Spatial Reasoning and Data Displays

Susan VanderPlas, and Heike Hofmann, *Member, IEEE,*

Abstract—
Tone down the 'statistical' part ... statistical, to the wrong ears sounds complicated.

Graphics convey numerical information very efficiently, but rely on a different set of mental processes than tabular displays.

This study examines the demographic characteristics and visual skills associated with perception of graphical lineups. We conclude that lineups are essentially a classification test in a visual domain, and that performance on the lineup protocol is associated with general aptitude, rather than specific tasks such as card rotation and spatial manipulation. We also examine the possibility that specific graphical tasks may be associated with certain visual skills and conclude that more research is necessary to understand which visual skills are required in order to understand certain plot types.

1 Introduction

Data displays provide quick summaries of data, models, and results, but not all displays are equally good, nor is any data display equally useful to all viewers. Graphics utilize higher-bandwidth visual pathways to encode information [1], allowing viewers to quickly and intuitively relate multiple dimensions of numerical quantities. Well-designed graphics emphasize and present important features of the data while minimizing features of lesser importance, guiding the viewer towards conclusions that are meaningful in context and supported by the data while maximizing the information encoded in working memory. Under this framework, well-designed graphics reduce memory load and make more cognitive resources available for other tasks (such as drawing conclusions from the data), at the cost of depending on certain visuospatial reasoning abilities.

Many theories of graphical learning center around the difference between visual and verbal processing: the dual-coding theory emphasizes the utility of complementary information in both domains, while the visual argument hypothesis emphasizes that graphics are more efficient tools for providing data with spatial, temporal, or other implicit ordering, because the spatial dimension can be represented graphically in a more natural manner [2]. Both of these theories suggest spatial ability would impact a viewer's use of graphics, because spatial ability either influences cognitive resource allocation or affects the processing of spatial relationships between graphical elements. In addition, previous investigations into graphical learning and spatial ability have found relationships between spatial ability and the ability to read information from graphs [3]. However, mathematical ability, not spatial ability, was shown [4] to be associated with accuracy on a simple two-dimensional line graph. Spatial ability becomes more important when more complicated graphical displays are used in comparison tasks: the lower performance of individuals with low spatial ability on tests utilizing diagrams and graphs is attributed [5] to the fact that more cognitive resources are required to process the visual stimuli, which leaves fewer resources to make connections and draw conclusions from those stimuli. It is theorized that graphics are a form of "external cognition" [6] that guide, constrain, and facilitate cognitive behavior [7].

"Lineups" have recently been introduced [8], [9], [10] as a tool to evaluate the statistical significance of a graphical finding. Lineups are also useful in assessing the effectiveness of different graphical displays [11], [12]. Like their police counterpart, lineups consist of several distractor plots (of randomly generated data) and one target (the data plot).

Figure 1 shows a sample lineup of boxplots; participants are expected to identify the most different among the plots shown. In this example, sub-plot 4 is the target because of the noticeably different locations of the two boxplot medians.

Lineups provide a quantitative measurement of the effectiveness of a particular plot: if participants consistently identify the target plot rather than the

• *S. VanderPlas and H. Hofmann are with the Department of Statistics and Statistical Laboratory, Iowa State University, Ames, IA, 50011.*
E-mail: skoons, hofmann@iastate.edu

Manuscript received Month XX, 2015; revised Month XX, XXXX.

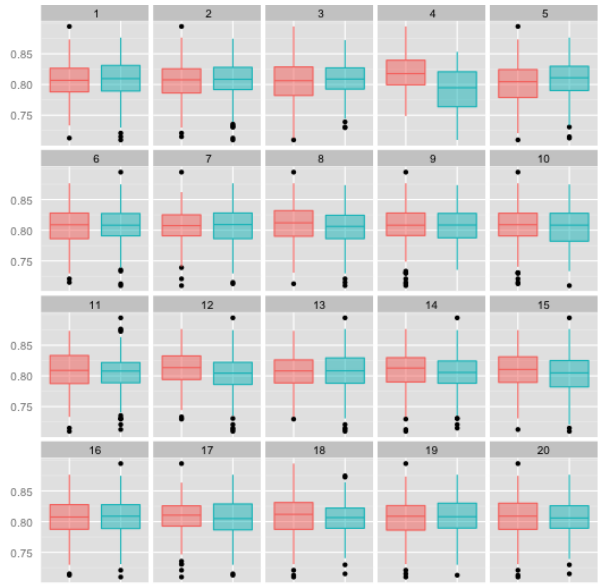


Fig. 1. Sample lineup of boxplots. Participants are instructed to choose the plot which appears most different from the others. In this lineup, plot 4 is the target plot, because the two groups have a large difference in medians.

randomly-generated distractors, the plot effectively shows the difference between real data and random noise. This removes much of the subjectivity from user evaluations of display effectiveness, and the procedure is simple enough that it does not generally require participants to be very familiar with data-based graphics. While previous research [3], [5] has examined the link between certain types of graphical perception and spatial skills, it is important to identify any additional visual skills participants utilize to complete the lineup task, as well as better understand demographic characteristics (math education, research experience, age, gender) which may impact performance [13].

This study is designed to compare lineup performace with visual aptitude and reasoning tests, examining the skills necessary to successfully evaluate lineups.

We compare lineup performace to the visual search task (VST), paper folding test, card rotation test, and figure classification test. The VST measures visual search speed [14], the paper folding and card rotation tests measures spatial manipulation ability, and the figure classification test measures inductive reasoning [15]; all of these skills are at least peripherally recruited during the lineup task, but some may dominate in predicting performance on the lineup task. We hope to facilitate comparison of the lineup task to known cognitive tests, inform the design of future studies, and better understand the perception of statistical lineups.

In section 2, we introduce the tests used in this study and describe how tests are scored. In section 3, we discuss the test scores, comparing established tests with previously validated results, examining demographic characteristics associated with test scores. We discuss multicollinearity in the data, and use principal components analysis and linear regression to draw some conclusions about the similarity between lineups and aptitude tests. Finally, in section 4, we discuss the implications of this study for the lineup protocol, and future work which could extend these results.

2 Methods

2.1 The Lineup Protocol

The lineup protocol [8], [9], [11] is a testing framework that allows researchers to quantify the statistical significance of a graphical finding with the same mathematical rigor as conventional hypothesis tests.

In a lineup test, the plot of the data is placed randomly among a set of, generally 19, distractor plots (or *null plots*) that are rendered from data generated by a model, without a signal (a null model). This sheet of charts is then shown to human observers, who are asked to identify the display that is "the most different". If observers identify the plot drawn from the actual data, this can be reasonably taken as evidence that the data it shows is different from the data of other plots. Let X be the number of observers (out of n) who identify the data plot from the lineup. Under the null hypothesis that the data plot is not different from the other plots, X has approximately a Binomial distribution [9], [10]. If k of the observers identify the data plot from the lineup, the probability $P(X \geq k)$ is the p-value of the corresponding visual test.

This aspect of lineups provides an important tool in the evaluation of graphical findings, in that we can aggregate responses from many individuals to evaluate a single graphic. In addition, however, we can aggregate an individual's score over several lineups to determine how well single individuals perform on the lineup task; data from previous lineup studies suggests that there is significant variation in individual performance as well. This study has the potential to explain some of that variance - if visual skills are important contributors, then perhaps we are simply measuring differences in spatial reasoning ability.

In order to evaluate an individual's overall lineup score, we would subtract a fraction of a point for each question answered incorrectly, and add a full point for each question answered correctly, as shown in 1.

$$\#\text{correct answers} - 1/(19) \cdot \#\text{wrong answers}. \quad (1)$$

This scoring scheme is chosen so that if participants are guessing, the expected score is 0.

Statistical lineups depend on the ability to search for a signal amid a set of distractors (visual search) and the ability to infer patterns from stimuli (pattern recognition). Depending on the choice of plot shown in the lineup, the task of identifying the most different plot might require additional abilities from participants, e.g. polar coordinates depend on the ability to mentally rotate stimuli (spatial rotation) and mentally manipulate graphs (spatial rotation and manipulation). By breaking the lineup task down into its components, we determine which visuospatial factors most strongly correlate with lineup performance, using carefully chosen cognitive tests to assess these aspects of visuospatial ability.

Demographic factors are known to impact lineup performance: country, education, and age affected score on lineup tests, and all of those factors plus gender had an effect on the amount of time spent on lineups [13]. In addition, lineup performance can be partially explained using statistical distance metrics [16], but these metrics do not completely succeed in predicting human performance, in part due to the difficulty of representing human visual ability algorithmically.

