```
## Warning:  attributes are not identical across measure variables;
they will be dropped
```

Spatial Reasoning and Statistical Graphics

Susan VanderPlas, Heike Hofmann

October 11, 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.

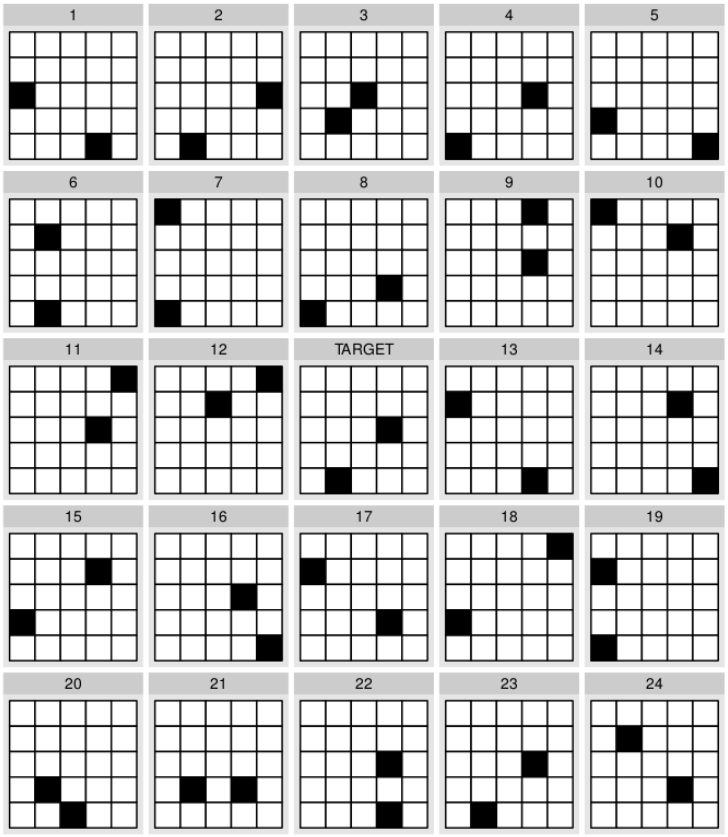

Figure 1: Visual Search Task. Participants are instructed to find the plot numbered 1-24 which matches the plot labeled "Target". Participants will complete up to 25 of these tasks in 5 minutes.

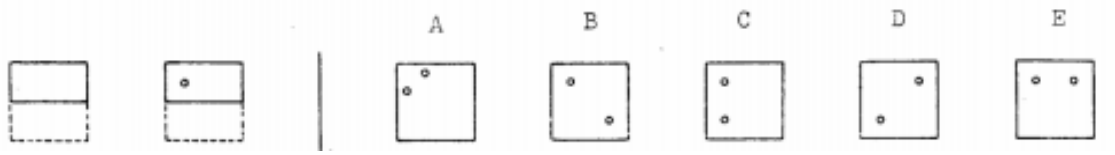

Figure 2: Paper Folding Task. Participants are instructed to pick the figure matching the sequence of steps shown in the left-hand figure. Participants will complete up to 20 of these tasks in 6 minutes.

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.

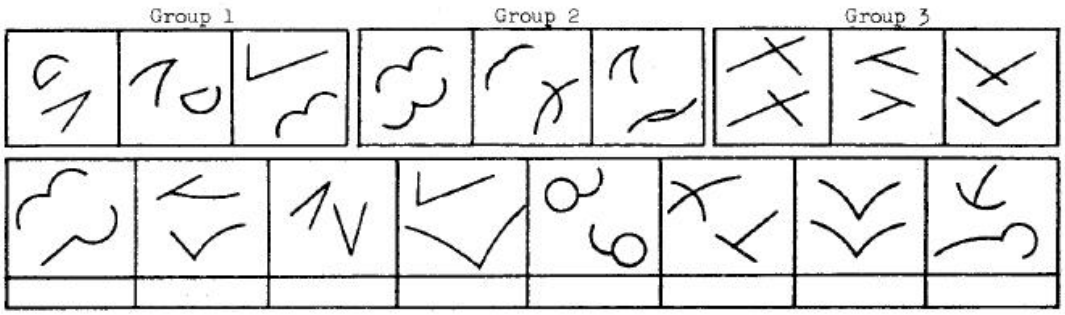

Figure 4: Figure Classification Task. Participants classify each figure in the second row as belonging to group 1, 2, or 3 (if applicable). Participants will complete up to 14 of these tasks (each consisting of 8 figures to classify) in 8 minutes.

