A Visuospatial Aptitude Tests

The Visual Search Task (VST), shown in Figure 1, is used to measure a person's ability to locate a target amid a field of distractors.

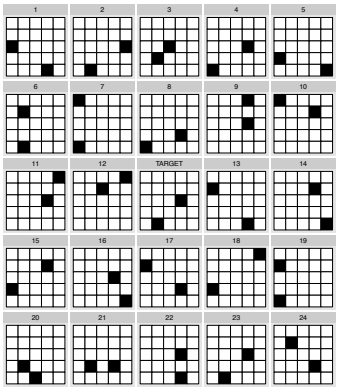
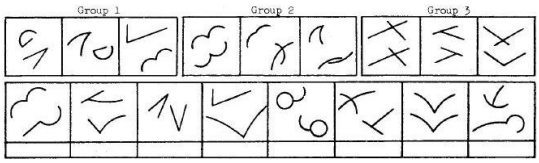


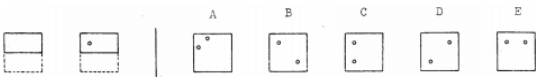
Fig. 1. Visual Search Task (VST). Participants are instructed to find the plot numbered 1-24 which matches the plot labeled "Target". Participants will complete up to 25 of these tasks in 5 minutes.

Figure 2 shows samples from the figure classification task, the card rotation task, and the paper folding task. All three tasks are part of the Kit of Factor-Referenced Cognitive Tests [1].



(a) Figure Classification Task. Participants classify each figure in the second row as belonging to one of the groups above. Participants complete up to 14 of these tasks (each consisting of 8 figures to classify) in 8 minutes.

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



(c) Paper Folding Task. Participants are instructed to pick the figure matching the sequence of steps shown on the left-hand side. Participants complete up to 20 of these tasks in 6 minutes.

Fig. 2. Visuospatial tests

B 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:

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

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

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

$$\begin{aligned}\mu_{\text{all}} &= (323 \cdot 114.9 + 294 \cdot 120.0)/(323+294) \\ &= 58.7 \\ \sigma_{\text{all}} &= \sqrt{\left(323 \cdot 27.8^2 + 294 \cdot 30^2\right)/(323+294)} \\ &= 14.4.\end{aligned}$$

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:

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

C Lineup Performance and Demographic Characteristics

Table A2 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 A2. 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

Table A2. Importance of principal components in an analysis of four tests of spatial ability: figure classification, paper folding, card rotation, and visual search.

	PC1	PC2	PC3	PC4
Standard deviation	1.61	0.81	0.73	0.49
Proportion of Variance	0.64	0.16	0.13	0.06
Cumulative Proportion	0.64	0.81	0.94	1.00

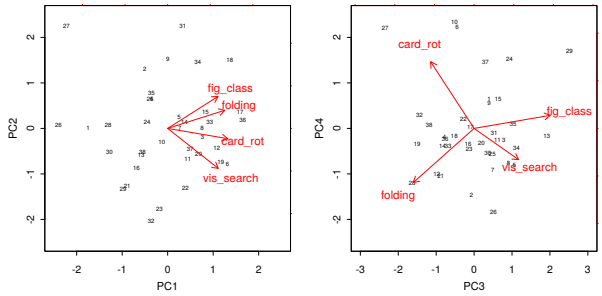


Fig. 3. Biplots of principal components 1-4 with observations. Principal component analysis was performed on the four cognitive tests used to understand the association between the cognitive skills required for these tests and the skills required for the lineup protocol.

D Principal Component Analysis of Visuo-Spatial Tests

D.1 Importance of principal components in an analysis of the four cognitive tests

Table A2 contains the proportion of the variance in the four cognitive tasks represented by each principal component. PC1 accounts for about 60% of the variance; Figure 3 and Table A2 confirm that PC1 is a measure of the similarity between all 4 tests; that is, a participant's general (or visual) aptitude. PC2 differentiates the figure classification test from the visual searching test, while PC3 differentiates these two from the paper folding test. PC4 is not particularly significant (it accounts for 5.9% of the variance), but it differentiates the card rotation task from the paper folding task.

Table A2. Rotation matrix for principal component analysis of the four cognitive tests (visual search, paper folding, card rotation, figure classification).

