

CHAPTER III

Results

Tests of repeated measures analysis of variance (ANOVA) are reported using both univariate and multivariate statistics. All ANOVA summary tables are presented in Appendix B. Planned contrasts were used to further examine certain pairs of graph. For exploratory purpose, the Bonferroni method was employed to compare all possible pairs of means. In order to verify these results assuming a binomial model, logistic regressions with exact tests and 95 percent confidence intervals of proportion were computed. Also, test item reliability over time was examined.

For a proper use of repeated measures ANOVA, the number of observations should be larger than $a + 10$, where a is the number of level of repeated measures (Maxwell & Delaney, 1990). This study was comprised of eighteen scenarios but there were only twenty-three subjects. Therefore, two separate within-subjects-repeated-measures ANOVAs were employed. The first ANOVA used proportion correct scores from Question 1 to 9 (outlier identification) as the outcome variables with sample size and graphical format as repeated measures factors while the second ANOVA used scores of Question 10 to 18 (relationship examination). This arrangement ensured the legitimate use of repeated measures because the number of observations (23) exceeds $a + 10$ (19).

Follow-up tests examining cell differences for significant omnibus test relied primarily on focused comparisons using dependent t-tests. Because too many dependent t-tests would inflate Type I error rate, only six pairs of means directly related to the hypotheses were compared. The alpha level of the correlated t-tests was set at .05. In addition, following Maxwell and Delaney's recommendation (1990), pairwise

comparisons of all cell means were examined in an exploratory mode with a Bonferroni correction for this within subjects design.

To assess the similarity of results assuming a binomial model compared to the normal random variable model of ANOVA, two logistic regressions were computed. For computing logistic regression models, categorical variables such as data size and graph type were converted to dummy codes. Subjects were stratified in the logistic regression models for verifying the results of repeated measures.

In addition, confidence intervals on proportion correct were calculated. However, confidence intervals assume independence of observations, which was not present. Therefore, the results should be interpreted cautiously. Further, unlike repeated measures ANOVA, logistic regressions with exact tests and 95 percent confidence intervals used scores of Test 1 and 2 separately instead of the average score of two tests.

Reliability of Test Questions

To assess test item stability over time, Chi-square tests of independence and Phi coefficients by individual item between the two tests were computed to check the response consistency. Results are listed in Table 2. With the exception of Question 11, the response patterns in Test 1 and 2 were not significantly different from each other at the alpha level of .05.

For the composite scores, a Pearson correlation coefficient relating scores of the two tests was used. The Pearson coefficient was .59 ($p = .0064$). The composite scores of the two tests are plotted in Figure 15 suggesting a linear relationship between the two test scores.

Results of Outlier Identification Tasks

Repeated measures ANOVA. For the repeated measures ANOVA, the response variables consisted of the mean correctness of each question across Test 1 and 2. As seen in Table 3, in both univariate and multivariate statistics the graphical format factor yielded a significant effect, as was the interaction between graph type and data size. The sample size effect was not significant.

Figure 16a shows the performance of different graph types across the three levels of data size. Within each sample size, the use of graphical format significantly influenced the level of performance i.e. performance increased as the graphical format changed from 2D scatterplot to 3D spin plot, and from 3D spin plot to 3D mesh plot. However, in the medium and large data sets, the score difference was slight between 3D spin and 3D mesh plots.

Figure 16b shows the data size effect across all three types of graph. For 3D spin and mesh plots, performance increased as the sample size increased, while this relationship reversed for 2D plots. This reversed effect decreased the contribution to the sample size effect by 3D spin and 3D mesh plots.

Focused contrast. To test the alignment framework, specific pairs of contrast were made according to the stated hypotheses (see Table 4). In the case of small sample size, a one-tailed correlated t-test comparing 3D mesh with 3D spin plots was significant with $t(22) = 3.03$, $p = .0031$. Because the mean of 3D spin plot was higher than that of 2D scatterplots (See Table 5a), the preceding result implies that the mean of 3D mesh plot was significantly higher than that of 2D plots.

For medium sample sizes, the difference between scores of 2D and 3D spin plots was examined and found to be significant with $t(22) = 3.94$, $p = .00035$. The mean response for 3D mesh graphs was higher than

that of 3D spin plots (see Table 5a), therefore, 3D mesh was also considered superior to 2D plots.

