Spatial Reasoning and Statistical Graphics

Susan VanderPlas, Heike Hofmann

October 20, 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.
- Voyer et al. (1995) completed a meta-analysis of spatial tasks and gender differences. They concluded that "rotation" tasks, such as the card rotation test, have robust sex differences across cohort, and while the sex differences seem to be declining for some tests, such as the card rotation test, the differences still exist for the time being. Test administration differences may account for the change. They also concluded that the paper-folding test showed no such sex difference, though this may be dependent on test scoring (a lower guessing penalty decreases the score difference).

  In the current analysis, we have scored the paper folding test out of 20. Alternate scoring procedures still need to be investigated... they cite a poster, but it's not available anywhere online as far as I can tell.
- Lowrie and Diezmann (2007) investigated students (9-10 yrs) ability to read mathematical graphs and their visual-spatial ability. They partition graphs into 6 "graphical languages" (some of these languages aren't statistical graphs - networks, venn diagrams, maps), and use multivariate analysis to associate spatial ability with some of the types of graphs studied.

  I can't understand their model results very well, so I'm not entirely sure what's going on here. They also cite a bunch of studies (including Voyer et al. (1995)) that I haven't had time to completely explore.
- Vekiri (2002) is lit review discussing the different theories about the utility of graphics in learning (dual coding theory, visual argument hypothesis, and conjoint retention hypothesis). The article contains a nice definition of graphics (though it might be too domain specific for statistical graphics, since it claims graphics have only one meaning and aren't prone to interpretation.) Depending on the context, "maps", "charts", and "graphs" here could fall into a loose classification as "statistical graphics", but in general "graphs" is probably accurate for most of the lineups we're testing: "[Referents are] quantitative data ... that enable viewers to compare and observe relations among variables".
  - **Dual Coding Theory**: people encode both words and pictures separately, so both are encoded for a graph.
    Supports the use of graphics, and there is support for dual-coding in working memory research

(pictures and words are processed separately but in parallel). In addition, since graphics depend on other abilities (spatial abilities), graphics that aren't properly designed create higher cognitive load and lower performance on graphical tasks. The review also specifically discusses visuospatial ability, citing a speculation from Mayer and Sims (1994) that students with low spatial ability perform worse with diagrams and graphs because more cognitive resources are required.

Lots of studies by Sweller cited throught the cognitive load discussion - I haven't gotten to that literature yet.
That said, that research has some potential to justify the use of lineups from a psychological perspective - if you can claim that types of graphs that have poor performance do so because they create higher cognitive load (cue psych research into why that would be true, cleveland & mcgill, etc.), you can justify the suckiness of those awful boxplot-jittered-things.

- **Visual Argument Hypothesis**: Graphics require less processing because their visuospatial properties encode some of the information through the arrangement of elements as well as the elements themselves. That is, graphics are a form of "external cognition" (Scaife and Rogers, 1996) that guide, constrain, and facilitate cognitive behavior(Zhang, 1997). This constraint reduces memory load and makes more cognitive resources available for other tasks; displays that clearly present information without deep processing requirements are more effective(Zhang and Norman, 1994). Graphical displays have higher search and computational efficiency than text (Larkin and Simon, 1987), and when graphics are designed according to gestalt principles of organization, where grouping lines up with visual chunks, gestalt heuristics can be used to more quickly encode information (Shah et al., 1999).

  I am working on making my way through all of this literature - right now, I'm re-citing stuff; I will print the papers out and add to this ASAP once I get toner into the printer.

- **Conjoint Retention Hypothesis** - based on dual coding theory, but claims that there are separate but interconnected memory codes for verbal and visual information. Assumes maps/graphs are encoded intact, with visuospatial properties preserved. This is pretty unlikely, and the literature is not so relevant to our cause. Spatial information is encoded, but words can actually interfere with this encoding; it is not "intact"

- 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

Besides pointing out what we asked participants to do, we also need some why, and also why in particular these tests were selected. part of the validity consideration can go in here - and the classification from the naval test of the tests we picked into different classes (maybe already with a hint that for our data these weren't actually orthogonal classes, because we found a really high correlation between card rotation and paper folding.)
I don't know if the tests are supposed to be orthogonal. It would be hard to construct a set of tests that was orthogonal, because the abilities themselves aren't easily separable - math reasoning and spatial reasoning are related, for instance, and verbal abilities are correlated with creativity too.