	PC1	PC2	PC3	PC4
card.rot	0.55	-0.19	-0.38	0.72
fig.class	0.46	0.58	0.66	0.14
folding	0.52	0.33	-0.53	-0.59
vis.search	0.46	-0.72	0.38	-0.34

Figure 3 shows that the first PC does not differentiate between any of the tasks; it might be best understood as a general aptitude factor. All of the remaining principal components distinguish between the cognitive tasks; PC2 and PC3 separate paper folding from visual search and from the lineup and figure classification tasks, while PC4 and PC5 mainly separate lineups from card rotation and figure classification. This separation allows us to compare the tasks which are similar from among the principal components. According to Table A2, the first three principal components account for 94.1% of the variance.

D.2 PCA of Cognitive Tests and Lineups

PC1 is essentially an average across all tests representing a general "visual intelligence" factor. Biplots of the remaining principal components are shown in Figure 4.

Figure classification is strongly related to lineups (PC2, PC3). Performance on the visual search task is also related to lineup performance (PC3). These two components highlight the shared demands

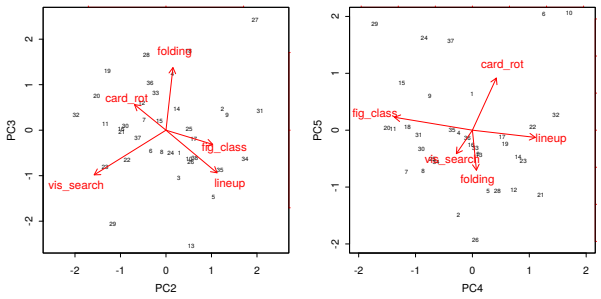


Fig. 4. Biplots of principal components 2-5 with observations. The lineup task appears to be most similar to the figure classification task, based on the plot of PC2 vs. PC3.

of the lineup task and the figure classification task: participants must establish categories from provided stimuli and then classify the stimuli accordingly.

The visual search task is also clearly important to lineup performance: PC3 captures the similarity between the visual search and lineup performance, and aspects of these tasks are negatively correlated with aspects of the paper folding and card rotation tasks within PC3. Paper folding does not seem to be strongly associated with lineup performance outside of the first principal component; card rotation is only positively associated with lineup performance in PC4.

PC4 captures the similarity between lineups and the card rotation task and separates this similarity from the figure classification task; this similarity does not account for much extra variance (10%), but it may be that only some lineups require spatial rotation skills. PC5 contains only 5% of the remaining variance, and is thus not of much interest, however, it seems to capture the relationship between the card rotation task and the paper folding and visual search tasks.

E Lineup Task Examples

E.1 Lineup Set 1

The experiment in the first lineup section examined the use of boxplots, density plots, histograms, and dotplots to compare two groups which vary in mean and sample size. The experiment was originally designed to explore the use of lineups to test plots of competing design[2]. This set of lineups consists of 20 plots selected from the plots used in the full experiment; each set of data is displayed with each of the four plot types.

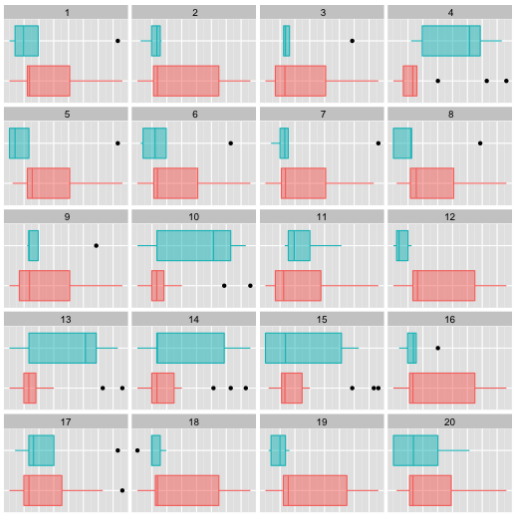


Fig. 5. Boxplots used to compare the two distributions.

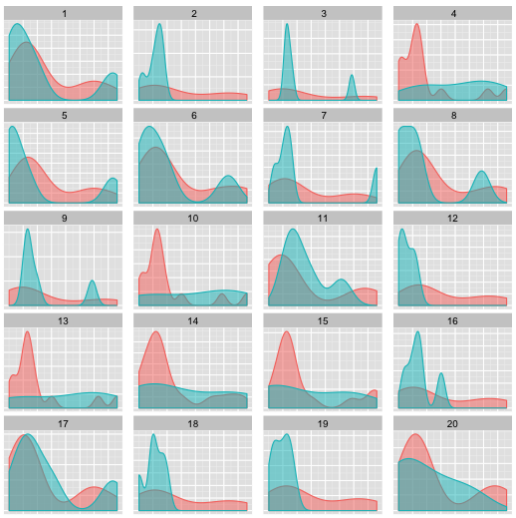


Fig. 6. Density plots used to compare the two distributions.

