

Two-Way Analysis of Variance (ANOVA)

An understanding of the one-way ANOVA is crucial to understanding the two-way ANOVA, so be sure that the concepts involved in the one-way ANOVA are clear. Important background information and review of concepts in ANOVA can be found in Ray Ch. 8, so be sure to read that chapter carefully.

The sort of experiment that produces data for analysis by a two-factor ANOVA is one in which there are two factors (independent variables). In Ray's example (p. 182 *ff.*), an experimenter is interested in assessing the impact of housing (the first factor) and feeding schedule (the second factor) on errors made in running a maze (the dependent variable). In this experiment, the housing factor can take on two levels (enriched or standard) and the feeding schedule can take on two levels (ad lib or once a day). Thus, this experiment is a 2x2 independent groups design, which means that there are 4 unique conditions to the experiment. Of the 40 mice in the experiment, 20 are randomly assigned to the enriched housing and 20 are assigned to the standard housing. Of the 20 mice assigned to the enriched housing, 10 are fed ad lib and 10 are fed once a day. Likewise, of the 20 mice in the standard housing, 10 are fed ad lib and 10 are fed once a day. Schematically, the design would look like the table below:

	Enriched Housing	Standard Housing
Ad Lib Feeding	$n = 10$	$n = 10$
Once a Day Feeding	$n = 10$	$n = 10$

Of course, we could conduct two separate experiments with our 40 mice (or think of this experiment as two separate one-way independent groups analyses). For instance, we could put 40 mice into a single factor experiment, with 20 exposed to enriched housing and 20 exposed to standard housing. We would be testing the simple $H_0: \mu_{\text{Enriched}} = \mu_{\text{Standard}}$. Were we to do so, our source table would look like this:

Source	SS	df	MS	F
Housing	640	1	640	33.13
Error	734	38	19.3	
Total	1374	39		

These results would lead us to reject H_0 and conclude that there was a significant effect of housing, $F(1,38) = 33.13$, $MSE = 19.3$, $p < .05$. The mice in the enriched environment make significantly fewer errors ($M = 10$) than those in the standard environment ($M = 18$).

Alternatively, we could imagine the 40 mice in a different second single factor experiment, with 20 exposed to ad lib feeding and 20 exposed to once a day feeding. In this case, we would be testing the simple $H_0: \mu_{\text{AdHoc}} = \mu_{\text{Once}}$. Were we to do so, our source table would look like this:

Source	SS	df	MS	F
Feeding	0	1	0	0
Error	1374	38	36.2	
Total	1374	39		

These results would lead us to retain H_0 and conclude that there was no significant effect of feeding, $F(1,38) = 0$, $MSE = 36.2$, $p > .05$. The mice fed once a day did not differ in number of errors ($M = 14$) compared to those fed on an ad hoc basis ($M = 14$).

The advantage of a two-factor design is that not only can we assess the independent impact of our two factors (as in the two separate single-factor designs), but also we can assess the interaction of the two factors in their effect on the DV. Thus, with the same data we would be able to test three different null hypotheses:

Null Hypothesis	Alternative Hypothesis
$H_0: \mu_{\text{enriched housing}} = \mu_{\text{standard housing}}$	$H_1: \text{Not } H_0$
$H_0: \mu_{\text{ad lib feeding}} = \mu_{\text{once-a-day feeding}}$	$H_1: \text{Not } H_0$
$H_0: \text{no interaction between the two factors}$	$H_1: \text{Not } H_0$

The concept of interaction is a difficult one, but it is essential that you come to grasp the concept. Here's one definition, of an interaction: *An interaction occurs when the effect of one of the factors is not the same across all levels of the other factor.* (Does that make sense to you, even after re-reading it several times?) Read the portion of Ray's chapter on interaction (p. 182 - 195), and we'll return to a discussion of interaction effects shortly. Before doing so, however, let's complete the analysis of the data provided by Ray.

First of all, let's look at the summary table of the mean number of errors made by each group of mice.

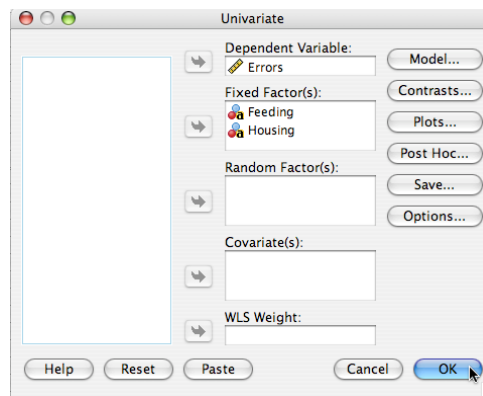
	Enriched Housing	Standard Housing	Marginal Means (feeding)
Ad Lib Feeding	6	22	14
Once a Day Feeding	14	14	14
Marginal Means (housing)	10	18	

Which two means do we compare to test the null hypothesis about the Housing factor? You should see that we would compare the marginal means for housing (10 vs. 18), because 10 represents the mean error score for all of the mice who were raised in the enriched housing and 18 represents the mean error score for all of the mice who were raised in standard housing. Likewise, to test the null hypothesis regarding Feeding, we would compare the

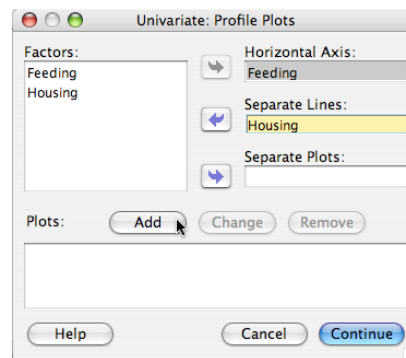
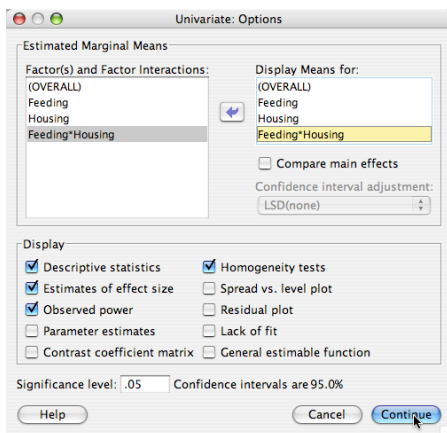
marginal means for feeding (14 vs. 14). On the face of it, it would certainly appear that there is no difference in number of errors whether we chose to feed the mice in an ad lib fashion or once a day. Well, let's actually compute the ANOVA using SPSS and then see how we would interpret the results.

You would need 3 columns to enter your data as seen below on the left (but only a portion of the data is shown). You would need a column for each factor (just as you would for a one-way ANOVA) and then use unique names or numbers to define the levels of the factor. In this case, I've chosen to use names as labels for the levels of the two IVs, so the first two columns are string variables. (This approach wouldn't work with One-Way ANOVA in SPSS, but is just fine for the procedure you'll use for a two-way ANOVA.) The final column holds the Error scores for each mouse. So the first mouse is raised in Enriched Housing and received Ad Lib Feeding, which leads that mouse to make 6 errors on the maze.

	Feeding	Housing	Errors
1	Ad lib	Enriched	6.00
2	Ad lib	Enriched	4.00
3	Ad lib	Enriched	8.00
4	Ad lib	Enriched	5.00
5	Ad lib	Enriched	7.00
6	Ad lib	Enriched	7.00
7	Ad lib	Enriched	5.00
8	Ad lib	Enriched	6.00
9	Ad lib	Enriched	4.00
10	Ad lib	Enriched	8.00
11	Ad lib	Standard	21.00
12	Ad lib	Standard	23.00
13	Ad lib	Standard	22.00
14	Ad lib	Standard	22.00
15	Ad lib	Standard	24.00
16	Ad lib	Standard	20.00
17	Ad lib	Standard	20.00
18	Ad lib	Standard	24.00
19	Ad lib	Standard	22.00
20	Ad lib	Standard	22.00
21	Once	Enriched	13.00
22	Once	Enriched	15.00
23	Once	Enriched	15.00
24	Once	Enriched	13.00
25	Once	Enriched	18.00
26	Once	Enriched	16.00
27	Once	Enriched	10.00
28	Once	Enriched	12.00



Once your data are entered, choose **General Linear Model** and then **Univariate...** from the **Analyze** menu. Doing so will prompt the window seen above on the right. I've dragged *Errors* into the **Dependent variable** slot and both *Feeding* and *Housing* into the **Fixed Factor(s)** slot. Before proceeding, however, you need to make some additional choices. On the right of the window, you'll notice buttons for **Options...** and **Plots...** Clicking on those buttons will reveal the windows seen below.



