One-Way Analysis of Variance (ANOVA)

Although the *t*-test is a useful statistic, it is limited to testing hypotheses about two conditions or levels. The analysis of variance (ANOVA) was developed to allow a researcher to test hypotheses about two or more conditions. Thus, the *t*-test is a proper subset of ANOVA and you would achieve identical results if applying either an ANOVA or a *t*-test to an experiment in which there were two groups. Essentially, then, one could learn to compute ANOVAs and never compute another *t*-test for the rest of one's life.

As with the *t*-test, you are computing a statistic (an *F*-ratio) to test the assertion that the means of the populations of participants given the particular treatments of your experiment will all perform in a similar fashion on the dependent variable. That is, you are testing H_0: $\mu_1 = \mu_2 = \mu_3 = ...$, typically with the hopes that you will be able to reject H_0 to provide evidence that the alternative hypothesis (H_1: Not H_0) is more likely. To test H_0, you take a sample of participants and randomly assign them to the levels of your factor (independent variable). You then look to see if the sample means for your various conditions differ sufficiently that you would be led to reject H_0. As always, we reject H_0 when we obtain a test statistic that is so unlikely that it would occur by chance when H_0 is true only 5% of the time (i.e., $\alpha = .05$, the probability of a Type I error).

The *F*-ratio is a ratio of two variances, called Mean Squares. In the numerator of the *F*-ratio is the $MS_{Treatment}$ and in the denominator is the MS_{Error}. Obviously, your *F*-ratio will become larger as the $MS_{Treatment}$ becomes increasingly larger than the MS_{Error}.

Independent Groups Designs (Between Groups)
In an independent groups (between participants) design you would have two or more levels of a single independent variable (IV) or factor. You also would have a single dependent variable (DV). Each participant would be exposed to only one of the levels of the IV. In an experiment with 3 levels of the IV (generally A_1, A_2, A_3, or, for a more specific example, Drug A, Drug B, Placebo), each condition would produce a mean score. If H_0 is true, the only reason that the group means would differ is that the participants within the groups might differ (that is, there would be Individual Differences and some Random Variability). If H_0 is not true, then there would also be Treatment Effects superimposed on the variability of the group means. Further, within each condition there would be variability among the scores of the participants in that condition, all of whom had been treated alike. Thus, the only reason that people within a group would differ is due to Individual Differences and Random Variability. Conceptually, then, the *F*-ratio would be:

$$F = \frac{\text{Treatment Effects} + [\text{Individual Differences} + \text{Random Effects}]}{\text{Individual Differences} + \text{Random Effects}}$$

In computing the numerator of the *F*-ratio ($MS_{Treatment}$ or $MS_{Between}$), one is actually calculating the variance of the group means and then multiplying that variance by the sample size (*n*). As should be obvious, the group means should differ the most when there are Treatment Effects present in the experiment. On the other hand, if the Individual Differences and/or Random Effects are quite large within the population, then you would expect that the $MS_{Treatment}$

would also be large. So how can you tell if the numerator values are due to the presence of Treatment Effects or if there are just large Individual Differences and Random Effects present, causing the group means to differ? What you need is an estimate of how variable the scores might be in the population when no Treatment Effects are present. That's the role of the MS_{Error}.

You compute the denominator of the F-ratio (MS_{Error} or MS_{Within}) by calculating the variance of each group separately and then computing the mean variance of the groups. Because all of the participants within each group are treated alike, averaging over the variances gives us a good estimate of the population variance (σ^2). That is, you have an estimate of the variability in scores that would be present in the absence of Treatment Effects. By creating a ratio of the $MS_{Treatment}$ to the MS_{Error}, you get a sense of the extent to which Treatment Effects are present. In the absence of Treatment Effects (i.e., H_0 is true), the F-ratio will be near 1.0. To the extent that Treatment Effects are larger, you would obtain F-ratios that are increasingly larger than 1.0.

A variance, or Mean Square (MS), is computed by dividing a Sum of Square (SS) by the appropriate degrees of freedom (df). Typically, computation of the F-ratio will be laid out in a Source Table illustrating the component SS and df for both $MS_{Treatment}$ and MS_{Error}, as well as the ultimate F-ratio. If you keep in mind how MSs might be computed, you should be able to determine the appropriate df for each of the MSs. For example, the $MS_{Treatment}$ comes from the variance of the group means, so if there are a group means, that variance would have a-1 df. Similarly, because you are averaging over the variances for each of your groups to get the MS_{Error}, you would have n-1 df for each group and a groups, so your df for MS_{Error} would be $a(n$ - $1)$.

For example, if you have an independent groups ANOVA with 5 levels and 10 participants per condition, you would have a total of 50 participants in the experiment. There would be 4 $df_{Treatment}$ and 45 df_{Error}. If you had 8 levels of your IV and 15 participants per condition, there would be 120 participants in your experiment. There would be 7 $df_{Treatment}$ and 112 df_{Error}.

A Source Table similar to the one produced by SPSS appears below. From looking at the df in the Source Table, you should be able to tell me how many conditions there were in the experiment and how many participants were in each condition.

SOURCE	SS	df	MS	F
Treatment (Between Groups)	100	4	25	5
Error (Within Groups)	475	95	5	
Total	575	99		

You should recognize that this Source Table is from an experiment with 5 conditions and 20 participants per condition. Thus, H_0 would be $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. You would be led to reject H_0 if the F-ratio is so large that it would only be obtained with a probability of .05 or less when H_0 is true. In your stats class, you would have looked up a critical value of F ($F_{Critical}$) using 4 and 95 df, which would be about 2.49. In other words, when H_0 is true, you would obtain F-ratios of 2.49 or larger only about 5% of the time. Using the logic of hypothesis testing, upon obtaining an unlikely F (2.49 or larger), we would be less inclined to believe that H_0 was true and would place

more belief in H_1. What could you then conclude on the basis of the results of this analysis? Note that your conclusion would be very modest: At least one of the 5 group means differed from one of the other group means. You could have gotten this large F-ratio because of a number of different patterns of results, including a result in which each of the 5 means differed from each of the other means. The point is that after the overall analysis (often called an omnibus F test) you are left in an ambiguous situation unless there were only two conditions in your experiment.

To eliminate the ambiguity about the results of your experiment, it is necessary to compute what are called post hoc or subsequent analyses (also called protected t-tests because they always involve two means and because they have a reduced chance of making a Type I error). There are a number of different approaches to post hoc analyses, but we will rely upon Tukey's HSD to conduct post hoc tests. To conduct post hoc analyses you must have the means of each of the conditions and you need to look up a new value called q (the Studentized-range statistic). A table of Studentized-range values is below (and you'll also receive it as a handout).

