2017

A Flexible and Customizable Method for Assessing Cognitive Abilities

Andrea Civelli
University of Arkansas

Cary Deck
Chapman University, deck@chapman.edu

A Flexible and Customizable Method for Assessing Cognitive Abilities

Comments

Working Paper 17-09

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

Raven's Progressive Matrices are a broadly used tool for measuring cognitive ability. This paper develops and validates a set of nonverbal puzzles that can be viewed as an extension of or substitute for the well-known Ravens tasks. Specifically, we describe the characteristics of our puzzles and provide a calibration of the matrices in terms of response accuracy and response time as a function of these characteristics. Then we directly compare within-subject performance on our puzzles and Ravens tasks. Finally, we replicate a previous experimental paper, substituting our puzzles for the Ravens matrices, and show the two tools have similar predictive success. Our approach offers several benefits due to the relatively large number of novel puzzles of a given difficulty level that can be generated.

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, it has been widely applied in research in Psychology, Behavioral Economics, and Neuroscience when an assessment of the subjects' cognitive ability is desired. For example, Burks, Carpenter, Goette, and Rustichini (2009) show that RPM performance is correlated with taking calculated risks, social awareness, and job perseverance. Benito-Ostolaza, Hernandez, and Sanchis-Llopis (2016) show that those who score highly on the RPM behave more strategically and Al-Ubaydli, Jones, and Weel (2016) reports 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 test is based on 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 attractions 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. Second, the RPM cannot finely partition people of similar, but distinct, abilities the way it could if it involved hundreds of puzzles calibrated to the respondents. 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 expand the set of Ravens like tasks.[1] This paper reports a similar effort to expand the set of puzzles available to researchers. Specifically, the advantages of our approach include:

1. Allowing researchers to measure cognitive ability without relying on specific puzzles that respondents are likely to have seen previously.

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

3. Providing researchers a real effort task that can be calibrated across individuals.

4. Providing researchers with a set of uniform real effort tasks for use in situations where subjects need to be able to differentiate their effort.

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

[1] In personal correspondence, Matzen, Benz, Dixon, Posey, Kroger, and Speed (2010) indicated the software they developed to generate Ravens like puzzles is not available for distribution.

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

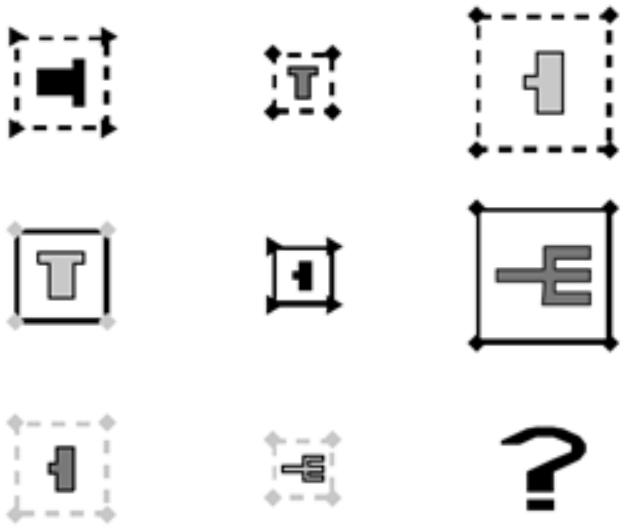
The puzzles are 3×3 graphical matrices in which each entry contains an image with specific attributes. The bottom right image in the matrix is hidden by a question mark. Subjects must understand the patterns followed by the attributes of the images and identify the correct image to complete the puzzle 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.

