

Spatial Reasoning and Statistical Graphics

Susan VanderPlas, Heike Hofmann

October 16, 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](#).

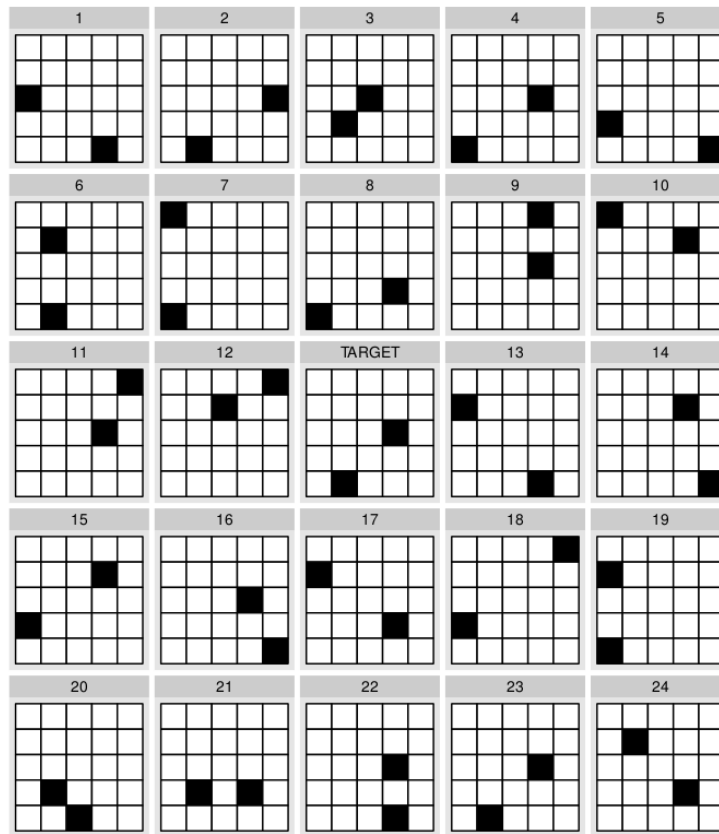


Figure 1: Visual Search Task. Participants are instructed to find the plot numbered 1-24 which matches the plot labeled "Target". Participants will complete up to 25 of these tasks in 5 minutes.



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](#)) and include some null lineups (i.e. lineups without a target plot). 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.