Note in the Options window on the left that I've chosen a number of options by clicking on appropriate boxes. Moreover, so that I'll be able to see means for the interaction as well as the main effects, at the top of the window I've moved variables from the left to the right. As you'll see later, choosing Descriptive statistics will give me the information for each of the four conditions (the interaction means), but otherwise I'd need to compute the marginal means. Choosing to display all the means may make my life easier later on.

Note in the Plot window on the right that I've placed the Feeding variable on the X axis and the Housing variable as separate lines in the graph. The cursor arrow is poised above the **Add** button and you need to actually click on that button to produce the graph. Then, click on the **Continue** button.

Below you'll find the Source Table and the Descriptive statistics.

Tests of Between-Subjects Effects

Dependent Variable: Errors								
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	1280.000 ^a	3	426.667	163.404	.000	.932	490.213	1.000
Intercept	7840.000	1	7840.000	3002.553	.000	.988	3002.553	1.000
Feeding	.000	1	.000	.000	1.000	.000	.000	.050
Housing	640.000	1	640.000	245.106	.000	.872	245.106	1.000
Feeding * Housing	640.000	1	640.000	245.106	.000	.872	245.106	1.000
Error	94.000	36	2.611					
Total	9214.000	40						
Corrected Total	1374.000	39						

a. R Squared = .932 (Adjusted R Squared = .926)

b. Computed using alpha = .05

Descriptive Statistics

Dependent Variable: Errors				
Feeding	Housing	Mean	Std. Deviation	N
Ad lib	Enriched	6.0000	1.49071	10
	Standard	22.0000	1.41421	10
	Total	14.0000	8.32877	20
Once	Enriched	14.0000	2.21108	10
	Standard	14.0000	1.15470	10
	Total	14.0000	1.71679	20
Total	Enriched	10.0000	4.49561	20
	Standard	18.0000	4.29198	20
	Total	14.0000	5.93555	40

First of all, you should note that the source table is fairly complex. That's because it includes a lot of information that you can readily ignore. First of all, you need look only at the rows for the two main effects, the interaction, the error term, and the total. Thus, you can ignore the rows labeled Corrected Model, Intercept, and Total. To the right of the source table, Partial Eta Squared is an effect size measure (closer to 1.0 is better) and an estimate of Power (closer to 1.0 is better). (You can ignore the Noncentrality Parameter). Looking at the Significance levels tells us that we can reject two of our null hypotheses, while retaining the remaining null hypothesis. Because the p -value for Housing is less than .05, we would reject H_0 for Housing. Because the p -value for Feeding is greater than .05, we would retain H_0 for Feeding. Finally, because the p -value for the AB (Housing x Feeding) interaction is less than .05, we would reject that H_0 .

Examining the output shows the Levene test, which indicates that there is little reason to be concerned that the data violate the homogeneity of variance assumption ($p > .05$). As a result, it seems reasonable to use the standard alpha-level of .05.

Levene's Test of Equality of Error Variances^a

Dependent Variable: Errors				
F	df1	df2	Sig.	
1.117	3	36	.355	

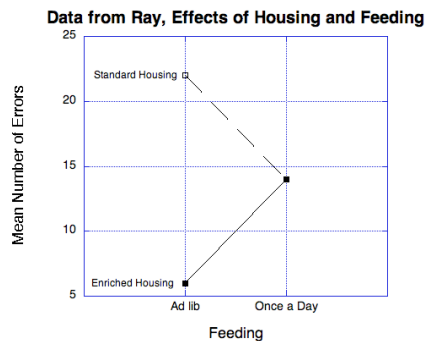
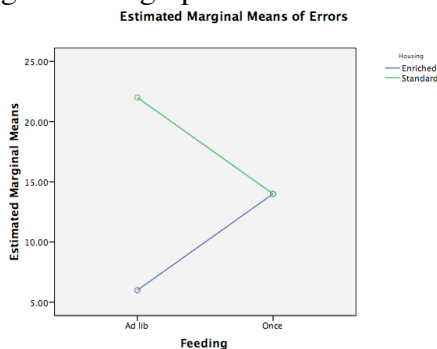
Tests the null hypothesis that the error variance of the dependent variable is equal across groups.
 a. Design: Intercept + Feeding + Housing + Feeding * Housing

In a multi-factor ANOVA where interactions are present, we need to concentrate first on explaining the interaction, so let's do so here. What has produced the significant interaction between Housing and Feeding? To understand the source of the interaction, we need to look at the means that are unique to the 4 conditions (that is, those in the interior of our means table). For instance, what happens to the error scores for mice in the Enriched Housing when we compare those getting Ad Lib Feeding with those getting Once-a-Day Feeding? As you should be able to see, the error scores are greater ($M = 14$) when the mice are fed once a day compared to error scores for mice fed on an ad lib basis ($M = 6$). So when raised in an enriched environment, mice do better when fed on an ad lib basis. Is that also true of mice raised in a standard environment? No, it is not. As you can see, for mice raised in the standard housing, errors were *higher* when they were fed on an ad lib basis ($M = 22$) compared to the once-a-day feeding ($M = 14$). This is what we mean by the effects of one factor not being the same at all levels of the other factor. In this case, error scores are *higher* for ad lib feeding compared to once-a-day feeding for mice in a standard environment **but** error scores are *lower* for ad lib feeding compared to once-a-day feeding for mice in the enriched environment. (Note that in describing an interaction, you'll typically say "but," "however," "on the other hand," etc.)

So far, all that we've done is to look at the means to determine the pattern of the interaction. To determine the statistical significance of these differences, we need to compute Tukey's HSD:

$$HSD = q \sqrt{\frac{MS_{Error}}{n}} = 3.8 \sqrt{\frac{2.611}{10}} = 1.94$$

Before we use Tukey's HSD, however, let's look at the interaction graphically. When an interaction is present, the lines used to connect the conditions in a graph will not be parallel, as seen in the graph below left. Unfortunately, SPSS uses color lines, which don't come out well on a black-and-white copy. Thus, I've used another piece of software (Kaleidagraph) to generate a graph that comes out equally well in color or black-and-white.



Can you see how to translate the table of means into the graphs seen above? One factor (Feeding) is shown on the x-axis, while the other factor (Housing) is shown within the body

of the graph by using different symbols. The y-axis is used to show scores on the DV (Errors). Lines are used to connect the means for the two levels of the Housing factor. That is, one line connects the two means for Standard Housing (22 and 14) and another line connects the two means for Enriched Housing (6 and 14). The fact that those two lines are not parallel is an indication that there is an interaction between the two factors.

If we were to use Tukey's HSD to analyze the means portrayed in the graph (or in the table), we would arrive at the same conclusion we'd arrived at simply by eye-balling the data. (That won't always be the case.) That is, we could look at the effect of feeding under Enriched Housing (called a simple effect), where we would learn that the error scores are greater ($M = 14$) when the mice are fed once a day compared to error scores for mice fed on an ad lib basis ($M = 6$), because the difference ($14 - 6 = 8$) is greater than the critical mean difference of 1.94. Examining the other simple effect (effect of feeding at Standard Housing) we find that when mice are raised in Standard Housing, the errors were *higher* when they were fed on an ad lib basis ($M = 22$) compared to the once-a-day feeding ($M = 14$), because the difference is greater than the HSD critical mean difference of 1.94.

It would also be possible to consider the interaction from a different perspective. That is, you could look at the simple effect of housing at Ad Lib feeding (finding that mice make significantly more errors under Standard Housing compared to Enriched Housing) and the simple effect of housing at Once-a-Day Feeding (finding that mice make an equal number of errors regardless of type of housing). Note that you would be telling a different "story," but it would still be an "interaction" story. That is, you could say that when mice are fed on an ad lib basis they make significantly more errors when they live in standard housing compared to enriched housing. *However*, when they are fed on a once-a-day basis, the mice make equivalent errors.