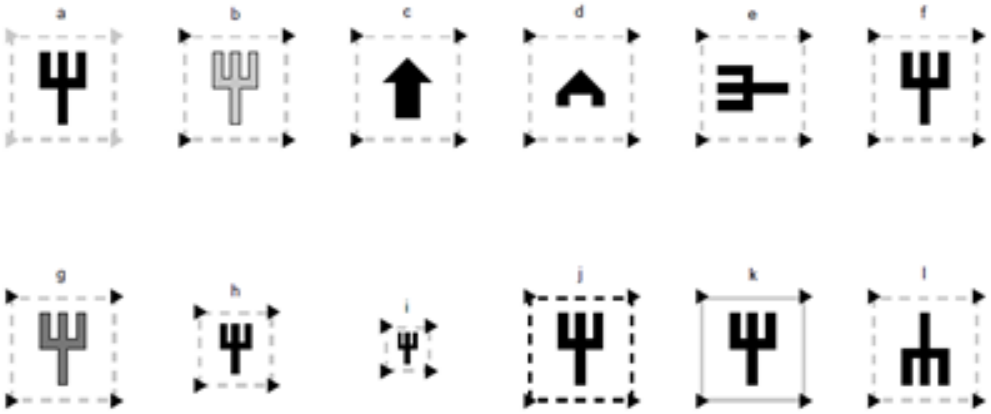
The difficulty level of a matrix is defined by the number of attributes allowed to change; hence we have up to 6 levels, although this could be extended. Each attribute can take one of four possible values, except for the shape which has 15 values. As shown by Figure 1 panel (c), each scheme re-arranges 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 along the column, the shading varies along the main-diagonal, the orientation varies 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

(a) Puzzle



(b) Solution set



(c) Patterns schemes

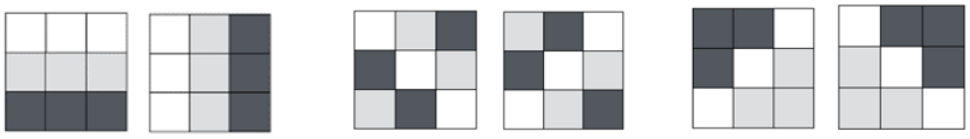


Figure 1: An example of matrix puzzle and a possible solution set.

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.[2]

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 a fourth experiment, which was designed for independent purposes but employed our matrices to measure the cognitive ability of the subjects.

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 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.

[2]The software is freely available from the authors' web-page (available at this link) where additional details are provided.

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.

The additional data is obtained from another experiment 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 Deck, Lee, and Nayga, 2017). 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.[3]

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 desirable property of the task in that it helps identify cutoffs for the subjects' types.

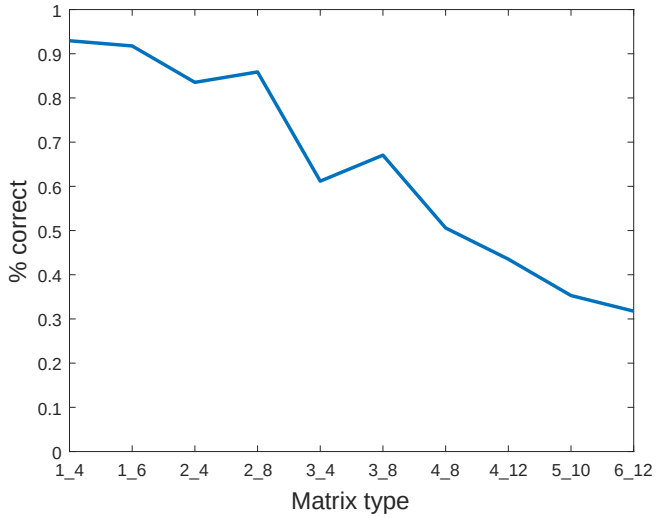
We also observe that the number of options has a less pronounced impact on the performance of the subjects, and it is usually less decisive for the difficulty of the puzzles than an increase in number of varying attributes. More options, for instance, seem to increase the time for a solution at matrix level 2 and 3, but are associated with a slight increase in accuracy of subjects' responses. The opposite occurs at matrix level 4.

