Analysis of the UCL Immersion Experiment

Stephen Jones, Statistician, The Boeing Company

In the first section I will describe the test plan for the experiment, outlining some of the distinctive characteristics of the test plan that influences the analysis. In the second section I will describe the analysis of the time to completion test data. In the third section I will describe the analysis of the number of errors test data. Finally I make a brief comment about the frame rate data. (The analyses are a "first-cut". It will give you an overview of what is there but there are some aspects of the data that I have not had time to look at that we could examine later.)

(1) Experimental Design

The experiment is set up to investigate the effect of three different display interfaces on the ability of subjects to recreate 3-D models. The three displays used in the experiment were a physical replica, a computer-generated model on a desktop workstation screen, and a computer-generated model using a head-mounted display (hmd). There were three different puzzles used in the experiment, denoted by the letters A, B, and C.

One of the issues that arises in tests with human subjects is that the variability from person to person is large compared with the expected differences in the "treatment" of interest (in this case display methodologies). So frequently a large number of subjects is required to have any reasonable expectation of being able to detect differences among the factor of interest. One way to mitigate against the large inherent variability due to different subjects is to design the experiment so that each subject is tested with all of the conditions of the factor of interest. For this test this would involve each of the subjects completing a puzzle with each of the different display interfaces. The analysis would then be conducted by comparing the performance using the three display interfaces within each of the subjects, thus effectively eliminating the large variability between subjects. With this experimental design the comparisons of the primary objective, namely those between the interfaces, are estimated more precisely than if each subject only uses one interface. If each subject uses only one interface then the comparisons among the interfaces are made between the subjects and so it includes the variability due to the different subjects.

Clearly, it would be inadvisable to have each subject complete the same puzzle with each of the displays because any learning effect could potentially bias the results. To overcome this effect, three puzzles were available of approximately equivalent complexity. The test plan required each subject to complete each of the puzzles once using each of the three displays.

The use of experimental designs where each subject receives two or more different treatments (in our case, displays and models) have a history that goes back to agricultural trials in the 1860's. This type of design has been used extensively in medical and pharmaceutical research. A typical application would be an experiment with a drug and a

placebo where half of the subjects receive the drug first and then after a suitably chosen period of time cross over to the placebo. The remaining subjects receive the placebo first and then cross over to the drug. Such trials are called **cross-over** or change-over trials. Cross-over trials have also been applied in many scientific areas, such as veterinary research, sports science, and psychological experiments.

As indicated above, the advantage of a cross-over trial is that the treatments are compared within subjects; that is, every subject provides a direct comparison of the treatments that is not contaminated by differences between the subjects. This is in contrast to an experiment where each subject is randomly assigned to one of the treatments being compared and estimates of the differences between the treatment differences are obtained from comparisons between the subjects.

There are two related potential disadvantages with cross-over designs. They are *potential* disadvantages because they can be overcome virtually entirely if three or more periods are used in an appropriate manner. The first disadvantage is the possibility that the effect of a treatment given in one period might affect the performance in subsequent periods. The interpretation of this carry-over effect, as it is known, depends on the particular experiment. For our test a possible interpretation is that it is a psychological effect in that the attitude or confidence as the subject starts the puzzle in one period depends on their experiences with the puzzle they completed in the previous period.

A second, similar potential disadvantage is that the effect of a treatment might be different dependent on the period. This would imply a treatment-period interaction. For our test a possible interpretation would be that there is differential learning due to the sequence a subject uses to complete the models. Another interpretation is that fatigue affects the subject in the latter periods differently depending on the display or puzzle.

In this study it was feasible to have each subject complete each of the three models with each of the display interfaces. There are thirty-six possible sequences for the three puzzles and three displays. Thirty-six subjects were available for the study. Subjects were randomly assigned to one of the thirty-six possible sequences. In this test each subject assembled four puzzles, with the puzzle/display given in the first period repeated in period 4.

