Spatial Reasoning and Statistical Graphics

Susan VanderPlas, Heike Hofmann

October 17, 2014

1 Introduction

Relevant literature:

- Shah and Carpenter (1995) showed that spatial ability was not correlated with accuracy on a simple two-dimensional line graph description task, but that mathematical ability was correlated with accuracy.
- Just and Carpenter (1985) showed that high-spatial-ability viewers used different rotation strategies than low-spatial-ability viewers when asked to whether three-dimensional alphabet cubes were the same.
- Hofmann et al. (2012) for lineup stimuli and general lineup performance

Lineups depend on the ability to search for a signal amid distractors (Visual Search Task) and the ability to infer patterns from stimuli (Pattern Recognition task). Some lineups (polar coords) also depend on the ability to mentally rotate stimuli (spatial rotation task) and mentally manipulate graphs (paper folding task). By breaking the lineup task down into component parts, we can correlate lineup performance with similar cognitive factor tests to determine where additional variation in skill level factors into performance differences. In addition, we can correlate previous experiences (science-based major, research experience, Auto-CAD skills) with performance to explore the effect that participant experience has on lineup performance.

2 Methods

Participants will complete the following tasks (sample pictures included, full stimuli set will be added to the appendix once testing is complete). Tasks are designed so that participants are under time pressure; they are not expected to complete all of the problems in each section. This provides more discrimination between high scorers and prevents score compression at the top of the range.

- Visual Search Task: designed to test participants' ability to find a target stimulus in a field of distractors. An example is shown in figure 1.
- Paper Folding Task: tests participants' ability to visualize and mentally manipulate figures in three dimensions. Associated with the ability to extrapolate symmetry and reflection over multiple steps. An example is shown in figure 2.
- Card Rotation Task: tests participant's ability to rotate objects in two dimensions to distinguish between left-hand and right-hand versions of the same figure. Tests spatial reasoning ability and mental rotation skills. An example is shown in figure 3.
- Figure Classification Task: tests participant's ability to extrapolate rules from provided figures. This task is associated with visual reasoning capabilities and we expect that it should correlate with the ability to pick out a signal plot from a lineup. An example is shown in figure 4.

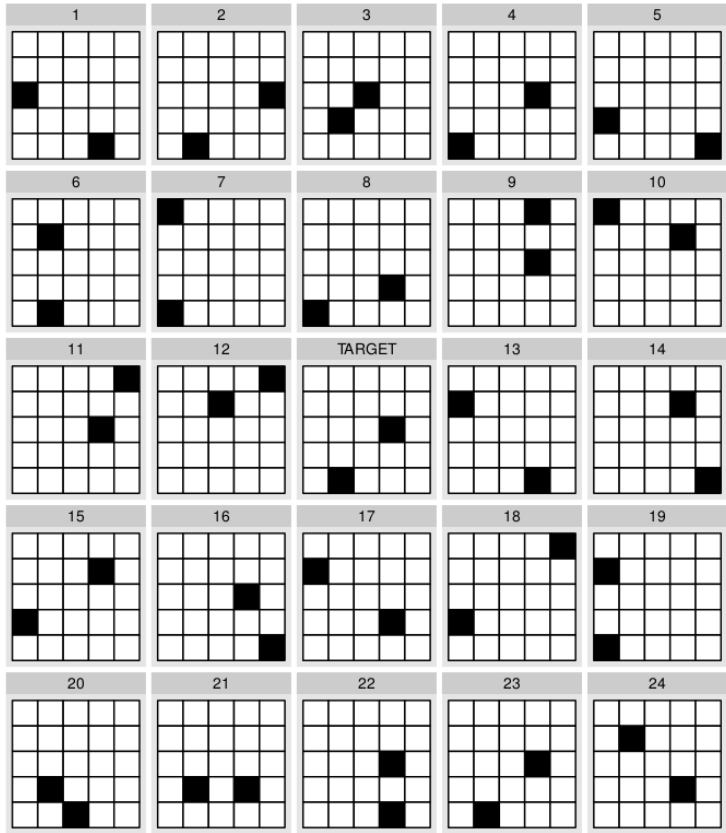

Figure 1: Visual Search Task. Participants are instructed to find the plot numbered 1-24 which matches the plot labeled "Target". Participants will complete up to 25 of these tasks in 5 minutes.