Now that we've gotten a good understanding of the source of the interaction, we can look at the two main effects (independent effects of each factor). First of all, would you be willing to conclude that there is no effect of the Feeding factor simply because there is no main effect for that factor? You should see that although there is no difference at all between the means for the two levels of that factor (both are 14) that is *not* an indication that there is no effect of Feeding. In fact, Feeding is a very influential factor, but it works very differently in each of the two environments examined in this experiment, and the effects cancel one another out. So although there is no main effect for Feeding (because the P-value for the main effect is $> .05$), the presence of the significant interaction tells us that Feeding is, in fact, an influential factor.

So let's look at the other main effect (Housing). On the basis of the significant main effect, you might be tempted to conclude that Standard Housing ($M = 18$) leads to significantly more errors than Enriched Housing ($M = 10$). Would you be justified in that global conclusion? What about mice who are fed once a day? Once again, you should see that the presence of the interaction would lead us to qualify our interpretation of a main effect. Mice who are fed on an ad lib basis *do* make far fewer errors when raised in Enriched Housing compared to those raised in Standard Housing, but that difference disappears when mice are fed once a day.

A complete interpretation of the outcome of this experiment, then, would hinge on the interpretation of the interaction. When a significant interaction is present, if you explain the interaction well, you will have made great strides toward interpreting the outcome of the experiment. In this case, in your Results section, you might say something like:

There was a significant main effect of housing, $F(1,36) = 245.106$, $MSE = 2.611$, $p < .001$. There was also a significant interaction between housing and feeding, $F(1,36) = 245.106$, $p < .001$. Post hoc analyses using Tukey's HSD indicated that mice raised in an enriched environment make fewer errors on a maze when fed on an ad lib basis ($M = 6$) compared to mice fed once a day ($M = 14$). However, mice raised in a standard environment make fewer errors on a maze when fed once a day ($M = 14$) compared to mice fed on an ad lib basis ($M = 22$). So, to enhance learning, your feeding should depend on the mouse's housing.

In your Discussion section, you'd want to talk about *why* you think those results emerged. That is, you might hypothesize that when the mice are fed on an ad lib basis, they don't worry about where their next meal is coming from, so they have more time to explore and benefit from the enriched environment. Because the standard environment is so boring, those in that condition simply sit around waiting for their next meal. On the other hand, mice fed once a day may spend a portion of the day hungry and focused more on trying to find food. As a result, they don't take advantage of the enriched environment, but they move around both environments looking for food.

Another (more complex) Example

The experiment described above is the simplest possible two-way design (2x2). You should also be able to analyze and interpret experiments in which the two factors might each contain more than two levels (e.g., 2x3, 2x4, 3x3, 3x4). In the above example, you may be able to completely interpret the outcome of the experiment without recourse to post hoc tests (if the interaction is not significant). However, with increasingly complex designs (more than two levels of either or both factors) you would have to use Tukey's test to completely analyze the data.

For example, Schachter, Christenfeld, Ravina and Bilous (1991) studied the presence of speech fillers (um, ah,...) in faculty from different disciplines (Natural Science, Social Science, and Humanities). [You should note that this factor is a non-manipulated participant characteristic.] Schachter et al. thought that when lecturing in a classroom setting, faculty in more "precise" disciplines would be inclined to use fewer fillers. However, when being interviewed (e.g., about their ongoing work with graduate students), Schachter et al. thought that there would be little difference among the faculty. I'm going to make up a little data set (below left) that would be consistent with the results they obtained and then analyze it using SPSS and then interpret the results. Just so you see how the data would be input in SPSS, below right is a portion of the data file. For the Situation under which the data are collected, I've used the labels Lecture and Interview. For the Discipline, I've used the abbreviated labels Nat Sci, Soc Sci, and Human.

Lecture			Interview		
Nat Sci	Soc Sci	Hum	Nat Sci	Soc Sci	Hum
1	4	4	5	4	5
2	4	5	5	5	5
1	4	4	5	5	6
1	4	5	6	5	5
2	4	5	6	5	6

	Situation	Discipline	Fillers
1	Lecture	Nat Sci	1.00
2	Lecture	Nat Sci	2.00
3	Lecture	Nat Sci	1.00
4	Lecture	Nat Sci	1.00
5	Lecture	Nat Sci	2.00
6	Lecture	Soc Sci	4.00
7	Lecture	Soc Sci	4.00
8	Lecture	Soc Sci	4.00
9	Lecture	Soc Sci	4.00
10	Lecture	Soc Sci	4.00
11	Lecture	Human	4.00
12	Lecture	Human	5.00
13	Lecture	Human	4.00
14	Lecture	Human	5.00
15	Lecture	Human	5.00
16	Interview	Nat Sci	5.00
17	Interview	Nat Sci	5.00
18	Interview	Nat Sci	5.00
19	Interview	Nat Sci	6.00
20	Interview	Nat Sci	6.00
21	Interview	Soc Sci	4.00
22	Interview	Soc Sci	5.00
23	Interview	Soc Sci	5.00

Below you can also see the source table and the means table for the ANOVA. (I've also included a graph.)

Tests of Between-Subjects Effects

Dependent Variable: Fillers

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	56.267 ^a	5	11.253	48.229	.000	.909	241.143	1.000
Intercept	546.133	1	546.133	2340.571	.000	.990	2340.571	1.000
Situation	26.133	1	26.133	112.000	.000	.824	112.000	1.000
Discipline	13.067	2	6.533	28.000	.000	.700	56.000	1.000
Situation * Discipline	17.067	2	8.533	36.571	.000	.753	73.143	1.000
Error	5.600	24	.233					
Total	608.000	30						
Corrected Total	61.867	29						

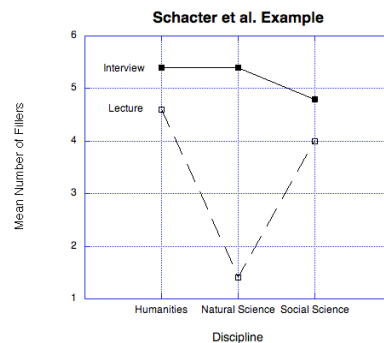
a. R Squared = .909 (Adjusted R Squared = .891)

b. Computed using alpha = .05

Descriptive Statistics

Dependent Variable: Fillers

Situation	Discipline	Mean	Std. Deviation	N
Interview	Human	5.4000	.54772	5
	Nat Sci	5.4000	.54772	5
	Soc Sci	4.8000	.44721	5
	Total	5.2000	.56061	15
Lecture	Human	4.6000	.54772	5
	Nat Sci	1.4000	.54772	5
	Soc Sci	4.0000	.00000	5
	Total	3.3333	1.49603	15
Total	Human	5.0000	.66667	10
	Nat Sci	3.4000	2.17051	10
	Soc Sci	4.4000	.51640	10
	Total	4.2667	1.46059	30



Note that because there are only two levels of the Situation factor, you would be able to reject the H_0 and conclude that there were significantly more fillers in the Interview than in the Lecture. But because there is a significant interaction, we would probably want to hold off on that interpretation until we'd explained the interaction, so let's try to do so now.

To see which of the 6 particular means differed, we would have to compute Tukey's HSD. We know what MS_{Error} is (.233) and we know what n is (5), so we only have to look up q to compute HSD. In this case, because there are 6 unique conditions, we would look up q with 6 treatments and 24 df_{Error} (so $q = 4.37$).

$$HSD = q\sqrt{\frac{MS_{Error}}{n}} = 4.37\sqrt{\frac{.233}{5}} = .94$$

Armed with that value, we can now interpret the interaction. First, looking at the Interview Situation, none of the means for the three Disciplines (5.4, 5.4, and 4.8) are significantly different, because their differences are all less than .94. However, looking at the Lecture Situation, both the Social Science ($M = 4.0$) and Humanities ($M = 4.6$) faculty used significantly more fillers than the Natural Science faculty ($M = 1.4$), although they don't differ from one another.

Another way to look at the source of the interaction is to compare each discipline at the two different situations. That is, we would be examining the simple effects of Situation at each level of Discipline. Using this approach, we can see that faculty in the Natural Sciences use significantly more fillers when in the Interview Situation ($M = 5.4$) than in the Lecture Situation, ($M = 1.4$) but faculty in the Social Sciences or Humanities don't differ in the number of fillers present in the Interview Situation or in the Lecture Situation (although these differences appear to be almost significant, and would become significant with more power). Notice that whichever approach we take, we get a sense of the source of the interaction. That is, the effects of Situation are not the same at all levels of Discipline (or the effects of Discipline are not the same at all levels of Situation).