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.[4]

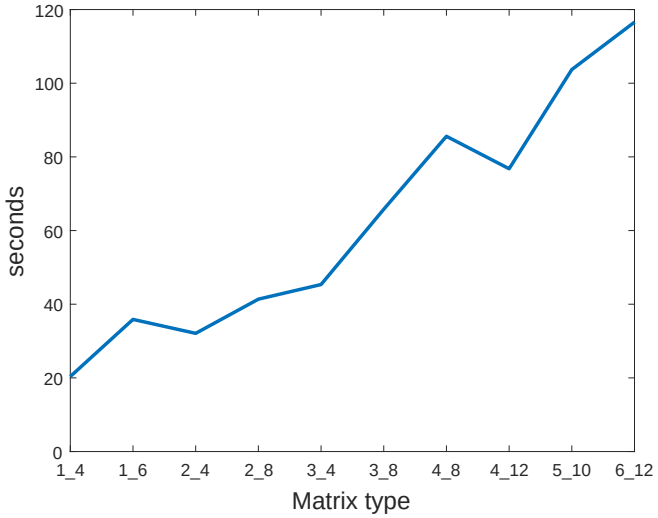
We quantify these effects more formally with the regressions estimates reported in Table 1, where the top panel refers to Experiment 1 and the bottom one to Experiment 2 (basic and mutation options respectively). In both cases, the effect of the number of varying attributes is negative and statistically strongly significant. On the contrary, the effect of the number of

[3]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.

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

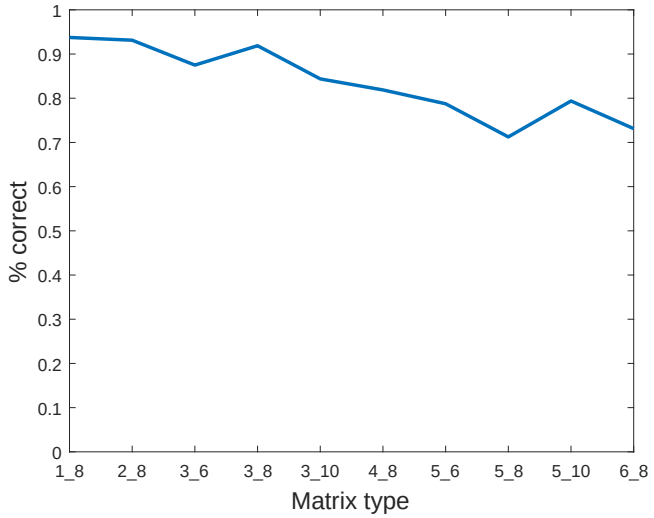


(a) Correct answers.

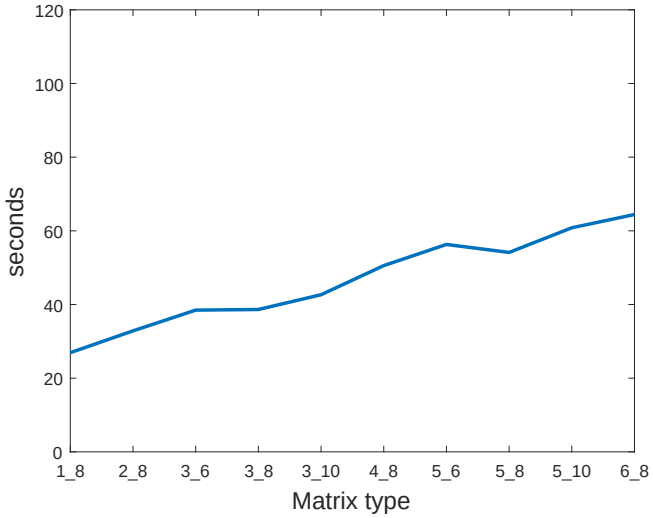


(b) Time.

Figure 2: Experiment 1: Average responses by matrix type (difficulty level). 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.



(a) Correct answers.



(b) Time.

Figure 3: Experiment 2: Average responses by matrix type (difficulty level). 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.