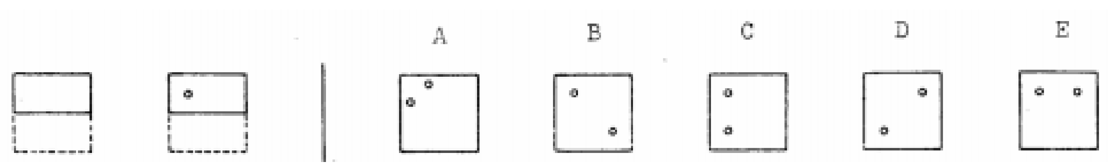

Figure 2: Paper Folding Task. Participants are instructed to pick the figure matching the sequence of steps shown in the left-hand figure. Participants will complete up to 20 of these tasks in 6 minutes.

Figure 3: Card Rotation Task. Participants mark each figure on the right hand side as either the same or different than the figure on the left hand side of the dividing line. Participants will complete up to 20 of these tasks (each consisting of 8 figures) in 6 minutes.

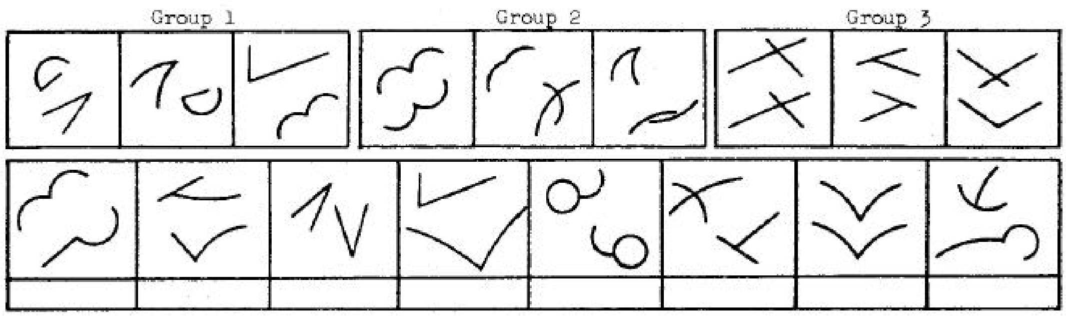

Figure 4: Figure Classification Task. Participants classify each figure in the second row as belonging to group 1, 2, or 3 (if applicable). Participants will complete up to 14 of these tasks (each consisting of 8 figures to classify) in 8 minutes.

Between cognitive tasks, participants will also complete three blocks of 20 lineups each. These lineups have been previously tested (Hofmann et al., 2012) . Participants have 5 minutes to complete each block of 20 lineups. Figure 5 shows a sample lineup of box plots.

In addition to these tests, participants will 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.

