Spatial Reasoning and Statistical Graphics

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

Abstract—The abstract goes here.

Index Terms—Computer Society, IEEEtran, journal, LaTeX, paper, template.

1 Introduction

THIS demo file is intended to serve as a "starter file" for IEEE Computer Society journal papers produced under LaTeX using IEEEtran.cls version 1.8a and later.

Next few steps for the analysis: (1) I think we should add a linear model for lineup scores that includes the demographics of participants as well; (2) could you do the scoring of lineups by type of plot used and investigate the relationship between those types of lineups and the other visual tests ... not sure that that will pay off, but it's interesting. (3) in terms of why we would use lineups in a test comparison ... lineups give us a way to say 'this is correct' which regular graphics don't. So we should put a spin sentence on at the intro and the back that we are suggesting to use lineups as a tool to not only evaluate individuals (which we can see is possible) but also as a tool to evaluate charts ... which we've done before ... and tie it into the regular testing literature ...

Intro about statistical graphics, lineups, ...

1.1 Spatial Reasoning and Statistical Graphics

Statistical graphics provide quick summaries of data, models, and results, but these displays may not be equally useful to all viewers. Mathematical ability is important, even for the simplest graphs: mathematical ability, not spatial ability, was shown [1] to be associated with accuracy on a simple two-dimensional line graph. Spatial ability becomes more important, however, 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 [2] 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" [3] that guide, constrain, and facilitate cognitive behavior [4]. This constraint reduces memory load and makes more cognitive resources available for other tasks, but also implicates visual ability as an important factor in graph comprehension.

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 [5]. 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 [6].

1.2 The Lineup Protocol

More lineup introduction here

Statistical lineups [7], [8], [9] depend on the ability to search for a signal amid distractors (visual search) and the ability to infer patterns from stimuli (pattern recognition). Some lineups (polar coords) also 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 component parts, we may be able to determine which visuospatial factors most strongly

- *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.

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 [10]. In addition, lineup performance can be partially explained using statistical distance metrics [11], but these metrics do not completely succeed in predicting human performance, in part due to the difficulty of representing human visual ability algorithmically. In this paper, we examine some demographic characteristics as well as characteristics such as spatial and pattern recognition ability with the goal of understanding the skills necessary for accurate lineup performance as well as underlying variation in lineup performance.

2 Methods

2.1 Measures of visuospatial ability

Participants are asked to complete several cognitive tests designed to measure spatial and reasoning ability. Tasks are times 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 1, 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 [12], [13], [14]; however, similar tasks have been used to measure cognitive performance in a variety of situations, for example under the influence of drugs in [15]. 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 [16]). An example is shown in figure 2a.

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.