	Experiment 1					
	Accuracy				Time	
	lin.	*lin.*	*logit*	*logit*	*lin.*	*lin.*
Attributes	-0.134	-0.134	0.504	0.461	14.752	14.752
	(0.012)***	(0.014)***	(0.000)***	(0.000)***	(2.420)***	(2.050)***
Options	-0.002	-0.002	0.998	0.996	2.598	2.598
	(0.005)	(0.008)	(0.946)	(0.93)	(0.536)***	(0.733)***
Sbj. Dumm.	N	Y	N	Y	N	Y
SE Clustered/ Robust	CL	R	CL	R	CL	R
Obs.	850	850	850	850	850	850
R^2	0.21	0.29	0.17	0.25	0.22	0.31
	Experiment 2					
	Accuracy				Time	
	lin.	*lin.*	*logit*	*logit*	*lin.*	*lin.*
Attributes	-0.048	-0.048	0.686	0.657	7.858	7.858
	(0.007)***	(0.006)***	(0.000)***	(0.000)***	(0.845)***	(0.475)***
Options	-0.005	-0.005	0.964	0.959	0.905	0.905
	(0.008)	(0.008)	(0.489)	(0.486)	(0.733)	(0.673)
Sbj. Dumm.	N	Y	N	Y	N	Y
SE Clustered/ Robust	CL	R	CL	R	CL	R
Obs.	1600	1600	1600	1600	1600	1600
R^2	0.04	0.13	0.04	0.16	0.12	0.24

Table 1: Effects of matrix levels and number of options on response accuracy and execution time of the subjects in Experiment 1 and 2. The top panel uses the basic solution methodology, while the bottom panel uses the mutation approach. Clustered by subject/Robust standard errors in parenthesis; significance at 1%, 5%, and 10% level is respectively indicated by *, **, and ***. For the logit model, odds ratios, p-values, and adjusted R^2 are reported.

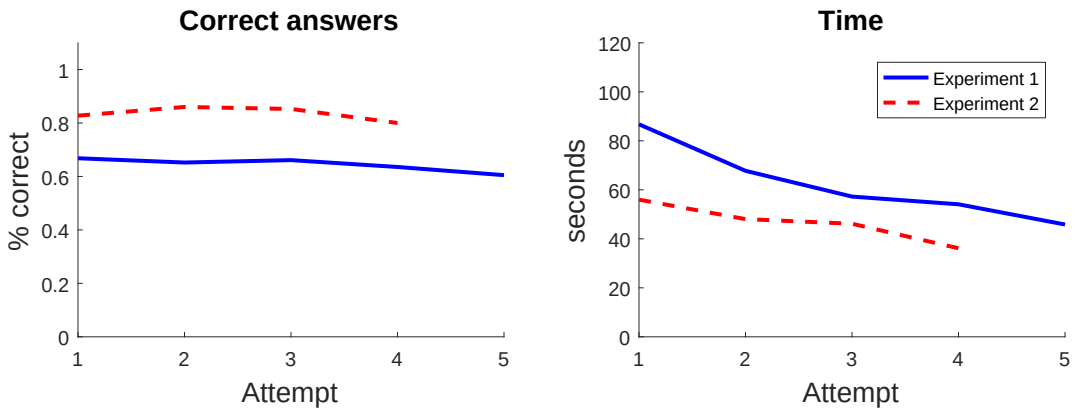


Figure 4: Average correct responses and time by attempt sequence for the two solution methodologies. In Experiment 1 each matrix difficulty is encountered five times, while we have four repetitions in Experiment 2.

options is negative, but very small and never 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 Attributes coefficient of the first logit model in Table 1 tells us that increasing the matrix level by 1 unit makes a correct answer of the puzzles 2 times less likely.