Once we've interpreted the interaction, you should see that the main effect for Discipline is of little interest. Although we could use a Tukey's HSD to see which Disciplines differ in the number of fillers present, the analysis of the interaction showed us that the Disciplines differed *only* in the Lecture Situation and not in the Interview Situation, so we should be reluctant to make any overall statement about differences among the Disciplines. Note that the same logic might lead us to be uncomfortable stating that there were significantly more fillers in the Interview Situation compared to the Lecture Situation (but the fact that the differences were almost significant might lead us to conclude that with a larger sample size we might well find that faculty use significantly more fillers in an Interview Situation compared to a Lecture Situation).

Thus, for this experiment, it appears that the results are entirely consistent with the predictions of Schacter et al., which essentially predicted an interaction. One might report the results as:

There was a significant main effect of discipline, $F(2,24) = 28.0$, $MSE = .233$, $p < .001$. There was also a significant main effect of situation, $F(1,24) = 112.0$, $p < .001$. There was also a significant interaction between discipline and situation, $F(2,24) = 36.571$, $p < .001$. Post hoc analyses using Tukey's HSD indicated that in the Interview Situation, the mean number of fillers for the three Disciplines do not differ ($M = 5.4$, 5.4, and 4.8 for Humanities, Social Sciences and Natural Sciences respectively). However, in the Lecture Situation, both the Social Sciences ($M = 4.0$) and Humanities ($M = 4.6$) faculty used significantly more fillers than the Natural Science faculty ($M = 1.4$), although they do not differ from one another.

How would you talk about these results in a Discussion section?

Making Sure That You Understand *df*

Suppose that there were more than five participants per condition, how would the *df* change? Answer the questions below to ensure that you understand the computation of *df* for a two-factor ANOVA.

<p>Conditions stay the same, but $n = 10$</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center;">Source</th> <th style="text-align: center;"><i>df</i></th> </tr> </thead> <tbody> <tr><td>Discipline</td><td></td></tr> <tr><td>Situation</td><td></td></tr> <tr><td>Situation x Discipline</td><td></td></tr> <tr><td>Error</td><td></td></tr> </tbody> </table>	Source	<i>df</i>	Discipline		Situation		Situation x Discipline		Error		<p>Introduce a 3rd Situation (e.g., party), with $n = 15$</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center;">Source</th> <th style="text-align: center;"><i>df</i></th> </tr> </thead> <tbody> <tr><td>Discipline</td><td></td></tr> <tr><td>Situation</td><td></td></tr> <tr><td>Situation x Discipline</td><td></td></tr> <tr><td>Error</td><td></td></tr> </tbody> </table>	Source	<i>df</i>	Discipline		Situation		Situation x Discipline		Error	
Source	<i>df</i>																				
Discipline																					
Situation																					
Situation x Discipline																					
Error																					
Source	<i>df</i>																				
Discipline																					
Situation																					
Situation x Discipline																					
Error																					
<p>Introduce a 4th Discipline, with $n = 10$</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center;">Source</th> <th style="text-align: center;"><i>df</i></th> </tr> </thead> <tbody> <tr><td>Discipline</td><td></td></tr> <tr><td>Situation</td><td></td></tr> <tr><td>Situation x Discipline</td><td></td></tr> <tr><td>Error</td><td></td></tr> </tbody> </table>	Source	<i>df</i>	Discipline		Situation		Situation x Discipline		Error		<p>Introduce a 4th Disc & 3rd Situation, with $n = 20$</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center;">Source</th> <th style="text-align: center;"><i>df</i></th> </tr> </thead> <tbody> <tr><td>Discipline</td><td></td></tr> <tr><td>Situation</td><td></td></tr> <tr><td>Situation x Discipline</td><td></td></tr> <tr><td>Error</td><td></td></tr> </tbody> </table>	Source	<i>df</i>	Discipline		Situation		Situation x Discipline		Error	
Source	<i>df</i>																				
Discipline																					
Situation																					
Situation x Discipline																					
Error																					
Source	<i>df</i>																				
Discipline																					
Situation																					
Situation x Discipline																					
Error																					

Two-Factor Independent Groups Designs:

A researcher was interested in the impact of a particular drug (Smart-O) on rats' performance in a maze. She decided to run an independent groups design, comparing Smart-O with a placebo. She also thought that the type of maze (simple vs. complex) might have an impact, so she introduced this second factor into the design — producing a 2x2 independent groups design. Her budget was pretty flush, so she decided to run 25 rats in each condition. She chose to use the number of errors the rats made (going down blind alleys) as her dependent variable. On completion of the study, she ran an analysis of the data, but absent-mindedly left her output where the rats could get to it and they nibbled away parts of the source table. Generate the missing parts of the table.

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>
Drug (Drug vs. Placebo)	10			
Maze (Simple vs. Complex)	20			
Drug x Maze				
Error (Within)	192			
Total	262			

Dr. Ignatius Pooh (Iggy to his friends) is interested in studying revulsion. He thinks that people will be reluctant to eat foods that they know are perfectly fine if the foods are formed into unpalatable shapes. In order to test his theories, he presents participants with chocolate formed either to look like dog feces or in a normal bar. He measures the time (in minutes) it takes the participants to consume the chocolate as his DV. Suppose that Dr. Pooh was also interested in the role of “presentation,” so for half of the participants in each of the two conditions the chocolate is presented to the participants on a plate that is spotlessly clean and for the other half of the participants the chocolate is presented on a plate that is dirty (smudged and caked with dried-on food remains). Thus, this is a 2x2 independent groups design, and Dr. Pooh runs a total of 200 participants, with 50 assigned randomly to each of the 4 conditions. Given the results shown below (and knowing that $F_{crit} = 3.89$), complete the source table and interpret the results of this study as completely as you can. How might you talk about these results in a Discussion section?

	Dog Feces		Normal Bar	
	Clean Plate	Dirty Plate	Clean Plate	Dirty Plate
Mean	8	10	1	5
Variance	5	5	5	5

Source	SS	df	MS	F
Chocolate Shape				
Plate Type			25	
Shape x Plate	25			
Error (Within)				
Total	1050			

Pierroutsakos and DeLoache (2003) were interested in the development of pictorial competence in children. They use the term pictorial competence to “encompass the many factors that are involved in perceiving, interpreting, understanding, and using pictures (DeLoache, Pierroutsakos, & Uttal, 2003, p. 115).” These researchers observe infants as they explore pictures of familiar objects. Unlike adults, infants are not content to look at the pictures, but will typically attempt to manipulate the pictured object.

In the data depicted below, 9-month-old infants were shown photographs and line drawings of familiar objects. The photographs and line drawings were either in black-and-white or in color. The dependent variable is the number of manual behaviors the infant exhibits toward the picture. Complete the following source table, which depicts results consistent with those of Pierroutsakos and DeLoache, and then analyze the results as completely as you can. Make an effort to interpret the results as you would in a Discussion section.

Descriptive Statistics

Dependent Variable: Manual Behav

Pict Type	Color	Mean	Std. Deviation	N
Drawing	B&W	.9000	.56765	10
	Color	2.7000	.94868	10
	Total	1.8000	1.19649	20
Photo	B&W	3.0000	.81650	10
	Color	5.0000	.81650	10
	Total	4.0000	1.29777	20
Total	B&W	1.9500	1.27630	20
	Color	3.8500	1.46089	20
	Total	2.9000	1.66102	40

Tests of Between-Subjects Effects

Dependent Variable: Manual Behav

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model					.000	.786	132.417	1.000
Intercept					.000	.936	526.539	1.000
PictType				75.75	.000	.678	75.757	1.000
Color			36.100	56.50	.000	.611	56.504	1.000
PictType * Color			.100		.695	.004	.157	.067
Error								
Total								
Corrected Total								

a. R Squared = .786 (Adjusted R Squared = .768)

b. Computed using alpha = .05

Catharsis and Aggression: Effects of Annoyance and Retaliation on Aggressive Behavior (Doob & Wood, 1972)

Doob & Wood (1972) conducted an experiment reported in “Catharsis and aggression: Effects of annoyance and retaliation on aggressive behavior.” As they state, “The catharsis hypothesis of aggression usually refers to a decrease in aggression after the expression of aggression. The assumption as stated by Freud is that there is a certain amount of aggression that has to be expressed, and that once this has happened, there is less left to be expressed later on.”