Participants will complete the following tasks (sample pictures included, full stimuli set will be added to the appendix once we are sure there is no need for follow-up experiments). 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. The visual search task is similar in concept to lineups: it tests one's ability to find the target plot. Historically, it has been used as a measure of brain damage(Goldstein et al., 1973; DeMita et al., 1981; Moerland et al., 1986); however, similar tasks, have been used to measure cognitive performance in a variety of situations (under the influence of drugs, for example, in Anderson and Revelle (1983)). The similarity to lineup protocol as well as the simiplicity of the test and its' lack of color justify the slight deviation from forms of visual search tasks typically used in normal populations.

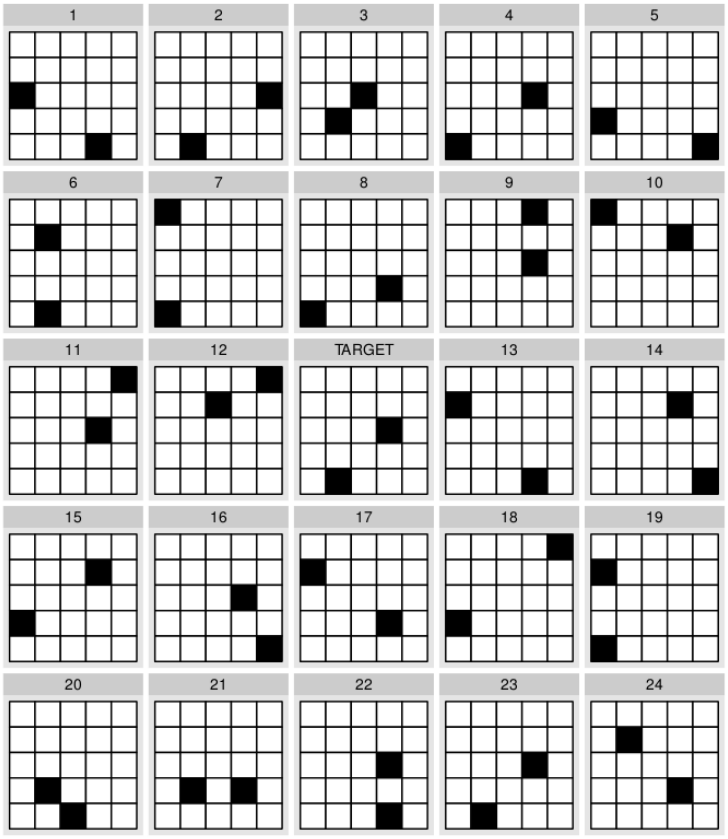

Figure 1: Visual Search Task. Participants are instructed to find the plot numbered 1-24 which matches the plot labeled "Target". Participants will complete up to 25 of these tasks in 5 minutes.

- 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

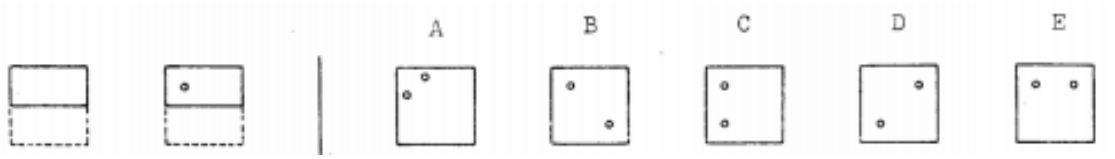

Figure 2: Paper Folding Task. Participants are instructed to pick the figure matching the sequence of steps shown in the left-hand figure. Participants will complete up to 20 of these tasks in 6 minutes.

mental rotation skills. An example is shown in figure 3.

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 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.