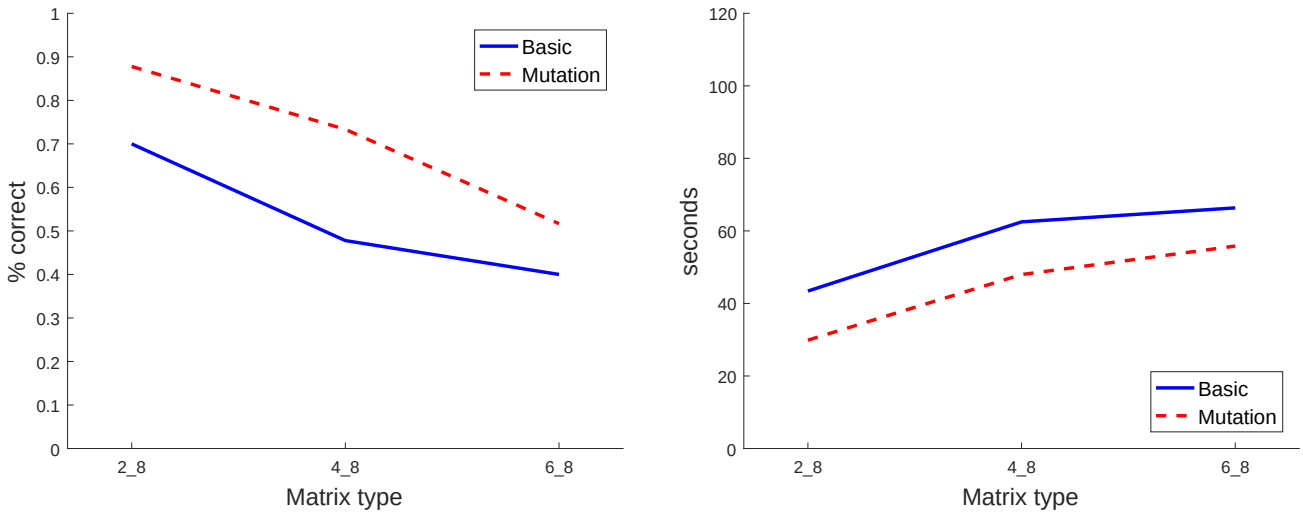
The effects for the basic solution methodology are about 2 times bigger than those for the mutation options. The same analysis holds for the execution time, except for the coefficient of the number of options in Experiment 1, but not Experiment 2, which is now significant. Despite affecting execution time, however, the number of options does not affect the degree of difficulty of the puzzles. This result is also consistent with the evidence discussed below for the learning over attempts (see Figure 4).

Overall, the evidence from Experiments 1 and 2 suggest 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.

The ordering of the matrices experienced by a subject does not affect performance.[5] However, we do observe a specific form of learning as illustrated in Figure 4. In the two panels of this Figure we compute the average percentage of correct answers and the average time to solve a matrix by attempt sequence, that is pooling together across subjects and matrix types all the observations for the n*th* repetition of each matrix encountered by the subjects. We see that the subjects get faster as they repeat the same type of puzzle more times, but their precision does not substantially improve. This 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 move now to experiment 3, in which 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

[5]For sake of brevity these results are not reported here, but are available from the authors upon request.



(a) Correct answers.

(b) Time.

Figure 5: Experiment 3: Average responses by matrix type (difficulty level). 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.

	Experiment 1					
	Accuracy				Time	
	lin.	*lin.*	*logit*	*logit*	*lin.*	*lin.*
Attributes	-0.083	-0.083	0.683	0.619	6.109	6.109
	(0.009)***	(0.008)***	(0.000)***	(0.000)***	(0.891)***	(0.555)***
Mutation	0.183	0.183	2.36	2.939	-12.865	-12.865
	(0.022)***	(0.026)***	(0.000)***	(0.000)***	(2.214)***	(2.170)***
Raven-9	-0	-0.067	1	0.724	10.698	-4.573
	(0.068)	(0.12)	(0.999)	(0.568)	(6.889)	(8.814)
Sbj. Dumm.	N	Y	N	Y	N	Y
SE Clustered/ Robust	CL	R	CL	R	CL	R
Obs.	1080	1080	1080	1080	1080	1080
R^2	0.11	0.28	0.09	0.24	0.09	0.33

Table 2: Effects of matrix levels and solution methodology on response accuracy and execution time of the subjects in Experiment 3. The dummy "mutation" takes value 1 when the mutation approach is adopted. Clustered by subject/Robust standard errors in parenthesis; significance at 1%, 5%, and 10% level is respectively indicated by *, **, and ***. For the logit model, odds ratios, p-values, and adjusted R^2 are reported.

than the mutation methodology. Table 2 illustrates that this conclusion is also supported econometrically. The Table replicates the regressions estimated in Table 1 with the options variable removed since the number of options is fixed. A dummy is added for the mutation solution methodology, as is 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.