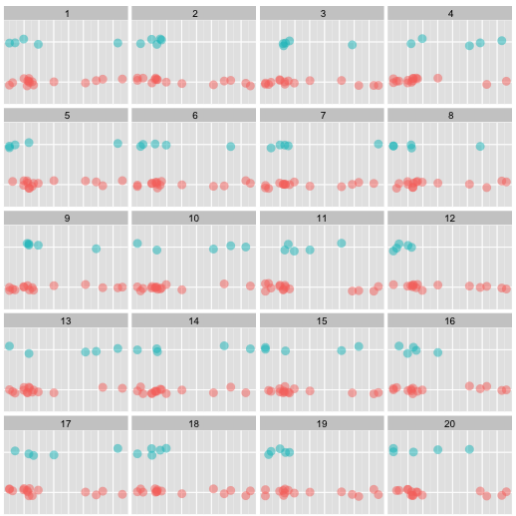


Fig. 7. Dotplots used to compare the two distributions.

Fig. 8. Histograms used to compare the two distributions.

E.2 Lineup Set 2

The second lineup section also explored two groups of data, this time comparing boxplots, bee swarm boxplots, boxplots with overlaid jittered data, and violin plots. Participants were much more accurate in this experiment than in the experiment described previously, because of the types of plots compared as well as the underlying data distributions.

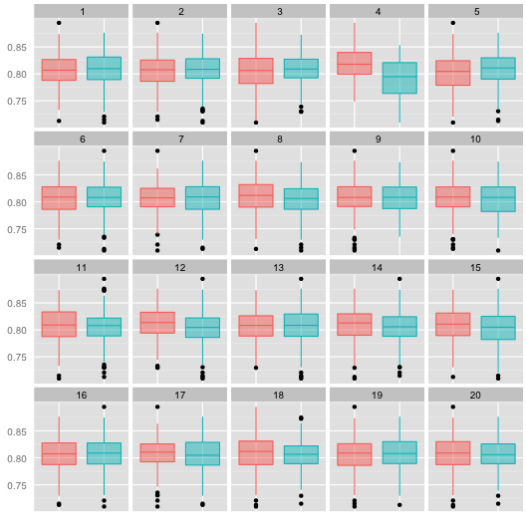


Fig. 9. Boxplots.

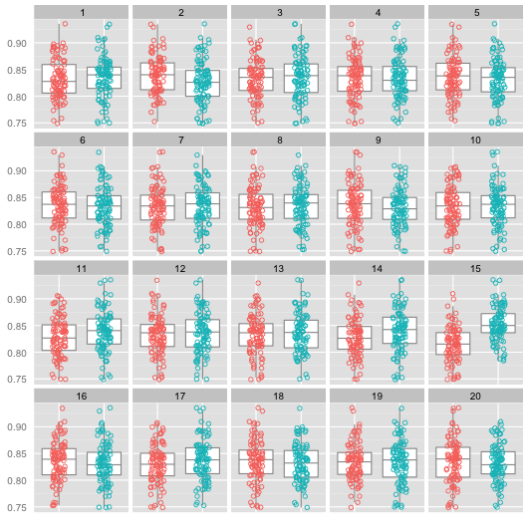


Fig. 10. Boxplots with jittered points.

E.3 Lineup Set 3

The final lineup section explored QQ-plots from various model simulations, using reference lines, acceptance bands, and rotation to determine which plots allowed participants to most effectively identify violations of normality. Rotated QQ-plots showed lower performance because participants were able to more accurately compare acceptance bands to residuals, and thus could identify that the reference bands were too liberal. As a result, performance was somewhat lower for rotated plots, even though participants were more accurate when comparing the residuals to the reference bands.

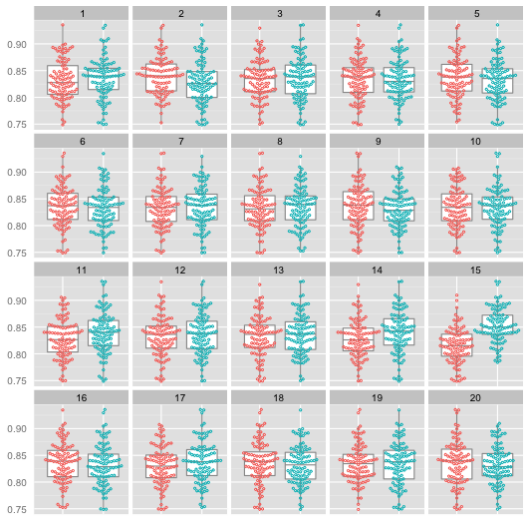


Fig. 11. Bee swarm boxplots.