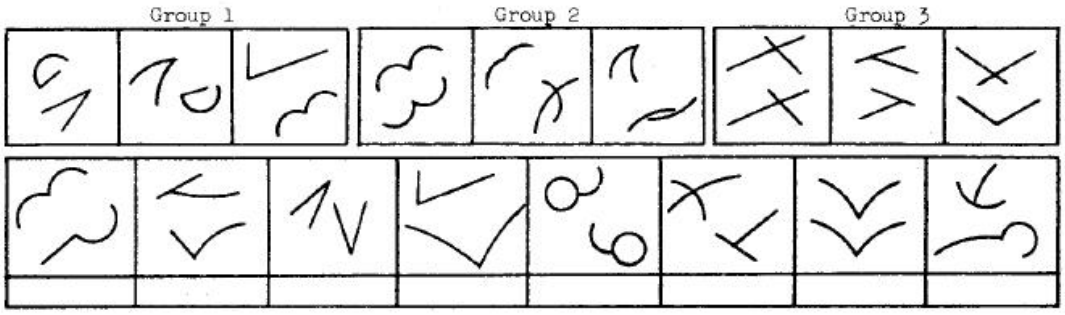

Figure 4: Figure Classification Task. Participants classify each figure in the second row as belonging to group 1, 2, or 3 (if applicable). Participants will complete up to 14 of these tasks (each consisting of 8 figures to classify) in 8 minutes.

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.

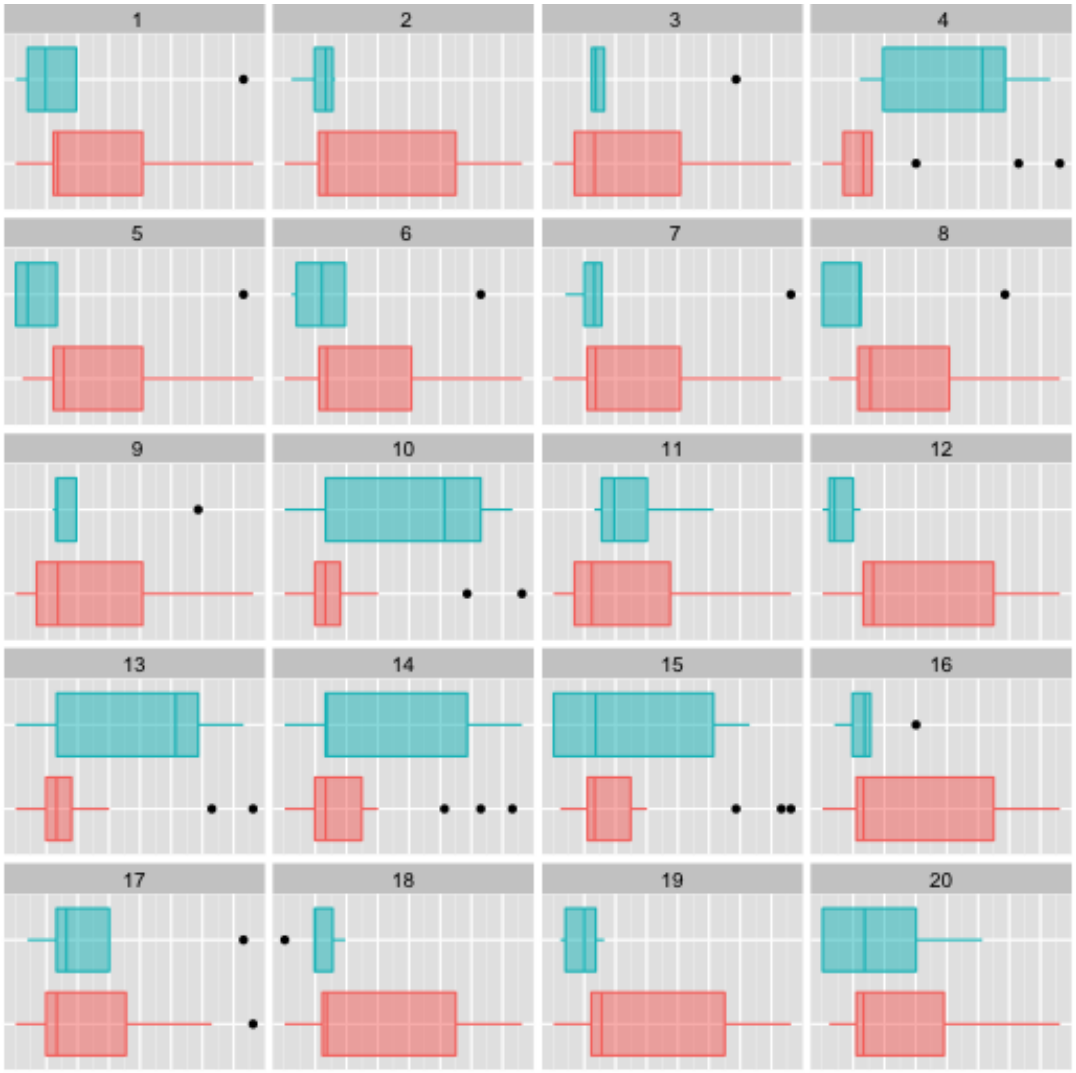

Figure 5: A sample lineup. Participants are instructed to choose the plot which appears most different from the others. In this lineup, plot 13 is the target.