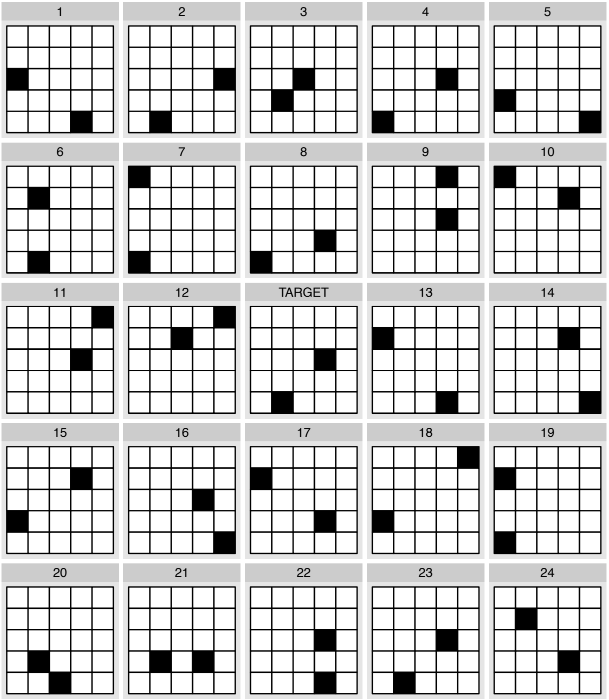
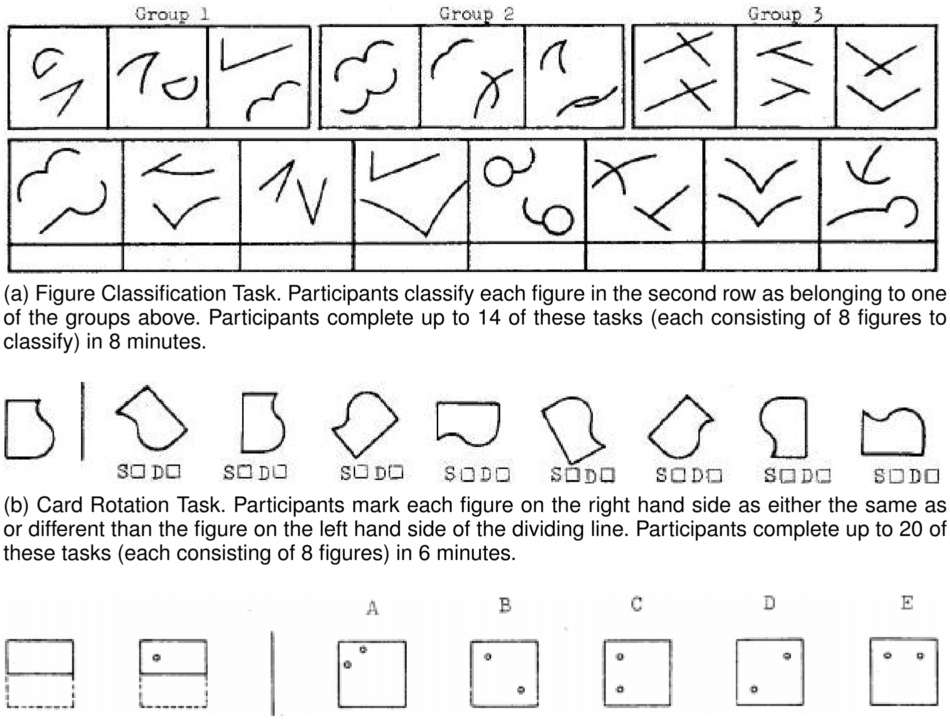


Fig. 1. Visual Search Task. Participants are instructed to find the plot numbered 1-24 which matches the plot labeled "Target". Participants will complete up to 25 of these tasks in 5 minutes.

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 [16], though it does require that participants have both mental rotation ability and short-term visual memory. An example is shown in figure 2b. The card rotation test is often used in studies investigating the effect of visual ability on the use of visual aids [2] and statistical graphs [6] 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 2c. It is classified as part of the visualization factor in [16], 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



(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. 2. Visuospatial tests

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 will also complete three blocks of 20 lineups each, assembled from previous studies [7], [17]. Participants have 5 minutes to complete each block of 20 lineups. Figure 3 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.

2.2 Test Scoring

All test results were scored so that random guessing produces an expected value of 0; therefore each ques-

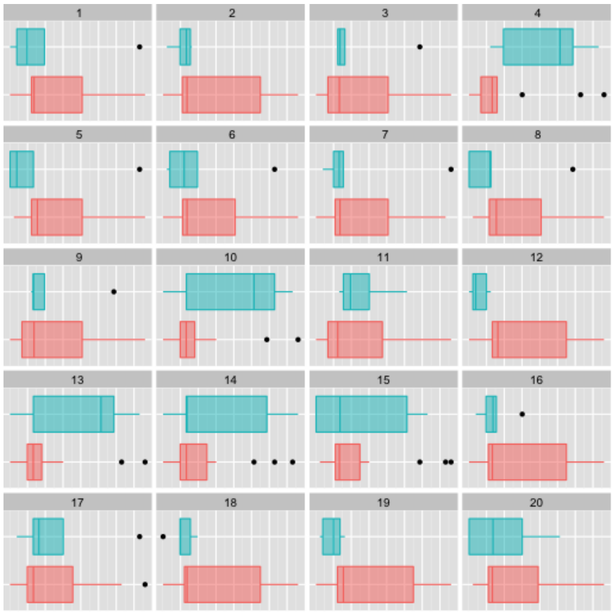


Fig. 3. Sample lineup of boxplots. Participants are instructed to choose the plot which appears most different from the others. In this lineup, plot 13 is the target.

tion 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

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

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 [16], [18] from which parts of this test are drawn. The advantage of using tests from the Kit of Factor Referenced Cognitive tests [16] is that the tests are extremely well studied (including an extensive meta-analysis in [19] of the spatial tests we are using in this study) and comparison data are available from the validation of these factors [2], [20], [21] and previous versions of the kit [22].

