Appendix: Spatial Reasoning and Data Displays

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

A Scaling Scores

To calculate "scaled" comparison scores between tests which included different numbers of test sections (as shown in Table 1), we scaled the mean in direct proportion to the number of questions (thus, if there were two sections of equivalent size, and the reference score included only one of those sections, we multiplied the reported mean score by two). The variance calculation is a bit more complicated: In the case described in the main text, where the reference section contained half of the questions, the variance is multiplied by two, causing the standard deviation to be multiplied by approximately 1.41.

This scaling gets slightly more complicated for scores which have two sub-groups, as with the figure classification test, which separately sumarizes male and female participants' scores. To get a single unified score with standard deviation, we completed the following calculations:

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

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

Here μ_F and μ_M are the mean scores for females and males, respectively; N_F and N_M are the number of female and male participants, and σ_F^2 and σ_M^2 are the variances in scores for females and males.

Substituting in the provided numbers, we get

$$\begin{aligned}\mu_{\text{all}} &= (323\cdot 114.9 + 294\cdot 120.0)/(323+294)\\ &= 58.7\end{aligned}$$

$$\begin{aligned}\sigma_{\text{all}} &= \sqrt{\left(323\cdot 27.8^2 + 294\cdot 30^2\right)/(323+294)}\\ &= 14.4.\end{aligned}$$

Whenever participants in two studies were not exposed to the same number of questions, the resulting scores are not comparable: both overall scores and their standard deviations are different. We can achieve comparability by scaling the scores accordingly. For example, in order to account for the fact that ISU students took only part I of two parts to the figure classification test (and thus completed half of the questions), we adjust the transformation as follows:

$$\begin{aligned}\mu_{\text{part I}} &= 1/2\cdot\mu_{\text{all}}\\ \sigma_{\text{part I}} &= 1/\sqrt{2}\cdot\sigma_{\text{all}}\end{aligned}$$

B Lineup Performance and Demographic Characteristics

Table A0 provides the results of a sequence of linear models fit to the lineup data. Each row in the table represents a single model, with one predictor variable (a factor with two or more levels). Due to sample size considerations, multiple testing corrections were not performed; in addition, the independent variables are correlated: in our sample, males are more likely to have completed Calculus 1, but are also more likely to spend time playing video games. As such, a model including two or more of the significant predictor variables shows all included variables to be nonsignificant. To better understand the effects of these variables, a larger study is necessary.

Table A0. Participant demographics' impact on lineup score. The table below shows each single demographic variable's association with lineup score. STEM major, completion of Calculus I, time spent playing video games, and gender all show some association with score on statistical lineups.

Variable	DF	MeanSq	F	p.val
STEM Major	1	401.517	14.44	0.001
Calculus 1	1	204.569	6.15	0.018
Video Game Hours	3	108.847	3.44	0.028
Sex	1	140.844	4.02	0.053
Art Skills	4	75.891	2.28	0.082
Verbal Skills	3	60.220	1.68	0.191
STEM Research	1	59.670	1.60	0.214
AutoCAD	1	50.893	1.36	0.252
Age	1	34.434	0.91	0.348
Math Skills	3	37.039	0.98	0.416
Statistics Class	1	9.062	0.23	0.631

C Lineup Plot Types

We can also compare participants' performance on specific types of lineup plots compared with their scores on the visual aptitude tests, for instance, accuracy on lineups which require mental rotation may be related to performance on the card rotation task.

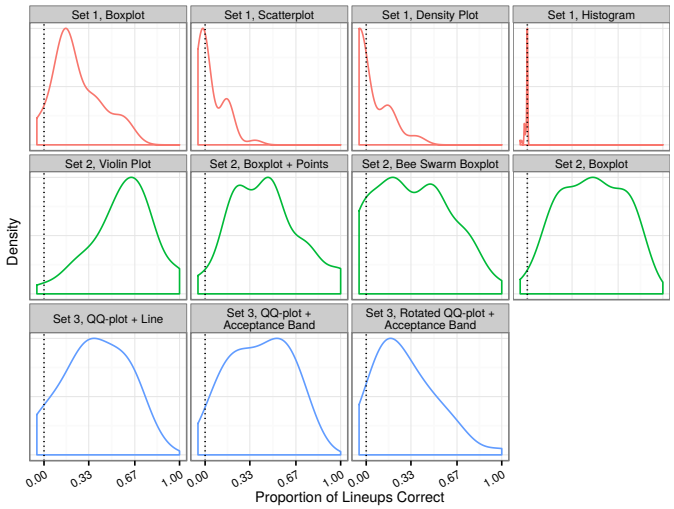

Fig. 1. Density plots of scaled scores for different types of lineups. For the same experiment (shown by line color), certain types of plots are more difficult to read and are associated with lower participant scores.