One of the most useful features of the lineup protocol is that it allows researchers to conclusively determine which graphics show certain features more conclusively by providing an experimental protocol for comparing graphics based on the accuracy of user conclusions. In addition, lineups provide researchers with a rigorous framework for determining whether a specific graph shows a real, statistically significant effect by comparing a target plot with plots formed using permutations of the same data, providing a randomization test protocol for graphics. As a result, lineups are a useful and innovative tool for evaluating charts; on an individual level, they can also be used to evaluate a specific participant's perceptual reasoning ability in the context of statistical graphics.

2.2 Measures of visuospatial ability

Participants are asked to complete several cognitive tests designed to measure spatial and reasoning ability. Tasks are timed such that participants are under pressure to complete; participants are not expected to finish all of the problems in each section. This allows for a better discrimination between scores and prevents score compression at the top of the range.

The **visual searching task** (VST), shown in figure 7, is designed to test a participant's ability to find a target stimulus in a field of distractors, thus making the visual search task similar in concept to lineups. Historically, visual search has been used as a measure of brain damage [14], [17], [18]; however, similar tasks have been used to measure cognitive performance in a variety of situations, for example under the influence of drugs in [19]. The similarity to the lineup protocol as well as the simplicity of the test and its' lack of color justify the slight deviation from forms of visual search tasks typically used in normal populations.

The **figure classification task** tests a participant's ability to extrapolate rules from provided figures. This task is associated with inductive reasoning ablities (factor I in [15]). An example is shown in figure 8a.

The figure classification test requires the same type of reasoning as the lineups: participants must determine the rules from the provided classes, and extrapolate from those rules to classify new figures. In lineups, participants must determine the rules based on the panels appearing in the lineup; they must then identify the plot which does not conform. As such, the figure classification test has content validity in relation to lineup performance: it is measuring similar underlying criteria.

The **card rotation test** measures a participant's ability to rotate objects in two dimensions in order to distinguish between left-hand and right-hand versions of the same figure. It tests mental rotation skills, and is classified as a test of spatial orientation in [15], though it does require that participants have both mental rotation ability and short-term visual memory. An example is shown in figure 8b. The card rotation test is often used in studies investigating the effect of visual ability on the use of visual aids [5] and statistical graphs [3] in education.

Two-dimensional comparisons are an important component of lineup performance. In some lineup situations, these comparisons sometimes involve translation, but in other lineups, rotation is required. Lineups also require visual short-term memory, so the additional factor measured implicitly by this test does not reduce its potential relevance to lineup performance.

The **paper folding test** measures participants' ability to visualize and mentally manipulate figures in three dimensions. A sample question from the test is shown in figure 8c. It is classified as part of the visualization factor in [15], which differs from the spatial

orientation factor because it requires participants to visualize, manipulate, and transform the figure mentally, which makes it a more complex and demanding task than simple rotation. The paper folding test is associated with the ability to extrapolate symmetry and reflection over multiple steps. Lineups require similar manipulations in two-dimensional space, and also require the ability to perform complex spatial manipulations mentally; for instance, comparing the interquartile range of two boxplots as well as their relative alignment to a similar set of two boxplots in another panel.

Between cognitive tasks, participants were also asked to complete three blocks of 20 lineups each, assembled from previous studies [10], [11]. Participants have 5 minutes to complete each block of 20 lineups. Figure 1 shows a sample lineup of box plots.

In addition to these tests, participants were asked to complete a questionnaire which includes questions about colorblindness, mathematical background, self-perceived verbal/mathematical/artistic skills, time spent playing video games, and undergraduate major. These questions are designed to assess different factors which may influence a participant's skill at reading graphs and performing spatial tasks.

2.3 Test Scoring

All test results were scored so that random guessing produces an expected value of 0; therefore each question answered correctly contributes to the score by 1, while a wrong answer is scored by $-1/(k-1)$, where k is the total number of possible answers to the question. Thus, for a test consisting of multiple choice questions with k suggested answers with a single correct answer each, the score is calculated as

$$\#\text{correct answers} \; - 1/(k-1) \cdot \#\text{wrong answers}. \quad (2)$$

This allows us to compare each participant's score in light of how many problems were attempted as well as the number of correct responses. Combining accuracy and speed into a single number does not only make a comparison of test scores easier, this scoring mechanism is also used on many standardized tests, such as the SAT and the battery of psychological tests [15], [20] from which parts of this test are drawn. The advantage of using tests from the Kit of Factor Referenced Cognitive tests [15] is that the tests are extremely well studied (including an extensive meta-analysis in [21] of the spatial tests we are using in this study) and comparison data are available from the validation of these factors [5], [22], [23] and previous versions of the kit [24].

3 Results

Results are based on an evaluation of 38 undergraduate students at Iowa State University. 61% of the participants were in STEM fields, the others were distributed relatively evenly between agriculture, business, and the social sciences. Students were evenly distributed by gender, and were between 18 and 24 years of age with only one exception. This is reasonably representative[1] of the university as a whole; in the fall 2014 semester, 26% of students were associated with the college of engineering, 24% were associated with the college of liberal arts and sciences, 15% were associated with the college of human sciences, 7% with the college of design, 13% with the business school, and 15% with the school of agriculture.

3.1 Comparison of Spatial Tests with Previously Validated Results

The card rotation, paper folding, and figure classification tests have been validated using different populations, many of which are demographically similar to Iowa State students (naval recruits, college students, late high-school students, and 9th grade students). We compare Iowa State students' unscaled scores in table 1, adjusting data from other populations to account for subpopulation structure and test length.

Table 1 shows mean scores and standard deviation for ISU students and other populations. Values have been adjusted to accommodate for differences in test procedures and sub-population structure; for instance, some data is reported for a single part of a two-part test, or results are reported for each gender separately (adjustment procedure is described in more detail in Appendix B). Once these adjustments have been completed, it is evident that Iowa State undergraduates scored at about the same level as other similar demographics. In fact, both means and standard deviations of ISU students' scores are similar to the comparison groups, which were chosen from available demographic groups based on population similarity.

Comparison population data was chosen to most closely match ISU undergraduate population demographics. Thus, if comparison data was available for 9th and 12th grade students, scores of Iowa State students were compared to scores of 12th grade students, who are closer in age to college students. When data was available from college students and Army enlistees, comparisons of scores were based on other college students, as college students are more likely to have a similar gender distribution to ISU students.

1. http://www.registrar.iastate.edu/sites/default/files/uploads/stats/university/F14summary.pdf

TABLE 1
Comparison of scores from Iowa State students and scores reported in [15]. Scaled scores are calculated based on information reported in the manual, scaled to account for differences in the number of questions answered during this experiment. Data shown are from the population most similar to ISU students, out of the data available. The Visual Search task [14], [17], [18] is not part of the Kit of Factor Referenced Cognitive Test data, and thus we do not have comparison data for the form used in this experiment.

	Card Rotation	Paper Folding	Figure Classification	Visual Search
ISU Students	83.4 (24.1)	12.4 (3.7)	57.0 (23.8)[2]	21.9 (2.3)
Scaled Scores	88.0 (34.8)	13.8 (4.5)	58.7 (14.4)[3]	–
Unscaled Scores	44.0 (24.6)[4]	13.8 (4.5)	M: 120.0 (30.0), F: 114.9 (27.8)	–
Population	approx. 550 male naval recruits	46 college students (1963 version)	suburban 11th & 12th grade students (288-300 males, 317-329 females)	

[2] ISU students took only Part I due to time constraints.
[3] Averages calculated assuming 294 males and 323 females.
[4] Data from Part I only.

Applying the grading protocol discussed in section 2.3, we see that the ranges of lineup and visuospatial test scores do not include zero; this indicates that we do not see random guessing from participants in any task. Figure 2 shows the range of possible scores and the observed score distribution. Participants' scores on the VST indicate score compression; that is, both participants with medium and high visual search abilities scored at the extremely high end of the spectrum. In future experiments, participants should be given less time (or more questions) to better differentiate participants with medium and high-ability.

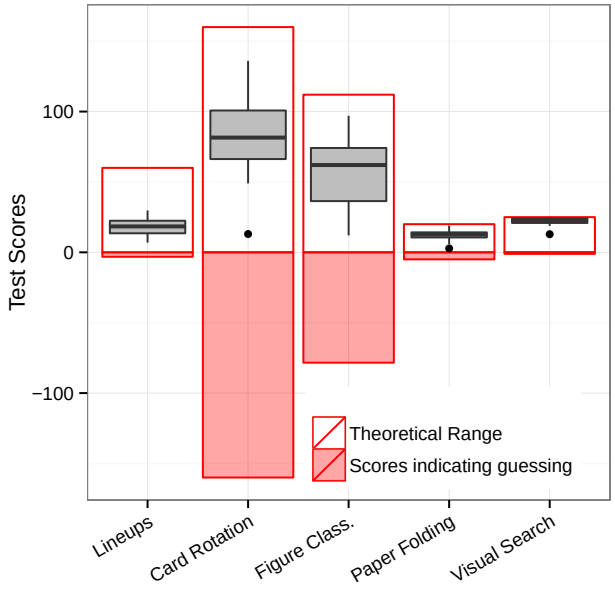


Fig. 2. Test scores for lineups and visuospatial tests. As none of the participants scored at or below zero, we can conclude that there is little evidence of random guessing. We also note the score compression that occurs on the Visual Search test; this indicates that most participants scored extremely high, and thus, participants' scores are not entirely representative of their ability.

