38
	(1.69)	(1.19)	(1.37)	(1.49)
Constant	19.67	11.68	16.74	13.91
	(5.52)***	(3.89)***	(4.47)***	(4.87)***
N. Subjects	40	40	40	40
R^2	.37	.22	.36	.31
Comparable column of Table 7 in Carpenter et al. (2013)		2	4	5

the 3 regressions.[17] We also find evidence that those with higher scores make lower guesses (column 1 of Table 6).

Finally, the 120 subjects who participated in the additional experiment on cognitive load (Deck, Jahedi, and Sheremeta, 2017) completed a survey in which the students self-reported their GPA. The correlation between GPA and performance on our puzzles in the absence of cognitive load was 0.20 (one-sided p-value for t-test = 0.014). The results of that experiment, as reported in Deck, Jahedi, and Sheremeta (2017), indicate that performance on our puzzles deteriorates significantly when respondents are placed under cognitive load in a similar manner to the deterioration in performance of basic arithmetic when under cognitive load. Interestingly, data from subjects not under cognitive load in Deck, Jahedi, and Sheremeta (2017) suggest there is no correlation in performance on our puzzles and risk taking ($\rho = 0.07$, two-sided p-value for t-test = 0.447) or any relationship between altruism and puzzle performance, which somewhat contradicts Burks, Carpenter, Goette, and Rustichini (2009).

[17] We also note that the top half of the subjects in Experiment 2 in terms of cognitive ability as measured by our matrices answered 0.95 out of 2 HIT15 questions correctly on average whereas the bottom half only answered 0.60 correctly on average. This difference is significant (one-sided p-value for two sample t-test = 0.027) and consistent with Carpenter, Graham, and Wolf (2013).

6 Conclusions

This paper describes the properties of a set of puzzles that can be viewed as behaviorally similar to the common Raven's Progressive Matrix test and subsequent work by Matzen, Benz, Dixon, Posey, Kroger, and Speed (2010). The RPM procedure, the tasks of Matzen, Benz, Dixon, Posey, Kroger, and Speed (2010), and our puzzles require respondents to select among a set of options the one that completes a logical relationship among a set of images arranged in a matrix. The difficulty of our puzzles is determined by the number of attributes that vary between images and the method for generating incorrect options, but the evidence suggests that the number of presented options does not impact difficulty.

We document how the accuracy rate declines with difficulty while response time increases with difficulty. We also show that performance on our puzzles and performance on the RPM are highly correlated. Further, we show that our puzzles yield similar predictive success to that of the Raven's test in a strategic setting. Thus, like the Raven's procedure, our puzzles can be used to type or classify subjects into different cognitive ability levels. The advantage of our approach is that one can generate a large number of distinct puzzles for a given level of difficulty, which expands the usefulness of these tools in experimental economics. For example, one can use our puzzles as a real effort task with varying levels of cognitive difficulty while maintaining physical consistency (as in the Rational Inattention experiments of Civelli, Deck, LeBlanc, and Tutino, 2018) or as a way to repeatedly measure performance under different circumstances (as in the comparison of cognitive load techniques by Deck, Jahedi, and Sheremeta, 2017).[18] However, some words of caution are in order. First, the evidence offered in support of our puzzles is drawn from subject pools that are standard in experimental economics - undergraduate students at universities in the western hemisphere. The degree to which these puzzles would have similar properties among other populations remains an open question. Further, we do not explore how performance varies with respondent demographics. Also, our subjects were incentivized to answer our puzzles correctly, whereas participants are often not compensated when the RPM is used to measure cognitive ability.

[18] Future research comparing boredom, fatigue, and ease of implementation of different real effort tasks would be valuable.

References

Al-Ubaydli, O., G. Jones, and J. Weel (2016): "Average player traits as predictors of cooperation in a repeated prisoner's dilemma," *Journal of Behavioral and Experimental Economics*, 64, 50–60.

Benito-Ostolaza, J., P. Hernandez, and J. Sanchis-Llopis (2016): "Do individuals with higher cognitive ability play more strategically?," *Journal of Behavioral and Experimental Economics*, 64, 5–11.