3 Results

Results are based on an evaluation of 38 undergraduate students at Iowa State University.

Teensy overview of the demographics - this might end up in an appendix, but then we need to refer to from here. I'd like to know breakdown of men/women, classification, maybe an overview of different majors (not necessarily a breakdown, that might be leading to problems with identifiability)

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

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

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.

could we also refer to the fact that these tests have been long established (maybe with references to earlier tests or updates?) *Yes, definitely. I'll pull up some references for that too...*

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 (Voyer et al., 1995; Schaie et al., 1998; Hampson, 1990) and comparison data are available from the validation of these factors.

the well studied reference list needs to be a bit more extensive ... do you know this book by Michel Hersen?
I was working on it yesterday along with trying to validate the general idea that reading graphics requires certain spatial abilities. It's definitely a work in progress. I haven't seen that book before, but I'm not at all surprised it exists.

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	Visual Search
ISU Students	83.4 (24.1)	12.4 (3.7)	57 (23.8)[1]	21.9 (2.3)
Scaled Scores	88.0 (34.8)	13.8 (4.5)	58.7 (14.4)[2]	N/A
Unscaled Scores	44.0 (24.6)[3]	13.8 (4.5)	M: 120.0 (30.0), F: 114.9 (27.8)	N/A
Population	approx. 550 male naval recruits	46 college students (1963 version)	suburban 11th & 12th grade students (288-300 males, 317-329 females)	N/A

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. The Visual Search task (Goldstein et al., 1973; DeMita et al., 1981; Moerland et al., 1986) 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.

Table 1 shows mean scores and standard deviation for ISU students and other populations. Values have been adjusted to accommodate differences in test procedures: some data is reported for a single part of a

[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.

two-part test (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 populuation data was chosen in the following manner: if comparison data was available for 9th and 12th grade students, we have compared Iowa State students' scores with the 12th grade students, as they are 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? $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$ ok I got the first one - because it is $X_1 + X_2$ not $2X_1$. still working on understanding the second one, though ... 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: $\mathrm{Var}(X) = \mathrm{Var}(X_{M_1} + X_{M_2}) \cdot (294/617)^2 + \mathrm{Var}(X_{F_1} + X_{F_2}) \cdot (323/617)^2 = 14.4^2$

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

by a factor of $1/n\sqrt{k-1}$, where n is the number of questions in a test and k is the number of offered answers on each questions (out of which only one is correct). This scaling ensures that test scores across different types are comparable. All test scores are now normalized such that they have under the assumption of random guessing an expected value of zero and a standard deviation of 1. In particular, this leads to a theoretical range of a normalized test score $Z_{n,k}$ of $[-\sqrt{k-1}, \sqrt{k-1}]$.

Variance consideration of scaled test score: Let $X_{n,k}$ be a participant's (unscaled) score on a test consisting of n questions with k answers each, out of which only one is correct.

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

The scaled test score $Z_{n,k}$ is $1/n\sqrt{(k-1)}X_{n,k}$ and the statement for the variance of $Z_{n,k}$ follows directly.

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.

OK, so on to the next steps ... we would like to figure out, how the lineups play into the established test scores. I think we need to talk about how to best talk through the visual search score. We have a lot of information on the other three scores - is there an established test in the naval test that is 'like' the visual search? - i.e. could we use those scores to compare to?
so outline for the next few steps:

1. some general words on behavior of lineup scores by themselves - I think the whole discussion that is at the moment at the end of the paper should move up here. The levels of the video gaming are a bit messed up - they should be ordered according to number of games played rather than alphabetically :) - and I think that then the relationship actually is a monotonic one ... In any case, sort the figures in the table according to their p-values in the table. Gender shows up as significantly different in my table of p values, but in the text you say that it is not. The table with the t-tests should probably go into an appendix ....
2. figure 8 with either three or four scatterplots depending on how the visual search scores fit in
3. a paragraph on discussing the multicollinearity between the test scores based on correlations (either, again, including visual search or not) - this should lead to the conclusion that we cannot directly use a linear model of lineup in test scores.
4. PCA of test scores, with a bit of discussion
5. linear model of lineup scores in PCA with discussion and tying it back to raw scores (figure classification really is the best predictor for lineups, but it only explains some of the variance).