(2) Analysis of the Data: Completion Time

Initial plots of the data are shown in the attached Figures. In this plot the box indicates the range of the middle 50% of the times, that is from the 25^{th} to the 75^{th} percentile. The horizontal line in the box is the median. The whiskers extend from the box to the 10^{th} and 90^{th} percentile. The stars indicate points outside that range.

This plot indicates that there are differences between the displays. Shorter times are associated with the real display interface, whereas there does not appear to be large

differences between the times for the workstation and the hmd. Furthermore, there do not appear to be any large differences between the puzzles.

As described above, the experimental design can be viewed as a two-tiered design; one design within the subjects and one that is between the subjects. The analysis discussed below takes account of this structure.

The analysis involves fitting an empirical model to the data. In all of the models that we will consider the response of interest, say completion time for the puzzle, is modeled as a linear function of various parameters. These parameters include an overall mean level, coefficients that measure the effects of the factors in the experiment (such as display, puzzle, gender, period, SAT), and a term that measures the random deviation between the response and the model. The coefficients for the factors fall into two categories, main effects that measure how the setting of a single factor changes the response level from the mean, and interaction effects that measure how the main effect of one factor is changed by the setting of a second factor.

Although a complete model that combines both the between and within subject analysis can be written, for clarity I will initially consider the between-subjects model.

(a) Analysis Between Subjects

Let p_i be the response (average completion time for the puzzles) for subject *i* (*i*=1, ... , 36). Now p_i can be written in terms gender, SAT, age and other factors associated with subject *i*. Thus,

$$p_i = \mu + \alpha_1 (\text{SAT})_i + \alpha_2 (\text{Age})_i + \alpha_3 (\text{Status})_i + \alpha_4 (\text{Computer Use})_i + g_{i} + \varepsilon_j \quad (1)$$

where
μ is the overall mean completion time,
α_1 measures the effect SAT,
α_2 measures the effect Age,
α_3 measures the effect Status,
α_4 measures the effect Computer Use,
g_i is the effect of the gender (*l* = male or female),
ε_j is the random error for subject *i*,.

At this stage it is worth making a comment about the continuous variables that are subject-unique, such as SAT, Status, Age and Computer Use. A correlation analysis of these four subject variables shows strong correlations between Age and Status and between Status and Computer Use. These correlations make it difficult to separately identify the effect these variables. So in the analysis, the variable for Status was not added to the model. Furthermore, two of the subjects did not report their age. Putting the Age variable in the model has the effect of omitting data from these two subjects. An initial analysis with the variables Age, SAT and Computer Use indicated that Age did not have a significant effect on completion time when the other variables are included in the model. So the Age variable has been excluded from the analysis.

The between-subjects analysis uses the data from the first three periods. The analysis is summarized in the Analysis of Variance (ANOVA) Table shown below. The p-values given in the last column are a measure of the significance of the indicated factor. A common interpretation of a p-value is that it is the probability that data at least as extreme as collected in the experiment could have been due to random chance. This implies that if you concluded that the indicated factor has a significant effect on the response then you would be making a mistake (and the truth is that the factor has no effect on the response) 100p % of the time. A traditional cut-off point for a p-value is 0.05, implying that you are risk of incorrectly stating that a factor is significant is 5% or 1 in 20.

The ANOVA Table indicates an effect of marginal significance due to SAT (as indicated by a p-value slightly more than 0.05). The coefficient for SAT (the estimate for α_1 in the above model) is –2.8, indicating that every single point increase in SAT results in a decrease in the average time to completion of 2.80 seconds. It is noted that there are no

significant differences due to Computer Use (after adjusting for SAT). Also there is no significant difference in completion time between males and females.

Table 1. ANOVA Table for Between Subjects: Completion Time

Source	df	Sums of Squares	Mean Squares	F Value	Pr > F
SAT	1	30988	30988	4.10	0.0514
Computer Use	1	9658	9658	1.28	0.2668
Gender	1	957	957	0.13	0.7244
Error	32	242003	7562		
Corrected Total	35	288190			

An analysis of the residuals from the model fit indicates that there is one subject (no. 28) for whom the model gives a poor fit - the predicted average completing time of 369 seconds is substantially less than the observed average of 155 seconds. If this observation is removed from the data, then the effect of SAT is more significant (p-value = 0.0072) and the coefficient for SAT is –3.7.