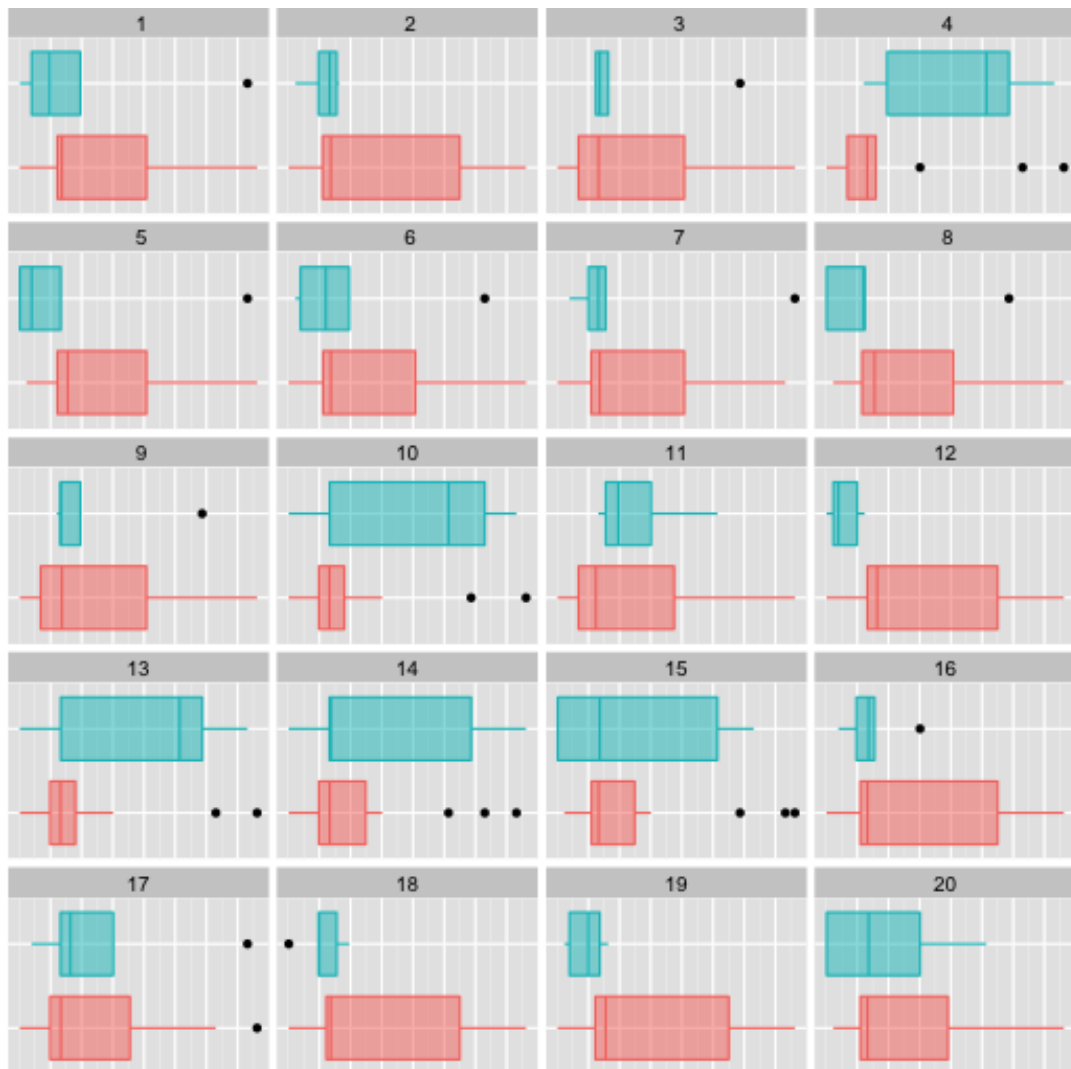


Figure 5: A sample lineup. Participants are instructed to choose the plot which appears most different from the others. In this lineup, plot 13 is the target.

3 Results

Results are based on an evaluation of 38 undergraduate students at Iowa State University. 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) which parts of this test are drawn from.

	Card Rotation	Paper Folding	Figure Classification
ISU Students	83.37 (24.1)	12.39 (3.7)	56.97 (23.8) ¹
Scaled Scores	88.0 (34.8)	13.8 (4.5)	58.67 (20.4) ²
Unscaled Scores	44.0 (24.6) ³	13.8 (4.5)	M: 120.0 (30.0), F: 114.9 (27.8)
Population	≈ 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 compared using the Kit of Factor-Referenced Cognitive Tests Ekstrom et al. (1976). 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

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

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

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

²Averages calculated assuming 294 males and 323 females.

³Data from Part I only.

⁴That is, if comparison data was available for 9th grade students and 12th grade students, we have compared Iowa State students’ scores with the 12th grade students.

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 1.33524 0.44730
## 2    card_rot 0.52105 0.15092
## 3    fig_class 0.60801 0.25354
## 4      folding 1.23947 0.36832
## 5 vis_search 4.19882 0.43247
```

```
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
```

using scaled version right now, should be changed to unscaled once the scores are internally scaled.

```
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
```

Just using the 4 cognitive tests and ignoring the lineups...

```
pca.cog.tests <- prcomp(ans.summary[,c("card_rot", "fig_class", "folding", "vis_search")])
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
```

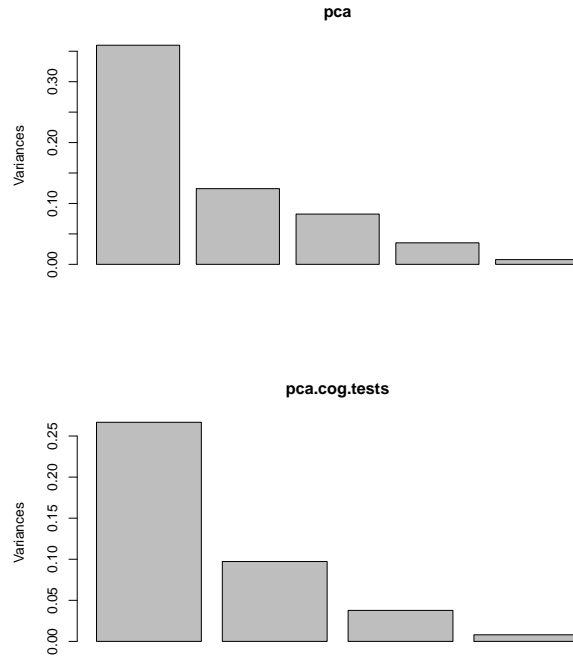


Figure 6: Scree plot of principle component analysis of performance on the different test batteries.