Bilker, W., J. Hansen, C. Brensinger, J. Richard, R. Gur, and R. Gur (2012): "Development of Abbreviated Nine-item Forms of the Ravens Standard Progressive Matrices Test," *Assessment*, 19(3), 354–369.

Burks, S., J. Carpenter, L. Goette, and A. Rustichini (2009): "Cognitive skills affect economic preferences, strategic behavior, and job attachment," *Proceedings of the National Academy of Sciences*, 106(19), 7745–7750.

Cameron, A. C., and D. L. Miller (2015): "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 50(1), 317–372.

Carpenter, J., M. Graham, and J. Wolf (2013): "Cognitive ability and strategic sophistication," *Games and Economic Behavior*, 80(C), 115–130.

Carpenter, P. A., M. A. Just, and P. Shell (1990): "What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test," *Psychological Review*, 97(3), 404–431.

Civelli, A., C. Deck, J. LeBlanc, and A. Tutino (2018): "Rational Inattention and Consumer Choices: An Experiment," University of Arkansas, Mimeo.

Condon, D., and W. Revelle (2014): "The international cognitive ability resource: Development and initial validation of a public-domain measure," *Intelligence*, 43(1), 52–64.

Cueva, C., and A. Rustichini (2015): "Is financial instability male-driven? Gender and cognitive skills in experimental asset markets," *Journal of Economic Behavior and Organization*, 119(C), 330–344.

Deck, C., S. Jahedi, and R. Sheremeta (2017): "The Effects of Different Cognitive Manipulations on Decision Making," Economic Science Institute, Working Paper.

Duttle, K. (2015): "Cognitive Skills and Confidence: Interrelations with Overestimation, Overplacement, and Overprecision," *Bulletin of Economic Research*, 68(s1), 42–55.

Flynn, J. R. (1987): "Massive IQ gains in 14 nations: What IQ tests really measure," *Psychological Bulletin*, 101, 171–191.

Lee, J. Y., R. M. Nayga, C. Deck, and A. Drichoutis (2018): "Cognitive ability and bidding behavior," University of Arkansas, Mimeo.

Matzen, L. E., Z. O. Benz, K. R. Dixon, J. Posey, J. K. Kroger, and A. E. Speed (2010): "Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties," *Behavior Research Methods*, 42(2), 525–541.

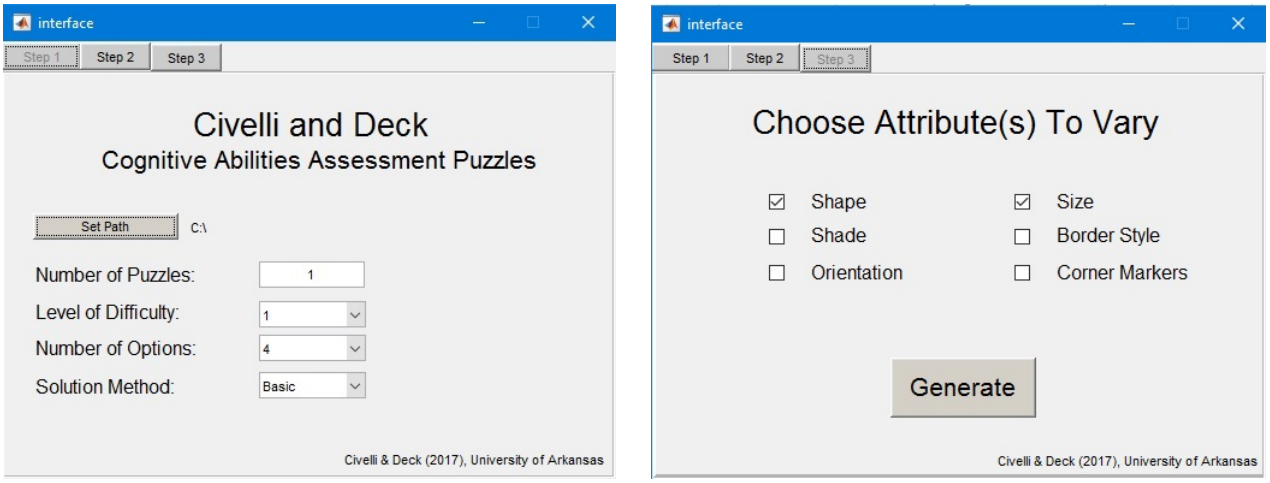
Raven, J., J. H. Court, and J. C. Raven (1998): *Manual for Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press.