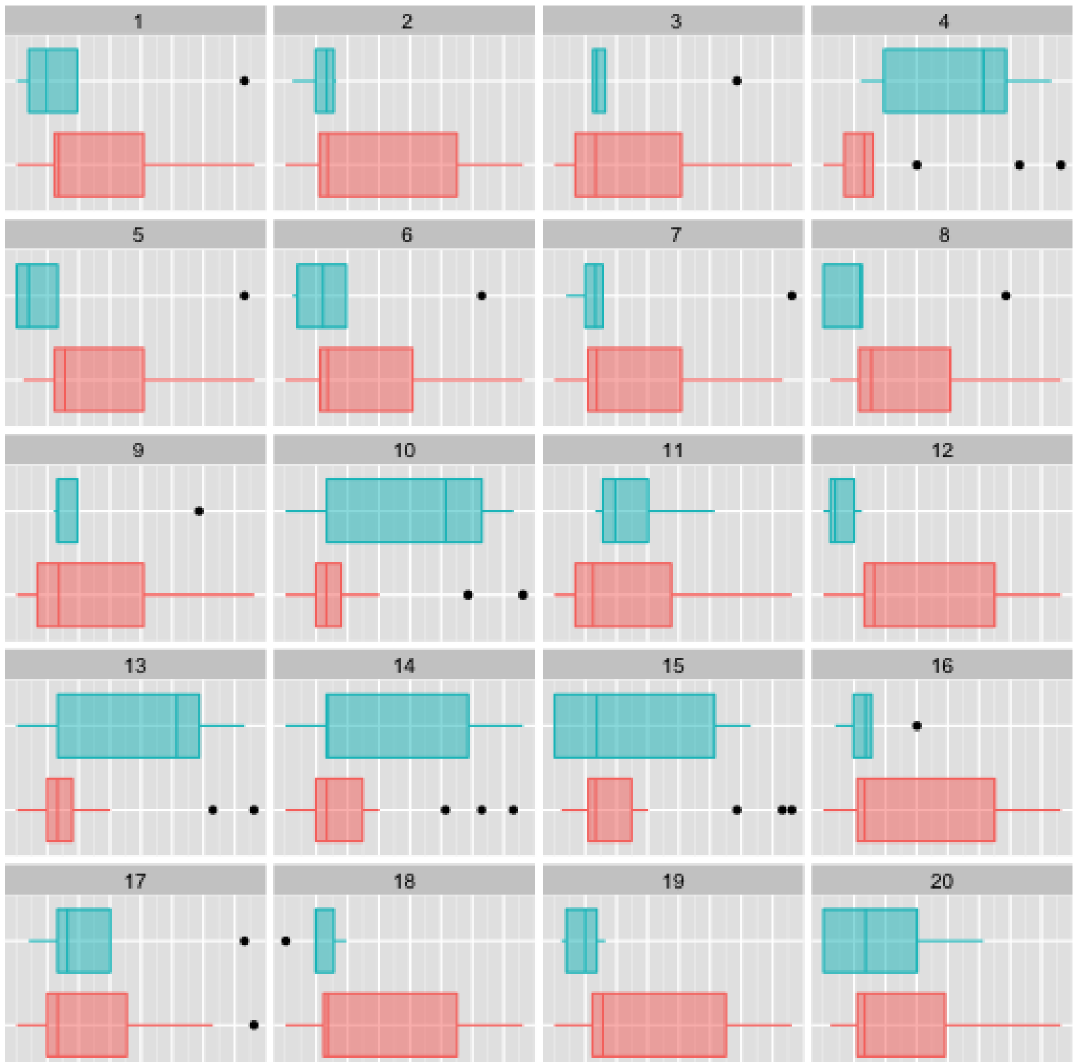

Figure 5: A sample lineup. Participants are instructed to choose the plot which appears most different from the others. In this lineup, plot 13 is the target.

3 Results

Results are based on an evaluation of 36 undergraduate students at Iowa State University. Scoring of all test results was done such that random guessing leads to 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

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

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 (Diamond and Evans, 1973; Ekstrom et al., 1976) from which parts of this test are drawn.

The advantage of using tests from the Kit of Factor Referenced Cognitive tests (Ekstrom et al., 1976) is that the tests are extremely well studied (**?**)

the bib file is not updated on the repo

and comparison data are available from the validation of these factors. 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).

	Card Rotation	Paper Folding	Figure Classification
ISU Students	82.8 (24.5)	12.4 (3.8)	57.5 (23.9)[1]
Scaled Scores	88.0 (34.8)	13.8 (4.5)	58.7 (14.4)[2]
Unscaled Scores	44.0 (24.6)[3]	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 11-12th grade students (288-300 males, 317-329 females)

Table 1: Comparison of scores from Iowa State students and scores reported in Ekstrom et al. (1976). 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.