TABLE B.5   THE STUDENTIZED RANGE STATISTIC (q)*

*The critical values for q corresponding to $\alpha = .05$ (lightface type) and $\alpha = .01$ (boldface type).

df FOR ERROR TERM	k = NUMBER OF TREATMENTS											
	2	3	4	5	6	7	8	9	10	11	12
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32
	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48	10.70
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65	6.79
	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.48
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43
	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55	8.71
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18
	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86	8.03	8.18
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87	5.98
	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49	7.65	7.78
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83
	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36	7.49
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71
	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13	7.25
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51	5.61
	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53
	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46
	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66	6.77
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40
	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55	6.66
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35
	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46	6.56
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21	5.31
	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38	6.48
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27
	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23
	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25	6.34
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20
	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.28
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10
	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02	6.11
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92	5.00
	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85	5.93
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82	4.90
	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60	5.69	5.76
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81
	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45	5.53	5.60
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71
	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.37	5.44
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.28	4.39	4.47	4.55	4.62
	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23	5.29

$$HSD = q\sqrt{\frac{MS_{Error}}{n}}$$

As seen at the bottom of the table, the formula to determine a critical mean difference is as follows:

$$HSD = q\sqrt{\frac{MS_{Error}}{n}}$$

So, for this example, your value of q would be about 3.94 (5 conditions across the top of the table and 95 df_{Error} down the side of the table). Thus, the HSD would be 1.97 for this analysis, so if two means differed by 1.97 or more, they would be considered to be significantly different. You would now have to subtract each mean from every other mean to see if the difference between the means was larger than 1.97. Two means that differed by less than 1.97 would be considered to be equal (i.e., testing H_0: $\mu_1 = \mu_2$, so retaining H_0 is like saying that the two means don't differ). After completing the post hoc analyses you would know exactly which means differed, so you would be in a position to completely interpret the outcome of the experiment.

It's increasingly the case that you would report the effect size of your treatment. There are a number of different measures one might use (with Jacob Cohen the progenitor of many). For our purposes, we'll rely on eta squared (η^2) and (later) partial eta squared. It's an easily computer measure:

$$\eta^2 = \frac{SS_{Treatment}}{SS_{Total}}$$

The measure should make intuitive sense: it's the proportion of total variability explained by (or due to) the treatment variability.

Using SPSS for Single-Factor Independent Groups ANOVA

Suppose that you are conducting an experiment to test the effectiveness of three pain-killers (Drug A, Drug B, and Drug C). You decide to include a control group that receives a Placebo. Thus, there are 4 conditions to this experiment. The DV is the amount of time (in seconds) that a participant can withstand a painfully hot stimulus. The data from your 12 participants ($n = 3$) are as follows:

| Placebo	Drug A	Drug B	Drug C
0	0	3	8
0	1	4	5
3	2	5	5

Input data as shown below left. Note that one column is used to indicate the level of the factor to which that particular participant was exposed and the other column contains that participant's score on the dependent variable. I had to label the type of treatment with a number, which is essential for *One-Way ANOVA* in SPSS. However, I also used value labels (entered in the *Variable View* window), e.g. 1 = Placebo, 2 = DrugA, 3 = DrugB, and 4 = DrugC.
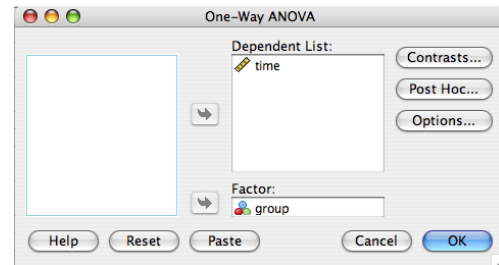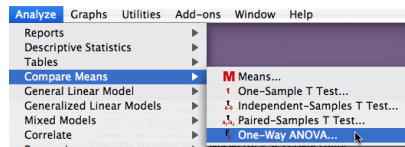
When the data are all entered, you have a couple of options for analysis. The simplest approach is to choose *One-Way ANOVA…* from *Compare Means* (above middle). Doing so will produce the window seen above right. Note that I've moved *time* to the *Dependent List* and *group* to the *Factor* window.

You should also note the three buttons to the right in the window (*Contrasts…*, *Post Hoc…*, *Options…*). We'll ignore the *Contrasts…* button for now. Clicking on the *Options…* button will produce the window below left. Note that I've chosen a couple of the options (Descriptive Statistics and Homogeneity of Variance Test). Clicking on the *Post Hoc…* button will produce the window seen below right. Note that I've checked the Tukey box.
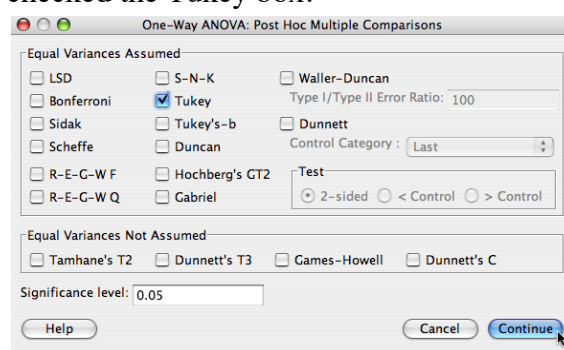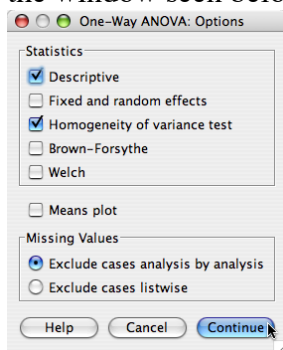
Okay, now you can click on the *OK* button, which will produce the output below. Let's look at each part of the output separately. The first part contains the descriptive statistics.

Descriptives

time

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
placebo	3	1.000	1.7321	1.0000	-3.303	5.303	.0	3.0
drug a	3	1.000	1.0000	.5774	-1.484	3.484	.0	2.0
drug b	3	4.000	1.0000	.5774	1.516	6.484	3.0	5.0
drug c	3	6.000	1.7321	1.0000	1.697	10.303	5.0	8.0
Total	12	3.000	2.5226	.7282	1.397	4.603	.0	8.0

The next part is the Levene test for homogeneity of variance. You should recall that homogeneity of variance is one assumption of ANOVA. This is the rare instance in which you don't want the results to be significant (because that would indicate heterogeneity of variance in your data). In this case it appears that you have no cause for concern. Were the Levene statistic significant, one antidote would be to adopt a more conservative alpha-level (e.g., .025 instead of .05).

Test of Homogeneity of Variances

time

Levene Statistic	df1	df2	Sig.
1.333	3	8	.330

The next part is the actual source table, as seen below. Each of the components of the source table should make sense to you.

ANOVA

time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	54.000	3	18.000	9.000	.006
Within Groups	16.000	8	2.000		
Total	70.000	11			

Of course, at this point, all that you know is that some difference exists among the four groups, because with an $F = 9.0$ and $p = .006$ (i.e., $< .05$), you can reject H_0: $\mu_{\text{drug a}} = \mu_{\text{drug b}} = \mu_{\text{drug c}} = \mu_{\text{placebo}}$.