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

```
|             lineup card_rot fig_class folding vis_search
| lineup     1.00000  0.50497   0.51216 0.47092    0.36275
| card_rot   0.50497  1.00000   0.47354 0.70471    0.60915
| fig_class  0.51216  0.47354   1.00000 0.53911    0.39673
| folding    0.47092  0.70471   0.53911 1.00000    0.40480
| vis_search 0.36275  0.60915   0.39673 0.40480    1.00000
```

```
pca <- prcomp(ans.summary[, c("lineup", "card_rot", "fig_class", "folding",
    "vis_search")], scale = F)
summary(pca)
```

```
| Importance of components:
|                           PC1   PC2   PC3   PC4    PC5
| Standard deviation     0.60 0.353 0.287 0.188 0.0881
| Proportion of Variance 0.59 0.204 0.135 0.058 0.0127
| Cumulative Proportion  0.59 0.794 0.929 0.987 1.0000
```

```
screeplot(pca)
pca$rotation
```

```
|                PC1       PC2      PC3       PC4       PC5
| lineup     0.60940  0.616776 -0.47011  0.158183 -0.046896
| card_rot   0.19480 -0.057084  0.11515  0.101039  0.967123
| fig_class  0.29018  0.074811  0.15510 -0.940998  0.025810
| folding    0.46501  0.070647  0.81000  0.276633 -0.214832
| vis_search 0.53875 -0.778294 -0.29256  0.052605 -0.125118
```