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 2012 semester, 23% of students were associated with the college of engineering, 23% were associated with the college of liberal arts and sciences, 23% were associated with the college of human sciences, 12% with the business school, and 14% 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 A). 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.

Applying the grading protocol discussed in section 2.2, 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 4 shows the range of possible scores and the observed score distribution.

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 [10]; our participant population is very homogeneous, which allows us to explore factors such as educational background and skills on performance in lineup tests.

Figure 5 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 [1], 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

[1] http://www.ir.iastate.edu/PDFfiles/2012-2013%20Student%20Profile.pdf

TABLE 1

Comparison of scores from Iowa State students and scores reported in [16]. 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 [12], [13], [14] 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.

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 [19] and performance on spatial tests is correlated with phase of the menstrual cycle in women [23]. 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 B.

3.3 Understanding Visual Abilities and Lineup Performance

Results from the visuospatial tests used in this experiment are highly correlated, as shown in figure 6; 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,

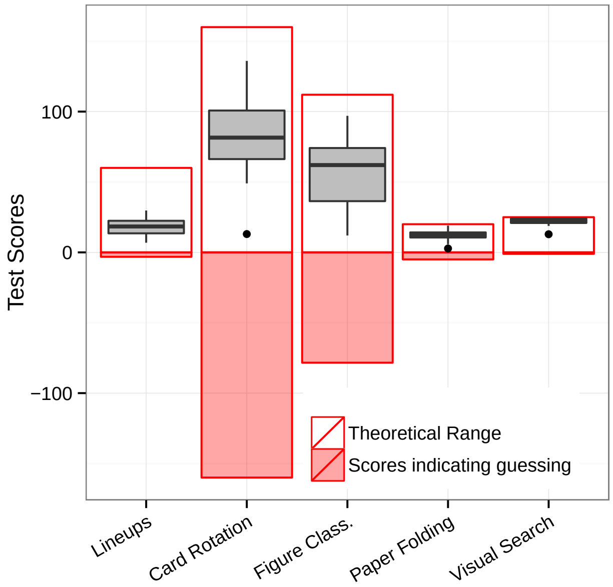


Fig. 4. Test scores for lineups and visuospatial tests. As none of the participants scored at or below zero, we can conclude that there is little evidence of random guessing. We also note the score compression that occurs on the Visual Search test; this indicates that most participants scored extremely high, and thus, participants' scores are not entirely representative of their ability.