The next part of the output helps to resolve ambiguities by showing the results of the Tukey's HSD. Asterisks indicate significant effects (with *Sig.* values $< .05$).

Multiple Comparisons

time
Tukey HSD

(I) group	(J) group	Mean Difference (I–J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
placebo	drug a	.0000	1.1547	1.000	-3.698	3.698
	drug b	-3.0000	1.1547	.117	-6.698	.698
	drug c	-5.0000*	1.1547	.011	-8.698	-1.302
drug a	placebo	.0000	1.1547	1.000	-3.698	3.698
	drug b	-3.0000	1.1547	.117	-6.698	.698
	drug c	-5.0000*	1.1547	.011	-8.698	-1.302
drug b	placebo	3.0000	1.1547	.117	-.698	6.698
	drug a	3.0000	1.1547	.117	-.698	6.698
	drug c	-2.0000	1.1547	.369	-5.698	1.698
drug c	placebo	5.0000*	1.1547	.011	1.302	8.698
	drug a	5.0000*	1.1547	.011	1.302	8.698
	drug b	2.0000	1.1547	.369	-1.698	5.698

*. The mean difference is significant at the 0.05 level.

As you can see, Drug C leads to greater tolerance than Drug A and the Placebo, with no other differences significant. If you didn't have a computer available, of course, you could still compute the Tukey-Kramer post hoc tests using the critical mean difference formula, as seen below. Of course, you'd arrive at the identical conclusions.

$$HSD = q\sqrt{\frac{MS_{Error}}{n}} = 4.53\sqrt{\frac{2.0}{3}} = 3.7$$

The *One-Way ANOVA* module has a lot of nifty features. However, it doesn't readily compute a measure of effect size. For that purpose, you can compute the one-way ANOVA using *General Linear Model->Univariate…* as seen below left. You'll next see the window in the middle below. Note that I've moved the two variables from the left window into the appropriate windows on the right (*time* to the *Dependent Variable* window and *group* to the *Fixed Factor(s)* window). Clicking on the *Options…* button yields the window on the right below. Note that I've checked the boxes for *Descriptive statistics*, *Estimates of effect size*, *Observed power*, and *Homogeneity tests*.
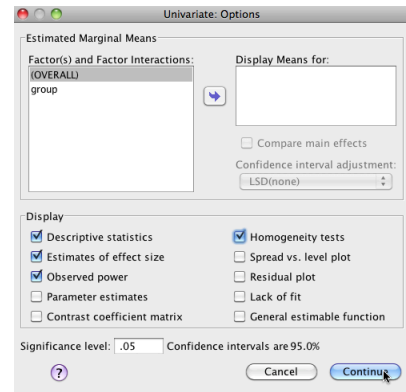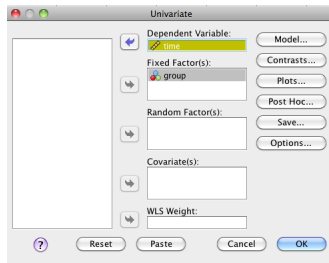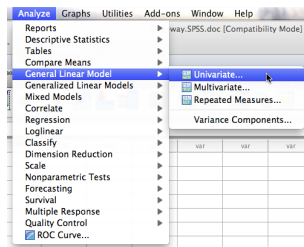
I could also choose Post Hoc tests, etc., as in *One-Way ANOVA*. However, just to illustrate the differences between the two approaches in SPSS, I'll focus on the differences. In the output below (left), note that the descriptive statistics are more basic than those found in the *One-Way ANOVA* output. The Levene test would be identical to that in the *One-Way ANOVA* output. However, the source table is quite a bit different. For our purposes, we can ignore the following lines: Corrected Model, Intercept, and Total. The group, Error, and Corrected Total lines are identical to the *One-Way ANOVA* up to the Sig. column. After that, however, you'll find Partial Eta Squared (which is just Eta Squared here), the Noncentrality Parameter (which you can ignore), and the Observed Power.

Descriptive Statistics

Dependent Variable:time

group	Mean	Std. Deviation	N
placebo	1.000	1.7321	3
drug a	1.000	1.0000	3
drug b	4.000	1.0000	3
drug c	6.000	1.7321	3
Total	3.000	2.5226	12

Tests of Between-Subjects Effects

Dependent Variable:time

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	54.000[a]	3	18.000	9.000	.006	.771	27.000	.936
Intercept	108.000	1	108.000	54.000	.000	.871	54.000	1.000
group	54.000	3	18.000	9.000	.006	.771	27.000	.936
Error	16.000	8	2.000					
Total	178.000	12						
Corrected Total	70.000	11						

a. R Squared = .771 (Adjusted R Squared = .686)
b. Computed using alpha =

You might report your results as follows:

There was a significant effect of Drug on the duration that people could endure a hot stimulus, $F(3, 8) = 9.0$, $MSE = 2.0$, $p = .006$, $\eta^2 = .771$. Post hoc tests using Tukey's HSD indicated that Drug C leads to greater tolerance ($M = 6.0$) than Drug A ($M = 1.0$) and the Placebo ($M = 1.0$), with no other differences significant.

Sample Problems

Darley & Latané (1968)

Drs. John Darley and Bibb Latané performed an experiment on bystander intervention ("Bystander Intervention in Emergencies: Diffusion of Responsibility"), driven in part by the Kitty Genovese situation. Participants thought they were in a communication study in which anonymity would be preserved by having each participant in a separate booth, with communication occurring over a telephone system. Thirty-nine participants were randomly assigned to one of three groups ($n = 13$): Group Size 2 (S & victim), Group Size 3 (S, victim, & 1 other), or Group Size 6 (S, victim, & 4 others). After a short period of time, one of the "participants" (the victim) made noises that were consistent with a seizure of some sort ("I-er-um-I think I-I need-er-if-if could-er-er-somebody er-er-er-er-er-er-er give me a little-er give me a little help here because-er-I-er-I'm-er-er-h-h-having a-a-a real problem..." — you get the picture). There were several DVs considered, but let's focus on the mean time in seconds to respond to the "victim's" cries for help. Below are the mean response latencies:

	Group Size 2	Group Size 3	Group Size 6
Mean Latency	52	93	166
Variance of Latency	3418	5418	7418

Given the information above:

1. state the null and alternative hypotheses

2. complete the source table below

3. analyze and interpret the results as completely as possible

4. indicate the sort of error that you might be making, what its probability is, and how serious you think it might be to make such an error

SOURCE	SS	df	MS	F
Group Size (Treatment)				8
Error				
Total				

It's not widely known, but another pair of investigators was also interested in bystander apathy—Dudley & Lamé. They did a study in which the participant thought that there were 5, 10, or 15 bystanders present in a communication study. The participant went into a booth as did the appropriate number of other "participants." (Each person in a different booth, of course.) The researchers measured the time (in minutes) before the participant would go to the aid of a fellow participant who appeared to be choking uncontrollably. First, interpret the results of this study as completely as you can. Can you detect any problems with the design of this study? What does it appear to be lacking? How would you redesign the study to correct the problem(s)? Be very explicit.