(b) Analysis Within Subjects

The analysis within subjects is conducted having adjusted for any between subject variation. Thus the first term in the model (after the mean) will be a term that measures subject-subject differences. Of course this terms includes the effect of gender, SAT, and Computer Use.

Let p_i be the random effect of subject i (i=1, ... , 36). Let $y_{ijkl,}$ be the completion time for the i-th person who used the j-th display, on the k-th puzzle, in period l. Then,

$$y_{ijkl} = \mu + p_i + \pi_j + m_k + (\tau m)_{jk} + \tau_l + \lambda_{jl} + \xi_{jjkl} \qquad (2)$$

where
μ is the overall mean completion time,
p_i is the subject effect,
m_k is the effect of the puzzle,
π_j is the effect of the display,
τ_l is the effect of the period,
λ_{il} is the carry-over effect,
$(\tau m)_{jk}$ is the interaction effect of the display and the puzzle,
ξ_{ijkl} is the random error.

The interaction $(\tau m)_{ip}$ measures how the effect of the interface is different depending on the puzzle. The carry-over effect , l_k, measures the effect of the previous puzzle on the

completion time of the current puzzle. We can also fit interaction terms from the between-models analysis by adding interactions involving gender..

The response that was modeled was the log(completion time). The justification for use of the log transformation of the data is that an underlying assumption of the model is that the errors, ξ_{ijkl} , should be independent and follow a normal distribution with constant variance. An analysis with the completion time is not appropriate since the assumption of normality is violated because completion time must always be positive and the constant variance assumption is violated because the variability increases as the completion time increases. In these situations a data transformation of the response will frequently yield an analysis that satisfies the assumptions; and a log transformation is indicated by an examination of some data plots.

The following table summarizes the statistical tests constructed from fitting the model in equation (2) given above

Table 2. Tests of Within-Subject Effects: Completion Time

Source	df	F ratio	p-value
Display	2	130.90	0.0000
Period	3	10.70	0.0000
Puzzle	2	8.95	0.0003
Display*Puzzle	4	1.08	0.3720
Display*Gender	4	0.62	0.5421
Puzzle*Gender	2	1.09	0.3402
Carry-Over	2	0.92	0.4029
Within-Subjects Error	91		

From Table 2 we conclude that there is overwhelming evidence that there are differences due to the displays, due to the periods, and due to the puzzles. There is no evidence of an interaction between the display and puzzle, or between display and gender. The following table gives the average values for some of the factors of interest on a logged scale, and back-transformed to a time scale of seconds. (These means are found by fitting a model without the non-significant interactions involving gender.)

Table 3. Table of Means

	Mean on Log Scale	Mean on Time Scale (secs)
Desktop	5.675	291
HMD	5.916	371
Real	5.142	171
Puzzle A	5.525	251
Puzzle B	5.700	299
Puzzle C	5.507	246
Period 1	5.607	272
Period 2	5.668	289
Period 3	5.633	279
Period 4	5.402	222

Further analysis of contrasts of interest is summarized in the following table.

Table 4. Analysis of Contrasts of Interest

Comparison	Estimate	p-value
Desktop - HMD	-0.241	0.0000
Desktop - Real	0.533	0.0000
HMD - Real	0.774	0.0000
Model A – Model B	-0.175	0.0014
Model A – Model C	0.018	0.7432
Model B – Model C	0.193	0.0005

This table indicates that both the HMD and desktop have significantly longer average completion times than the physical interface (the estimate is positive). Furthermore, there is a significant difference between HMD and desktop, with the HMD completion times being longer than the desktop times (on average 80 seconds longer). There is no difference in the average completion times for puzzles A and C, but completion times for puzzle B were significantly longer. Also there is no difference in the average completion times for puzzles in periods 1, 2, and 3, but completion times for period 4 were significantly shorter.