While the number of attributes again affects the matrix difficulty level, we find that on average the mutation solution is positively and very significantly associated with a higher accuracy of subjects answers across the board. Similarly, the mutation solution significantly reduces the average execution time reflecting the lower difficulty of this alternative solution set. On the contrary, the dummy for the 9-task RPM executed first is not significant in any specification neither for accuracy nor for execution time. The results for this second dummy is relevant for the discussion of the broader validity of our matrices in Section 5. Whether the 9-task RPM (unpaid task) 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 affect on performance in the unpaid RPM task itself as we explain in the next section.

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 are 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.[6]

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 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

[6]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.

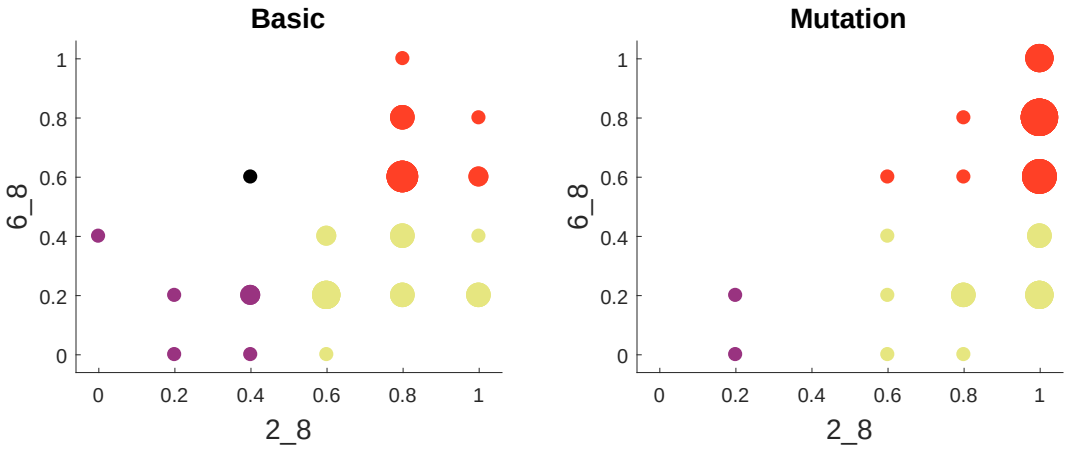


Figure 6: Percentage of correct answers for the easiest matrix in Experiment 3, type 2_8, versus the the percentage of correct answers for the hardest matrix, type 6_8. The size of the dots is proportional to the distribution of subjects across performance levels. The colors identify groups of subjects with similar abilities.

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.

	Own Guesses	*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)***
Comparable column of Table 7 in Carpenter et al. (2013)		2	4	5

Table 3: The p-values for the Constant are two-sided, but the p-values for Puzzle Score are one-sided given the results of 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 ***. For the logit model, odds ratios and adjusted R^2 are reported.

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-

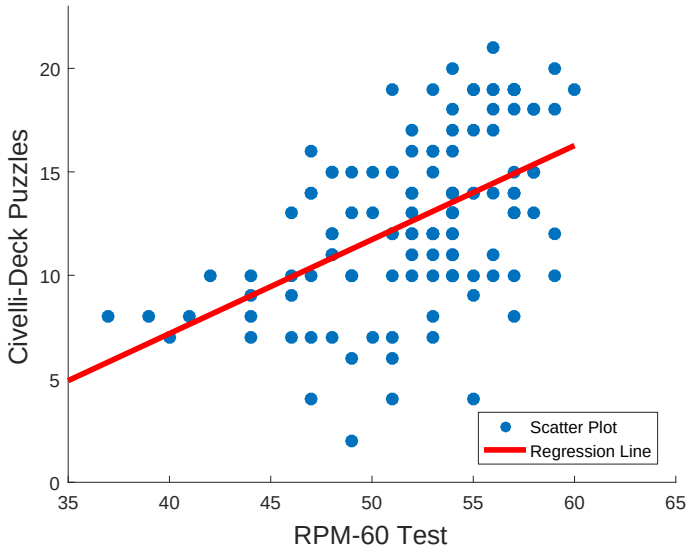


Figure 7: Additional experiment: Scatter plot of the RPM test total score (maximum of 60 points) Versus the total correct answers for our puzzles (maximum score 21 points). Regression line in red ($\beta = .45$ and $p\text{-}value < 0.001$).