Half the participants were first annoyed by the confederate (who made all sorts of nasty personal comments), and the other half were not annoyed by the confederate. Then one-third of each group: (1) shocked the confederate, who had been moved into an adjoining room, for errors in a learning task, (2) watched the experimenter shock the confederate, who had been moved into an adjoining room, for errors in a learning task, or (3) saw nothing, as the experimenter took the confederate into an adjoining room to administer the “learning test.” In the final phase of the experiment, participants judged the “creativity” of associations made by the confederate to words presented by the participant. If they thought that a response was uncreative they gave the boorish confederate a shock. They were told to increase the duration of the shock for increasingly uncreative responses. Confederates (who did not really receive any shocks) were cautioned by the experimenter not to comment if a shock was delivered, and then the experimenter left the room.

Doob & Wood were interested in testing the hypothesis that participants would experience a cathartic effect from observing the boorish woman (confederate) being shocked, or shocking her themselves. This would lead them to be less likely to shock the confederate in the final (judging creativity) phase of the experiment.

a. Besides running more participants, describe two very specific ways in which you could make this experiment more powerful.

b. The mean number of shocks delivered by the participants in the final phase of the experiment are shown below:

Condition	Participant Shocks	Experimenter Shocks	No Shock
Annoy	6.80	7.60	10.67
No Annoy	8.07	9.73	6.60

Graph the data, then tell me (just by looking at the results) what you expect an ANOVA to tell you about the data and why. Furthermore, tell me how you would interpret these results in terms of Doob & Wood’s hypothesis about catharsis. How would you interpret the results of the No Annoy condition?

c. Below is an ANOVA source table consistent with the results reported by Doob & Wood. Complete the portions that are missing, and then tell me everything you can about the experiment from looking at the source table and the means. Some of the information you will know exactly, and other information you can make reasonable inferences about. [F_{crit} for the interaction = 3.10]

Source	SS	df	MS	F
Confed behavior (Annoy vs. No Annoy)				1.0
Confed Shocked (No vs. Partic. vs. Exper.)				1.0
Interaction	18			4.5
Error				
Total		89		

The state superintendent of instruction asks the director of educational research to investigate differences in scores on a standardized teacher examination for senior education students majoring in the following subject areas: English, Mathematics, Physical Education, and Vocational Education. The following results are from a random sample of 32 graduating seniors (16 males and 16 females). Complete the source table and interpret the results as completely as you can. Then make sense of these results as you might in a Discussion section.

Descriptive Statistics

Dependent Variable: Score				
Major	Gender	Mean	Std. Deviation	N
English	Female	37.7500	1.70783	4
	Male	37.7500	2.62996	4
	Total	37.7500	2.05287	8
Math	Female	43.0000	5.59762	4
	Male	39.0000	5.29150	4
	Total	41.0000	5.47723	8
Phys-Ed	Female	35.5000	2.08167	4
	Male	34.7500	.95743	4
	Total	35.1250	1.55265	8
Voc-Ed	Female	27.7500	1.70783	4
	Male	29.7500	2.06155	4
	Total	28.7500	2.05287	8
Total	Female	36.0000	6.35610	16
	Male	35.3125	4.64354	16
	Total	35.6562	5.48670	32

Tests of Between-Subjects Effects

Dependent Variable: Score								
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model					.000	.738	67.511	1.000
Intercept					.000	.994	3989.421	1.000
Major	647.34				.000	.726	63.478	1.000
Gender	3.78				.548	.015	.371	.090
Major * Gender	37.34				.324	.132	3.662	.285
Error	244.75							
Total								
Corrected Total	933.21							

a. R Squared = .738 (Adjusted R Squared = .661)

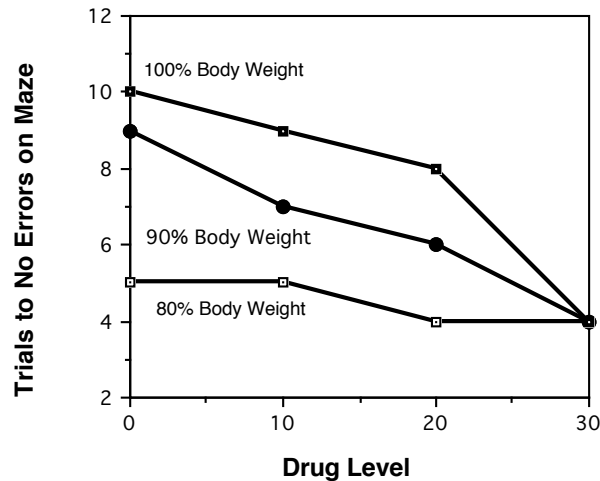
b. Computed using alpha = .05

The following is a summary of an experiment in problem solving:

The task involved a “lights and switches” apparatus (each light connected to several switches and each switch connected to several lights). The participant’s task was to figure out which switches had to be activated in order to have just the second light on the panel “on.” Two independent variables were involved in the experiment. One was the number of switches available to the participant (5, 7, or 9 switches); the other was the amount of money participants were told they could win by solving the problem very quickly (\$1, \$5, \$20). There were 10 participants in each of the cells and each participant was randomly assigned to one of the nine cells formed by the factorial combination of these two independent variables. Each participant’s time-to-solve the problem was measured in minutes. The numbers given here refer to the mean solution times for the nine cells. Assuming that all numerical differences are statistically significant, describe the results of the experiment to me in terms of main effects and the interaction. What df would be found in the source table for the components of variance (Switches, Reward, Switches x Reward, Within {or Error}, and Total)? How would you describe the outcome in words?

Amount of Money	Number of Switches		
	Five	Seven	Nine
\$1	14	17	20
\$5	10	17	24
\$20	8	17	26

An experimenter was interested in the effects of a memory-enhancing drug on the maze-completion behavior of rats. She gave 4 levels of the drug (0, 10, 20, 30) to 3 groups of rats that were food deprived to different levels of weight (80%, 90%, and 100% of free-feeding weight). She used a total of 120 rats, with an equal number in each condition. A graph of the data is seen below. The dependent variable is the number of trials it took for the rat to traverse the maze without making any errors.



Tell me what effects you would expect to find should you do an ANOVA on these data. Tell me the degrees of freedom you would use for the critical F you would use to test each of the effects. How would you interpret these results?

Suppose that you have a theory that self-esteem has an impact on a person's ability to solve problems. You randomly sample a large group of students and give them a test to measure self-esteem and are able to assign 40 participants to the Low Self-Esteem group, 40 participants to the Moderate Self-Esteem group, and 40 participants to the High Self-Esteem group. Suppose that you also believe that the difficulty of the problem will have an impact on problem-solving ability, so you assign a quarter of the participants in each of the 3 self-esteem groups to one of four different levels of problem complexity (Easy, Moderately Easy, Moderately Difficult, and Difficult).

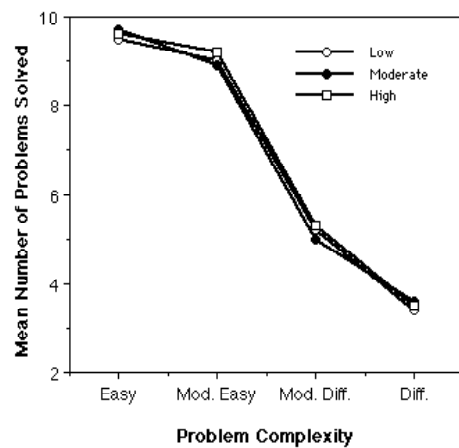
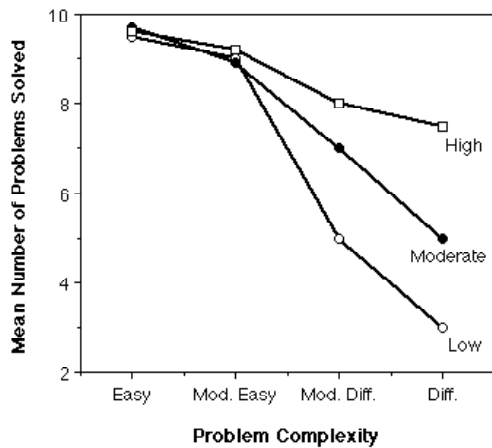
How would you describe this design?

How many participants are present in the study?

What df would you find in the source table analyzing the data?

Can you spot any limitations on your interpretation of the results of this study (i.e., any problems with your IVs)?