3.2 Lineup Performance and Demographic Characteristics

Previous work found a relationship between lineup performance and demographic factors such as education level, country of origin, and age [13]; our participant population is very homogeneous, which allows us to explore factors such as educational background and skills on performance in lineup tests.

Figure 3 shows participants' lineup scores in relationship to their responses in the questionnaire given at the beginning of the study; this allows us to explore effects of demographic characteristics (major, research experience, etc.) on test performance.

Completion of Calculus I is associated with increased performance on lineups; this may be related to general math education level, or it may be that success in both lineups and calculus requires certain visual skills. This association is consistent with findings in [4], which associated mathematical ability to performance on simple graph description tasks. There is also a significant relationship between hours of video games played per week and score on lineups, however, this association is not monotonic and the groups do not have equal sample size, so the conclusion may be suspect. There is a (nearly) significant difference between male and female performance on lineups; this is not particularly surprising, as men perform better on many spatial tests [21] and performance on spatial tests is correlated with phase of the menstrual cycle in women [25]. There is no significant difference in lineup performance for participants of different age, self-assessed skills in various domains, previous participation in math or science research, completion of a statistics class, or experience with AutoCAD. These demographic characteristics were chosen to account for life experience and personal skills which may have influenced the results. Statistical test results are available in appendix C.

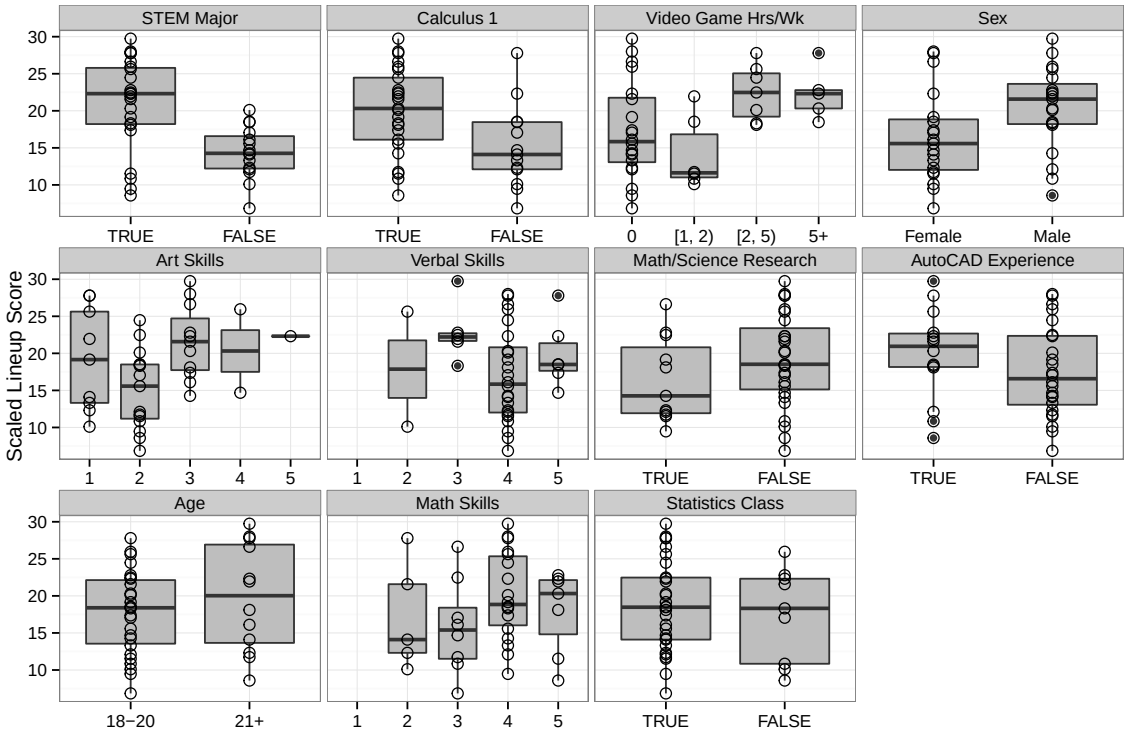


Fig. 3. Demographic characteristics of participants compared with lineup score. Categories are ordered by effect size; majoring in a STEM field, calculus completion, hours spent playing video games per week, and sex are all associated with a significant difference in lineup score.

3.3 Understanding Visual Abilities and Lineup Performance

Results from the visuospatial tests used in this experiment are highly correlated, as shown in figure 4; this is to be expected given that all of these tests are in some way measuring individuals' visual ability. What is of more interest to us is how other factors, such as e.g. general intelligence, mental processing speed, cognitive resources, motivation, and attention affect performance. In order to assess factors contributing to lineup performance, we first examine the separate dimensions measured by the battery of cognitive tests (other than lineups) using principal components analysis on the scaled test scores, then we examine all five tests using the same procedure.

A principal component analysis (PCA) of the four established visuo-spatial tests reveals that they all share a very strong first component, which explains about 64% of the total variability. PC1 is essentially an average across all tests representing a general "visual intelligence" factor. The other principal components span another two dimensions, while the last dimension is weak (at 6%). PC2 differentiates the figure classification test from the visual searching test, whereas PC3 differentiates these two tests from the paper folding test. More detailed results from the 4-test analysis are provided in Appendix D.

Incorporating the lineup task into the principal component analysis, we find the principal components to be fairly similar to the four-component analysis. Table 2 shows the importance of each principal component. From the distribution of the variance components, we see that the lineup test spans an additional dimension within the space of the four established tests.

TABLE 2
Importance of principal components, analyzing all five tests.

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.73	0.84	0.75	0.70	0.48
Proportion of Variance	0.60	0.14	0.11	0.10	0.05
Cumulative Proportion	0.60	0.74	0.85	0.95	1.00

From the rotation matrix (see Table 3) we see that the first principal component, PC1, is again essentially an average across all tests and accounts for 60.1% of the variance in the data. Biplots of the remaining components are provided in Appendix D.2.

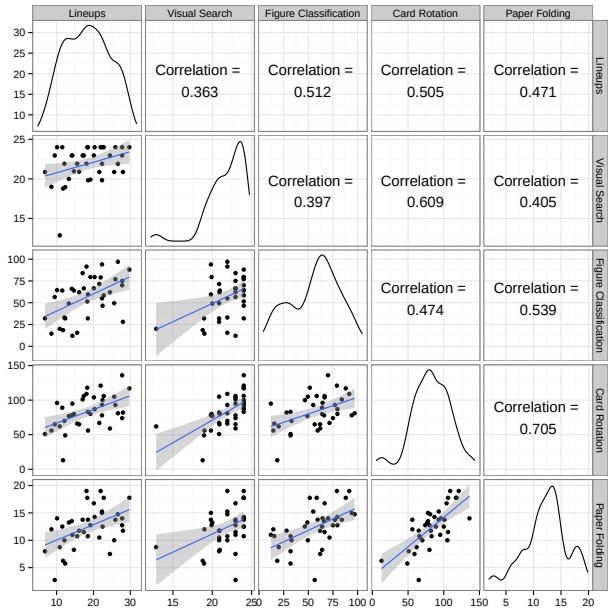


Fig. 4. Pairwise scatterplots of test scores. Lineup scores are most highly correlated with figure classification scores, and are also highly correlated with card rotation scores. Paper folding and card rotation scores are also highly correlated.

TABLE 3
PCA Rotation matrix for all five tests. The first principal component is essentially an average of all five tests.

	PC1	PC2	PC3	PC4	PC5
lineup	0.42	0.49	-0.46	0.60	-0.10
card.rot	0.50	-0.30	0.28	0.23	0.73
fig.class	0.43	0.45	-0.15	-0.75	0.18
folding	0.47	0.07	0.68	0.04	-0.56
vis.search	0.41	-0.69	-0.48	-0.15	-0.33

Figure classification is strongly related to lineups, and as in the four-component PCA, figure classification is strongly represented in the first two principal components. While lineups do span a separate dimension, the PCA suggests that they are most closely related to the figure classification task, and least related to the visual searching task.