$value = 0.009$). 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 $\rho = 0.57$ ($p\text{-}value = 0.010$).[7] 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. That the correlation is stronger for the shortened RPM and our puzzles with the mutation methodology may suggest the 9-task RPM is more suitable for binary identification of abilities, which is not surprising since this is a reduction of the full Raven's test.

The 120 subjects in the additional experiment 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 $\rho = 0.51$ ($p\text{-}value < 0.001$).[8] 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

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

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

Carpenter, Graham, and Wolf (2013). In Table 7 of Carpenter, Graham, and Wolf (2013), they 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 3 we report similar evidence for our subjects using our puzzles. For consistency with Carpenter, Graham, and Wolf (2013) we report the coefficients for gender, but suppress those for ethnicity and class standing.[9] In each case the coefficient on puzzle score has the anticipated sign and is significant in 2 of the 3 regressions.[10] We also find evidence that those higher scores make lower guesses (column 1 of Table 3).

6 Conclusions

This paper describes the properties of a set of puzzles that can be viewed as an extension of or an alternative to the common Raven's Progressive Matrix test. Both the Raven's procedure 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 not by the number of presented options.

We document how the accuracy rate declines with difficulty while the response time increases for our puzzles. 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, 2017) 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).

[9] 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).

[10] 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).

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.

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 (2017): "Rational Inattention and Consumer Choices: An Experiment," University of Arkansas, mimeo.

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.

Deck, C., J. Y. Lee, and R. M. Nayga (2017): "Cognitive ability and bidding behavior," University of Arkansas, mimeo.

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

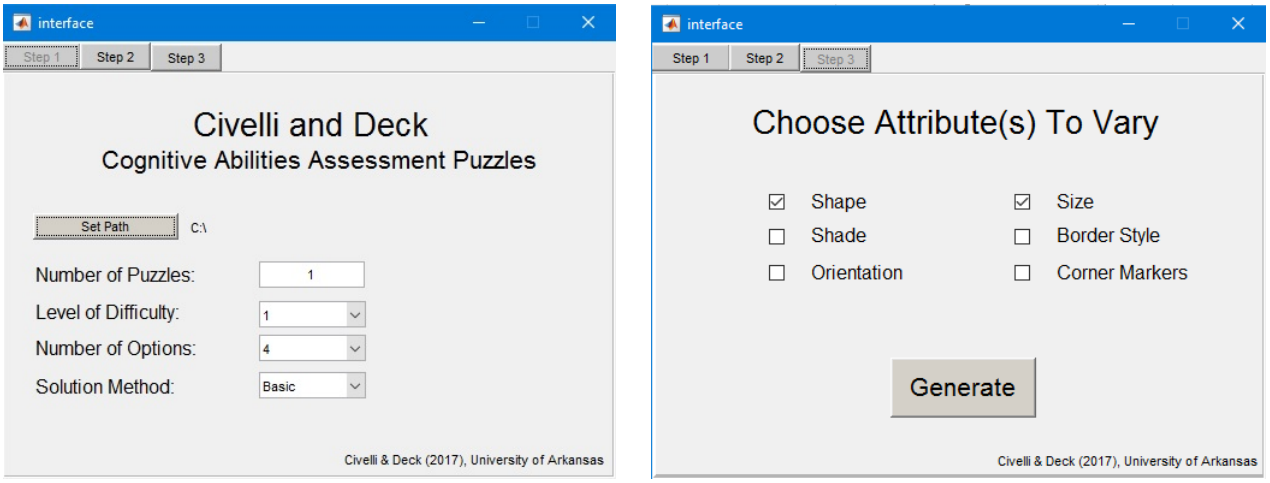
Matzen, L. E., Z. O. Benz, K. R. Dixon, J. Posey, J. K. Kroger, and A. E. Speed (2010): "Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties," *Behavior Research Methods*, 42(2), 525–541.