```
screepLOT(pca.cog.tests)
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

If test scores have the same ranges, we could include a mini table of means and standard deviations of the four tests, and discuss whether these findings are consistent to how people usually score (except for the females we have about the same population). *I'm not sure where to find the "normal scores" in the factor-referenced manual. I'll check through it again, but I don't recall seeing it.*

In figure 7, 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 8 shows participants' responses to the questionnaire given at the beginning of the study; these

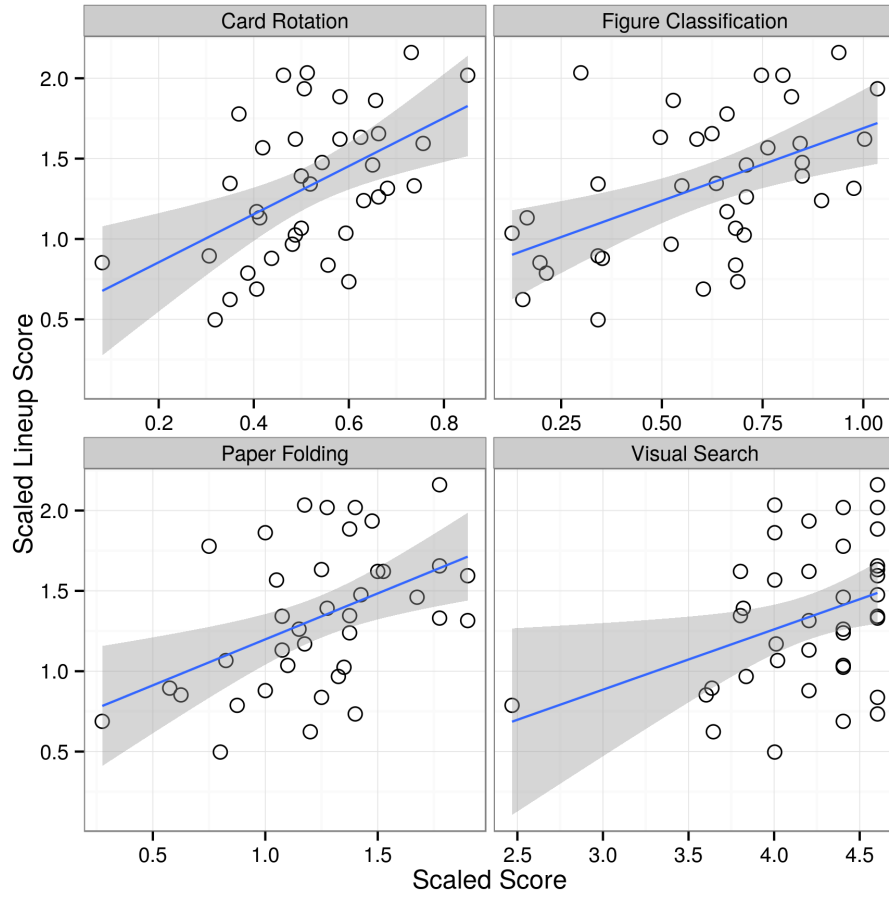


Figure 7: Scatterplots of all test scores compared to participants' scores in the lineup tests. There is a relatively strong positive correlation between lineup score and scores on visuospatial reasoning tests.