Table 1 shows mean scores and standard deviation for ISU students and other populations. Once values have been adjusted to accommodate different test procedures (some data is reported for a single part of a two-part test, for instance), 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.[4]

[1] ISU students took only Part I due to time constraints.

[2] Averages calculated assuming 294 males and 323 females.

[3] Data from Part I only.

[4] If comparison data was available for 9th and 12th grade students, we have compared Iowa State students' scores with the 12th grade students, to closer in age to college students. When data was available from college students and Army enlistees, we have compared ISU students to other college students, as college students are more likely to have similar gender distribution to ISU students.

Could you include some more details on how to get from the Unscaled scores to the scaled versions? I'm assuming that for the Card Rotation, the two parts have equal size, explaining the doubled score, but why not double the standard deviation? $\text{Var}(aX) = a^2\text{Var}(X)$ I just don't quite follow the code for the combined scores ... it would be worthwhile to include a bit more detail and maybe put that in an appendix to the paper. If you want to, you could instead extend the footnotes a bit for now. ... so just let me try this: Let's assume that X_M and X_F are the scores based on 294 men, and 323 women (are those midscore s of the intervals reported?). with corresponding s_M and s_F for the reported standard deviation of the full test. Then half of the test would have half of the scores with half of the standard deviation. we have a breakdown of 17 girls and 19 boys. $X = (X_M/2 \cdot 294 + X_F/2 \cdot 323)/(294 + 323) = 58.7$. For the Variance of X we then have: $\text{Var}(X) = \text{Var}(X_M)/2^2 \cdot (294/617)^2 + \text{Var}(X_F)/2^2 \cdot (323/617)^2 = 10.2^2$

As we have established that the results obtained for the ETS tests are similar to other studies, we will now compare the results to the lineups also tested in this study. To facilitate this goal, for the remainder of this analysis, we will scale the test results so that the ranges and units of test scores are comparable. Let $X_{n,k}$ be a participant's score on a test consisting of n questions with k answers each, out of which only one is correct. This leads to a theoretical range of $X_{n,k}$ of $[-n/(k-1), n]$ and, under an additional assumption of random guessing, a variance of

$$\begin{aligned}
\text{Var}(X_{n,k}) &= n^2\text{Var}(X_{1,k}) = \\
&= n^2(\underbrace{1/k \cdot 1^2}_{\text{correct answer}} + \underbrace{(-1/(k-1))^2 \cdot (k-1)/k}_{\text{wrong answer}}) = \\
&= n^2/(k-1).
\end{aligned}$$

The above consideration only assumes independence between questions, which is reasonable. While we only consider a test consisting of questions with the same number of choices, k, an extension to varied number of answers is trivial and has been done in the adjustment for the figure classification score. In order to ensure all tests have similar variance, we scaled test scores by the standard deviation under random guessing. This approach also allows us to compare the test scores using similar orders of magnitude.

```
library(plyr)
ldply(ans.summary[, c("lineup", "card_rot", "fig_class", "folding", "vis_search")],
    c(mean = mean, sd = sd))

|          .id   mean     sd
| 1     lineup 1.3179 0.4439
| 2   card_rot 0.5174 0.1533
| 3  fig_class 0.6138 0.2547
| 4    folding 1.2437 0.3782
| 5 vis_search 4.1986 0.4421
```

```
cor(ans.summary[, c("lineup", "card_rot", "fig_class", "folding", "vis_search")])

|            lineup card_rot fig_class folding vis_search
| lineup     1.0000   0.5355    0.6003  0.4955     0.4000
| card_rot   0.5355   1.0000    0.4787  0.7206     0.6074
| fig_class  0.6003   0.4787    1.0000  0.5489     0.3877
| folding    0.4955   0.7206    0.5489  1.0000     0.4087
| vis_search 0.4000   0.6074    0.3877  0.4087     1.0000

pca <- prcomp(ans.summary[, c("lineup", "card_rot", "fig_class", "folding",
```