This emphasizes the underpinnings of lineups: the test utilizes a visual medium, but is ultimately are a classification task presented in a graphical manner. Using lineups as a proxy for statistical significance tests is similar to using a classifier on pictoral data: while the data is presented "graphically", the participant is actually classifying the data based on underlying summary statistics.

3.4 Linear model of demographic factors

Note that all of the demographic variables in the survey are highy correlated, for example there is a high correlation between STEM majors and taking calculus. Similarly, the correlation between having taken a statistics class and having been involved in mathematics or statistics research is high. Only one student is doing research who has not taken a statistics course. A principal component of the five math/stats questions splits the variables into two main areas: the first principal component is an average of math skills, calculus 1 and STEM, while the second principal component is an average of having taken a statistics class and doing research. We therefore decided to use sums of these variables to come up with a separate math and a stats score. Note, that the correlation between the math and the stats score is almost zero.

We fit a linear model of lineup scores in the thus modified demographic variables and the test scores from the visuo-spatial tests, selecting the best model using AIC and stepwise backwards selection. The result is shown in table 4. Only two covariates stay in the model: PC1 and MATH, reflecting two dimensions of what affects lineup scores. We can think of PC1 as a measure of innate visual or intellectual ability, while the MATH score is a matter of both ability and training. The remaining principal components were not sufficiently associated with lineup score to be included in the model.

TABLE 4
Estimates and significances of a linear model of lineup scores.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	14.1192	1.9149	7.37	0.0000
PC1	1.7672	0.5230	3.38	0.0018
MATH	2.1246	0.8732	2.43	0.0202

3.5 Lineup Types

Each of the three sets of 20 lineups was taken from previous studies on different designs to investigate which plot type most effectively conveyed important characteristics of the underlying data set. Examples lineups for each experiment and more detailed explanations for each of these experiments are provided in Appendix E.

Figure 5 shows the correlations between the three lineup tasks and the figure classification, card rotation, and paper folding tasks. The visual search task is only slightly correlated with the three lineup tasks and is therefore omitted from this figure. Performance on lineup tasks 1 and 2, which dealt with the distributuion of two groups of numerical data, is most strongly correlated with the performance on the figure classification task, which measures general reasoning ability. Performance on lineup task 3, which investigated the potential to visually identify nonnormality in residual

QQ-plots, is more associated with the card rotation and paper folding tests, which measure visuospatial ability. This suggests that certain lineup tasks may require more visual ability than others; in the case of the QQ-lineups a successful evaluation needed participants to mentally rotate plots to compare vertical distances, requiring more mental manipulation than the first two lineup tasks.

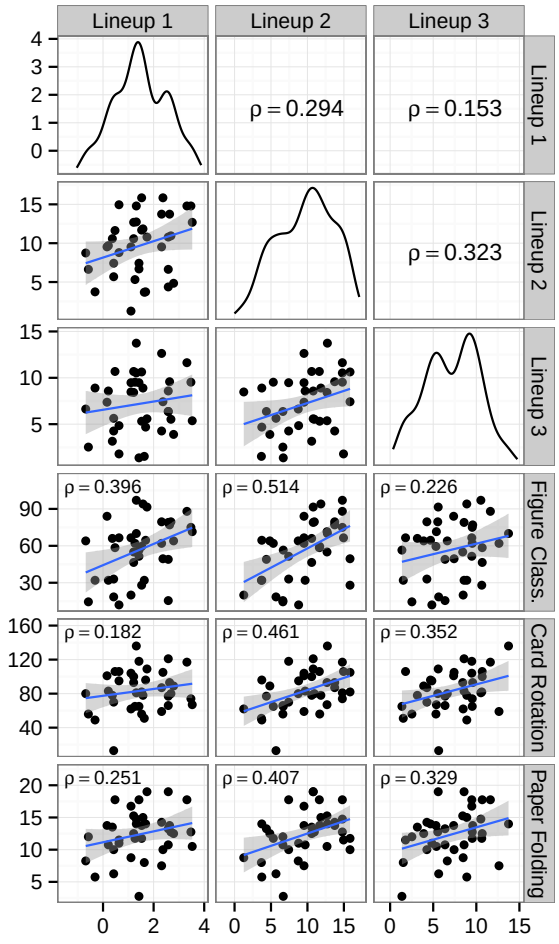


Fig. 5. Pairwise scatterplots of test scores, with lineups separated into the three lineup tasks. All lineup tasks are moderately correlated with the figure classification task, and while tasks 1 and 2 are most strongly correlated with figure classification, lineup task 3 is most strongly correlated with the card rotation and paper folding tasks.

In order to examine which lineup tasks are most closely associated with visual abilities tested in the aptitude portions of an experiment, we employ principal component analysis on participant scores averaged across each block of lineups.

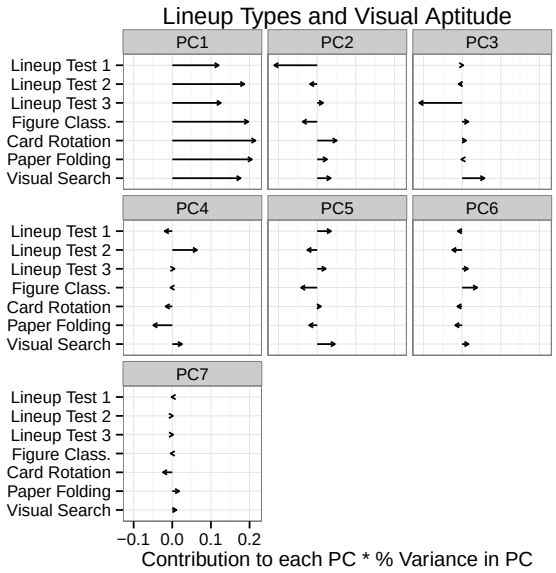
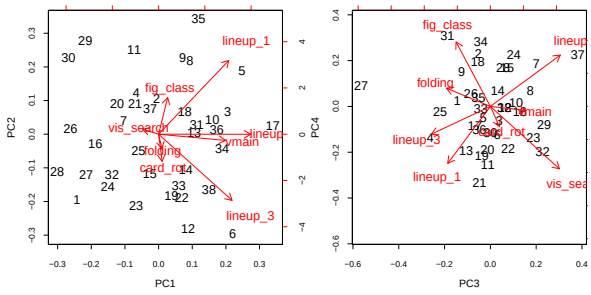


Fig. 6. Lineup block PCA rotation matrix in graphical form.

Figure 6 shows the relative contribution each input variable makes to each principal component, weighted by the importance of each PC. This allows us to consider the rotation matrix visually; for instance, we see that again, PC1 accounts for most of the variance and seems to represent general visual aptitude.

PC2 emphasizes the overlapping variation in performance on lineup test 1 and the figure classification test. PC3 emphasizes the additional variation in performance on lineup test 3 and the visual search test, while PC4 emphasizes the extra variability in performance on lineup test 2 and the paper folding test. All three lineups, plus the figure classification, paper folding, and visual search tests contribute to PC5. PC6 and PC7 jointly account for less than 10% of the variance in the data and do not display any distinct patterns in the loadings.

While lineups constitute a distinct principal component when aggregated into a single score, this PCA of the separate lineup types and the cognitive tests indicates that different lineup experiments exist in different principal component loadings. Overall, there is an additional principal component gained from separating the lineup blocks by experiment type. As lineup tasks 1 and 2 contained similar plot types, it is possible that those two tasks overlap in the component space while lineup task 3 is distinct.



The relationship between participant performance on different types of lineups (and different types of plots) and performance on tests of spatial ability bears further investigation; this study suggests that there may be an effect, but there is simply not enough variation to make definitive conclusions about the relationship between different measures of visual aptitude and performance on specific lineup tasks.

A larger study might not only address the question of which lineup tasks require certain visual skills, but could also address the use of different types of plots from a perceptual perspective. Preliminary results of performance on different types of plots are shown in appendix F, but a larger study is needed for definitive results. The advantage of the lineup protocol is that it allows us to not only consider individual performance but also to compare aggregate performance on different types of plots. Integrating information about the visual skills required for each type of plot provides information about the underlying perceptual skills and experience required to read different types of plots.

4 Discussion and Conclusions

Performance on lineups is strongly related to performance on tests of visual ability; however, this relationship is mediated by demographic factors such as major (STEM or not) and completion of calculus I. In addition to these demographic factors, many facets of intelligence are highly correlated; those who score higher on general aptitude tests may score higher on tests of visual ability (and may also score higher on lineup tests).

Despite these caveats, we have demonstrated that the general lineup task is most closely related to a classification task, rather than tests of spatial ability. This is an important verification of a tool that is useful for examining statistical graphics, as it emphasizes the idea that while the testing medium is graphical in nature, the task is in fact a classification task, where the viewer must determine the most important features of each plot and then identify which plot is different.