If your results came out as illustrated below, what would you expect to find in terms of the main effects and interactions and how would you interpret the results? (The left and right graphs illustrate the outcomes of two different experiments.)



A health educator suspects that the “days of discomfort” caused by common colds can be reduced by ingesting large doses of Vitamin C and visiting a sauna every day. Participants with new colds are randomly assigned to one of four different doses of Vitamin C (500, 1000, 1500, or 2000 milligrams) and to one of three different daily exposures to a sauna (0, .5, or 1 hour). The DV is the number of days of discomfort experienced by each of the participants. Complete the source table below and analyze and interpret the results of this study as completely as you can. Then tell me what your next step would be.

Descriptive Statistics

Dependent Variable: Days of Discomfort				
Dose of Vitamin C	Sauna	Mean	Std. Deviation	N
1000 mg	.5 Hr	3.6000	1.14018	5
	0 Hr	3.8000	.83666	5
	1 Hr	3.6000	1.14018	5
	Total	3.6667	.97590	15
1500 mg	.5 Hr	3.8000	.83666	5
	0 Hr	3.2000	.83666	5
	1 Hr	4.2000	.83666	5
	Total	3.7333	.88372	15
2000 mg	.5 Hr	3.0000	.70711	5
	0 Hr	3.6000	.54772	5
	1 Hr	3.6000	.54772	5
	Total	3.4000	.63246	15
500 mg	.5 Hr	4.2000	.83666	5
	0 Hr	4.0000	.70711	5
	1 Hr	4.4000	.54772	5
	Total	4.2000	.67612	15
Total	.5 Hr	3.6500	.93330	20
	0 Hr	3.6500	.74516	20
	1 Hr	3.9500	.82558	20
	Total	3.7500	.83615	60

Tests of Between-Subjects Effects

Dependent Variable: Days of Discomfort								
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model					.275	.224	13.875	.595
Intercept					.000	.963	1265.625	1.000
DoseofC	4.983				.071	.135	7.475	.582
Sauna	1.200				.413	.036	1.800	.196
DoseofC * Sauna	3.067				.600	.087	4.600	.273
Error	32.000							
Total								
Corrected Total	41.250							

a. R Squared = .224 (Adjusted R Squared = .046)

b. Computed using alpha = .05

A researcher was interested in examining the role of gender in the context of bystander apathy. To that end, she randomly assigned male and female college students to observe a man attacking a woman and the woman yelling, "Stop, I don't know you!" The situation is established so that the participants don't realize that the attack is part of the study. The situation is manipulated so that there are 0, 5, or 15 other people present when the attack takes place. The DV is the time it takes a participant to intervene in the altercation (number of minutes). If the participant hasn't intervened within 10 minutes, that participant receives a score of 10. Complete the analysis below and interpret the results of this study as completely as you can. How might you interpret these results in a Discussion section?

Descriptive Statistics

Dependent Variable: Time to Intervene

Gender	Number	Mean	Std. Deviation	N
Female	0	9.4000	1.07497	10
	15	9.5000	.97183	10
	5	9.9000	.31623	10
	Total	9.6000	.85501	30
Male	0	3.1000	1.19722	10
	15	9.2000	.78881	10
	5	6.7000	1.15950	10
	Total	6.3333	2.74595	30
Total	0	6.2500	3.41629	20
	15	9.3500	.87509	20
	5	8.3000	1.83819	20
	Total	7.9667	2.60356	60

Tests of Between-Subjects Effects

Dependent Variable: Time to Intervene

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model					.000	.874	374.500	1.000
Intercept					.000	.987	4080.071	1.000
Gender	160.067				.000	.761	171.500	1.000
Bystanders	99.433				.000	.664	106.536	1.000
Gender * Bystanders	90.033				.000	.641	96.464	1.000
Error	50.400							
Total								
Corrected Total	399.933							

a. R Squared = .874 (Adjusted R Squared = .862)

b. Computed using alpha = .05

Suppose that the same data had been analyzed as a one-way ANOVA on Number of Bystanders. What would that source table look like?

Source	SS	df	MS	F
Treatment				
Error				
Total				

Hmmm. There's an article with the intriguing title, "Why people fail to recognize their own incompetence" by Dunning, Johnson, Ehrlinger, and Kruger (2003). According to Confucius, "real knowledge is to know the extent of one's ignorance." Dunning, et al. (2003) asked students who were leaving an exam to judge how well they'd done on the exam. It turned out that students who performed the worst on the exam actually overestimated their performance and students who did the best on the exam were fairly accurate in their self-assessment (with a slight underestimation among the students with the best performance).

In one study, Kruger and Dunning (1999) gave additional information to some students, and that information had an impact on their judgments. Let's imagine a set of results that are consistent with their report. The dependent variable is the percent overestimation of a person's performance on an exam. So a score of zero is an accurate judgment. A positive score indicates overestimation and a negative score is an underestimation of one's performance. The students were divided into four groups based on their actual performance (Bottom Quartile, Second Quartile, Third Quartile, and Top Quartile). In addition, half of the students in each quartile were given a mini-lecture about the material after completing the exam (Add Info), but before making their judgments. The other half of each quartile was not given any additional information (No Info). Complete the source table below and interpret the results of this study as completely as you can. Interpret the results as you would in a Discussion section.

Descriptive Statistics

Dependent Variable: Estimate

Quartile	Additional Info	Mean	Std. Deviation	N
Bottom	Add Info	12.3000	3.16403	10
	No Info	30.4000	4.29987	10
	Total	21.3500	9.98565	20
Second	Add Info	9.3000	1.41814	10
	No Info	20.9000	2.07900	10
	Total	15.1000	6.19762	20
Third	Add Info	4.0000	1.15470	10
	No Info	3.7000	1.33749	10
	Total	3.8500	1.22582	20
Top	Add Info	-.3000	.82327	10
	No Info	-1.3000	.94868	10
	Total	-.8000	1.00525	20
Total	Add Info	6.3250	5.22512	40
	No Info	13.4250	13.18874	40
	Total	9.8750	10.58824	80

Tests of Between-Subjects Effects

Dependent Variable: Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model					.000	.960	1720.260	1.000
Intercept					.000	.956	1578.668	1.000
Quartile	6184.650				.000	.946	1251.531	1.000
AddInfo	1008.200				.000	.739	204.020	1.000
Quartile * AddInfo	1308.100				.000	.786	264.708	1.000
Error	355.800							
Total								
Corrected Total	8856.750							

a. R Squared = .960 (Adjusted R Squared = .956)

b. Computed using alpha = .05

Dr. Noah Weisser was interested in investigating the effects of sleep deprivation and alcohol on driving ability. To that end, he conducted a 3x3 independent groups study. Participants were given identical 32 oz. drinks that could contain 2, 4, or 8 ozs. of vodka (with the remainder of the glass filled with orange juice). Within each level of alcohol, one-third of the participants were deprived of sleep for 2 days, one-third were deprived of sleep for 3 days, and the final third were deprived of sleep for 4 days. Each of the participants operated a driving simulator for 30 minutes, during which the number of driving errors (objects hit, crossing into the oncoming lane, etc.) were recorded. Complete the analysis shown below and interpret the results as completely as you can, including the sort of interpretation that you'd include in a Discussion section.

Descriptive Statistics

Dependent Variable: Errors

Amt of ...	Hours of ...	Mean	Std. Deviation	N
2 oz.	2 days	1.3750	1.06066	8
	3 days	3.2500	1.03510	8
	4 days	6.3750	1.18773	8
	Total	3.6667	2.35292	24
4 oz.	2 days	3.5000	.92582	8
	3 days	6.8750	.99103	8
	4 days	7.7500	.70711	8
	Total	6.0417	2.05319	24
8 oz.	2 days	7.2500	1.28174	8
	3 days	9.5000	1.19523	8
	4 days	16.5000	3.07060	8
	Total	11.0833	4.47133	24
Total	2 days	4.0417	2.69426	24
	3 days	6.5417	2.81269	24
	4 days	10.2083	4.94297	24
	Total	6.9306	4.39695	72

Tests of Between-Subjects Effects

Dependent Variable: Errors

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model					.000	.906	606.716	1.000
Intercept					.000	.964	1687.325	1.000
Alcohol			344.264		.000	.842	335.932	1.000
SleepDep			230.889		.000	.781	225.301	1.000
Alcohol * SleepDep			23.306		.000	.419	45.483	1.000
Error			2.050					
Total								
Corrected Total								

a. R Squared = .906 (Adjusted R Squared = .894)

b. Computed using alpha = .05

Dr. Putz was interested in assessing the degree to which male and female schizophrenics might respond to a new drug that he had developed. The drug was intended to make them more sane (of course, as all Marx brothers devotees know, there is no sanity clause, right?), and so Dr. Putz measured their sanity on a 9-point scale with 9 being perfectly sane and 1 being perfectly insane (in this case, schizophrenic). He wasn't quite sure of the proper dosage, so he gave his participants four different dosage levels in a completely between (independent groups) 4 (dosage) x 2 (gender) design, with an equal number of participants per condition. Below is a table of mean sanity scores and a partially completed source table for the ANOVA that Dr. Putz performed. Complete the source table and interpret the results as completely as possible.