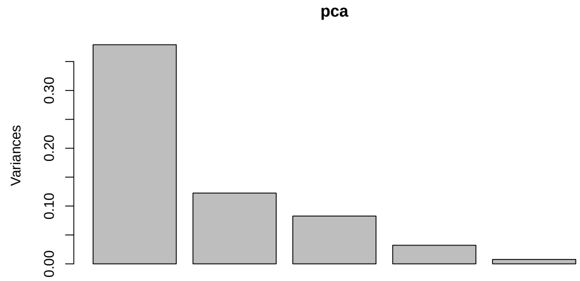

Figure 6: Scree plot of principle component analysis of performance on the different test batteries.

```
    "vis_search")], scale = F)
summary(pca)

| Importance of components:
|                          PC1   PC2   PC3    PC4    PC5
| Standard deviation     0.616 0.350 0.287 0.1793 0.0874
| Proportion of Variance 0.608 0.196 0.132 0.0515 0.0122
| Cumulative Proportion  0.608 0.804 0.936 0.9878 1.0000

screeplot(pca)
pca$rotation

|              PC1      PC2      PC3     PC4      PC5
| lineup     0.5985  0.54024 -0.53381  0.2487 -0.05576
| card_rot   0.1950 -0.03803  0.13031  0.1095  0.96518
| fig_class  0.2955  0.16141  0.05234 -0.9390  0.04612
| folding    0.4692  0.16319  0.81231  0.2101 -0.22185
| vis_search 0.5443 -0.80871 -0.18836  0.0160 -0.11820
```

```
# Just using the 4 cognitive tests and ignoring the lineups...
pca.cog.tests <- prcomp(ans.summary[, c("card_rot", "fig_class", "folding",
    "vis_search")], retx = T)
summary(pca.cog.tests)

| Importance of components:
|                          PC1   PC2    PC3    PC4
| Standard deviation     0.528 0.319 0.1947 0.0892
| Proportion of Variance 0.654 0.239 0.0888 0.0186
| Cumulative Proportion  0.654 0.893 0.9814 1.0000

pca.cog.tests$rotation

|              PC1      PC2      PC3       PC4
| card_rot   0.2322 -0.08273  0.09724  0.964256
| fig_class  0.3081 -0.25301 -0.91707 -0.003424
| folding    0.5599 -0.69764  0.38146 -0.233134
| vis_search 0.7333  0.66516  0.06335 -0.125890
```