Descriptives

TIME

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
5	10	2.80	.632	.200	2.35	3.25	2	4
10	10	3.40	.966	.306	2.71	4.09	2	5
15	10	3.40	.843	.267	2.80	4.00	2	5
Total	30	3.20	.847	.155	2.88	3.52	2	5

Test of Homogeneity of Variances

TIME

Levene Statistic	df1	df2	Sig.
1.391	2	27	.266

One-Way ANOVA output

ANOVA

TIME

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.400	2	1.200	1.761	.191
Within Groups	18.400	27	.681		
Total	20.800	29			

General Linear Model output

Tests of Between-Subjects Effects

Dependent Variable:time

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	2.400[a]	2	1.200	1.761	.191	.115	3.522	.336
Intercept	307.200	1	307.200	450.783	.000	.943	450.783	1.000
bystand	2.400	2	1.200	1.761	.191	.115	3.522	.336
Error	18.400	27	.681					
Total	328.000	30						
Corrected Total	20.800	29						

a. R Squared = .115 (Adjusted R Squared = .050)
b. Computed using alpha =

Festinger & Carlsmith (1959)

   Drs. Leon Festinger and James M. Carlsmith performed an experiment in cognitive dissonance ("Cognitive Consequences of Forced Compliance"). In it, they had three groups: (1) those in the *Control* condition performed a boring experiment and then were later asked several questions, including a rating of how enjoyable the task was; (2) those in the *One Dollar* condition performed the same boring experiment, and were asked to lie to the "next participant" (actually a confederate of the experimenter), and tell her that the experiment was "interesting, enjoyable, and lots of fun," for which they were paid \$1, and then asked to rate how enjoyable the task was; and (3) those in the *Twenty Dollar* condition, who were treated exactly like those in the One Dollar condition, except that they were paid \$20.

   Cognitive dissonance theory suggests that the *One Dollar* participants will rate the task as more enjoyable, because they are being asked to lie for only \$1 (a modest sum, even in the 1950's), and to minimize the dissonance created by lying, will change their opinion of the task. The *Twenty Dollar* participants, on the other hand, will have no reason to modify their interpretation of the task, because they would feel that \$20 was adequate compensation for the little "white lie" they were asked to tell. The *Control* condition showed how participants would rate the task if they were not asked to lie at all, and were paid no money. Thus, this was a simple independent random groups experiment with three groups.

   Thus, $H_0: \mu_{\$1} = \mu_{\$20} = \mu_{Control}$ and H_1: Not H_0

   The mean task ratings were -.45, +1.35, and -.05 for the *Control*, *One Dollar*, and *Twenty Dollar* conditions, respectively.

   Below is a partially completed source table consistent with the results of their experiment. Complete the source table, analyze the results as completely as possible and tell me what interpretation you would make of the results. You should also be able to tell me how many participants were in the experiment overall.

SOURCE	SS	df	MS	F
Treatment				9
Error		57	2	
Total				

Dr. Jane Picasso is a clinical psychologist interested in autism. She is convinced that autistic children will respond better to treatment if they are in familiar settings. To that end, she randomly assigns 50 autistic children to treatment under 5 different conditions (Home, Familiar Office, Familiar Clinic, New Unfamiliar Clinic, and New Unfamiliar Office for every treatment). The children undergo treatment twice a week for one month under these conditions and are then tested on their progress using a 10-point rating scale (1 = little progress and 10 = lots of progress). Below are incomplete analyses from her study. Complete the analyses and interpret the data as completely as you can.

Descriptives

PROGRESS

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
home	10	6.20	2.394	.757	4.49	7.91	2	9
familiar office	10	4.10	1.370	.433	3.12	5.08	2	6
familiar clinic	10	3.30	1.160	.367	2.47	4.13	1	5
new clinic	10	2.80	.789	.249	2.24	3.36	2	4
new office	10	1.40	.516	.163	1.03	1.77	1	2
Total	50	3.56	2.091	.296	2.97	4.15	1	9

Test of Homogeneity of Variances

progress

Levene Statistic	df1	df2	Sig.
6.534	4	45	.000

ANOVA

PROGRESS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups				15.96	.000
Within Groups			1.9		
Total					

Professor Isabel C. Ewe has taught for several years and, based on her experience, believes that students who tend to sit toward the front of the classroom perform better on exams than students who tend to sit toward the back of the classroom. To test her hypothesis, she collects data from several of her classes, recording the student's final grade (A = 4, B = 3, etc.) and the typical row in which the student sat throughout the term (Row 1 = Front and Row 8 = Back). Thus, she conceives of her study as a single-factor independent groups design with two levels of the IV (Front and Back). She decides to analyze her data on SPSS using ANOVA. Her output is shown below. Interpret the results of her study as completely as you can.

Descriptives

gpa

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
Front	20	3.4400	.27985	.06258	3.3090	3.5710	3.00	3.80
Back	20	3.0900	.31772	.07104	2.9413	3.2387	2.50	3.90
Total	40	3.2650	.34459	.05448	3.1548	3.3752	2.50	3.90

Test of Homogeneity of Variances

gpa

Levene Statistic	df1	df2	Sig.
.059	1	38	.810

GPA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.225	1	1.225	13.667	.001
Within Groups	3.406	38	.090		
Total	4.631	39			

Betz and Thomas (1979) have reported a distinct connection between personality and health. They identified three personality types who differ in their susceptibility to serious stress-related illness (heart attack, high blood pressure, etc.). The three personality types are *alphas* (who are cautious and steady), *betas* (who are carefree and outgoing) and *gammas* (who tend toward extremes of behavior, such as being overly cautious or very careless). For the analysis below, lower scores indicate poorer health. Complete the analysis and interpret the results.

Descriptives

Health

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
alpha	10	44.2000	9.80703	3.10125	37.1845	51.2155	21.00	56.00
beta	10	47.2000	7.37564	2.33238	41.9238	52.4762	36.00	57.00
gamma	10	36.8000	6.01480	1.90205	32.4973	41.1027	25.00	45.00
Total	30	42.7333	8.81587	1.60955	39.4414	46.0252	21.00	57.00

Test of Homogeneity of Variances

Health

Levene Statistic	df1	df2	Sig.
.627	2	27	.542

ANOVA

Health

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	573.0				.019
Within Groups	1680.8				
Total	2253.8				

Because old exams contain questions in which StatView was used for analysis, the last two examples use StatView. You'll note that the major difference is that the order of *df* and *SS* in the source table is reversed, relative to SPSS. Otherwise the differences are negligible.