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.

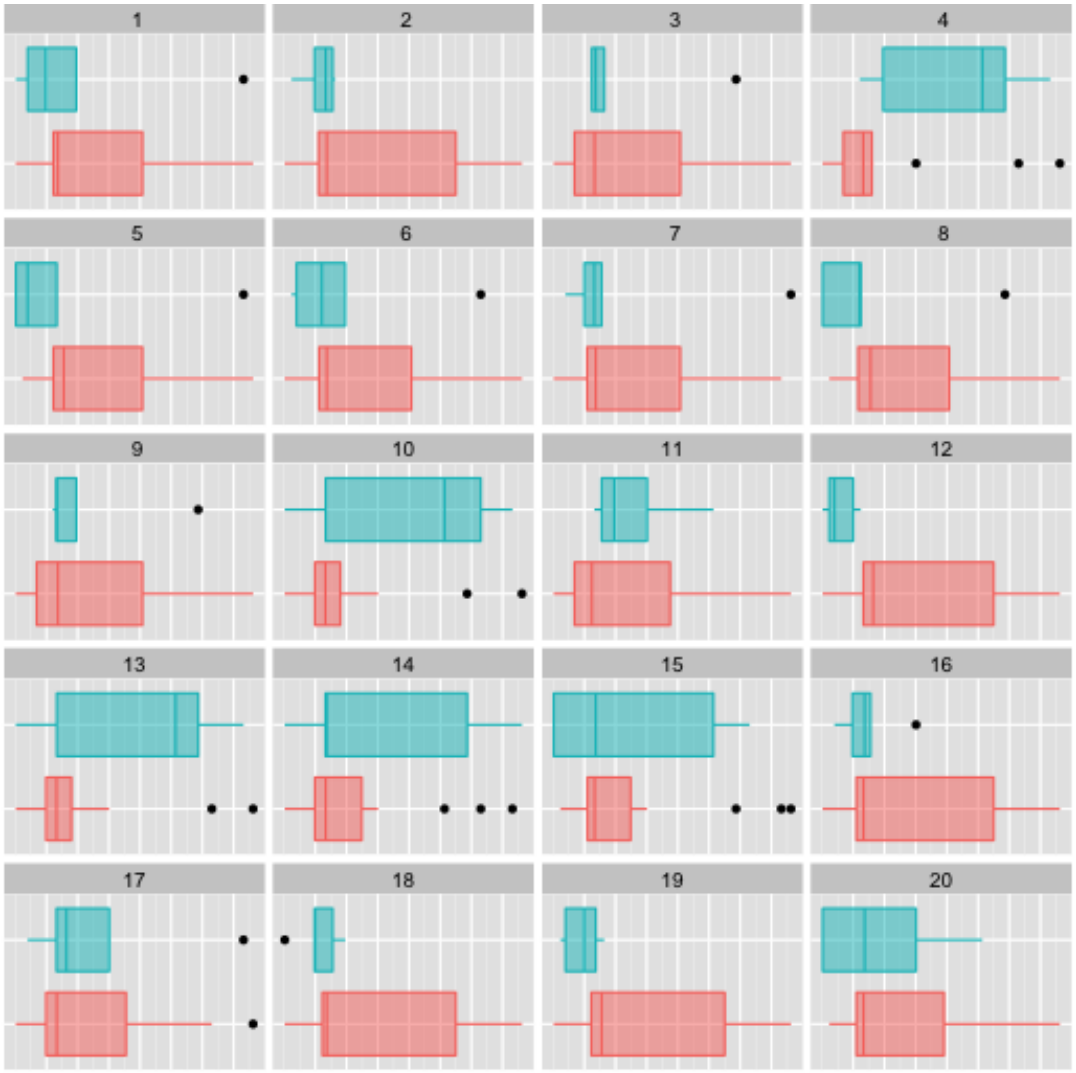

Figure 5: A sample lineup. Participants are instructed to choose the plot which appears most different from the others. In this lineup, plot 13 is the target.