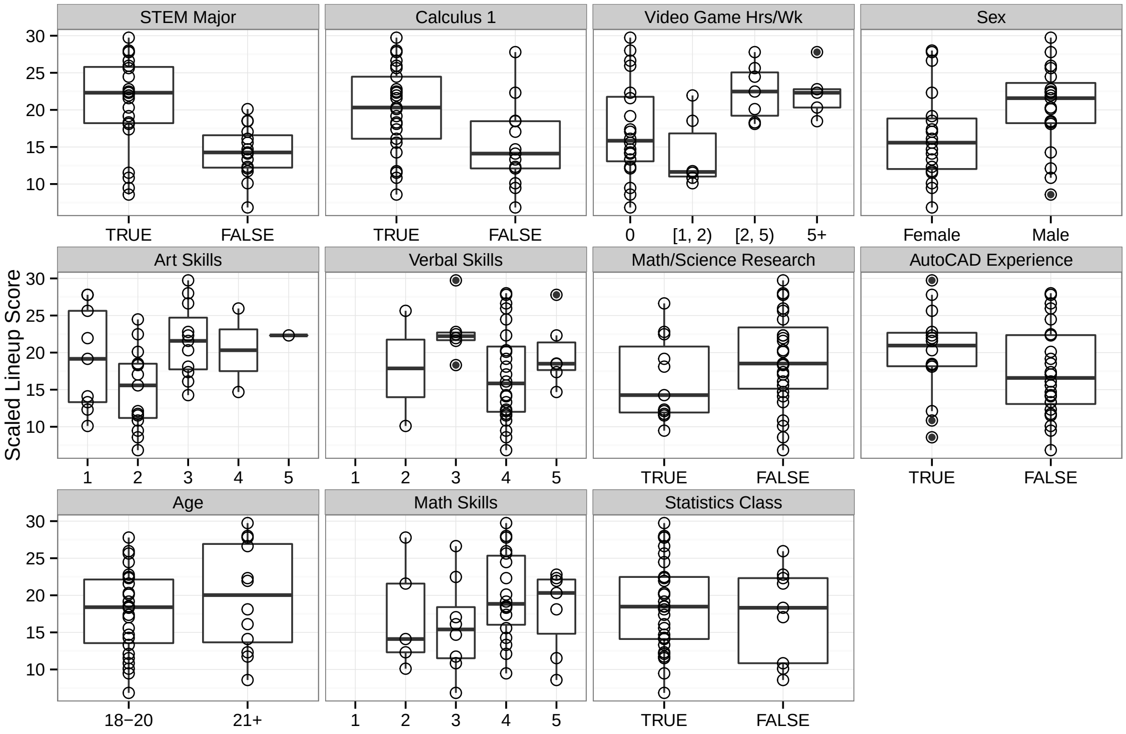


Fig. 5. Demographic characteristics of participants compared with lineup score. Categories are ordered by effect size; majoring in a STEM field, calculus completion, hours spent playing video games per week, and sex are all associated with a significant difference in lineup score.

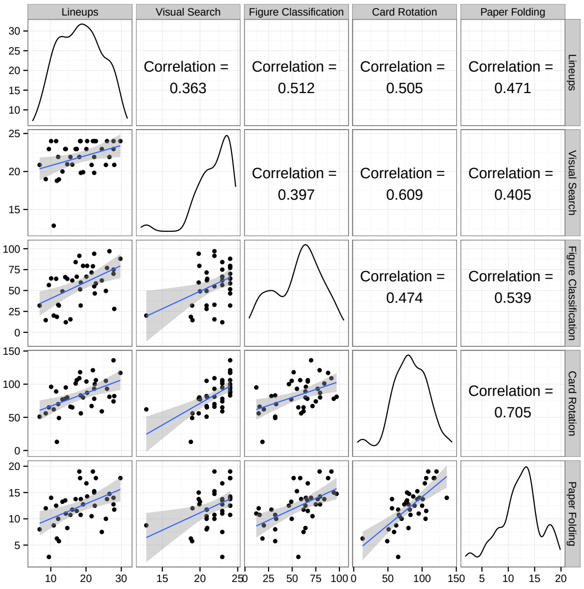


Fig. 6. Pairwise scatterplots of test scores. Lineup scores are most highly correlated with figure classification scores, and are also highly correlated with card rotation scores. Paper folding and card rotation scores are also highly correlated.

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, while PC3 differentiates these two tests from the paper folding test. More detailed results from the 4-test analysis are provided in Appendix C.

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

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

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 C.2.

Figure classification is strongly related to lineups (PC2, PC3), and as in the four-component PCA, figure classification is strongly represented in both principal components. Performance on the visual search task is also related to lineup performance (PC3), and is more strongly represented in PC3 in the five-component PCA than in the four-componet PCA. 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.

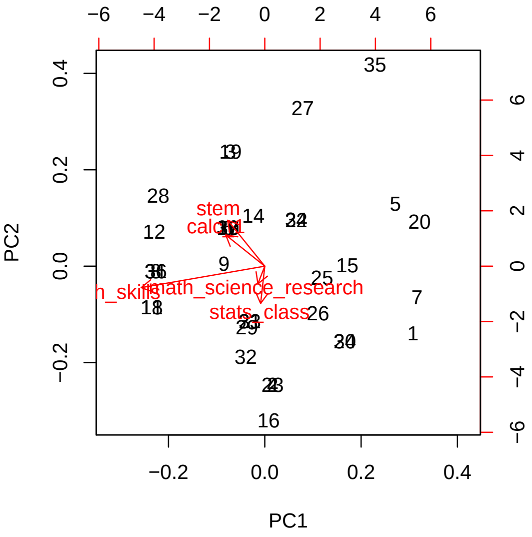
While lineups do span a separate dimension, but the PCA also suggests that they are most closely related to the figure classification task, and least related to the visual searching task.