In Chapter 8 of your textbook, Ray describes the five conditions found in the Bransford & Johnson (1972) memory study (first mentioned in Chapter 6). These conditions are: Complete-context picture before passage; Complete-context picture after passage; Partial-context picture before passage; Hearing passage once without picture; and Hearing passage twice without picture. Below is the Complete-context picture (as a recall cue for you).
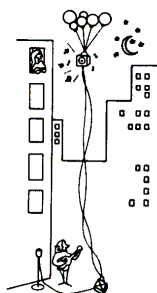
Suppose that the results for this study had turned out as seen below. Complete the source table and then analyze the study as completely as you can.

ANOVA Table for Recall

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Condition		301			<.0001	453.375	1.000
Residual		46					

Means Table for Recall
Effect: Condition

	Count	Mean	Std. Dev.	Std. Err.
Complete-after	15	2.933	.704	.182
Complete-before	15	7.667	1.113	.287
No pic – Hear twice	15	2.400	.737	.190
No pic-Hear once	15	2.267	.704	.182
Partial-before	15	4.133	.743	.192

A study has been replicated several times over several decades. In spite of presumed changes in awareness of Sexually Transmitted Diseases, the results come out in a remarkably similar fashion. In one condition of the study, a young woman goes up to a male college student and says, "I've been noticing you around campus for a few weeks and I'd really like to have sex with you. Would it be possible to make a date for tonight to have sex?" [The prototypical male response is, "Why wait until tonight?"] In the other condition of the study, a young man goes up to a female college student and says the same thing ("I've been noticing you…"). [The prototypical female response is not particularly encouraging. ☺] In no case are the people acquainted, so these requests come from perfect strangers.

Dr. Randi Mann is interested in conducting a variant of this study. She thinks that the nature of the request may be what's producing the extraordinarily different results. Thus, she decides to replicate the study, but have the request changed from sexual intercourse to accompanying the stranger to a concert. Thus, (on a night that the Dave Matthews Band was in town) the request might be, "My friend was supposed to join me at the Dave Matthews concert tonight, but she won't be able to come. She's already paid for the ticket, so if you came with me to the concert it wouldn't cost you anything. Would you join me?" A hidden camera records the interactions and three raters independently judge the person's response on a scale of 1 (Definitely won't go) to 5 (Definitely will go). The score used for each interaction is the mean of the three rater's responses. The IV is the gender of the person making the request (so Female means that a female asked a male to go to the concert).

Complete the source table below and interpret the results as completely as you can. What might be some limitations of this study?

ANOVA Table for Response

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Gender			42.0	47.3	<.0001	47.317	1.000
Residual							

Means Table for Response
Effect: Gender

|	Count	Mean	Std. Dev.	Std. Err.
Female	20	3.850	.988	.221
Male	20	1.800	.894	.200

Dependent Groups Design (Repeated Measures, Within Groups)

As was the case for *t*-tests, there is an ANOVA appropriate for both independent groups and dependent groups (repeated measures) designs. In choosing to make use of a repeated measures design, you have the advantage of increased power and efficiency over an independent groups design, coupled with the necessity to counterbalance the order in which participants are exposed to the treatments (and the possibility that you simply cannot make use of the repeated measures design in some cases of permanent changes).

Unlike the independent groups design, the dependent groups design uses the same participant in all the conditions, which means that there are no Individual Differences contributing to the differences observed among the treatment means. Conceptually, then, the *F*-ratio for a repeated measures design would be:

$$F = \frac{\text{Treatment Effects} + \text{Random Effects}}{\text{Random Effects}}$$

One way to think about what the ANOVA for repeated measures designs is doing is to imagine that the variability due to Individual Differences is being estimated by looking at the variability among participants and then removing it from the error term that would have been used in an independent groups design. First of all, think of the total variability in the DV as a piece of pie. As seen below, some of that variability is due to the Treatment Effect. Some of the variability is due to factors other than the treatment, which we collectively call error. As seen on the left below, the Error term for an Independent Groups ANOVA contains both Random Error and Individual Differences. For the Dependent Groups ANOVA (on the right below) the Error term is reduced because the Individual Differences have been removed. This is the source of the power found in repeated measures designs.
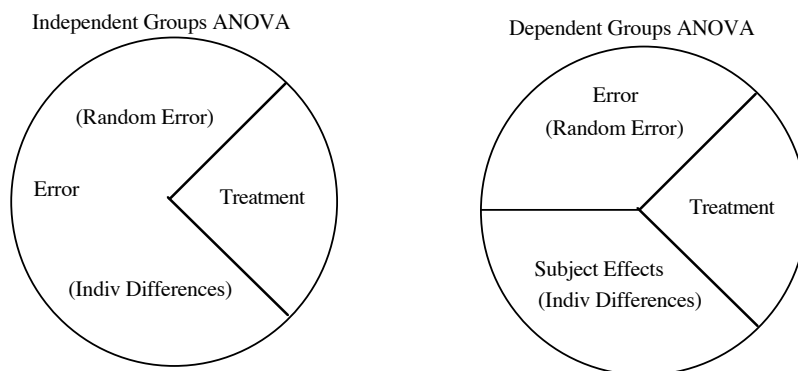
So, how does one go about removing the effects of differences among participants? Essentially, one treats participants as a factor in the analysis. That is, there will be a mean score for each participant and one computes the variance of the participants' mean responses to provide an estimate of the participant differences. Just as would be the case for any factor, when there is little variability among the participant means, you will have a problem (just like having no treatment effect). In this case, what it implies is that the Error Term will not be reduced by much as a result of removing the participant effects.

Suppose that for the ANOVA shown previously the data had actually come from a repeated measures design. That is, the exact same 100 scores were obtained, but instead of coming from 100 different participants, the scores came from only 20 different participants (note the efficiency of the repeated measures design... same amount of data from one-fifth the participants). The Source Table might look something like that shown below.

Source	SS	df	MS	F
Treatment	100	4	25	12.5
Error (Within)	475	95		
Subjects	323	19		
Error (Residual)	152	76	2	
Total	575	99		

Note the increase in power as the *F* jumps up to 12.5. What led to the increase in the *F*-ratio? As you can see in the Source Table, the original SS_{Error} was 475, but 323 of that was due to the Individual Differences found among the participants. So when that variability is removed, the new Error Term has a substantially reduced *SS*, leading to a new MS_{Error} of 2 (instead of 5).

There are a couple of other aspects of the repeated measures ANOVA Source Table to attend to. Note, for instance, that the $SS_{Treatment}$, $df_{Treatment}$, and $MS_{Treatment}$ do not change as a result of moving to a repeated measures design. If you understand what goes into those terms, it should be clear to you why they do not change. That is, the values of the group means would not change, nor does the number of conditions, so none of these terms should change. Note, further, that the new df_{Error} is smaller than those present in the original df_{Error}. In general, one would want to have the largest df_{Error} possible (can you say why?). In this instance, having fewer *df* in the Error Term doesn't seem to have hurt us. There are, however, some instances in which the loss of *df will* be disadvantageous. What would have happened to our *F*-ratio in this analysis if the $SS_{Subjects}$ had been 19 instead of 323? What would the new SS_{Error} become? What about the new MS_{Error}? The new *F*-ratio? Can you see how important it is that there is variability among our participants when we're using a repeated measures design?