When lineup tasks with different goals are viewed separately, there is some indication that different tasks are associated with different visual abilities. Lineup tasks 1 and 2 are quite similar, and are more associated with the figure classification task; lineup task 3, while still moderately correlated with the figure classification task, is also moderately correlated with the visuospatial ability tests (paper folding, card rotation). Future studies testing larger sets of lineups may be useful to understand which types of plots require additional visuospatial skills, as plots which appeal to a wider audience may be more successful when conveying important information.

In addition to this theoretical information, the figure classification test may be useful for pre-screening participants in future online lineup studies. Such studies often suffer from participants who do not take the task seriously, and internal verification questions, as well as pre-qualification tasks are often used to reduce extraneous variability. While it would be impractical to require participants to score well on several different tests, it would be reasonable to ask participants to pre-qualify for a task by completing a figure classification test. As the figure classification test is different from the lineup task, this would not bias participants' scores on the domain of interest, but would ensure a participant pool that is sufficiently motivated to complete the lineup questions.

The demographic results from this study indicate that in future lineup studies, it may be important to record information about participants' mathematical training, so that studies can be compared across participant pools with more reliability.

All results and data shown here were collected and analyzed in accordance with IRB # 13-581.

References

[1] A. D. Baddeley and G. Hitch, "Working memory," Psychology of learning and motivation, vol. 8, pp. 47–89, 1974.

[2] I. Vekiri, "What is the value of graphical displays in learning?" Educational Psychology Review, vol. 14, no. 3, pp. 261–312, 2002.

[3] T. Lowrie and C. M. Diezmann, "Solving graphics problems: Student performance in junior grades," The Journal of Educational Research, vol. 100, no. 6, pp. 369–378, 2007.

[4] P. Shah and P. A. Carpenter, "Conceptual limitations in comprehending line graphs." Journal of Experimental Psychology: General, vol. 124, no. 1, p. 43, 1995.

[5] R. E. Mayer and V. K. Sims, "For whom is a picture worth a thousand words? extensions of a dual-coding theory of multimedia learning." Journal of educational psychology, vol. 86, no. 3, p. 389, 1994.

[6] M. Scaife and Y. Rogers, "External cognition: how do graphical representations work?" International journal of human-computer studies, vol. 45, no. 2, pp. 185–213, 1996.

[7] J. Zhang, "The nature of external representations in problem solving," Cognitive science, vol. 21, no. 2, pp. 179–217, 1997.

[8] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham, "Statistical inference for exploratory data analysis and model diagnostics," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 367, no. 1906, pp. 4361–4383, 2009.

[9] H. Wickham, D. Cook, H. Hofmann, and A. Buja, "Graphical inference for infovis," Visualization and Computer Graphics, IEEE Transactions on, vol. 16, no. 6, pp. 973–979, 2010.

[10] M. Majumder, H. Hofmann, and D. Cook, "Validation of visual statistical inference, applied to linear models," Journal of the American Statistical Association, vol. 108, no. 503, pp. 942–956, 2013.

[11] H. Hofmann, L. Follett, M. Majumder, and D. Cook, "Graphical tests for power comparison of competing designs," IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 12, pp. 2441–2448, 2012.

[12] A. Loy, L. Follett, and H. Hofmann, "Variations of q-q plots – the power of our eyes!" The American Statistician, vol. tentatively accepted, pp. ???–???, 2015.

[13] M. Majumder, H. Hofmann, and D. Cook, "Human factors influencing visual statistical inference," 2014.

[14] G. Goldstein, R. B. Welch, P. M. Rennick, and C. H. Shelly, "The validity of a visual searching task as an indicator of brain damage." Journal of consulting and clinical psychology, vol. 41, no. 3, p. 434, 1973.

[15] R. B. Ekstrom, J. W. French, H. H. Harman, and D. Dermen, Manual for kit of factor-referenced cognitive tests, Educational Testing Service, Princeton, NJ, 1976.

[16] N. Roy Chowdhury, D. Cook, H. Hofmann, M. Majumder, and Y. Zhao, "Utilizing distance metrics on lineups to examine what people read from data plots," 2014.

[17] M. A. DeMita, J. H. Johnson, and K. E. Hansen, "The validity of a computerized visual searching task as an indicator of brain damage," Behavior Research Methods & Instrumentation, vol. 13, no. 4, pp. 592–594, 1981.

[18] M. Moerland, A. Aldenkamp, and W. Alpherts, "A neuropsychological test battery for the apple ll-e," International journal of man-machine studies, vol. 25, no. 4, pp. 453–467, 1986.

[19] K. J. Anderson and W. Revelle, "The interactive effects of caffeine, impulsivity and task demands on a visual search task," Personality and Individual Differences, vol. 4, no. 2, pp. 127–134, 1983.

[20] J. Diamond and W. Evans, "The correction for guessing," Review of Educational Research, pp. 181–191, 1973.

[21] D. Voyer, S. Voyer, and M. P. Bryden, "Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables." Psychological bulletin, vol. 117, no. 2, p. 250, 1995.

[22] K. W. Schaie, S. B. Maitland, S. L. Willis, and R. C. Intrieri, "Longitudinal invariance of adult psychometric ability factor structures across 7 years." Psychology and aging, vol. 13, no. 1, p. 8, 1998.

[23] E. Hampson, "Variations in sex-related cognitive abilities across the menstrual cycle," Brain and cognition, vol. 14, no. 1, pp. 26–43, 1990.

[24] J. W. French, R. B. Ekstrom, and L. A. Price, Kit of reference tests for cognitive factors, Educational Testing Service, Princeton, NJ, 1963.

[25] M. Hausmann, D. Slabbekoorn, S. H. Van Goozen, P. T. Cohen-Kettenis, and O. Güntürkün, "Sex hormones affect spatial abilities during the menstrual cycle." Behavioral neuroscience, vol. 114, no. 6, p. 1245, 2000.

APPENDIX A
VISUOSPATIAL APTITUDE TESTS

The Visual Search Task (VST), shown in figure 7, is used to measure a person's ability to locate a target amid a field of distractors.

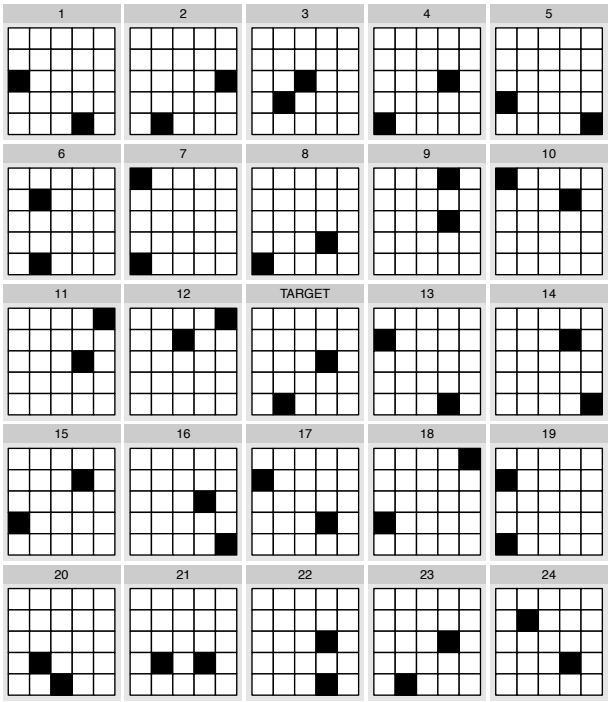
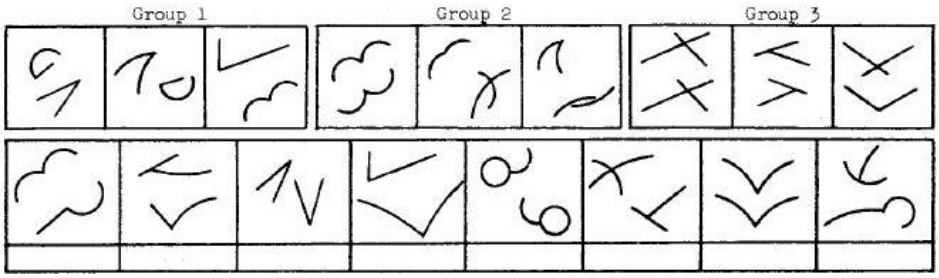


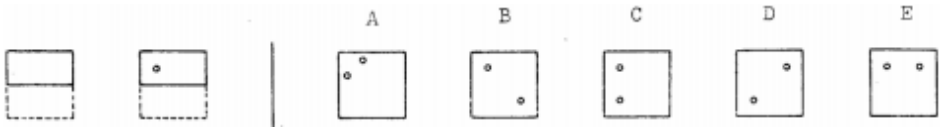
Fig. 7. Visual Search Task (VST). Participants are instructed to find the plot numbered 1-24 which matches the plot labeled "Target". Participants will complete up to 25 of these tasks in 5 minutes.