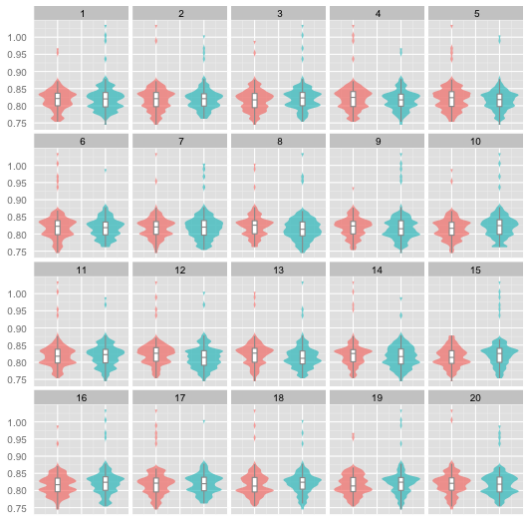


Fig. 12. Violin plots.

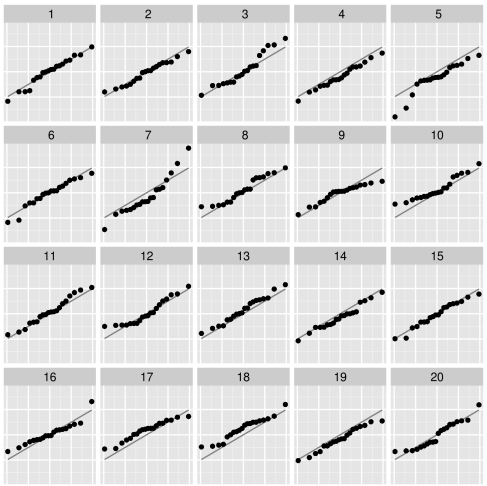


Fig. 13. QQ-plot with guide line.

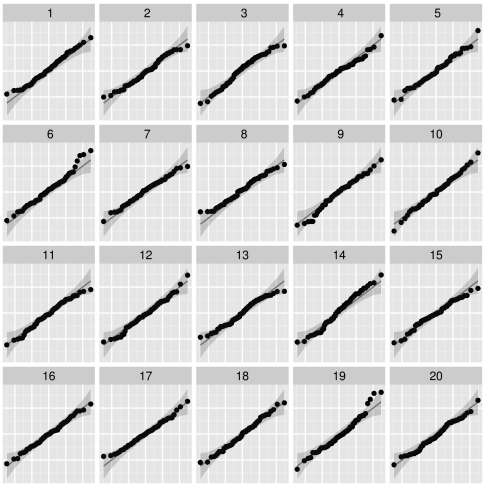


Fig. 14. QQ-plot with acceptance bands.

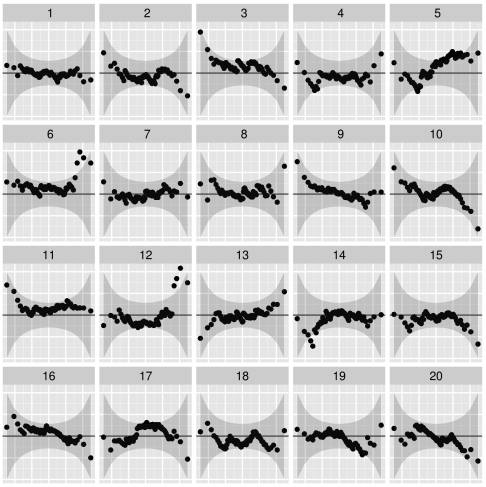


Fig. 15. QQ-plot rotated 45 degrees.

F 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.