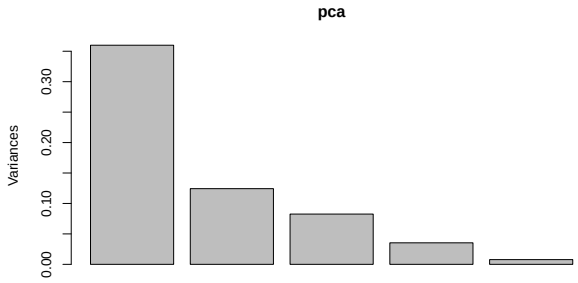

Figure 6: Scree plot of principle component analysis of performance on the different test batteries.

```
# 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     0.516 0.312 0.1943 0.0896
| Proportion of Variance 0.651 0.237 0.0921 0.0196
| Cumulative Proportion  0.651 0.888 0.9804 1.0000

pca.cog.tests$rotation

|                PC1       PC2       PC3        PC4
| card_rot   0.23206 -0.078194  0.095174  0.9648709
| fig_class  0.31620 -0.247444 -0.915835 -0.0057654
| folding    0.55259 -0.704563  0.382583 -0.2277389
| vis_search 0.73540  0.660491  0.076274 -0.1308659

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)	1.3352	0.0617	21.65	0.0000
PC1	0.4527	0.1210	3.74	0.0007
PC2	-0.2607	0.2004	-1.30	0.2023
PC3	-0.3952	0.3216	-1.23	0.2279
PC4	0.7103	0.6974	1.02	0.3158

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

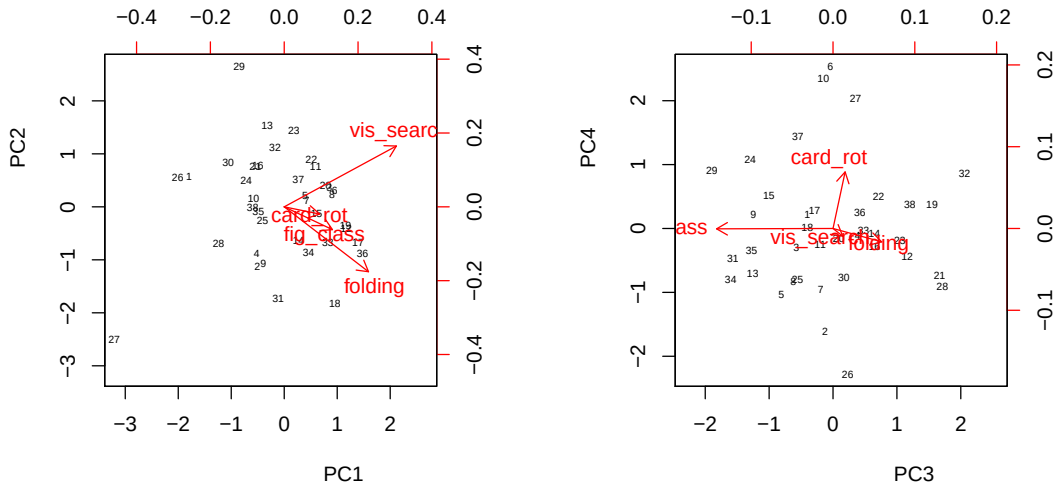

Figure 7: Plots of the principal components with observations. Visual Search and Paper Folding strongly contribute to both PC1 and PC2, while Figure Classification and Paper Folding strongly contributes to PC3 and Card Rotation strongly contributes to PC4

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.

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. There is a significant difference between male and female performance on lineups; this is not particularly surprising, since men perform better on many spatial tests (Voyer et al., 1995) and performance on spatial tests is correlated with phase of the menstrual cycle in women (Hausmann et al., 2000). 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.

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

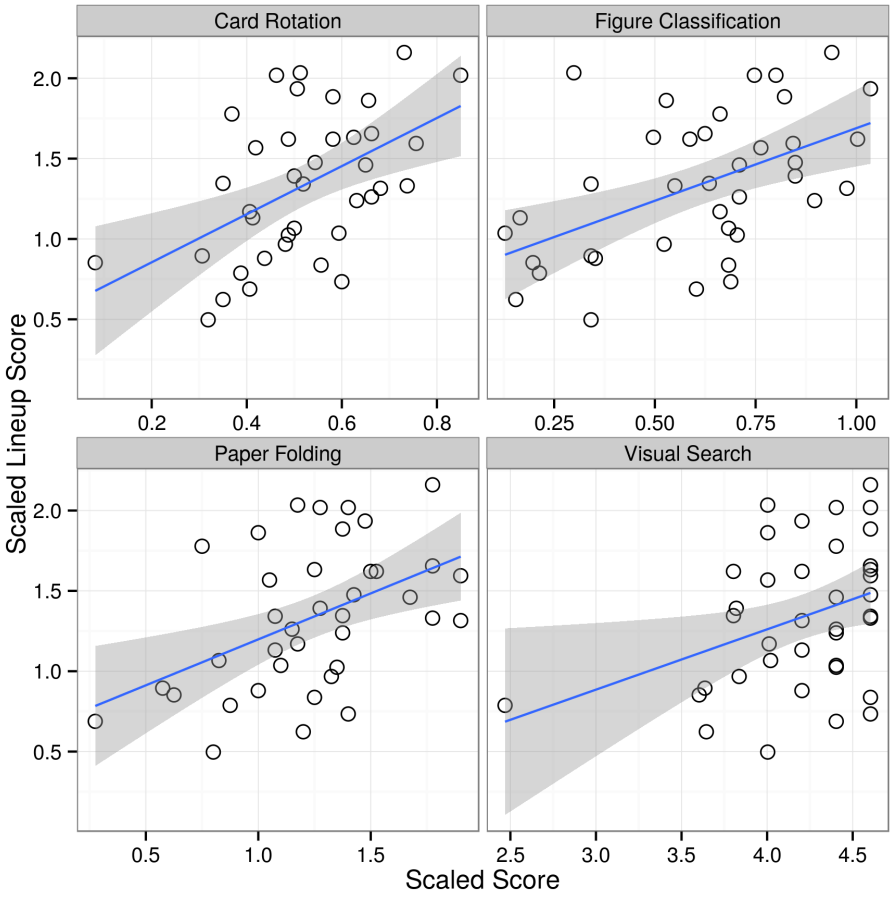

Figure 8: Scatterplots of all test scores compared to participants' scores in the lineup tests. There is a relatively strong positive correlation between lineup score and scores on visuospatial reasoning tests.

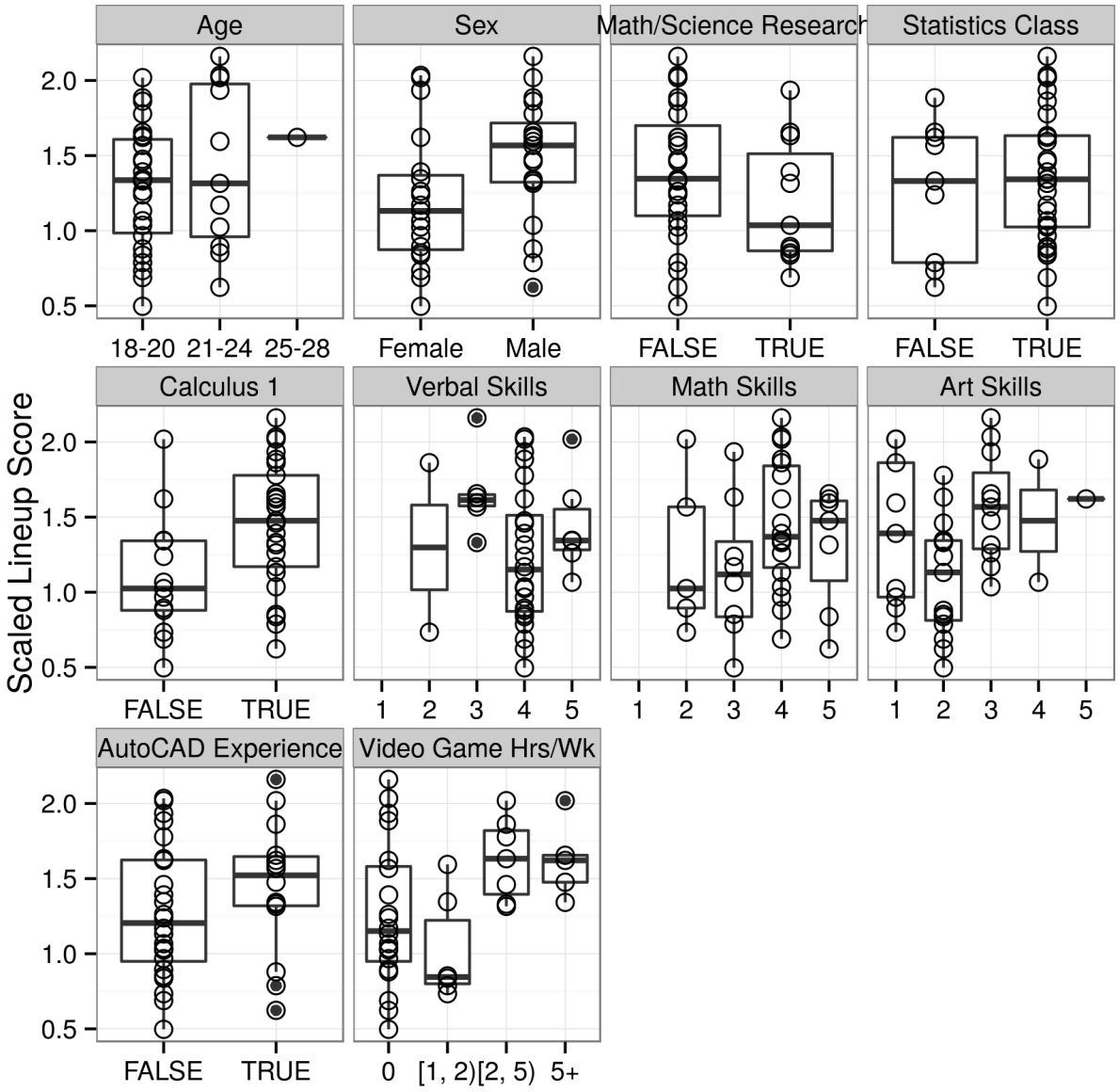

Figure 9: Sample demographic characteristics compared with lineup score. Sex (male), Calculus completion, and hours spent playing video games per week are all associated with a difference in lineup score.

Variable	DF	Sum.of.Squares	Mean.Squared	F.value	p.value