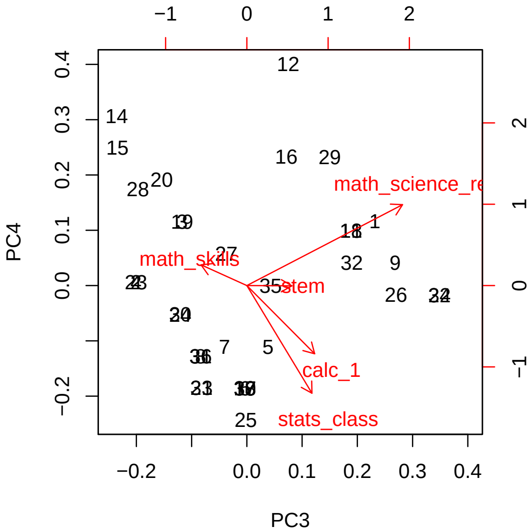
The additional dimension, PC4 in the five-component PCA, separates lineups from the figure classification task and incorporates the relatively weak correlation between performance on the card rotation task and lineup performance; this dimension accounts for 9.9% of the variance. PC5 is almost identical to PC4 in the four-component PCA; as before, it does not account for a significant portion of the variance in the data.

This emphasizes the underpinnings of lineups: they are a visual medium, but they 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

```
pca2 <- prcomp(ans.summary[, c("card_rot",
    "fig_class", "folding", "vis_search")],
    scale = T, retx = T)
ans.model <- data.frame(ans.summary,
    data.frame(pca2$x))
# exclude all variables that
# are not supposed to go into
# the linear model those are
# the unrotated tests; instead
# we have the PCs
excl <- c("id", "card_rot", "fig_class",
    "folding", "vis_search", "vidgame_hrs",
    "major1", "major2", "minor1",
    "minor2", "learning_disability",
    "colorblind", "epilepsy", "normal_vision")

ans.model$math_science_research <- ans.model$math_sci
    "y"
ans.model$stats_class <- ans.model$stats_class ==
    "y"
ans.model$calc_1 <- ans.model$calc_1 ==
    "y"
pcmath <- prcomp(ans.model[, c("math_science_research
    "stats_class", "calc_1", "math_skills",
    "stem")])
biplot(pcmath)
biplot(pcmath, 3:4)
# variables split into
# stats_class/math_science_research
# and calc_1/math_skills leave
# out math_skills in favor of
# calc_1 - for interpretability


xtabs(~stem + sex, data = ans.model)
# sad day!

xtabs(~stem + calc_1, data = ans.model)
# stem major is more associated
# with performance on lineups
# than calc 1...

xtabs(~stats_class + math_science_research,
    data = ans.model)
# only one kid is doing
# research without a stats
# class

# recode to sum of them
ans.model$statsP <- with(ans.model,
    stats_class + math_science_research)

excl <- c(excl, c("math_skills",
    "math_science_research", "stats_class"))
incl <- setdiff(names(ans.model),
    excl)

m0 <- lm(lineup ˜ ., data = ans.model[,
    incl])

# use AIC in backward selection
library(MASS)
m1 <- stepAIC(m0, direction = "both")
```





```
anova(m1)

| Analysis of Variance Table
|
| Response: lineup
|           Df Sum Sq Mean Sq
| age        1  34.43  34.434
| sex        1 159.10 159.096
| AutoCAD    1  23.83  23.830
| stem       1 284.48 284.478
| PC1        1 220.82 220.822
| Residuals 32 679.97  21.249
|           F value    Pr(>F)
| age        1.6205 0.2121921
| sex        7.4872 0.0100544
| AutoCAD    1.1215 0.2975289
| stem      13.3877 0.0009033
| PC1       10.3921 0.0029101
| Residuals
|
| age
| sex          *
| AutoCAD
| stem         ***
| PC1          **
| Residuals
| ---
| Signif. codes:
|   0 '***' 0.001 '**' 0.01
|   '*' 0.05 '.' 0.1 ' ' 1

summary(m1)

|
| Call:
```

```
| lm(formula = lineup ~ age + sex + AutoCAD + stem + PC1, data = ans.model[,
|     incl])
|
| Residuals:
|     Min      1Q  Median
| -9.2213 -2.9291  0.1439
|      3Q     Max
|  3.4071  8.0396
|
| Coefficients:
|             Estimate
| (Intercept)  14.3967
| age21+        2.4287
| sexm          3.6797
| AutoCADy     -3.8469
| stemTRUE      4.6150
| PC1           1.6588
|             Std. Error
| (Intercept)     1.4514
| age21+          1.6770
| sexm            2.4477
| AutoCADy        2.5027
| stemTRUE        1.7545
| PC1             0.5146
|             t value Pr(>|t|)
| (Intercept)   9.919 2.77e-11
| age21+        1.448  0.15727
| sexm          1.503  0.14256
| AutoCADy     -1.537  0.13411
| stemTRUE      2.630  0.01301
| PC1           3.224  0.00291
|
| (Intercept) ***
| age21+
| sexm
| AutoCADy
| stemTRUE    *
| PC1         **
| ---
| Signif. codes:
|   0 '***' 0.001 '**' 0.01
|   '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 4.61 on 32 degrees of freedom
| Multiple R-squared:  0.5152,Adjusted R-squared:  0.4395
| F-statistic: 6.802 on 5 and 32 DF,  p-value: 0.0002012
```