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.



(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

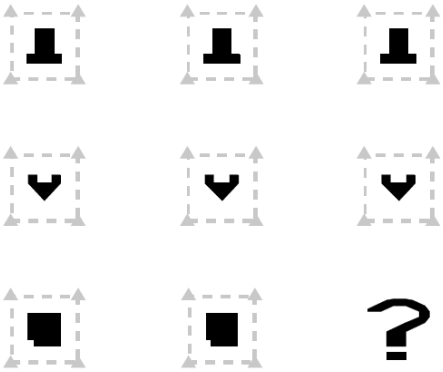
This Appendix shows the instructions provided to the subjects before engaging in the solution of our matrices and the questions used for the Carpenter, Graham, and Wolf (2013) game and for the basic demographic survey. These are report in the following pages:

Instructions

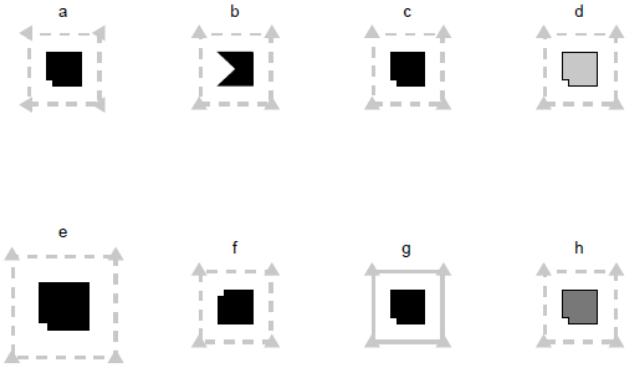
In this study you will be given 50 pattern problems to solve. For each one you answer correctly, you will earn $0.50. This amount will be added to the $7.00 you are receiving for participating in this study.

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.

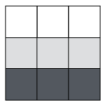
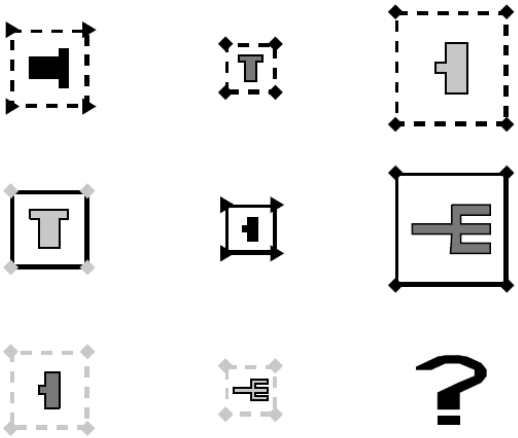


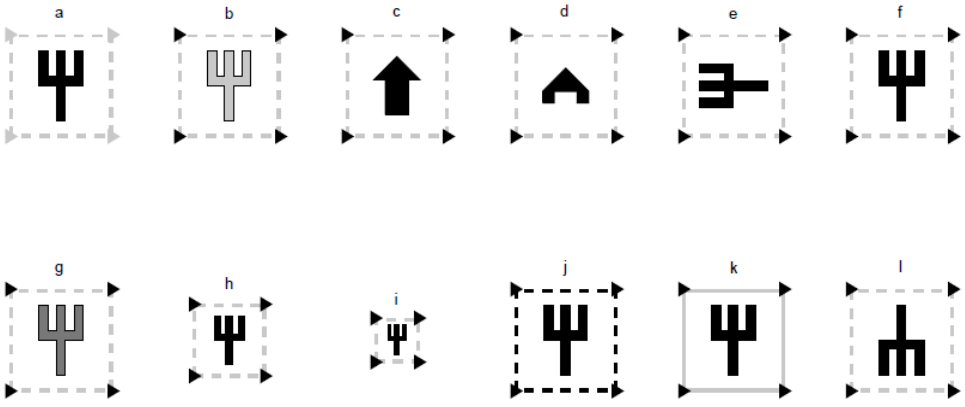
Image characteristics can also change by column, along the diagonal, and by the corner.

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? _____