Once you have your *F*-ratio for the repeated measures design, the rest of the analysis is identical to that for an independent groups design. That is, in this case you could reject H_0, but that doesn't help you to explain the outcome of the experiment because there are so many ways in which one might have rejected H_0. So once again it becomes imperative that we conduct subsequent analyses to determine the exact source of the differences.

The computation of Tukey's HSD would be straightforward, assuming that it was clear for the independent groups design. Here we would look up *q* with 5 conditions and 76 *df* and we would use 2 as the MS_{Error}. Thus *q* is about 3.96 and HSD = 1.25. So if two means differed by 1.25 or more, they would be significantly different. Two means that differed by less than 1.25 would be considered statistically equivalent.

SPSS for Repeated Measures ANOVA

Remember that in entering your data, you should consider a row to represent information about a person. Thus, you would enter as many columns (variables) as you have levels of your independent variable (with each person providing a score for each level). In this case, let's assume that we're looking at quiz scores at four different times (Time 1, Time 2, etc.) for five different people. The data would be entered as seen below left.

	time1	time2	time3	time4
1	1	2	2	5
2	2	2	3	4
3	1	3	3	4
4	3	4	4	6
5	4	5	4	8
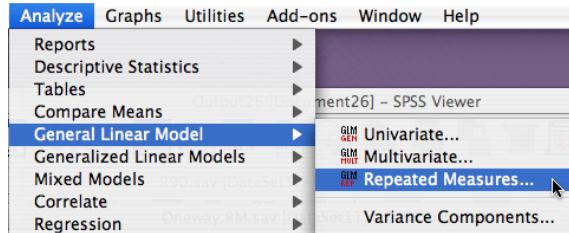
You can no longer use the **One-Way ANOVA** procedure. Instead, you must now use the **General Linear Model->Repeated Measures**, as seen above right. Doing so will produce the window seen below left. You should choose an appropriate name for your Within-Subject Factor Name (rather than the nondescript *factor1* that's the default). In this case, I've chosen time as the variable name and indicated that it has 4 levels. After clicking on the *Add* button, the variable will appear in the upper window, as seen below middle. Clicking on the *Define* button will open the next window (below right).
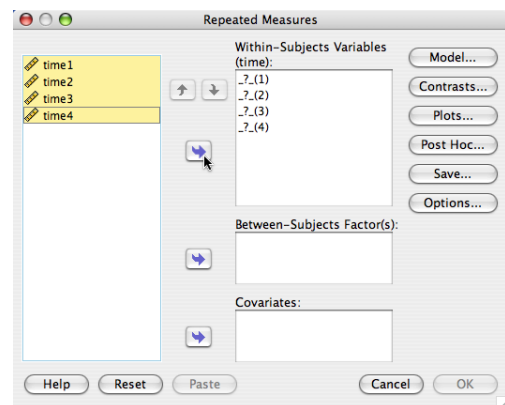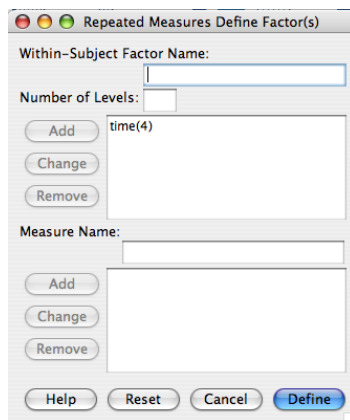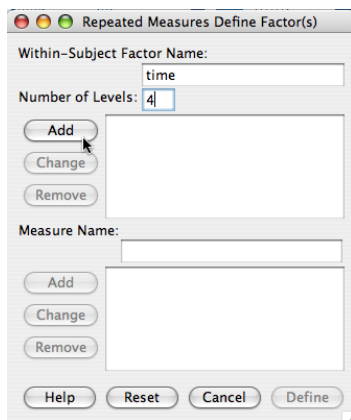
The next step is to tell SPSS which variables represent the levels of your independent variable. In this case the variables representing responses at each of the four levels of time are contiguous, so I've highlighted all four and will click on the arrow button to move them to the Within-Subjects Variables window on the right.

You will want to click on the Options… button before proceeding to the analysis. Doing so will open the window seen below, where you'll notice that I've chosen several options by checking the appropriate boxes. Next click on the *Continue* button and then on the *OK* button in the original window. Doing so will produce the appropriate analyses. For instance, below right are the descriptive statistics.
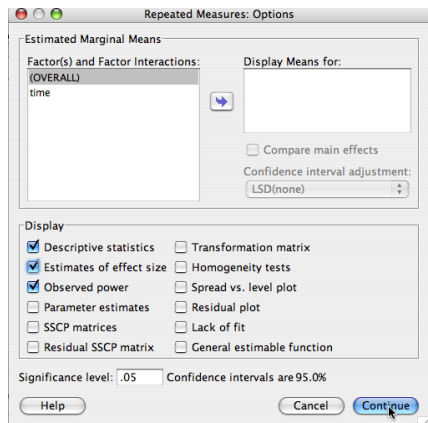
Descriptive Statistics

	Mean	Std. Deviation	N
time1	2.20	1.304	5
time2	3.20	1.304	5
time3	3.20	.837	5
time4	5.40	1.673	5

You'll also see a number of tables of information that you may safely ignore for this course (i.e., Multivariate Tests, Mauchly's Test of Sphericity, Tests of Within-Subjects Contrasts, Tests of Between-Subjects Effects). Instead, you should focus on the Tests of Within-Subjects Effects seen below. You will immediately note that it's a fairly complicated looking table. ☺

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
time	Sphericity Assumed	27.400	3	9.133	23.826	.000	.856	71.478	1.000
	Greenhouse–Geisser	27.400	1.658	16.523	23.826	.001	.856	39.511	.997
	Huynh–Feldt	27.400	2.687	10.198	23.826	.000	.856	64.015	1.000
	Lower-bound	27.400	1.000	27.400	23.826	.008	.856	23.826	.947
Error(time)	Sphericity Assumed	4.600	12	.383					
	Greenhouse–Geisser	4.600	6.633	.693					
	Huynh–Feldt	4.600	10.747	.428					
	Lower-bound	4.600	4.000	1.150					

a. Computed using alpha = .05

First, focus solely on the first lines of each box (Sphericity Assumed). The *SS*, *df*, and *MS* for these lines are crucial. The Treatment line will be labeled by the name of your Within-Subjects Variable (*time* in this case). The Error line will be appropriately labeled, but as always is the interaction of Treatment and Subject.