3.5 Lineup Types

results need to go here

4 Discussion and Conclusions

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

References

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

[2] 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.

[3] 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.

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

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

[6] 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.

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

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

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

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

[23] 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
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:

$$\mu_{\text{all}} = (N_F \mu_F + N_M \mu_M)/(N_F + N_M) \tag{2}$$

$$\sigma_{\text{all}} = \sqrt{\left(N_F \sigma_F^2 + N_M \sigma_M^2\right)/(N_F + N_M)}, \tag{3}$$

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

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

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 B
Lineup Performance and Demographic Characteristics

Table 4 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 much larger study would be required.

TABLE 4
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 C
Principal Component Analysis of Visuo-Spatial Tests

C.1 PCA of the Four Cognitive Tests

Table 5 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 7 and table 6 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

TABLE 6
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

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.

C.2 PCA of all Five Tests (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 8.

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.

TABLE 5
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

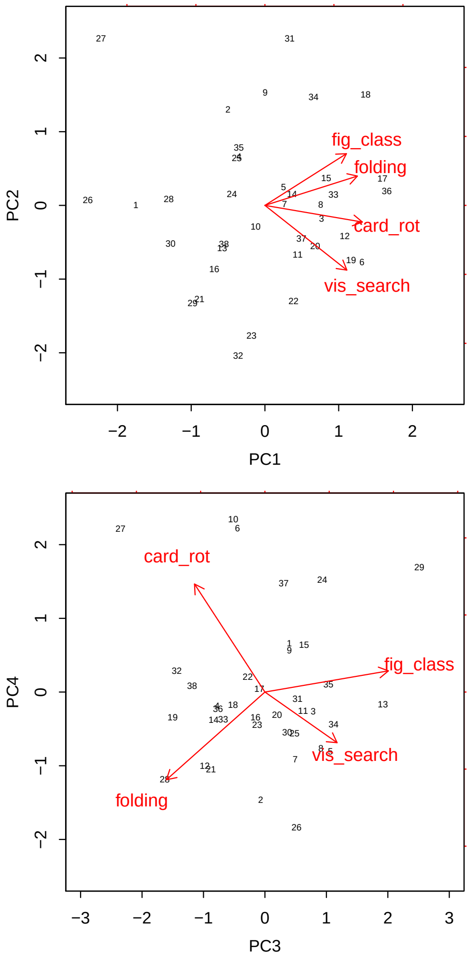


Fig. 7. Plots of principal components 1-4 with observations. PCA was performed on the four cognitive tests used to understand the cognitive demands of the lineup protocol.

Appendix D
Lineup Types

The two lineup sections which contain boxplots are shown as separate density curves, as they appear to be testing different things (outliers vs. difference in medians).