For large sample sizes, the one-tailed dependent t-test indicated that 3D spin plots significantly outperformed 2D graphs with $t(22) = 8.04$, $p = .00005$. Again, since the mean performance on 3D mesh plots was higher than 3D spin plots, 3D mesh plots were also more effective than 2D plots.

Pairwise tests. Under the condition of small sample size, mean performance with 3D mesh plots was significantly higher than with 3D spin and 2D graphs. Under the circumstances of medium and large sample sizes, although the means of 3D mesh and 3D spin were significantly greater than that of 2D plots, there was no significant difference between 3D mesh and 3D spin. Table 5a is a reminder of the factorial design. The highest mean in each set is shaded. The results are summarized in Table 5b.

Logistic regression with exact tests. A logistic regression with exact test using Test 1 scores found a significant graph effect for the medium sample size and the large sample size. However, no significant graph effect was found for the small data size. Exact tests using Test 2 scores yielded significant results across all three sample sizes. Summary of exact tests are reported in Table 6.

Confidence intervals. The confidence intervals of proportion regarding outlier detection are shown in Table 7 and Figure 17. As shown in Figure 17, the results here were comparable to those of repeated measures ANOVA illustrated in Figure 16a.

Patterns of Test 1 and 2 were slightly different. For the small sample size of Test 1, the confidence bands of three types of graphs overlapped, and are not distinguishable in terms of performance. For the small data size of Test 2, however, 3D mesh plots were superior to

2D scattergrams. For the medium data size of Test 1, 3D spin plots were superior to 2D plots. Nonetheless, for the medium data size of Test 2, performance difference between 3D spin plots and 2D graphs were trivial.

Results of Relationship Examination Tasks

Repeated measures ANOVA. For the task of relationship examination, significant effects were found for size, graph and the size*graph interaction. Univariate and multivariate statistics are reported in Table 8.

Figure 18a shows the performance of different graphical format across different levels of sample size. In small and medium sample sizes performance increased as the graphical format changed from 2D scatterplot to 3D spin plot, and from 3D spin plot to 3D mesh plot. In the large sample size, 2D and 3D spin plots were equally effective but 3D mesh plots had an advantage over the other two.

Figure 18b shows the sample size effect across different graph types. For 2D plots, mean scores in small and medium sample sizes were identical. For 3D spin plots, performance improved substantively as the sample size increased. Although scores still increased as the sample size increased for 3D mesh plot, score differences among the three data sizes were slight.

Focused contrast. In the small data set condition, 3D mesh was found to be significantly better than 3D spin plots by one-tailed dependent t-test with $t(22) = 5.13$, $p = .00001$. Because the mean for use of 3D spin plots was greater than that of 2D graphs (see Table 9a), 3D mesh plots should also be considered better than its 2D counterparts.

For medium sample sizes, the difference between 3D mesh and spin plots was found to be significant, $t(22) = 3.10$, $p = .0026$. Again, the fact that the mean of 3D spin plots was higher than that of 2D plots led to the conclusion that 3D mesh plots were also superior to 2D plots.

Last, for large sample sizes, a significant effect was found in the comparison between 3D mesh plots and 2D plots, $t(22) = 2.24$, $p = .0357$. Since means for performance with 3D spin plots and 2D plots were virtually identical, 3D mesh plots are also superior to 3D spin plots.

Pairwise tests. In small sample sizes, the mean score of 3D mesh plots was significantly higher than those of 2D plots and 3D spin plots. In medium sample sizes, again 3D mesh plots were superior to the other two formats, and 3D spin plots were better than 2D plots. However, in large sample sizes, there was no significant difference among the mean scores of the three display types, though 3D mesh plots led to the highest mean. In Table 9a the highest mean in each set is shaded. The summary results are presented in Table 9b.

Logistic regression with exact tests. Exact tests reported here, which were stratified by subjects, were analogous to those of repeated measures ANOVA. In both Test 1 and 2 significant results were found in small and medium sample sizes. However, in the large sample size of both tests the three graphical formats did not differ from each other. Summary of exact tests are reported in Table 10.

Confidence intervals. As shown in Table 11 and Figure 19, confidence intervals of proportion pertaining to relationship examination resemble those of repeated measures ANOVA. Here differences in effectiveness of graph type were congruent with those illustrated in Figure 18a. In the medium data size of Test 1, variation of performance using 3D spin plots and 2D graphs were indistinguishable. This pattern did not hold in Test 2, in which 3D spin plots had an advantage over 2D plots.