OK, so what's this sphericity stuff? For the repeated measures ANOVA a rough equivalent of the homogeneity of variance assumption for an independent groups analysis is called the sphericity assumption. Violations of that assumption will have a similar impact—a deceptively smaller *p*-value than is warranted. Although you might take an approach similar to that for independent groups analyses (using a more conservative alpha-level), there are actually statistical tests to correct for deviations from sphericity. The one we will use is the Greenhouse-Geisser correction. Essentially, we will simply replace the p-value (Sig.) for the Sphericity Assumed line with that from the Greenhouse-Geisser line. For all its complexity, then, with a bit of condensation the above table becomes:

Source	SS	df	MS	F	Sig.
Treatment (time)	27.4	3	9.133	23.826	.001
Error (time x subject)	4.6	12	.383		

As you might intuit, because the difference in Significance level isn't all that great (.000 vs. .001), in these made-up data there is not much violation of the sphericity assumption. The greater the violation, the greater the discrepancy between the significance levels.

As with the independent groups ANOVA, your significant main effect for the IV ($F = 23.826, p = .001$) is ambiguous when there are more than two levels of your IV. In this case, you would have rejected H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. In order to determine exactly which of your sample means are indicative of differences in the populations from which they were drawn, you need to compute Tukey's HSD. With four treatments and $df_{Error} = 12$, HSD would be:

$$HSD = q\sqrt{\frac{MS_{Error}}{n}} = 4.2\sqrt{\frac{.383}{5}} = 1.16$$

Thus, it appears that scores were better at Time 4 compared to Time 3, Time 2, and Time 1, but none of those means differed.

Referring briefly to the SPSS source table, you should notice the columns labeled Partial Eta Squared and Power. Partial Eta Squared is (of course) similar to Eta Squared (SS_{Effect} / SS_{Total}), so larger is better. However, instead of SS_{Total} in the denominator, Partial Eta Squared uses $SS_{Total} - SS_{Subjects}$ (which is also $SS_{Treatment} + SS_{Error}$). Power should make sense to you, and you'd like it to be greater than .80.

Here's how might you report these results:

There was a significant effect of time of quiz on scores on the quiz, $F(3,12) = 23.826$, $MSE = .383, p = .001, \eta^2 = .856$. Subsequent analyses using Tukey's HSD indicated that scores on the fourth quiz were significantly higher ($M = 5.4$) than those on the third ($M = 3.2$), second ($M = 3.2$), or first quiz ($M = 2.2$), none of which differed significantly.

Sample Problems

Complete the source tables for the following ANOVAs.
1.

Source	SS	df	MS	F
Subjects	49		7	
Treatment		4		9
Error (T x S)	28			
Total				

How many participants were in this study?
How many levels of the treatment were there?

2. This is a repeated measures experiment with 6 levels of treatment and $n = 25$.

Source	SS	df	MS	F
Subjects	144			
Treatment	50			
Error (T x S)				
Total	314			

In this experiment, suppose that $SS_{Subjects} = 24$. How would that change the outcome of the analysis?

A habituation task is often used to test the memory of infants. In the habituation procedure, a stimulus is shown for a brief period, and the researcher records how much time the infant spends looking at the stimulus. This same process is repeated again and again. If the infant begins to lose interest in the stimulus (decreases the time looking at it), the researcher can conclude that the infant "remembers" the earlier presentations and is demonstrating habituation to (boredom with?) an old familiar stimulus. Hypothetical data from a habituation experiment were analyzed to produce the analyses seen below. Infants were exposed to stimuli of Low, Medium, and High Complexity. In the course of the experimental session, infants saw each stimulus 6 times for 1 minute each, with the order of presentation varied in a counterbalanced fashion (e.g., LMH, LHM, MLH, etc.). The dependent variable of interest here is the time spent looking at each type of stimulus on the sixth (and last) presentation of each stimulus. Complete the source table and then analyze and interpret the data as completely as you can.

Descriptive Statistics

	Mean	Std. Deviation	N
Low	.74400	.116065	10
Medium	.89000	.073485	10
High	.96600	.048120	10

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
complex	Sphericity Assumed	.255				.000	.864	113.956	1.000
	Greenhouse-Geisser						.864	67.734	1.000
	Huynh-Feldt						.864	72.125	1.000
	Lower-bound						.864	56.978	1.000
Error(complex)	Sphericity Assumed	.040							
	Greenhouse-Geisser								
	Huynh-Feldt								
	Lower-bound								

a. Computed using alpha = .05

Hess has done work that suggests that pupil size is affected by emotional arousal. In order to ascertain if the type of arousal makes a difference, Dr. Bo Ring decides to measure pupil size under three different arousal conditions (Neutral, Pleasant, Aversive). Participants look at pictures that differ according to the condition (neutral = plain brick building, pleasant = young man and woman sharing an ice cream cone, aversive = graphic photograph of an automobile accident). The pupil size after viewing each photograph is measured in millimeters. The data are seen below:

Neutral	Pleasant	Aversive
4	8	3
3	7	7
2	7	3
3	8	4
3	6	7
5	7	4
4	8	4
3	7	6
5	6	5
2	7	6

a. Due to miscommunication between Dr. Ring and her research assistant Igor, she has no idea if the study was run as an independent groups design or as a repeated measures design. Igor has left the lab in a cloud of controversy amidst a whole range of allegations (involving improper use of lab alcohol, monkeys, and videotapes), but that's another story entirely. With no immediate hope of clarifying the way in which the data were collected, Dr. Ring decides that the most reasonable strategy would be to analyze the data as an independent groups design. Her logic should make sense to you, so explain to me why she would be smart to analyze the data as though they were collected in an independent groups design.

b. OK, Dr. Ring analyzes the data using SPSS, but she's a notorious slob and has dropped donut filling over portions of the source table. But she's not just a slob, she's a frugal slob. You've just been hired to replace her research assistant (so stay away from the alcohol, monkeys, etc.). She knows how sophisticated you are, so she wants you to figure out the ANOVA source table without running the data through the computer again. Of course, you can easily do so with the information remaining in the source table, so do so now. Then, to truly ingratiate yourself to your new employer, analyze the data as completely as you can.

Descriptives

pupil size

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
neutral	10	3.4000	1.07497	.33993	2.6310	4.1690	2.00	5.00
pleasant	10	7.1000	.73786	.23333	6.5722	7.6278	6.00	8.00
aversive	10	4.9000	1.52388	.48189	3.8099	5.9901	3.00	7.00
Total	30	5.1333	1.90703	.34818	4.4212	5.8454	2.00	8.00

Test of Homogeneity of Variances

pupil size

Levene Statistic	df1	df2	Sig.
4.529	2	27	.020

ANOVA

pupil size

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups			34.633		.000
Within Groups	36.200				
Total					

c. As you noodle around the lab, having completed the analyses that Dr. Ring wanted, you stumble across the original data from the study and realize that Igor had actually run the study as a repeated measures design. This revelation strikes you as a combination of good news and bad news. What's the typical good news about running a repeated measures design? In this particular case, what might strike you as bad news in the design of this repeated measures study?

d. Now that you know that the data were collected as a repeated measures design, you re-compute the analyses appropriately, as seen below. [Notice that the means, etc., aren't displayed below, because they would not change from the earlier analysis, right? Note, also, that other parts of the source table don't change as you move to a repeated measures analysis.] Once again, analyze these data as completely as you can. Given what you know about the relationship between independent groups analyses and repeated measures analyses, the results might surprise you somewhat. However, you can readily explain the anomaly, right?