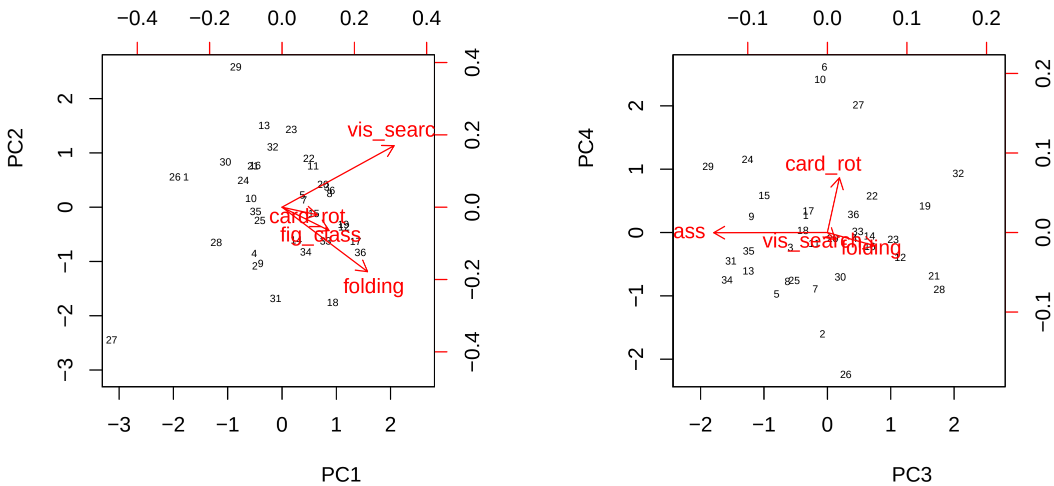

Figure 7: Plots of the principal components with observations. Visual Search and Paper Folding strongly contribute to both PC1 and PC2, while Figure Classification and Paper Folding strongly contributes to PC3 and Card Rotation strongly contributes to PC4

```
biplot(pca.cog.tests, choices = 1:2, pc.biplot = T, cex = c(0.5, 1), adj = 0.75)
biplot(pca.cog.tests, choices = 3:4, pc.biplot = T, cex = c(0.5, 1), adj = 0.75)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3179	0.0587	22.46	0.0000
PC1	0.4786	0.1127	4.25	0.0002
PC2	-0.2549	0.1863	-1.37	0.1811
PC3	-0.5808	0.3057	-1.90	0.0668
PC4	0.7073	0.6674	1.06	0.2974

Table 2: Lineup score, as predicted by principal components. Only the first principal component is a significant predictor of lineup score.

In figure 8, we see that participant performance on lineups is positively correlated with performance on card rotation, figure classification, and paper folding tasks. This suggests that skills associated with visual reasoning ability are related to lineup performance. As participants must use the same skills in lineups (mental rotation, classification and determining categorization schemes, and multi-step spatial reasoning) as in the factor-referenced tests, this is not particularly surprising. In addition,there seems to be some positive relationship between a participant's score on the visual search task and their score on lineups: the visual search task represents a baseline of a participant's ability to find a matching pattern, while lineups require that task as well as the ability to determine what the pattern is for a particular graph. Even excluding the one low visual search score that is a high-leverage point, there seems to be a positive relationship between a participant's score on lineups and their score for visual search.