Conclusion: stacked bar plots are awful - the same as guessing.

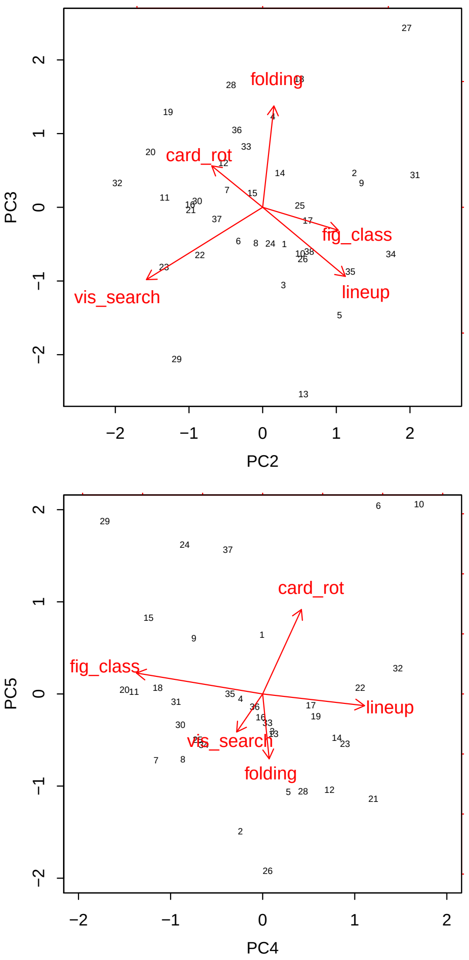


Fig. 8. Plots of principal components 2-5 with observations.