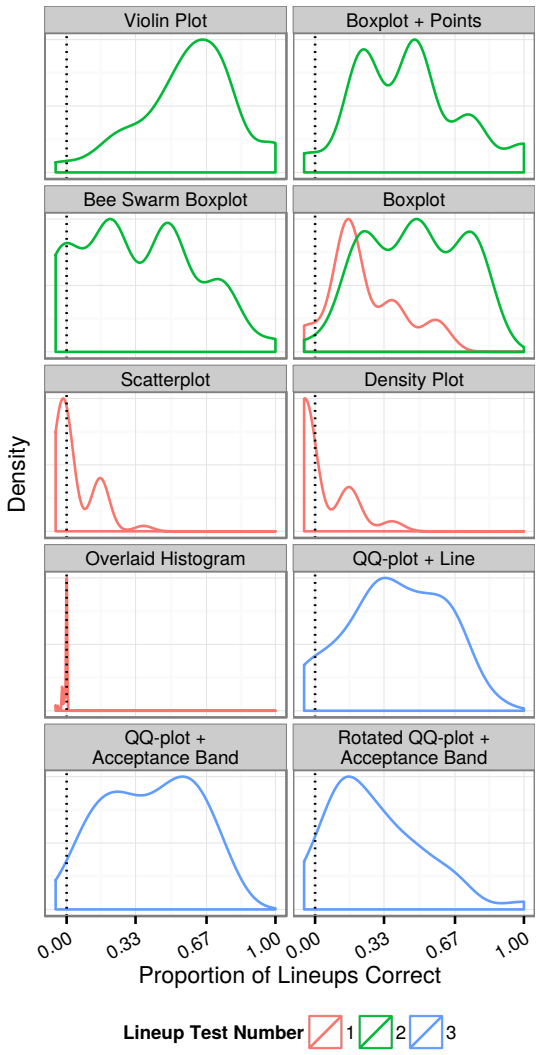


Fig. 16. Density plots of scaled scores for different types of lineups. For the same experiment (shown by line color), certain types of plots are more difficult to read and are associated with lower participant scores.

Figure 16 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), 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 17 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 card rotation task is much more significantly associated with the QQ-plots tasks compared to the other tasks. In addition, 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 17 are simply not sufficient to uncover any specific relationship between reasoning ability and lineup task.

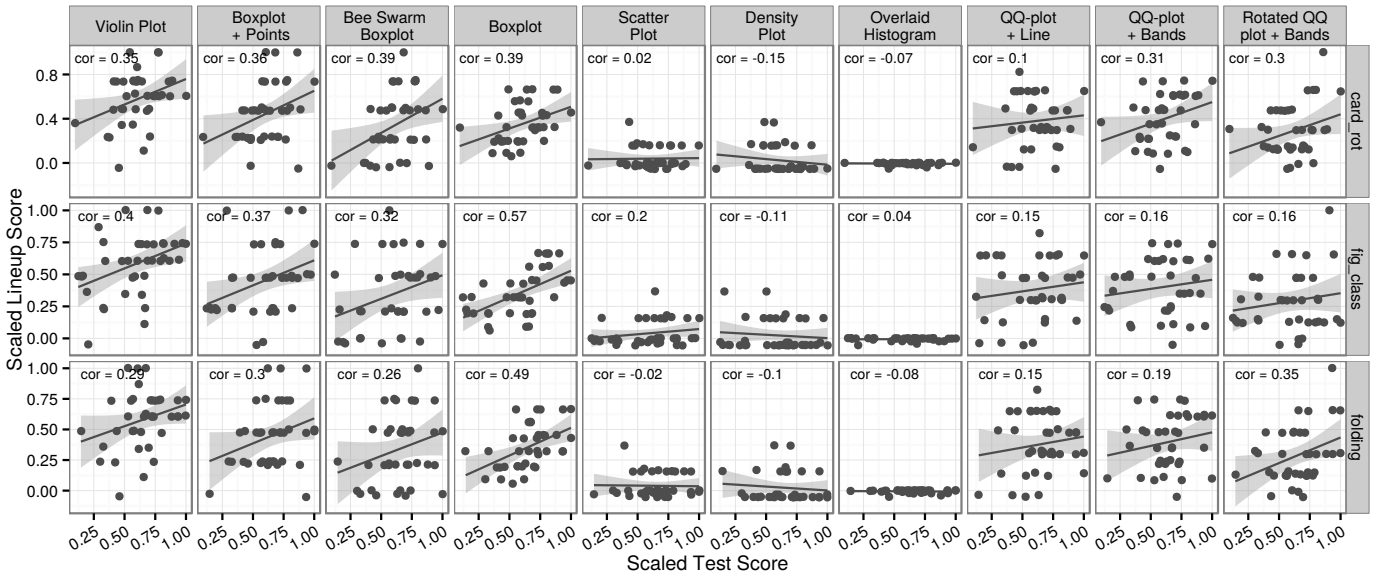


Fig. 17. Scatterplots of scaled lineup scores by aptitude test scores. There is some indication that different types of lineup tasks may utilize different visual skills; for instance, QQ-plots with confidence bands may require more skill at mental rotation than QQ-plots without the bands.

REFERENCES

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

[2] H. Hofmann, L. Follett, M. Majumder, and D. Cook. Graphical tests for power comparison of competing designs. IEEE Transactions on Visualization and Computer Graphics, 18(12):2441–2448, 2012. E.1