3 Results

Results are based on an evaluation of 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; thus for a test consisting of multiple choice questions with k suggested answers with a single correct answer each, the score is calculated as

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

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

Additionally, we have to ensure that the ranges and units of test scores are comparable. Assume n questions with k choices (including one correct answer) each. This leads to a theoretical range of $[-n/(k-1), n]$ and, under an additional assumption of random guessing, a variance of

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

We might need to expand this a bit to account for the figure classification section, which has some questions with 2 answers and some with 3. Or we could leave it as an exercise for the reader :)

So are we scaling things by the variance/std dev. under random guessing? Or by the theoretical test range? Right now, I'm assuming that we're scaling between 0 and 100 by test range $[-a, b]$. So we take the guessing-fixed score and add a, then divide by $(a+b)$ and multiply by 100.

The next step is an overview of the possible ranges of the test scores in our study, because that is driven by the number of tests shown. It would be good to scale everything, so that we have the same theoretical range. That will actually allow a comparison across different tests.

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

```
##          .id   mean      sd
## 1     lineup 33.722  9.6756
## 2   card_rot 75.868  7.6631
## 3  fig_class 72.757 11.9356
## 4    folding 69.750 15.1287
## 5 vis_search 88.065  8.8338
```

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).

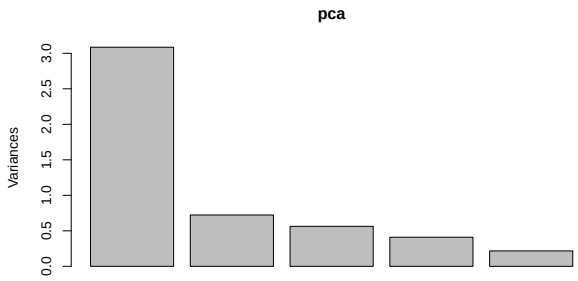

Figure 6: Scree plot of principle component analysis of performance on the different test batteries.

Split the next sentence up into two paragraph: what is shown in each of the figures, what are the immediate conclusions? chuck out the gaming hours from figure 7, that's a demograph, so move it to the other plot (I am aware that nine demographics are easier to display than ten, but content trumps).
Actually, they were split up like that because of continuous vs. categorical. It's not easy to facet when you have different x scales. I've removed video games entirely for the time being. I agree it didn't really belong there, but I can't get it to work with the facetting, either. I will probably just bin it eventually.

What happened to the arts skills? I seem to have a positive association now - why is my data different from yours?
I have the same updated results you do. I see that Stephanie changed the data file (but didn't add any more rows)... is it possible something changed there? Otherwise, maybe my dropbox didn't update correctly the first time? Or maybe it was figure caching?

Results are presented graphically in figures 7 and 8.

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

```
|             lineup card_rot fig_class folding vis_search
| lineup     1.00000  0.53553   0.60032 0.49547    0.40003
| card_rot   0.53553  1.00000   0.47870 0.72063    0.60739
| fig_class  0.60032  0.47870   1.00000 0.54890    0.38775
| folding    0.49547  0.72063   0.54890 1.00000    0.40867
| vis_search 0.40003  0.60739   0.38775 0.40867    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=T)
summary(pca)
```

```
| Importance of components:
|                           PC1   PC2   PC3   PC4    PC5
| Standard deviation     1.757 0.850 0.751 0.640 0.4662
| Proportion of Variance 0.617 0.145 0.113 0.082 0.0435
| Cumulative Proportion  0.617 0.762 0.875 0.957 1.0000
```

```
screeplot(pca)
```