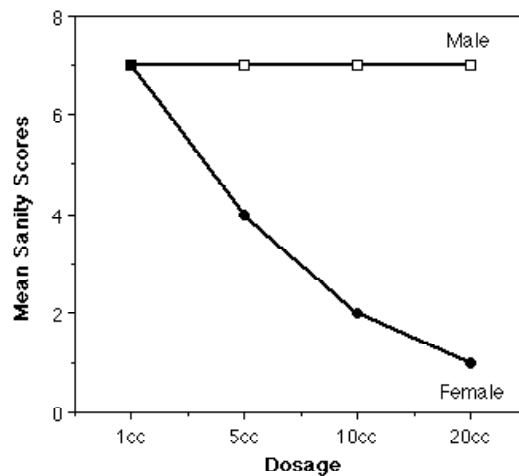
	Dosage			
	1cc	5cc	10cc	20cc
Male	6	6	7	9
Female	3	3	4	6

Source	SS	df	MS	F	F _{Crit}
Dosage				6	2.65
Gender				8	3.89
Dosage x Gender				1	2.65
Error (Within)			2		
Total	394				

How many participants were in each cell (condition) of Dr. Putz's experiment? Also, showing only the actual value (no computation required), what would be your best estimate of the variance of sanity scores in the population from which these samples were selected?

Dr. Putz was interested in re-running his experiment as a completely within (repeated measures) design. Can you think of *two* really good reasons why this might be problematic?

Suppose that Dr. Putz's results had turned out as seen in the figure below. What would you expect to find in a source table analyzing these data (in terms of main effects and the interaction)? How would you interpret the results?



Here are a couple of problems using StatView, so that you can make the translation to the older problems you'll find on old exams.

Gangestad, Simpson, Cousins, Garver-Apgar, and Christensen (2004) studied women over the course of their menstrual cycles to determine if they had a preference for male behavioral displays. I'll reconstruct their study as a two-factor independent groups design, while retaining the basic message of their article. Women watched a videotape of a male being interviewed. Half of the women saw the male respond to a question about himself ("Please tell me about yourself, including who you are, what you like to do, etc."). The other half of the women watched a videotape in which a male responded to a competitor for a date with a young woman (detailing why she should prefer to go on a date with him). For each video, one-third of the women responded on Day 3 of their menstrual cycle (a low fertility day). One-third of the women responded on Day 11 of their menstrual cycle (a high fertility day). Another third of the women responded on Day 21 of their menstrual cycle (a low fertility day). The dependent variable is a rating (on a 5-pt scale) by the women of the attractiveness of the male on a short-term basis. High scores indicate that the males were judged to be attractive for short-term sexual affairs. Complete the source table below and analyze these data as completely as you can. How might you interpret these results in a Discussion section?

Means Table for Attractiveness
Effect: Day of Cycle * Video

	Count	Mean	Std. Dev.	Std. Err.
Day 11, Competitive	15	4.200	.941	.243
Day 11, Pers Info	15	2.600	.828	.214
Day 21, Competitive	15	2.533	1.060	.274
Day 21, Pers Info	15	2.333	.617	.159
Day 3, Competitive	15	2.800	.862	.223
Day 3, Pers Info	15	2.533	.743	.192

ANOVA Table for Attractiveness

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Day of Cycle			7.6		<.0001	20.954	.992
Video			10.6		.0002	14.656	.979
Day of Cycle * Video		9.3			.0025	12.841	.906
Residual		61.2					

In the Dutton & Aron bridge study, elevated physiological arousal of the male participants was presumed to be the root cause of a greater attraction to a young woman waiting at the end of the high rope bridge as compared to the perceived attraction from the non-elevated physiological arousal that followed crossing a stable concrete bridge. But maybe the perceived attraction had to do with elements other than physiological arousal. Dutton & Aron also used only male participants. As an extension of their study, suppose that male and female participants (a non-manipulated “independent variable”) are given one of three levels of epinephrine [None (saline), Small, Large] to produce varying levels of physiological arousal. After 5 minutes (to allow the drug to have its effect), participants were shown a picture of the face a person who had previously been judged by other people to be moderately attractive (4 on a 7-point scale of attractiveness). For the male participants, the picture was of a female face. For the female participants, the picture was of a male face. Each participant rated the attractiveness of the target face on a 7-point scale. Complete the analysis of this 2x3 independent groups design and interpret the results as completely as you can, including the sort of explanation you’d include in a Discussion section.

Means Table for Attraction Score

Effect: Gender * Epinephrine

	Count	Mean	Std. Dev.	Std. Err.
Female, Large	8	4.875	1.126	.398
Female, None	8	4.000	.926	.327
Female, Small	8	4.500	1.195	.423
Male, Large	8	6.250	.707	.250
Male, None	8	4.000	1.309	.463
Male, Small	8	5.250	1.035	.366

ANOVA Table for Attraction Score

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Gender				5.282	.0266	5.282	.608
Epinephrine				8.608	.0007	17.217	.967
Gender * Epinephrine				1.663	.2018	3.326	.319
Residual			1.140				

Two-Factor Repeated Measures Designs and Two-Factor Mixed Designs:

Although we will not actually learn to compute analyses of two-factor repeated measures designs or two-factor mixed designs, it is important that you become familiar with the basic notions involved in these types of designs.

Counterbalancing Completely Within Designs: Implications for Number of Participants

First of all, you should recognize that regardless of the number of factors, if any factor is a repeated measure, we must counterbalance that factor. As you have already learned, when dealing with a repeated factor, the number of levels involved is crucial. When we are counterbalancing 4 or fewer levels, we should definitely use complete counterbalancing. When we're dealing with 6 or more levels, we would most likely use incomplete counterbalancing. With 5 levels, we could easily go either way.

That's clear for the single-factor repeated measures design, but what about the two-factor repeated measures design? Suppose that we're dealing with a simple 2x2 repeated measures design. What does this design imply for the participant's experience? As you can see in the figure below, the 2x2 repeated measures design implies that the participant will actually provide a response to each of 4 unique combinations of the two factors:

	a1	a2
b1		
b2		

OK, so how would you go about counterbalancing this two-factor design? The trick is to squish the two-factor design to make it look like a single-factor design. In this case, we could translate the 2x2 repeated measures design into a single-factor design with 4 levels. Thus, we would use complete counterbalancing, which would mean that we'd need to run multiples of 24 participants. If we wanted to have a minimum of 30 scores per cell (for reasons of power), we would need to run 48 participants. Can you see why? Note that the 48 participants would generate 48 scores in each of the 4 cells, for a total of 192 pieces of data. That's the efficiency of the repeated measures design.

Suppose that we had a 3x3 completely within (repeated measures) design, as seen below:

	a1	a2	a3
b1			
b2			
b3			

How would you go about counterbalancing the repeated factors in this design? First, recognize that in this design a participant would experience 9 unique conditions. Thus, we could think of it as a single-factor repeated measures design with 9 levels. That's way too many to completely counterbalance, so we would use incomplete counterbalancing. Because of the odd number of levels, we would run in multiples of 18 participants. If we wanted to

have a minimum of 30 scores per cell, we would need to run 36 participants. Once again, note the efficiency of the repeated measures design. The 36 participants would generate a score in each of the 9 cells, for a total of 324 pieces of data.

Counterbalancing Mixed Designs: Implications for Number of Participants

In a mixed design, one factor is between (independent groups) and the other is within (repeated measures). Once you recognize the implications of such a design, the process of counterbalancing should also be clear.