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 or completion of a statistics class. 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). In addition, completion of Calculus I is associated with increased performance on lineups (though completion of calculus is also associated with sex). AutoCAD experience is also not significantly associated with lineup performance; there is a difference in the medians, but it does not rise to the level of significance. 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 may be at least partially a result of the large difference in performance due to sex.

```
t.test(ans.summary$lineup[ans.summary$sex=="m"], ans.summary$lineup[ans.summary$sex=="f"])

|
| Welch Two Sample t-test
|
| data:  ans.summary$lineup[ans.summary$sex == "m"] and ans.summary$lineup[ans.summary$sex == "f"]
| t = 2.0046, df = 35.809, p-value = 0.05261
| alternative hypothesis: true difference in means is not equal to 0
| 95 percent confidence interval:
|  -0.0033311  0.5627835
| sample estimates:
| mean of x mean of y
|      1.4751      1.1954
```

```
t.test(ans.summary$lineup[ans.summary$calc_1=="y"], ans.summary$lineup[ans.summary$calc_1=="n"])

|
| Welch Two Sample t-test
|
| data:  ans.summary$lineup[ans.summary$calc_1 == "y"] and ans.summary$lineup[ans.summary$calc_1 == "n"]
| t = 2.5, df = 24.988, p-value = 0.01935
| alternative hypothesis: true difference in means is not equal to 0
| 95 percent confidence interval:
|  0.062591 0.648010
| sample estimates:
| mean of x mean of y
|      1.4568      1.1015
```

```
ans.summary$vidgame_hrs_factor_new <- factor(ans.summary$vidgame_hrs_factor, ordered=FALSE)
summary(lm(data=ans.summary, lineup~vidgame_hrs_factor_new))

|
| Call:
| lm(formula = lineup ~ vidgame_hrs_factor_new, data = ans.summary)
|
| Residuals:
```