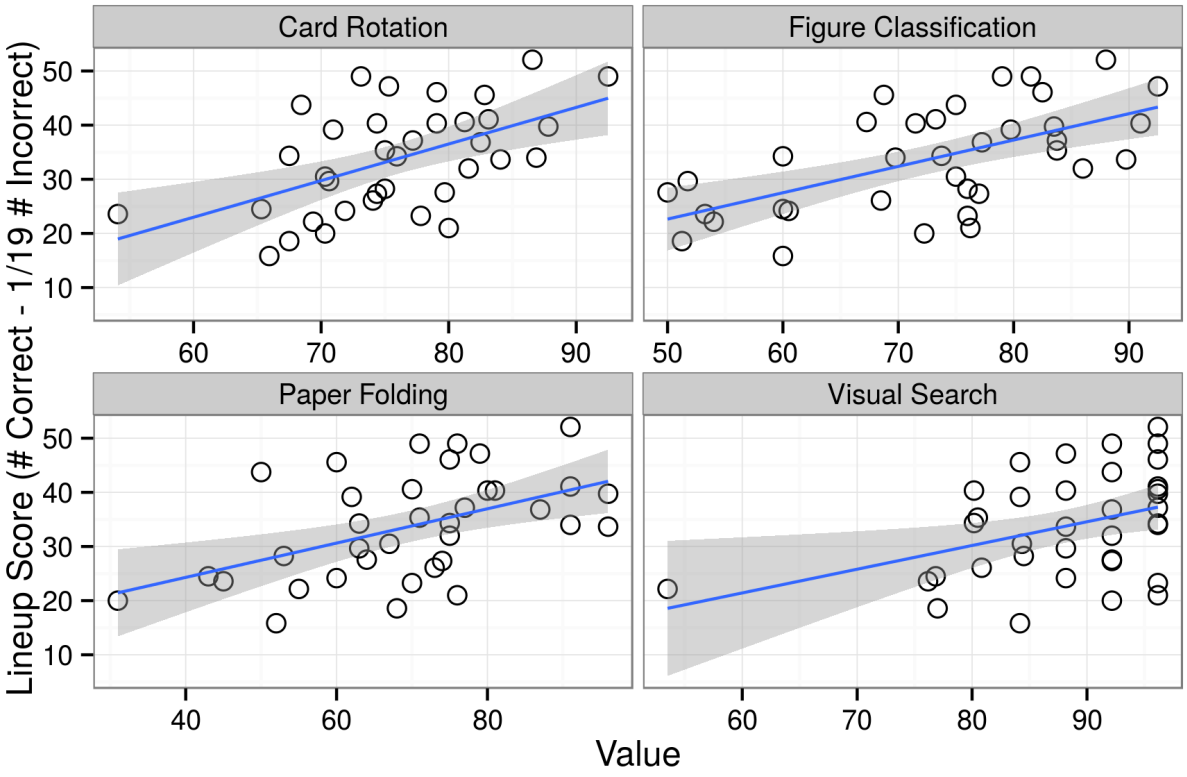

Figure 7: Preliminary results of continuous variables compared to lineup score.

I did change the lineup scoring - for some reason, I was averaging before (and I hadn't included the third battery of lineups b/c of scoring issues). That explains the scale, but not the change in art skills association.

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.