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
arousal	Sphericity Assumed	69.267	2	34.633	20.284	.000	.693	40.568	1.000
	Greenhouse–Geisser	69.267	1.630	42.495	20.284	.000	.693	33.063	.999
	Huynh–Feldt	69.267	1.940	35.699	20.284	.000	.693	39.357	1.000
	Lower–bound	69.267	1.000	69.267	20.284	.001	.693	20.284	.979
Error(arousal)	Sphericity Assumed	30.733	18	1.707					
	Greenhouse–Geisser	30.733	14.670	2.095					
	Huynh–Feldt	30.733	17.463	1.760					
	Lower–bound	30.733	9.000	3.415					

a. Computed using alpha = .05

Professor Phyllis Tyne was interested in doing some research on visual search. She decides to have her participants search for a target geometric shape (e.g., square or circle) of a particular size presented among distractor geometric figures that are (1) equal in area to the target and similar in shape, (2) equal in area but dissimilar in shape, (3) unequal in area to the target but similar in shape, (4) unequal in area to the target and dissimilar in shape, and finally (5) no distractor stimuli at all (target only). The stimuli are to be presented on a computer display, with the participant pointing to the target on the screen (which the computer immediately detects and records the time in seconds to locate the target). The location of the target and distractors on the screen is determined randomly.

a. What would you conclude is the purpose of the final (control) condition—Target only?

b. As Dr. Tyne's consultant, you are asked to describe for her how this experiment would be conducted as an independent groups design and as a repeated measures design. Briefly describe how the experiment would be conducted under both types of designs, being sure to indicate specifics of design differences between them. Present Dr. Tyne (and me) with the advantages and disadvantages of both designs for her specific research, and then make a recommendation as to which design you would use.

c. Suppose that the study was conducted as an independent groups design and the results were as seen below. Complete the source table and interpret the results as completely as you can.

Descriptives

Time

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
Area =, Shape =	10	1.71000	.412176	.130341	1.41515	2.00485	1.000	2.500
Area =, Shape not =	10	1.40000	.169967	.053748	1.27841	1.52159	1.200	1.700
Area not =, Shape =	10	.90000	.200000	.063246	.75693	1.04307	.600	1.200
Area not =, Shape not =	10	.20000	.105409	.033333	.12459	.27541	.100	.400
Target only	10	.22500	.088976	.028137	.16135	.28865	.100	.400
Total	50	.88700	.651984	.092204	.70171	1.07229	.100	2.500

Test of Homogeneity of Variances

Time

Levene Statistic	df1	df2	Sig.
5.427	4	45	.001

ANOVA

Time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups				89.7	.000
Within Groups			.05		
Total					

d. Suppose that the study had been conducted as a repeated measures design. (Note that the data are from different participants than those in the independent groups experiment!) Complete the source table below and interpret the results of the study as completely as you can. [10 pts]

Descriptive Statistics

	Mean	Std. Deviation	N
Area =, Shape =	1.71000	.166333	10
Area =, Shape Not =	1.59500	.130064	10
Area Not =, Shape =	.81000	.179196	10
Area Not =, Shape No…	.25000	.117851	10
Target Only	.18000	.125167	10

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
distract	Sphericity Assumed	20.877				.000	.985	2364.944	1.000
	Greenhouse–Geisser					.000	.985	1256.141	1.000
	Huynh–Feldt					.000	.985	1655.022	1.000
	Lower–bound					.000	.985	591.236	1.000
Error(distract)	Sphericity Assumed			.009					
	Greenhouse–Geisser								
	Huynh–Feldt								
	Lower–bound								

a. Computed using alpha = .05

Dr. Ty Pest is a human factors psychologist who is interested in testing the impact of different keyboard layouts on typing speed. To that end, he chooses four different conditions: a normal keyboard with normal (QWERTY) keyboard layout, a normal keyboard with the Dvorak layout of keys (supposedly a more efficient layout), an ergonomic (split) keyboard with normal keyboard layout, and an ergonomic (split) keyboard with the Dvorak layout of keys. He chooses as his participants secretaries who have at least 5 years of typing experience. The DV is the number of words per minute of typing of complex textual material in 15 minutes.

Dr. Pest should use counterbalancing in his experiment, of course. What kind would he use and how many different orders would that generate?

Below are the incomplete analyses of his experiment. Complete the analyses and interpret the results as completely as you can.

Descriptive Statistics

	Mean	Std. Deviation	N
normal-QWERTY	74.0000	11.73788	10
normal-Dvorak	43.0000	12.73665	10
ergonomic-QWERTY	70.5000	13.00641	10
ergonomic-Dvorak	35.0000	12.69296	10

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
keyboard	Sphericity Assumed	11436.875				.000	.835	136.824	1.000
	Greenhouse-Geisser					.000	.835	77.736	1.000
	Huynh-Feldt					.000	.835	94.050	1.000
	Lower-bound					.000	.835	45.608	1.000
Error(keyboard)	Sphericity Assumed	2256.875							
	Greenhouse-Geisser								
	Huynh-Feldt								
	Lower-bound								

a. Computed using alpha = .05

Once again, here is an example using StatView. You should familiarize yourself with the differences in the output between SPSS and StatView so that you can make sense of the StatView problems you encounter.

It is at least part of the folklore that repeated experience with the Graduate Record Examination (GRE) leads to better scores, even without any intervening study. We obtain ten participants and give them the GRE verbal exam every Saturday morning for three weeks. The data are as follows:

Participant	1st	2nd	3rd
1	550	575	580
2	440	440	470
3	610	630	610
4	650	670	670
5	400	460	450
6	700	680	710
7	490	510	515
8	580	550	590
9	550	590	600
10	600	610	630

Complete the source table below and interpret the data as completely as you can. Note that the nature of this design does not lend itself to counterbalancing of the time of the GRE, because the time is critical (first must be first, etc.). [For extra credit, can you think of how the design would likely involve the counterbalancing of some aspect of the study?] As you can see in the source table, the $SS_{Subject}$ is quite large. Using the data above to make your point, tell me where the $SS_{Subject}$ comes from. What kind of error might you be making in your decision about H_0 and what is its probability?

ANOVA Table for Week

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject		195980.000					
Category for Week		3271.667			.0046	14.741	.904
Category for Week * Subject		3995.000					

Means Table for Week
Effect: Category for Week

	Count	Mean	Std. Dev.	Std. Err.
Week 1	10	557.000	92.622	29.290
Week 2	10	571.500	82.126	25.971
Week 3	10	582.500	83.041	26.260