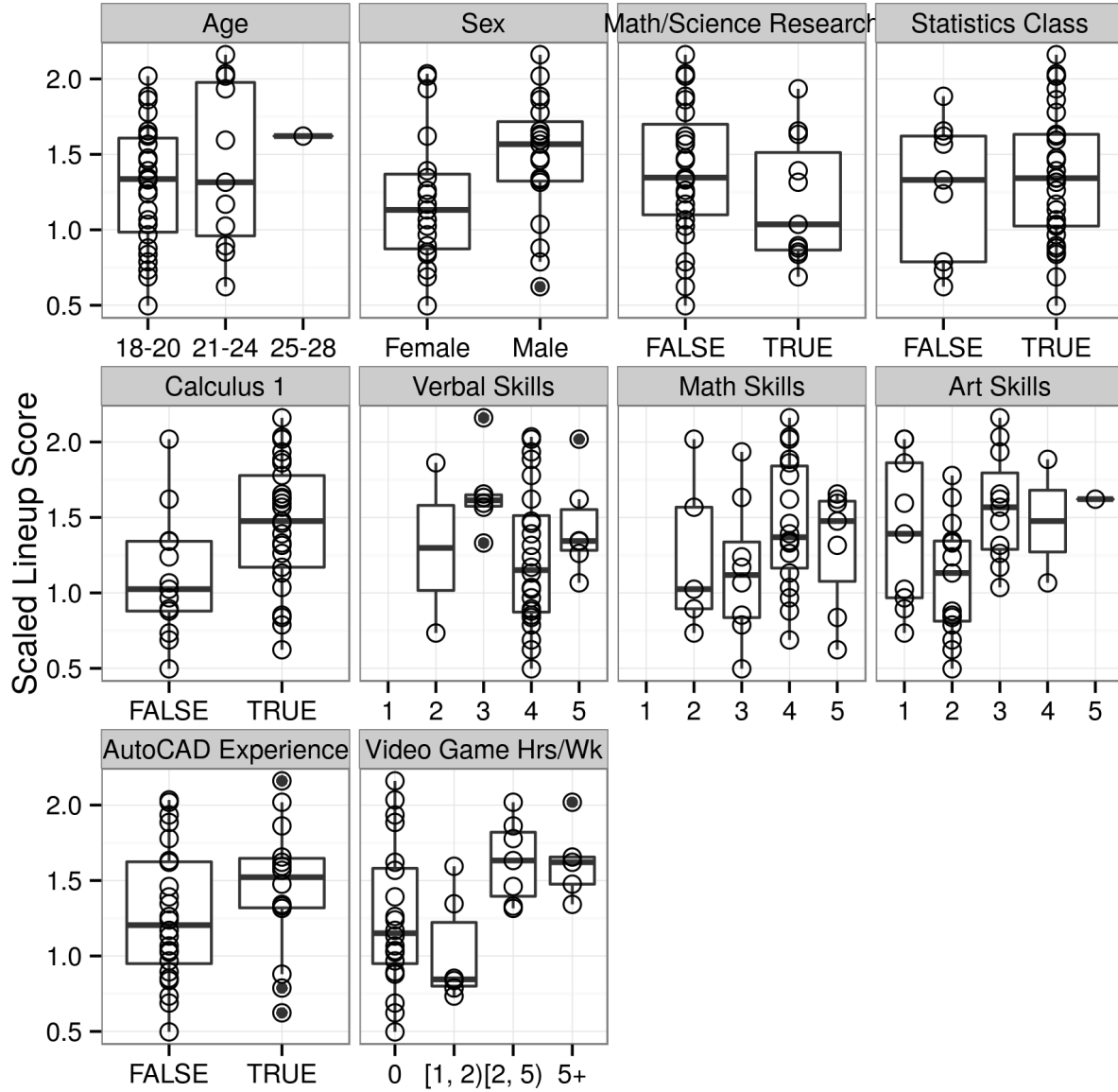


Figure 8: Sample demographic characteristics compared with lineup score. Sex (male), Calculus completion, and hours spent playing video games per week are all associated with a difference in lineup score.

```
|      Min      1Q  Median      3Q      Max
| -0.757 -0.270 -0.103  0.294  0.907
|
| Coefficients:
|               Estimate Std. Error t value Pr(>|t|)
| (Intercept)          1.2538    0.0914   13.72  2e-15 ***
| vidgame_hrs_factor_new[1, 2) -0.2284    0.1902   -1.20  0.238
| vidgame_hrs_factor_new[2, 5)  0.3745    0.1795    2.09  0.044 *
| vidgame_hrs_factor_new5+      0.3690    0.2044    1.81  0.080 .
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 0.409 on 34 degrees of freedom
| Multiple R-squared:  0.233, Adjusted R-squared:  0.165
| F-statistic: 3.44 on 3 and 34 DF,  p-value: 0.0275
|
# summary(lm(data=ans.summary, lineup~vidgame_hrs))
```

```
summary(lm(data=ans.summary, lineup~sex+calc_1+vidgame_hrs_factor_new))
|
| Call:
| lm(formula = lineup ~ sex + calc_1 + vidgame_hrs_factor_new,
|     data = ans.summary)
|
| Residuals:
|      Min      1Q  Median      3Q      Max
| -0.7206 -0.2844 -0.0781  0.2076  0.8165
|
| Coefficients:
|               Estimate Std. Error t value Pr(>|t|)
| (Intercept)          1.0667    0.1197    8.91  3.5e-10 ***
| sexm                 -0.0693    0.1676   -0.41  0.682
| calc_1y               0.3465    0.1480    2.34  0.026 *
| vidgame_hrs_factor_new[1, 2) -0.2492    0.1806   -1.38  0.177
| vidgame_hrs_factor_new[2, 5)  0.3241    0.1904    1.70  0.098 .
| vidgame_hrs_factor_new5+      0.4175    0.2265    1.84  0.075 .
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 0.388 on 32 degrees of freedom
| Multiple R-squared:  0.351, Adjusted R-squared:  0.249
| F-statistic: 3.46 on 5 and 32 DF,  p-value: 0.0131
```

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

Appendix

T-tests of results for Hillary and Stephanie:

```
t.test(ans.summary$card_rot[1:18], ans.summary$card_rot[-c(1:18)])

##
##  Welch Two Sample t-test
##
## data:  ans.summary$card_rot[1:18] and ans.summary$card_rot[-c(1:18)]
## t = 1.0631, df = 35.992, p-value = 0.2948
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.046928  0.150331
## sample estimates:
## mean of x mean of y
##  0.54826  0.49656

t.test(ans.summary$folding[1:18], ans.summary$folding[-c(1:18)])

##
##  Welch Two Sample t-test
##
## data:  ans.summary$folding[1:18] and ans.summary$folding[-c(1:18)]
## t = 1.2677, df = 35.934, p-value = 0.2131
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.089571  0.388183
## sample estimates:
## mean of x mean of y
##  1.3181  1.1687

t.test(ans.summary$lineup[1:18], ans.summary$lineup[-c(1:18)])
```

```
##
## Welch Two Sample t-test
##
## data:  ans.summary$lineup[1:18] and ans.summary$lineup[-c(1:18)]
## t = 1.7447, df = 35.661, p-value = 0.08965
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.040166  0.533537
## sample estimates:
## mean of x mean of y
##    1.4651    1.2184

t.test(ans.summary$vis_search[1:18], ans.summary$vis_search[-c(1:18)])

##
## Welch Two Sample t-test
##
## data:  ans.summary$vis_search[1:18] and ans.summary$vis_search[-c(1:18)]
## t = 1.0542, df = 32.86, p-value = 0.2995
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.13466  0.42417
## sample estimates:
## mean of x mean of y
##    4.2750    4.1303
```