Figure 9 shows participants' responses to the questionnaire given at the beginning of the study; these demographic questions allow us to compare the participants in our study to the undergraduate population of Iowa State as well as to explore relationships between demographic characteristics (major, research experience, etc.) and score on various sections of this test.

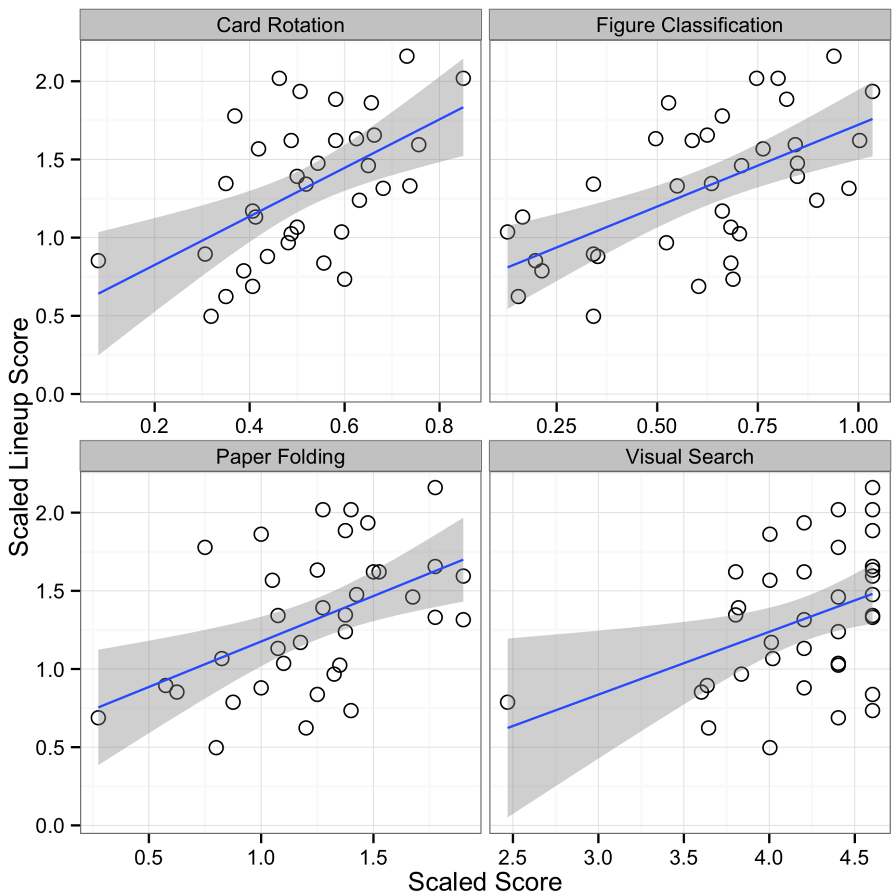

Figure 8: Scatterplots of all test scores compared to participants' scores in the lineup tests. There is a relatively strong positive correlation between lineup score and scores on visuospatial reasoning tests.

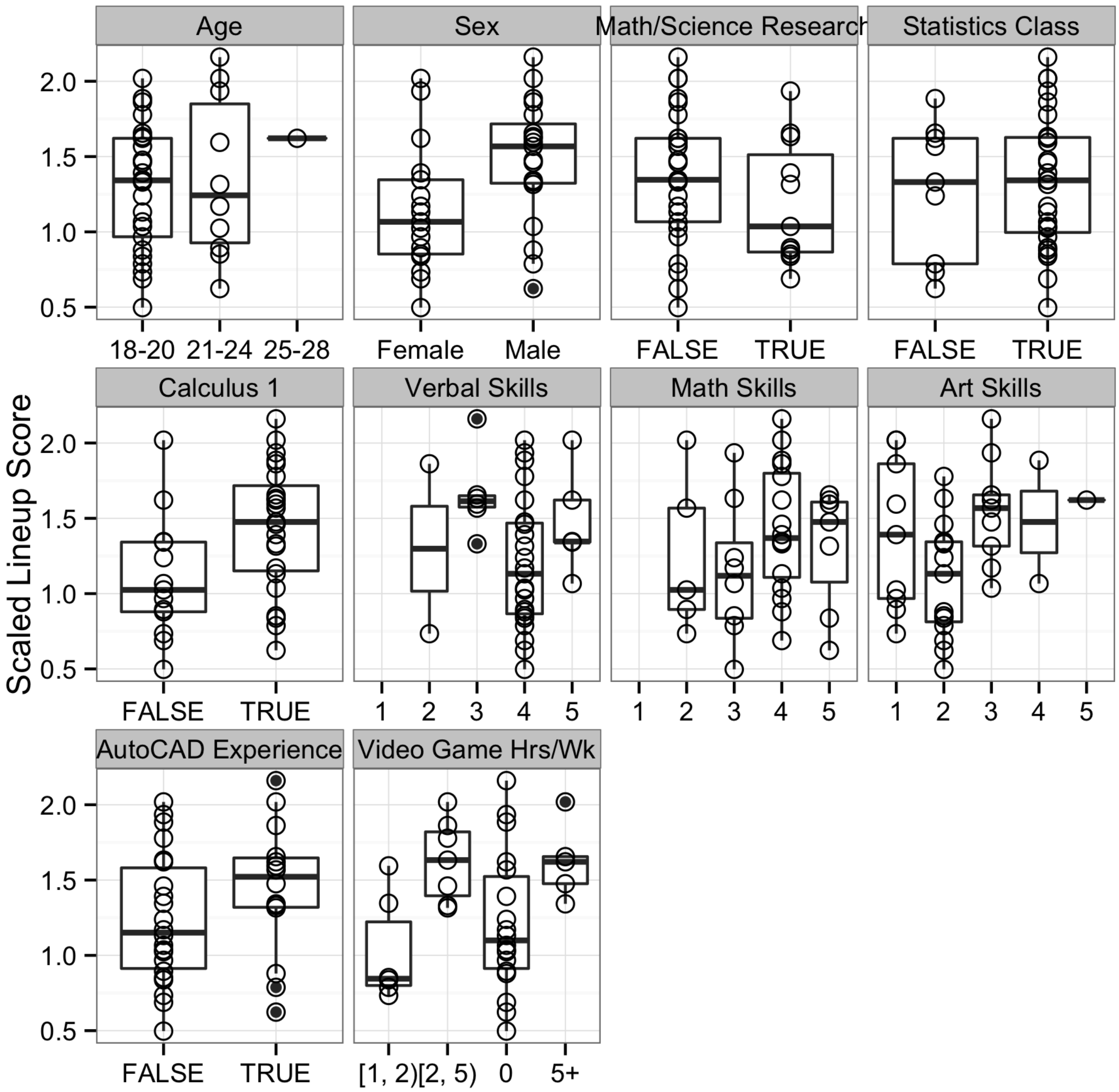

Figure 9: Sample demographic characteristics compared with lineup score. Sex (male), Calculus completion, and hours spent playing video games per week are all associated with a difference in lineup score.

There is no significant difference in lineup performance for participants of different age, self-assessed skill rating, previous participation in math or science research, completion of a statistics class, or experience with AutoCAD. 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 result is consistent with (Shah and Carpenter, 1995), who found an association between mathematical ability and performance on simple graph description tasks.

There is also a significant association 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.

Variable	DF	Sum.of.Squares	Mean.Squared	F.value	p.value
Video Game Hrs/Wk	3	1.862	0.621	3.94	0.017