An analysis of the within subject residuals revealed one unusual observation – subject 7 in period 1 (real display, puzzle A) had a completion time that was substantially longer than that predicted by the model. Removing this observation and repeating the analysis did not result in any other substantive changes to the conclusions.

(3) Analysis of the Data: Error Rates

I will now give a brief overview of the analysis of error rates. Similar models to those that were fit to completion time were used in the analysis of error rates.

(a) Analysis Between Subjects

The between-subjects analysis uses the data from the first three periods. The response that was modeled was the number of errors. There is some question as to whether this is the appropriate metric, but as a "first-cut" the analysis was conducted without any transformation. If I knew more about how the number of errors was determined I might be able to come up with a more appropriate metric. As with completion rates, Age did not appear to be an important variable when SAT and Computer Use are in the model and so it is not included in the analysis reported below. The analysis is summarized in the Analysis of Variance (ANOVA) Table shown below.

The ANOVA Table indicates a marginal effect due to SAT (as indicated by a p-value just slightly more than 0.05). The coefficient for SAT is negative, indicating that as the SAT increases the time to completion decreases. Specifically, the coefficient is –0.03, indicating that every single point increase in SAT results in a decrease in the number of errors of 0.03. It is noted that there are no significant differences due to Computer Use (after adjusting for SAT). Also there is no significant difference in completion time between males and females.

Table 5. ANOVA Table for Between Subjects: Number of Errors

Source	df	Sums of Squares	Mean Squares	F Value	Pr > F
SAT	1	3.480	3.480	3.61	0.0665
Computer Use	1	0.927	0.927	0.96	0.3341
Gender	1	2.378	2.378	2.47	0.1261
Error	32	30.859	0.964		
Corrected Total	35	41.173			

Subject 17 has substantially more errors than would be predicted by the model. Removing this subject from the analysis resulted in a more significant effect due to SAT, but no substantial changes to the conclusions.

(b) Analysis Within Subjects

As before, the analysis within subjects is conducted having adjusted for any between subject variation. The following table summarizes the statistical tests constructed.

Table 6. Tests of Within-Subject Effects: Completion Time

Source	df	F ratio	p-value
Display	2	12.00	0.0001
Period	3	0.33	0.8044
Puzzle	2	5.14	0.0077
Display*Puzzle	4	0.87	0.4866
Display*Gender	4	0.61	0.5453
Puzzle*Gender	2	0.46	0.6333
Carry-Over	2	0.13	0.8774
Within-Subjects Error	91		

From Table 6 we conclude that there is overwhelming evidence that there are differences due to the displays, and due to the puzzles. There is no evidence of an interaction between the display and puzzle, or between display and gender. The following table gives the average values for some of the factors of interest. (These means are found by fitting a model without the non-significant interactions involving gender.)

Table 7. Table of Means

	Mean Number of Errors
Desktop	1.35
HMD	1.67
Real	0.52
Puzzle A	0.79
Puzzle B	1.65
Puzzle C	1.10
Period 1	1.31
Period 2	1.22
Period 3	1.08
Period 4	1.11

Further analysis of contrasts of interest is summarized in the following table.

Table 8. Analysis of Contrasts of Interest

Comparison	Estimate	p-value

Desktop - HMD	-0.32	0.1832
Desktop - Real	0.84	0.0006
HMD - Real	1.15	0.0001
Model A – Model B	-0.86	0.0015
Model A – Model C	-0.31	0.2499
Model B – Model C	0.55	0.0384

This table indicates that both the HMD and desktop have significantly more errors than the physical interface (the estimate is positive). The difference in error rate between HMD and desktop is not significant. There is no difference in the error rates for puzzles A and C, but error rates for puzzle B were significantly higher.

(4) Analysis of the Effect of Frame Rate

From the data collected at the end of this test, there is no evidence of an effect due to frame rates.