Appendix

A The Software to Generate the Matrices

We illustrate here the main features of the software tool to generate customized sets of matrices. The software has been developed in MATLAB v2016a and it is supported by an intuitive and convenient interface that facilitates the interaction with the underlying code for users of any level of familiarity with MATLAB. The interface can be deployed as a standalone self-executable application after installing the MATLAB Compiler Runtime 64-bit (available for free from the MATLAB Runtime webpage). The interface can be run as a regular .m script from the MATLAB platform as well, if preferred by more expert MATLAB users. This modality clearly gives access to the source files of the code too.

This software is available for academic purposes only and the license agreement requires users to cite this paper in any projects that utilize the the software. The software can be downloaded from the authors' web-page.



(a) Selection of Main Features. **(b)** Selection of the Attributes.

Figure A1: Screen-shots from the MATLAB interface for the generation of our puzzles.

Figure A1 illustrates two screen-shots from two steps of the interface, in which the user is asked to select some of the features of the matrices to be generated.

B Experimental Designs and Details

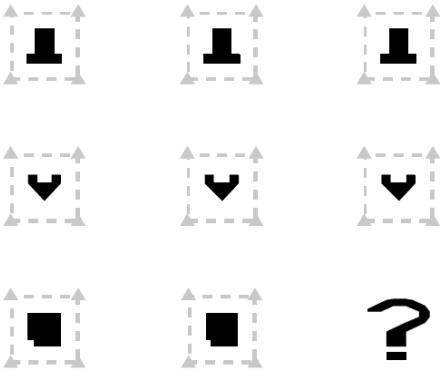
This Appendix shows the instructions provided to the subjects before engaging in the solution of our matrices and the questions used for the Carpenter, Graham, and Wolf (2013) game and for the basic demographic survey. These are report in the following pages:

Instructions

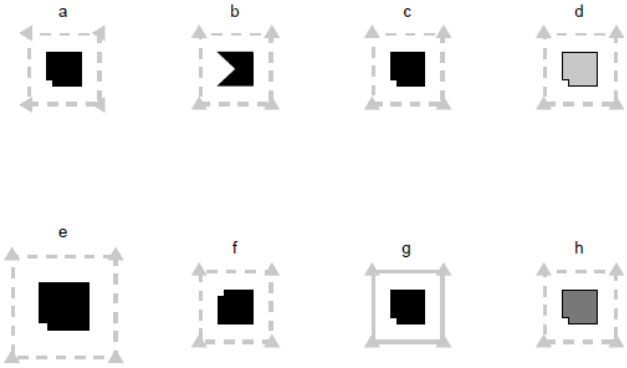
In this study you will be given 50 pattern problems to solve. For each one you answer correctly, you will earn $0.50. This amount will be added to the $7.00 you are receiving for participating in this study.

So what are pattern problems? A pattern problem is a 3x3 table of images that are arranged in a pattern, with the image in the lower right corner removed. You will need to identify the missing image.

Some pattern problems are relatively easy, like the one pictured to the right. In this example, the shape is the same on each row but different from row to row. Notice that the border around the shape is the same for every image.



You will be given multiple possible correct answers and have to identify the letter for the correct one. If you were given the options below, the correct answer would be “c” because it has the right shape and border.



Option “a” is incorrect because the little triangles on the border point in the wrong direction. Option “b” is incorrect because it has the wrong shape. Option “d” is incorrect because the shape is the wrong color. Option “e” is incorrect because the shape is the wrong size. Option “f” is incorrect because the shape is turned the wrong way. Option “g” is incorrect because the border is not dashed. Option “f” is incorrect because the shape is the wrong color.

As you can tell from the preceding example, images can differ in lots of dimensions: shape, size, color, direction, border style, border corner marker.

Image characteristics can change in the table in several ways. In the example on the previous page, the change was from row to row, like the image to the right.