Figure 8 shows samples from the figure classification task, the card rotation task, and the paper folding task. All three tasks are part of the Kit of Factor-Referenced Cognitive Tests [15].



(a) Figure Classification Task. Participants classify each figure in the second row as belonging to one of the groups above. Participants complete up to 14 of these tasks (each consisting of 8 figures to classify) in 8 minutes.

(b) Card Rotation Task. Participants mark each figure on the right hand side as either the same as or different than the figure on the left hand side of the dividing line. Participants complete up to 20 of these tasks (each consisting of 8 figures) in 6 minutes.



(c) Paper Folding Task. Participants are instructed to pick the figure matching the sequence of steps shown on the left-hand side. Participants complete up to 20 of these tasks in 6 minutes.

Fig. 8. Visuospatial tests

APPENDIX B
SCALING SCORES

To calculate "scaled" comparison scores between tests which included different numbers of test sections (as shown in table 1), we scaled the mean in direct proportion to the number of questions (thus, if there were two sections of equivalent size, and the reference score included only one of those sections, we multiplied the reported mean score by two). The variance calculation is a bit more complicated: In the case described in the main text, where the reference section contained half of the questions, the variance is multiplied by two, causing the standard deviation to be multiplied by approximately 1.41.

This scaling gets slightly more complicated for scores which have two sub-groups, as with the figure classification test, which separately sumarizes male and female participants' scores. To get a single unified score with standard deviation, we completed the following calculations:

$$\mu_{\text{all}} = (N_F\mu_F + N_M\mu_M)/(N_F + N_M) \quad (3)$$

$$\sigma_{\text{all}} = \sqrt{(N_F\sigma_F^2 + N_M\sigma_M^2)\,/(N_F + N_M)}, \quad (4)$$

where μ_F and μ_M are the mean of female and male scores respectively, N_F and N_M are the number of participants in each group, and σ_F^2 and σ_M^2 are the variance of each group. Substituting in the provided numbers, we get

$$\begin{aligned}\mu_{\text{all}} &= (323\cdot 114.9 + 294\cdot 120.0)\,/(323+294)\\ &= 58.7\\ \sigma_{\text{all}} &= \sqrt{(323(27.8)^2 + 294(30)^2)\,/(323+294)}\\ &= 14.4.\end{aligned}$$

Whenever participants in two studies were not exposed to the same number of questions, the resulting scores are not comparable: both overall scores and their standard deviations are different. We can achieve comparability by scaling the scores accordingly. For example, in order to account for the fact that ISU students took only part I of two parts to the figure classification test (and thus completed half of the questions), we adjust the transformation as follows:

$$\begin{aligned}\mu_{\text{part I}} &= 1/2 \cdot \mu_{\text{all}}\\ \sigma_{\text{part I}} &= 1/\sqrt{2} \cdot \sigma_{\text{all}}\end{aligned}$$

APPENDIX C
LINEUP PERFORMANCE AND DEMOGRAPHIC CHARACTERISTICS

'Results' of lineup score - te caption in the table is too generic. give more details.

Table 5 provides the results of a sequence of linear models fit to the lineup data. Each row in the table represents a single model, with one predictor variable (a factor with two or more levels). Due to sample size considerations, multiple testing corrections were not performed; in addition, the independent variables are correlated: in our sample, males are more likely to have completed Calculus 1, but are also more likely to spend time playing video games. As such, a model including two or more of the significant predictor variables shows all included variables to be nonsignificant. To better understand the effects of these variables, a larger study would be required.

TABLE 5
Results of lineup score modeled by single demographic variables.

Variable	DF	MeanSq	F	p.val
STEM Major	1	401.517	14.44	0.001
Calculus 1	1	204.569	6.15	0.018
Video Game hrs	3	108.847	3.44	0.028
Sex	1	140.844	4.02	0.053
Art Skills	4	75.891	2.28	0.082
Verbal Skills	3	60.220	1.68	0.191
STEM Research	1	59.670	1.60	0.214
AutoCAD	1	50.893	1.36	0.252
Age	1	34.434	0.91	0.348
Math Skills	3	37.039	0.98	0.416
Statistics Class	1	9.062	0.23	0.631

APPENDIX D
PRINCIPAL COMPONENT ANALYSIS OF VISUO-SPATIAL TESTS

D.1 PCA of the Four Cognitive Tests

'Importance of principal components *in an analysis* of four tests' – be more precise: you mean the PCA. Give some details as to how importance is measured.

TABLE 6
Importance of principal components in an analysis of four tests of spatial ability: figure classification, paper folding, card rotation, and visual search.

	PC1	PC2	PC3	PC4
Standard deviation	1.61	0.81	0.73	0.49
Proportion of Variance	0.64	0.16	0.13	0.06
Cumulative Proportion	0.64	0.81	0.94	1.00

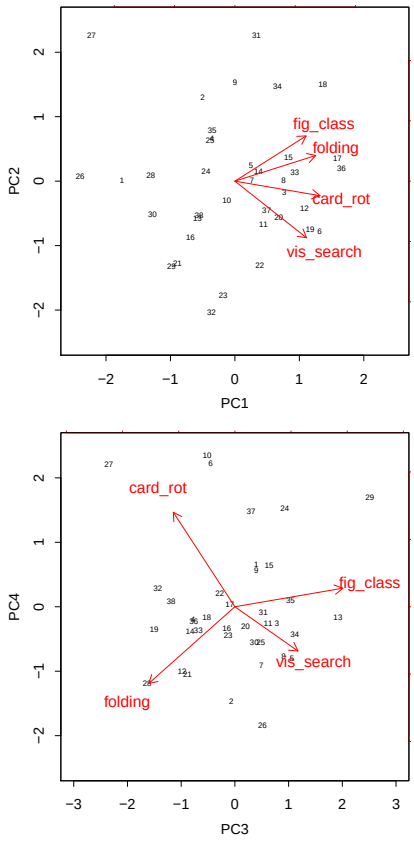


Fig. 9. Plots of principal components 1-4 with observations. PCA was performed on the four cognitive tests used to understand the cognitive demands of the lineup protocol.

Table 6 contains the importance of each resultant PC and the proportion of the variance each PC represents.

PC1 accounts for about 60% of the variance; figure 9 and table 7 confirm that PC1 is a measure of the similarity between all 4 tests; that is, a participant's general (or visual) aptitude. PC2 differentiates the figure classification test from the visual searching test, while PC3 differentiates these two from the paper folding test. PC4 is not particularly significant (it accounts for 5.9% of the variance), but it differentiates the card rotation task from the paper folding task.

TABLE 7
PCA Rotation matrix for the four cognitive tests.

	PC1	PC2	PC3	PC4
card.rot	0.55	-0.19	-0.38	0.72
fig.class	0.46	0.58	0.66	0.14
folding	0.52	0.33	-0.53	-0.59
vis.search	0.46	-0.72	0.38	-0.34

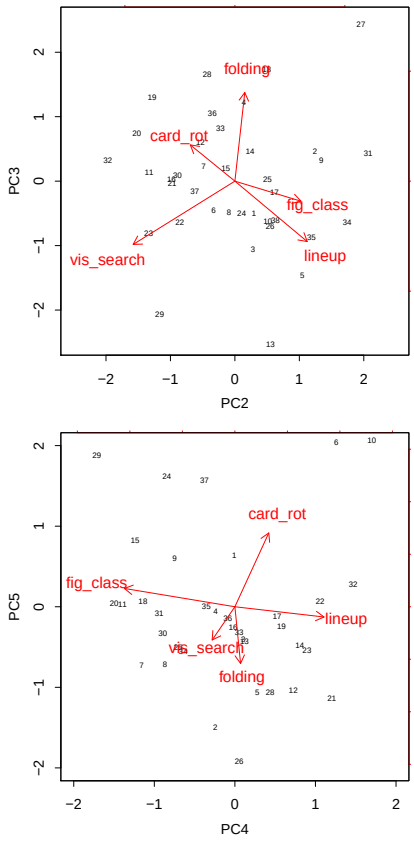


Fig. 10. Plots of principal components 2-5 with observations.

figure 9 and 10 are Biplots, say so in the caption - in the discussion of the figures and the tables in the text, please think again of adding the sentence summarising what is shown.

D.2 PCA of Cognitive Tests and Lineups

PC1 is essentially an average across all tests representing a general "visual intelligence" factor. Biplots of the remaining principal components are shown in figure 10.