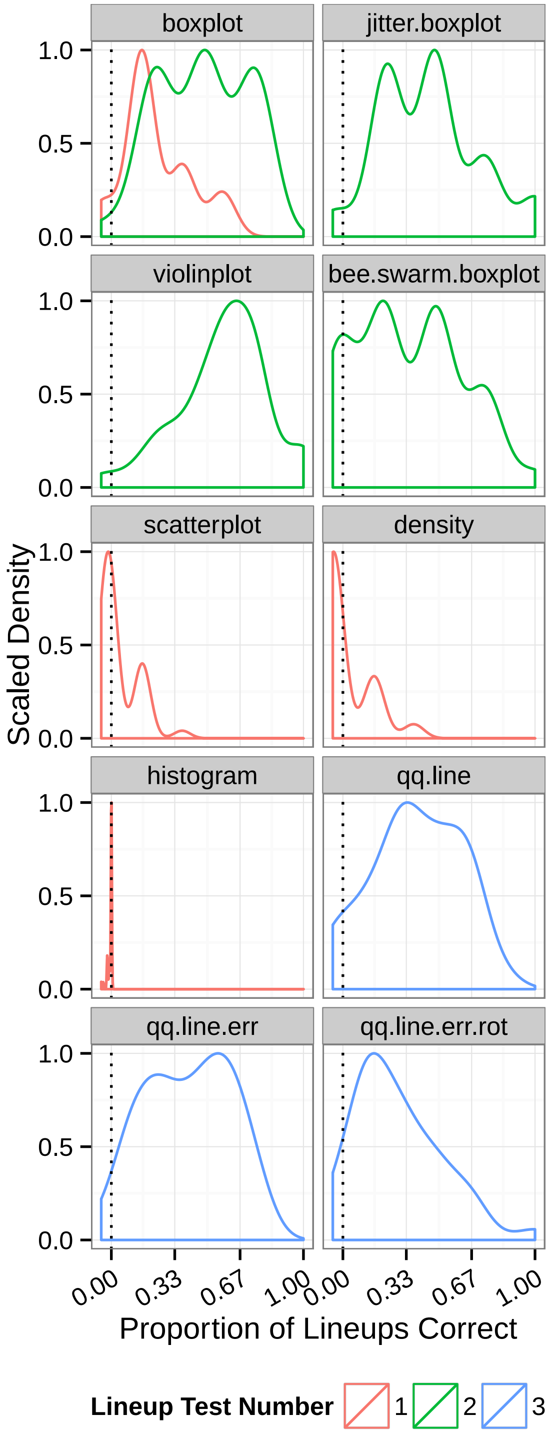


Fig. 9. Plot of scaled scores for different types of lineups

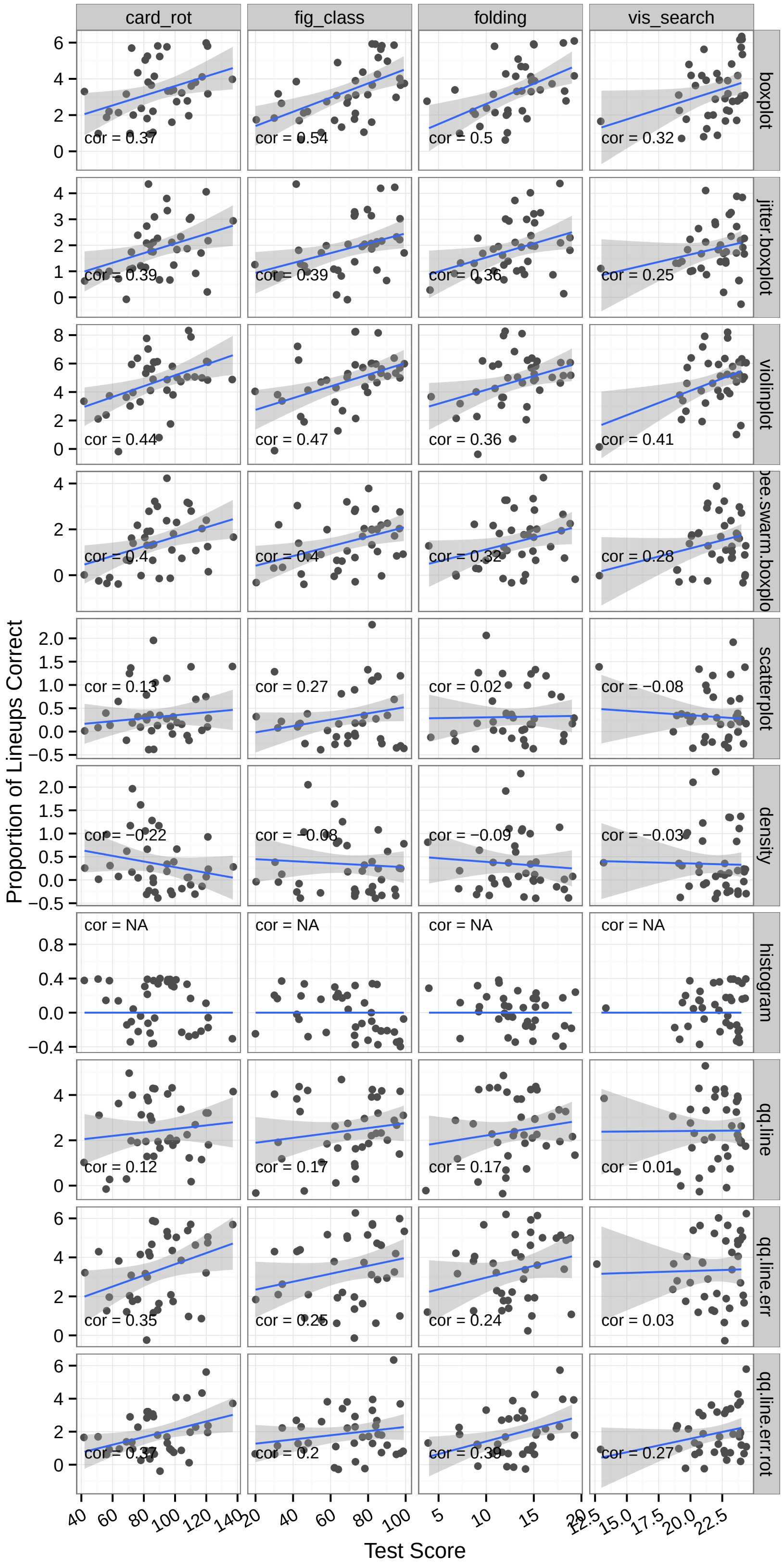


Fig. 10. Plot of scaled lineup scores by aptitude test scores