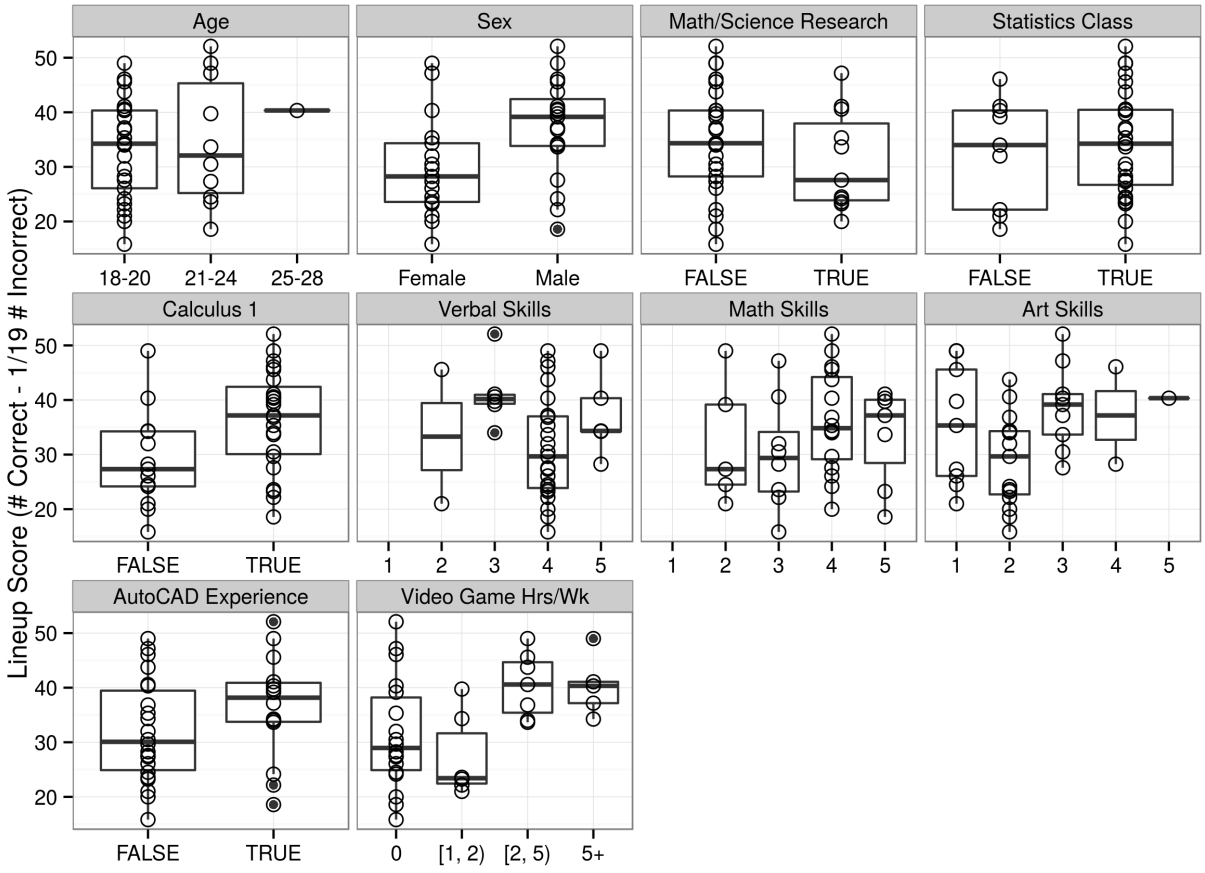

Figure 8: Preliminary results of categorical variables compared to lineup score.

Figure 8 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 little 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 = 2.3918, df = 33.454, p-value = 0.02254
| alternative hypothesis: true difference in means is not equal to 0
| 95 percent confidence interval:
|   1.0872 13.4267
| sample estimates:
| mean of x mean of y
|    37.149    29.892
```

```
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 = 2.3495, df = 25.529, p-value = 0.02682
| alternative hypothesis: true difference in means is not equal to 0
| 95 percent confidence interval:
|   0.91771 13.84483
| sample estimates:
| mean of x mean of y
|    36.388    29.006
```

```
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:
|    Min     1Q Median     3Q    Max
| -15.54  -5.49  -2.41   5.57  20.71
|
| Coefficients:
|                              Estimate Std. Error t value Pr(>|t|)
| (Intercept)                     31.37       2.04   15.39  2.4e-16 ***
| vidgame_hrs_factor_new[1, 2)    -4.02       4.08   -0.99    0.331
```

```
| vidgame_hrs_factor_new[2, 5)     9.12       3.85    2.37    0.024 *
| vidgame_hrs_factor_new5+         9.00       4.37    2.06    0.048 *
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 8.65 on 32 degrees of freedom
| Multiple R-squared:  0.27,Adjusted R-squared:  0.202
| F-statistic: 3.94 on 3 and 32 DF,  p-value: 0.0168
```

```
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
| -15.40  -5.88  -1.85   4.46  18.10
|
| Coefficients:
|                              Estimate Std. Error t value Pr(>|t|)
| (Intercept)                    27.886      2.594   10.75  8.3e-12 ***
| sexm                           -0.434      3.760   -0.12    0.909
| calc_1y                         6.532      3.342    1.95    0.060 .
| vidgame_hrs_factor_new[1, 2)   -4.749      3.938   -1.21    0.237
| vidgame_hrs_factor_new[2, 5)    7.375      4.098    1.80    0.082 .
| vidgame_hrs_factor_new5+        8.995      4.862    1.85    0.074 .
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 8.32 on 30 degrees of freedom
| Multiple R-squared:  0.367,Adjusted R-squared:  0.261
| F-statistic: 3.48 on 5 and 30 DF,  p-value: 0.0135
```

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 = 1.2181, df = 33.289, p-value = 0.2317
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.0693  8.2499
## sample estimates:
## mean of x mean of y
##    77.413    74.323

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 = 1.1856, df = 32.641, p-value = 0.2443
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.2608 16.1496
## sample estimates:
## mean of x mean of y
##    72.722    66.778

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 = 2.0822, df = 33.941, p-value = 0.04493
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.15354 12.67980
## sample estimates:
## mean of x mean of y
##    36.931    30.514
```

```r
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 = 1.0389, df = 28.345, p-value = 0.3077
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.9660  9.0772
## sample estimates:
## mean of x mean of y
##    89.593    86.537
```