Figure 1 compares performance on each different type of plot. The x axis shows scaled score, the y axis shows the density of participant scores. As two different lineup tasks utilized boxplots to test different qualities of the distribution of data (outliers vs. difference in medians),

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 25 Oct. 2015. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

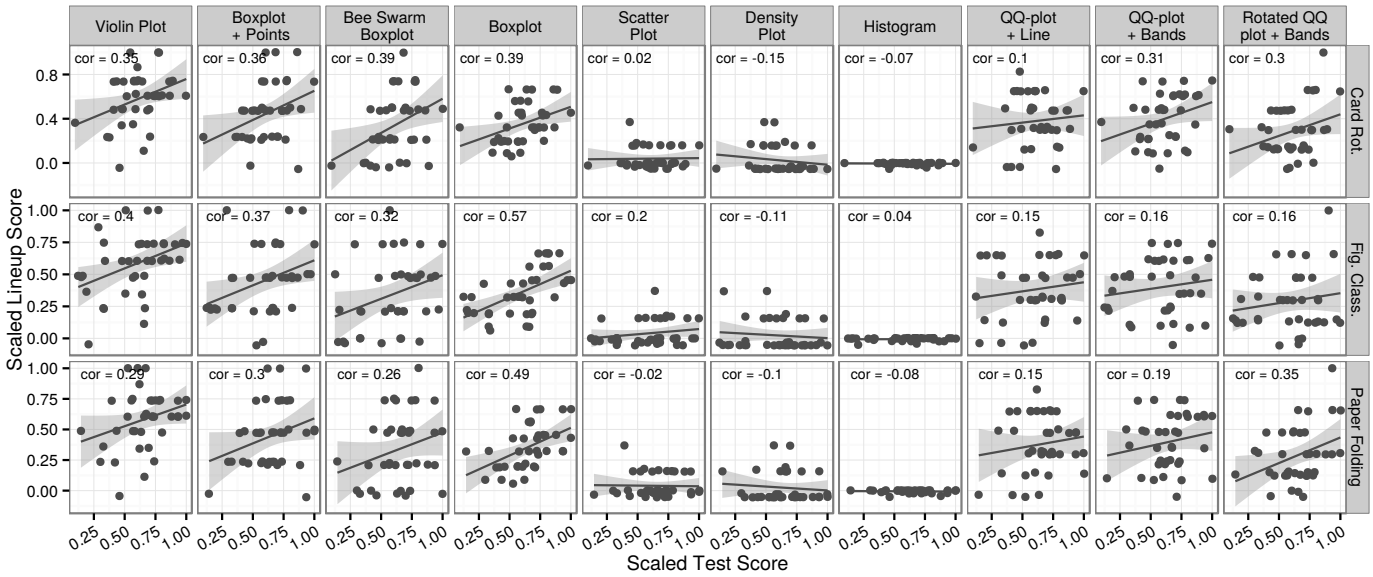

Fig. 2. Scatterplots of scaled lineup scores by aptitude test scores. There is some indication that different types of lineup tasks may utilize different visual skills; for instance, QQ-plots with confidence bands may require more skill at mental rotation than QQ-plots without the bands.

different tasks are shown as different colors, so that accuracy on tasks which are shown in blue can be compared to other blue density curves.

Figure 2 shows the association between scaled score on each type of lineup and score on the visual reasoning tests. Sample size for each plot type is fairly small - between 5 and 10 plots per individual, so there is low power for systematic inference, but we can establish that the QQ plots are more strongly associated with the card rotation task than with the figure classification task, particularly when confidence bands are included. Rotated QQ-plots seem to be much more asssociated with the paper folding task scores than other QQ-plot tasks; this may be because they require more visual manipulation than other QQ-plots.

For comparison, the correlation between general lineup score (non-subdivided) and the card rotation test score was 0.505, the correlation between general lineup score and the figure classification test was 0.512, and the correlation between lineup score and the paper folding test was 0.471.While we can compare the correlation strength between tasks, it is clear that the correlation between the score on any single lineup type and a particular visual aptitude score is lower than the overall relationship that we attribute to visual ability. Additional data is imperative to understand the reasoning required for specific types of plots - it is likely that the 5-10 trials per participant presented in each chart in Figure 2 are simply not sufficient to uncover any specific relationship between reasoning ability and lineup task.