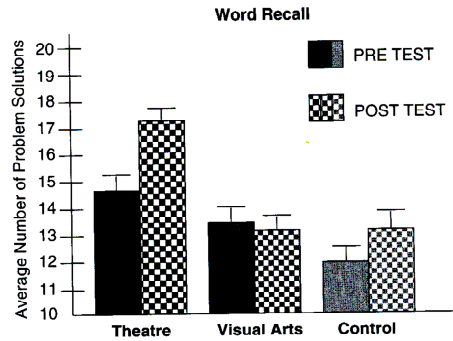
Suppose, for instance, that we were dealing with a simple 2x2 mixed design, with Factor A between and Factor B within. You need to recognize that an easy way to think about the design is to see it as a series of two separate repeated measures experiments, each with two levels. That is, for level 1 of Factor A, you would be conducting a repeated measures experiment with two levels. Then, for level 2 of Factor A, you would be doing the same thing. With only two levels of the repeated factor, you could get away with as few as 4 total participants (2 to counterbalance the two levels of B for level 1 of Factor A and 2 to counterbalance the two levels of B for level 2 of Factor A). If I wanted to get a minimum of 30 scores in each cell, I'd need to actually run 60 participants. I could run 30 people through level 1 of Factor A (15 with the ordering b1 -> b2 and 15 with the ordering b2 -> b1). I would need to run another 30 people who receive level 2 of Factor A (15 with b1 -> b2 and 15 with b2 -> b1).

Think of a 3x3 design. Suppose that Factor A is a between-groups factor and Factor B is a within-groups factor. Thus, I would think of the overall study as a series of 3 repeated measures experiments with 3 levels of the repeated factor. With 3 levels, I could completely counterbalance with 6 orders (3!). The minimum number of participants I could run in this study would be 18 (6 participants, one for each order, at each level of the between factor). Suppose that I wanted to have a minimum of 30 scores in each cell? I would need to run a total of 90 participants, 30 participants would allow me to completely counterbalance the 6 levels of Factor B for level 1 of Factor A, then I would need another 30 participants for level 2 of Factor A and another 30 participants for level 3 of Factor A.

OK, suppose that you're dealing with a 4x5 design. Suppose, furthermore, that you want to have a minimum of 30 scores per cell.

- a. If both factors are repeated measures, how many participants would you need in all?
- b. If Factor A was between and Factor B was within, how many participants would you need in all?
- c. If Factor B was between and Factor A was within, how many participants would you need in all?

Noice and Noice (2006) recently published an article entitled “What studies of actors and acting can tell us about memory and cognitive functioning.” In their abstract they state: “The art of acting has been defined as the ability to live truthfully under imaginary circumstances...we first discuss how large amounts of dialogue, learned in a very short period, can be reproduced in real time with complete spontaneity.” In a 2x3 mixed design (Noice, Noice, & Staines, 2004), older adults (65 to 90 years of age) agreed to participate in a study involving instruction to improve cognition and were randomly assigned to one of three conditions. In the Theatre condition, participants were given acting lessons that didn’t focus specifically on memory. In the Visual Arts condition, participants were given an art appreciation class. In the Control condition, participants were given no treatment. Participants received a memory test before and after treatment. What effects would you expect to emerge from the analysis (main effects and interaction)? Interpret the results of the Word Recall test as though you were writing a Discussion section.



Dr. Jaime Klutz was interested in studying developmental differences in problem solving, so he conducted the following experiment. He had two groups of children (5-year-olds and 7-year-olds) solve two different types of puzzles from the EZ Puzzle company (10-piece puzzle and 30-piece puzzle). He used a mixed design with participant's age as a between (independent groups) variable and type of puzzle as a within (repeated measures) variable.

a. If he wanted to have 30 data points per cell (or condition), how many participants would he need in all?

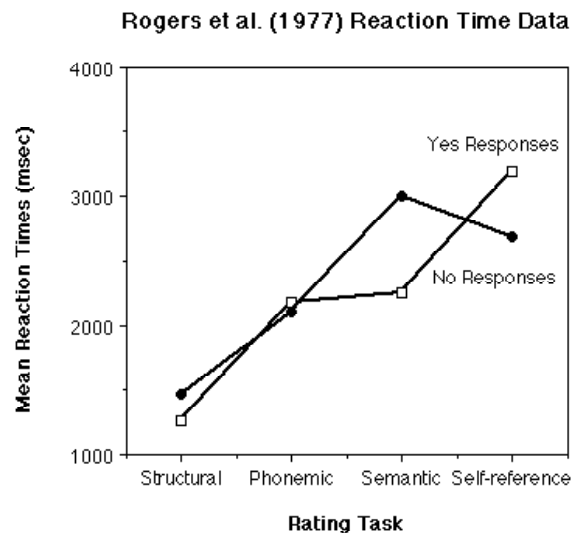
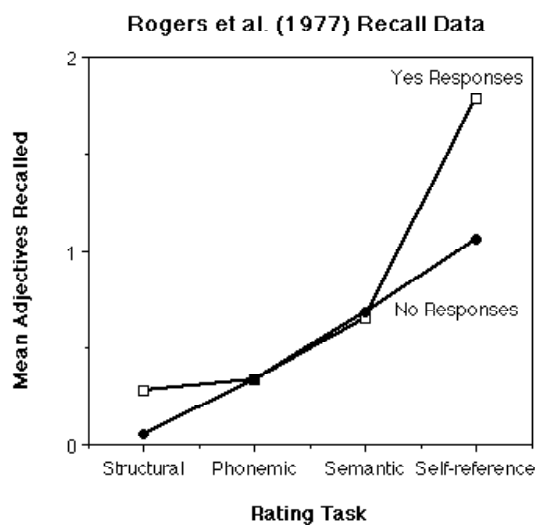
b. The children from both age groups came into the laboratory in a random fashion and were then asked to solve the 10-piece puzzle followed by the 30-piece puzzle. As his dependent variable, Dr. Klutz used the time it took participants to solve the puzzles. Dr. Klutz was disappointed to discover that there were absolutely no significant differences in his study. On average, both age groups solved the 10-piece puzzle in 20 seconds (2 secs per piece) and the 30-piece puzzle in 60 seconds (2 secs per piece). He really thought that his study was going to show significant results, and in a very despondent state he comes to ask your advice. What would you tell the good Dr. Klutz about his experiment?

c. Dr. Klutz brightens considerably after hearing your advice. He thanks you and indicates that he's given the experiment some further thought and has decided that it would be more powerful to do the study as a completely within (repeated measures) design. Not only would it be more powerful, but it would be more efficient. How many participants would he actually need? What do you think about his plan?

Rogers, T.B., Kuiper, N.A., and Kirker, W.S. (1977)
Self-reference and the Encoding of Personal Information

Rogers et al. were interested in the effect of self-reference on memory. They devised a two factor repeated measures study (Experiment 1) in which participants answered one of 4 different questions about the 40 target adjectives presented. The four questions defined the four conditions of the orienting task factor (Structural, Phonemic, Semantic, and Self-Reference). For the Structural condition, participants rated whether or not the 10 target adjectives were in big letters or not (5 were in big letters and 5 were in small letters). In the Phonemic condition, participants rated whether or not the 10 adjectives rhymed with a word (5 rhymed with the adjective and 5 did not). In the Semantic condition, participants rated whether or not each of the 10 adjectives meant the same as a word (5 were synonyms of the adjective and 5 were unrelated). Finally, in the Self-reference condition, participants rated whether or not the adjective described themselves (some number were thought to describe the participant and some were thought not to describe the participant). The items were presented such that in the 40 items seen by the participant, the particular conditions varied randomly, so that in the early part of the list some of the items were Semantic-Yes, some were Phonemic-No, some were Structural-Yes, some were Phonemic-Yes, etc. The participant's responses and reaction times to respond were recorded. After the acquisition phase was the test phase of the experiment in which each participant was given a blank piece of paper and asked to recall as many of the 40 adjectives as possible.

In the analysis, Rogers et al. created a two-factor analysis with both factors within participants (the four different orienting tasks at acquisition and whether participants had responded "yes" or "no" to an adjective). In effect, then, there were two DVs (items recalled and reaction time). Graphs of the actual results of their experiment are seen below.



Simply given the graphs, what would you have predicted the outcomes of the ANOVAs to be? How would you interpret the results of the two analyses?