Calculus 1	1	1.080	1.080	6.15	0.018
Video Game Hrs/Wk	3	1.723	0.574	3.44	0.028
Sex	1	0.743	0.743	4.02	0.053
Art Skills	4	1.602	0.401	2.28	0.082
Verbal Skills	3	0.953	0.318	1.68	0.191
Math/Science Research	1	0.315	0.315	1.60	0.214
AutoCAD Experience	1	0.269	0.269	1.36	0.252
Math Skills	3	0.586	0.195	0.98	0.416
Age	2	0.219	0.109	0.53	0.592
Statistics Class	1	0.048	0.048	0.23	0.631

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.

References

Anderson, K. J. and Revelle, W. (1983). The interactive effects of caffeine, impulsivity and task demands on a visual search task. Personality and Individual Differences, 4(2):127–134.

DeMita, M. A., Johnson, J. H., and Hansen, K. E. (1981). The validity of a computerized visual searching task as an indicator of brain damage. Behavior Research Methods & Instrumentation, 13(4):592–594.

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.

Goldstein, G., Welch, R. B., Rennick, P. M., and Shelly, C. H. (1973). The validity of a visual searching task as an indicator of brain damage. Journal of consulting and clinical psychology, 41(3):434.

Hampson, E. (1990). Variations in sex-related cognitive abilities across the menstrual cycle. Brain and cognition, 14(1):26–43.

Hausmann, M., Slabbekoorn, D., Van Goozen, S. H., Cohen-Kettenis, P. T., and Güntürkün, O. (2000). Sex hormones affect spatial abilities during the menstrual cycle. Behavioral neuroscience, 114(6):1245.

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.

Larkin, J. H. and Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. Cognitive science, 11(1):65–100.

Lowrie, T. and Diezmann, C. M. (2007). Solving graphics problems: Student performance in junior grades. The Journal of Educational Research, 100(6):369–378.

Mayer, R. E. and Sims, V. K. (1994). For whom is a picture worth a thousand words? extensions of a dual-coding theory of multimedia learning. Journal of educational psychology, 86(3):389.

Moerland, M., Aldenkamp, A., and Alpherts, W. (1986). A neuropsychological test battery for the apple ll-e. International journal of man-machine studies, 25(4):453–467.

Scaife, M. and Rogers, Y. (1996). External cognition: how do graphical representations work? International journal of human-computer studies, 45(2):185–213.