Sex	1	0.995	0.995	5.73	0.022
Calculus 1	1	0.953	0.953	5.45	0.026
Verbal Skills	3	1.161	0.387	2.16	0.112
Art Skills	4	1.412	0.353	1.99	0.120
AutoCAD Experience	1	0.372	0.372	1.94	0.173
Math/Science Research	1	0.248	0.248	1.27	0.268
Math Skills	3	0.464	0.155	0.77	0.520
Statistics Class	1	0.026	0.026	0.13	0.723
Age	2	0.129	0.065	0.32	0.732

Table 3: Model results of each demographic variable compared with lineup score. Multiple testing issues aside, it appears that very few demographic variables (if any) are significantly associated with score on lineups among Iowa State undergraduate students.

All results and data shown here are done in accordance with IRB # 13-581.

References

Diamond, J. and Evans, W. (1973). The correction for guessing. Review of Educational Research, pages 181–191.

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

Hofmann, H., Follett, L., Majumder, M., and Cook, D. (2012). Graphical tests for power comparison of competing designs. Visualization and Computer Graphics, IEEE Transactions on, 18(12):2441–2448.

Just, M. A. and Carpenter, P. A. (1985). Cognitive coordinate systems: accounts of mental rotation and individual differences in spatial ability. Psychological review, 92(2):137.

Shah, P. and Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. Journal of Experimental Psychology: General, 124(1):43.

Appendix

T-tests of results for Hillary and Stephanie:

Variable	Mean.Hillary	Mean.Stephanie	t	df	p-value	Diff.LB	Diff.UB
Card Rotation	0.55	0.49	1.22	33.29	0.23	-0.04	0.16
Paper Folding	1.32	1.17	1.19	32.64	0.24	-0.11	0.40
Figure Classification	0.70	0.53	2.00	28.67	0.05	-0.00	0.33
Lineups	1.47	1.17	2.08	33.94	0.04	0.01	0.58