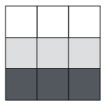
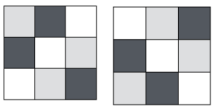
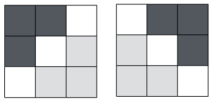


Image characteristics can also change by column,

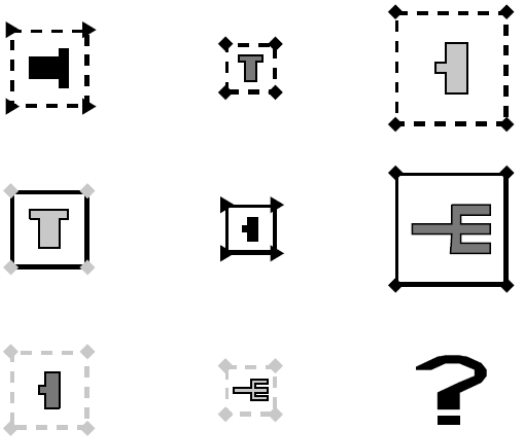
along the diagonal,



and by the corner.

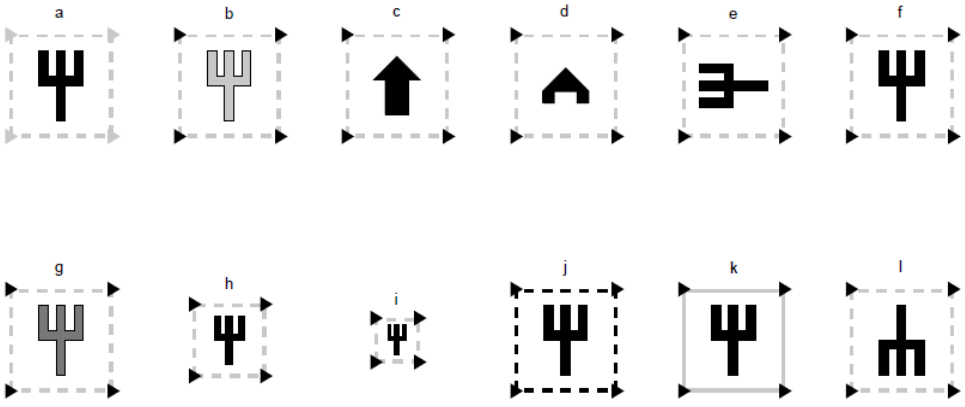


So the problems can be very very difficult, like the example below. It is OK to guess and you should keep in mind that you have 50 pattern problems to complete.



In this example, the missing image should have a light gray dashed border, since this characteristic changes by the row. The border markers should be black triangles since this characteristic changes by the corner. The shape should be large since this changes by the column. The shape should be a trident because this changes by the corner. The shape should be solid black because this changes on the diagonal. The three pointed side of the trident should be facing up as the direction changes by the diagonal. Therefore, option "f" is the correct answer.

Please raise your hand when you are ready to start or if you have any questions.



Subject #_____

First Paid Task ***[this is the Carpenter, Graham, and Wolf, 2013, game in the manuscript]***

You will guess a number between 0 and 20 up to two decimal places. The person who wins the game is the person who picks the number that ends up being closest to one-half the average of the guesses from all the 9 other participants. This winner will receive $10 in addition to his or her other earnings. There is also a second way to win. The person who most accurately predicts the distribution of guesses will win another $10.

Your Guess: _____

Your prediction of the other 9 guesses:
_____, _____, _____, _____, _____, _____, _____, _____, _____,

Survey

Consider the following two-person game: There is a "basket" in which people place "points". The two players take turns placing 1, 2, or 3 points in the basket. The person who places the 15th point in the basket wins a prize. Say you are playing and want to win the prize.

Q1. If you go first, how many points will you place in the basket?
Please pick one of the answers: 1 2 3

Q3. If you go second and the other player has already put 2 points in the basket on her first turn, how many would you put in?
Please pick one of the answers: 1 2 3

Q4. What is you sex? _____

Q5. Which is your class standing?
Freshman Sophomore Junior Senior Graduate-Student Not a Student

Q6. What is your ethnicity? _____