Figure classification is strongly related to lineups (PC2, PC3). Performance on the visual search task is also related to lineup performance (PC3). These two components highlight the shared demands of the lineup task and the figure classification task: participants must establish categories from provided stimuli and then classify the stimuli accordingly.

The visual search task is also clearly important to lineup performance: PC3 captures the similarity between the visual search and lineup performance, and aspects of these tasks are negatively correlated with aspects of the paper folding and card rotation tasks within PC3. Paper folding does not seem to be strongly associated with lineup performance outside of the first principal component; card rotation is only positively associated with lineup performance in PC4.

PC4 captures the similarity between lineups and the card rotation task and separates this similarity from the figure classification task; this similarity does not account for much extra variance (10%), but it may be that only some lineups require spatial rotation skills. PC5 contains only 5% of the remaining variance, and is thus not of much interest, however, it seems to capture the relationship between the card rotation task and the paper folding and visual search tasks.

Appendix E
Lineup Task Examples

E.1 Lineup Set 1

The experiment in the first lineup section examined the use of boxplots, density plots, histograms, and dotplots to compare two groups which vary in mean and sample size. The experiment was originally designed to explore the use of lineups to test plots of competing design [11]. This set of lineups consists of 20 plots selected from the plots used in the full experiment; each set of data is displayed with each of the four plot types.

Describe performance - need to get data from original experiment.

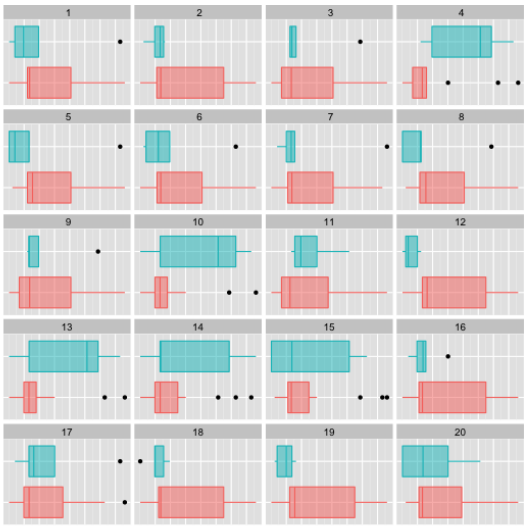


Fig. 11. Boxplots used to compare the two distributions.

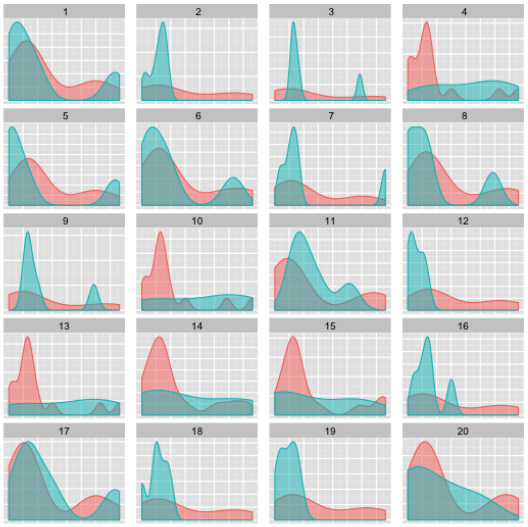


Fig. 12. Density plots used to compare the two distributions.

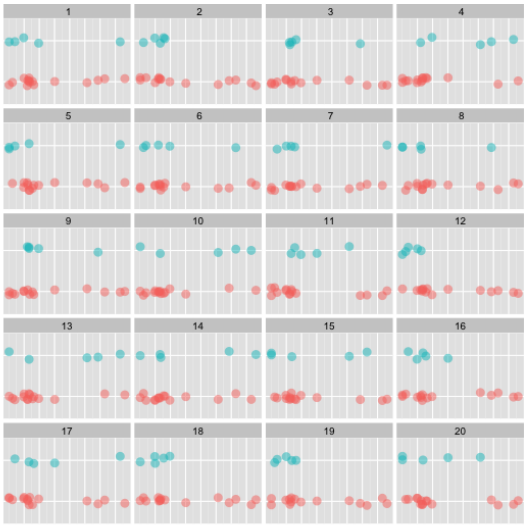


Fig. 13. Dotplots used to compare the two distributions.

Fig. 14. Histograms used to compare the two distributions.

E.2 Lineup Set 2

The second lineup section also explored two groups of data, this time comparing boxplots, bee swarm boxplots, boxplots with overlaid jittered data, and violin plots. Participants were much more accurate in this experiment than in the experiment described previously, because of the types of plots compared as well as the underlying data distributions.

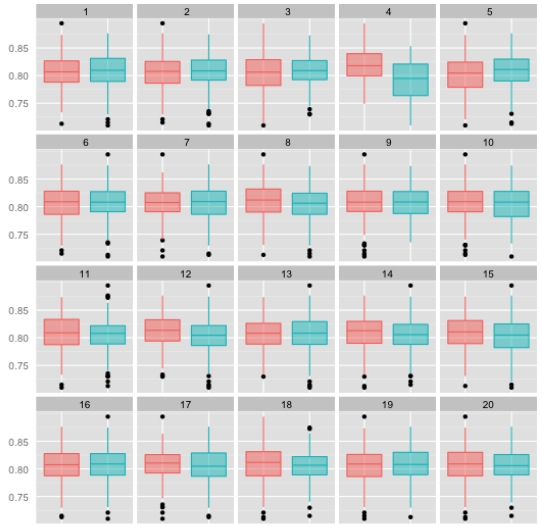


Fig. 15. Boxplots.

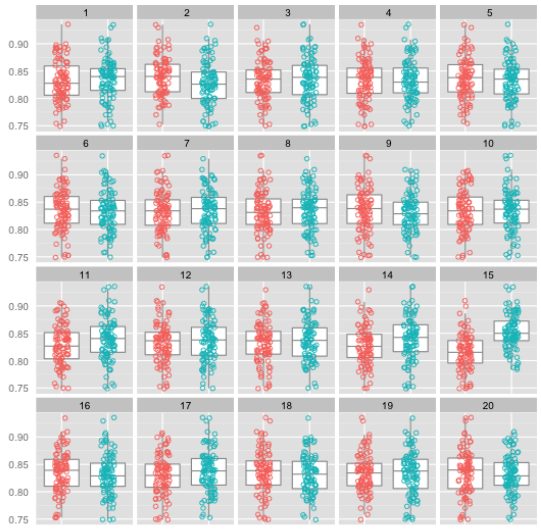


Fig. 16. Boxplots with jittered points.

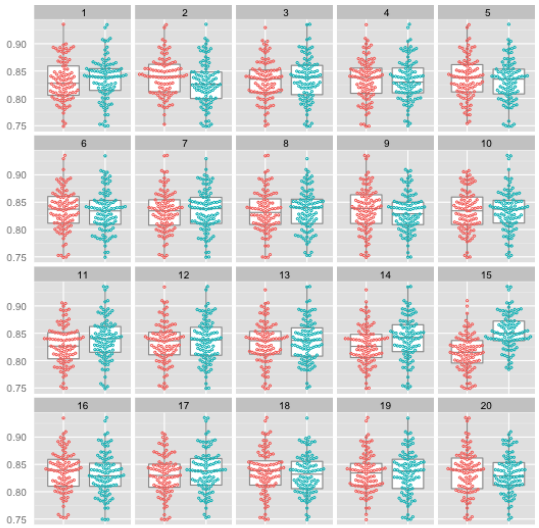


Fig. 17. Bee swarm boxplots.

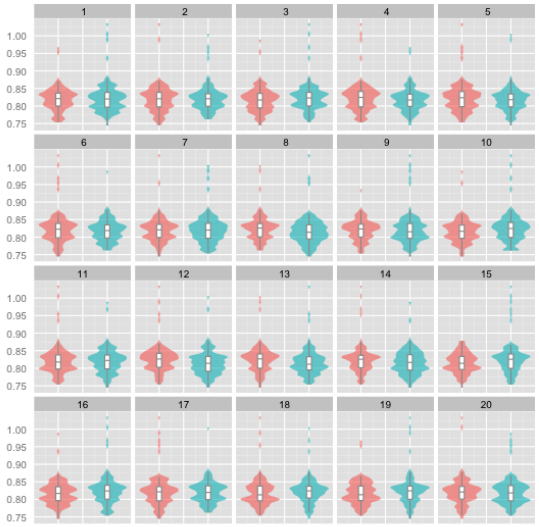


Fig. 18. Violin plots.

E.3 Lineup Set 3

The final lineup section explored QQ-plots from various model simulations, using reference lines, acceptance bands, and rotation to determine which plots allowed participants to most effectively identify violations of normality. Rotated qq plots showed lower performance because participants could more accurately compare acceptance bands to residuals, and thus could identify that the reference bands were too liberal. As a result, performance was somewhat lower for rotated plots, even though participants could

more accurately compare the residuals to the reference bands.