Schaie, K. W., Maitland, S. B., Willis, S. L., and Intrieri, R. C. (1998). Longitudinal invariance of adult psychometric ability factor structures across 7 years. Psychology and aging, 13(1):8.

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

Shah, P., Mayer, R. E., and Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. Journal of Educational Psychology, 91(4):690.

Vekiri, I. (2002). What is the value of graphical displays in learning? Educational Psychology Review, 14(3):261–312.

Voyer, D., Voyer, S., and Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. Psychological bulletin, 117(2):250.

Zhang, J. (1997). The nature of external representations in problem solving. Cognitive science, 21(2):179–217.

Zhang, J. and Norman, D. A. (1994). Representations in distributed cognitive tasks. Cognitive science, 18(1):87–122.

Appendix

A Examining potential testing biases

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	0.50	1.06	35.99	0.29	-0.05	0.15
Paper Folding	1.32	1.17	1.27	35.93	0.21	-0.09	0.39
Figure Classification	0.70	0.53	2.16	32.86	0.04	0.01	0.32
Lineups	1.47	1.22	1.74	35.66	0.09	-0.04	0.53

B Scaling Scores

To calculate "scaled" comparison scores between tests which included different numbers of sections, 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 above, 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. To get a single unified score with standard deviation, we did the following calculations:

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

$$\sigma_{all} = \sqrt{(N_F \sigma_F^2 + N_M \sigma_M^2)/(N_F + N_M)} \quad (3)$$

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

	Card Rotation	Paper Folding	Figure Classification
ISU Students	83.4 (24.1)	12.4 (3.7)	57 (23.8)[4]
Scaled Scores	88.0 (34.8)	13.8 (4.5)	58.7 (14.4)[5]
Unscaled Scores	44.0 (24.6)[6]	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)

Table 4: 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.

Then, 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 adjusted the transformation as follows

$$\mu_{all} = 1/2(\mu_F N_F + \mu_M N_M)/(N_F + N_M) \quad (4)$$

$$\sigma_{all} = \sqrt{(N_F \sigma_F^2 + N_M \sigma_M^2)/(2(N_F + N_M))} \quad (5)$$