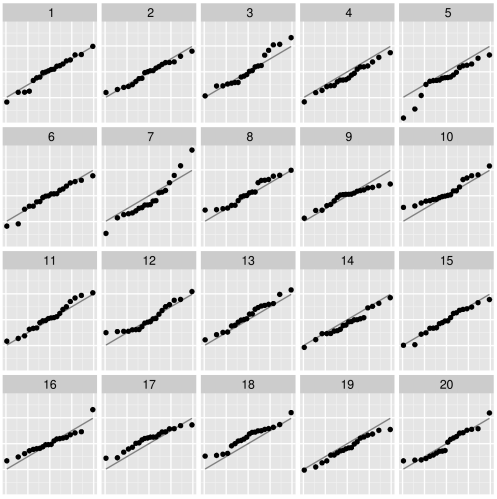


Fig. 19. QQ Plot with guide line.

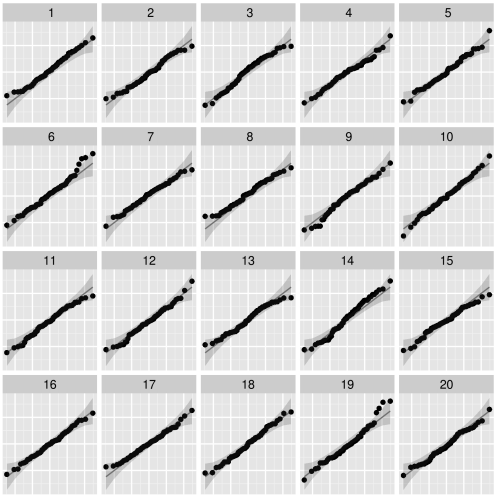


Fig. 20. QQ Plot with acceptance bands.

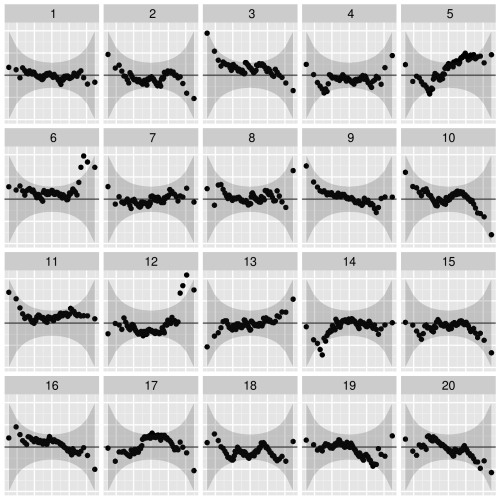


Fig. 21. QQ Plot rotated 45 degrees.

APPENDIX F
LINEUP PLOT TYPES

We can also compare participants' performance on specific types of lineup plots compared with their scores on the visual aptitude tests, for instance, accuracy on lineups which require mental rotation may be related to performance on the card rotation task.

Figure 22 compares performance on each different type of plot. As two different lineup tasks utilized boxplots to test different qualities of the distribution of data (outliers vs. difference in medians), different tasks are shown as different colors, so that accuracy on tasks which are shown in blue can be compared to other blue density curves.

Figure 23 shows the association between scaled score on each type of lineup and score on the visual reasoning tests. Sample size for each plot type is fairly small - between 5 and 10 plots per individual, so there is low power for systematic inference, but we can establish that the card rotation task is much more significantly associated with the QQ plots tasks compared to the other tasks. In addition, rotated qq plots seem to be much more asssociated with the paper folding task scores than other qq plot tasks; this may be because they require more visual manipulation than other qq plots.

For comparison, the correlation between general lineup score (non-subdivided) and the card rotation test score was 0.505, the correlation between general lineup score and the figure classification test was 0.512, and the correlation between lineup score and the paper folding test was 0.471.While we can compare

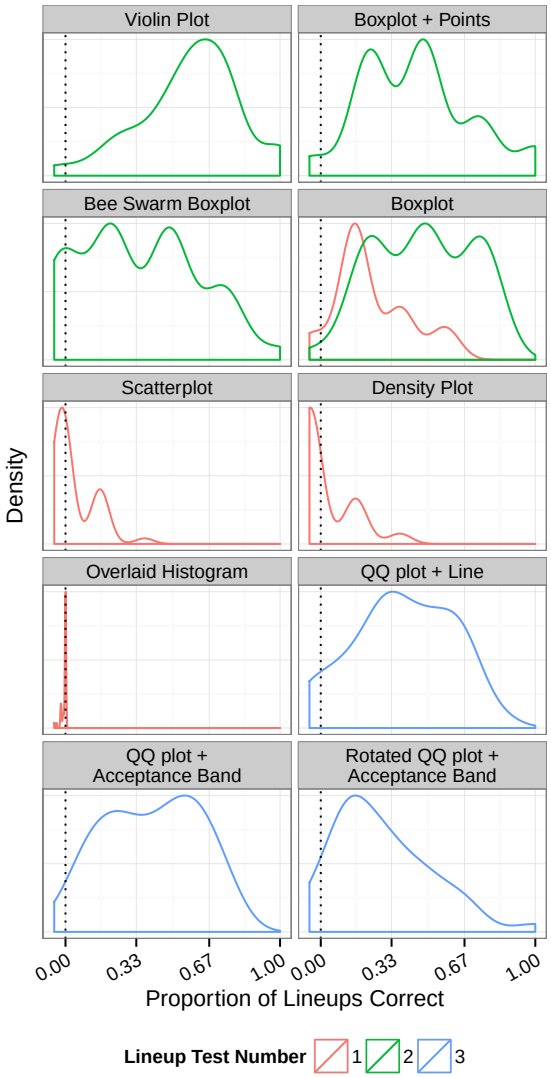


Fig. 22. Plot of scaled scores for different types of lineups

the correlation strength between tasks, it is clear that the correlation between the score on any single lineup type and a particular visual aptitude score is lower than the overall relationship that we attribute to visual ability. Additional data is imperative to understand the reasoning required for specific types of plots - it is likely that the 5-10 trials per participant presented in each chart in figure 23 are simply not sufficient to uncover any specific relationship between reasoning ability and lineup task.

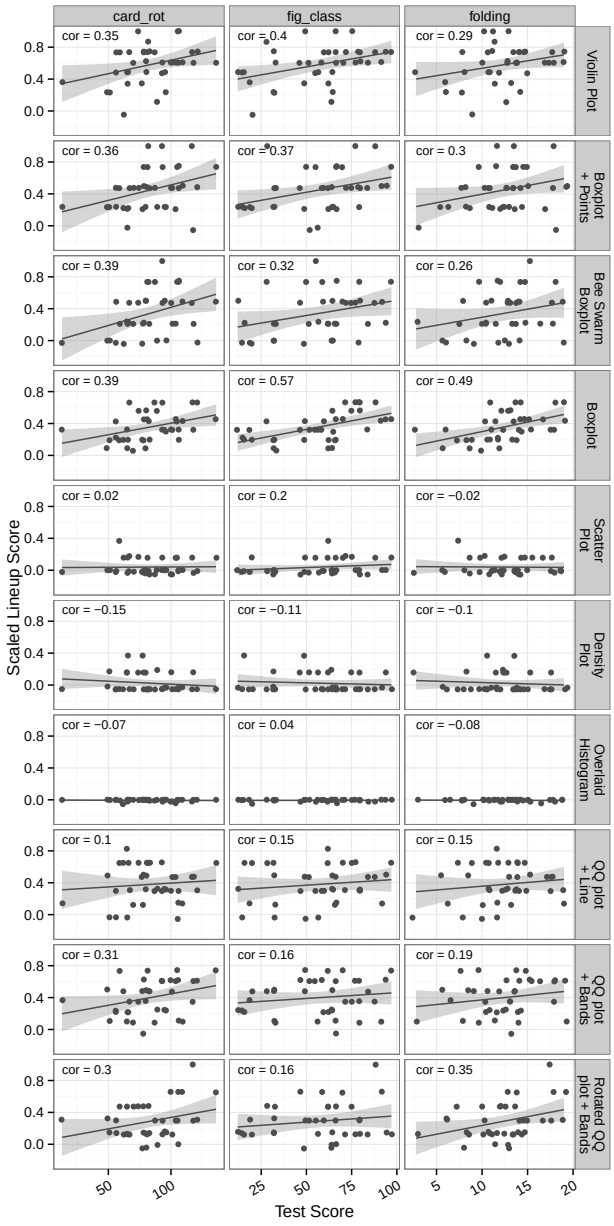


Fig. 23. Plot of scaled lineup scores by aptitude test scores.