

1. We've been talking about confounds, so first tell me how you would define a confound, including the kind of validity attacked by a confound. Then, as you know, I've indicated that a study in which one examines a non-manipulated characteristic of a participant is not a true experiment. One reason for that position is that such a study is potentially confounded. Imagine, if you will, a study that looks at the impact of intelligence on learning a list of non-words (e.g., the CVC stimuli used by Ebbinghaus). Participants fall into one of three intelligence groups (Average Intelligence, Above Average Intelligence, and High Intelligence), though you need not worry about the operational definition of those terms to answer the question. Let's suppose that the results indicate that people in the High Intelligence group exhibited significantly better memory than those in the Average Intelligence group. Why might you describe such a study as confounded? Be *very* explicit! [10 pts.]

**A confound is a design problem in an experiment that opens the door to multiple interpretations of any effects that emerge. Thus, confounds undermine internal validity. By their very nature, studies that employ a non-manipulated characteristic of the participants are confounded. Consider the following results:**

	Average IQ	Above-average IQ	High IQ
Mean Recall Score	18	25	29

**As long as you can hypothesize a reasonable third variable, you will have articulated the confound in such a study. That is, one *could* reasonably conclude that people with High IQ recall significantly more words than people with Average IQ. However, one could not conclude that the difference was due to the IQ difference. For example, people with High IQ may tend to read more routinely than people with Average IQ, which may lead them to better recall the words. Or, it may be that people with High IQ tend to do better in school, which then rewards them for remembering words, so that they then do better than people with Average IQ, who may not do as well in school.**

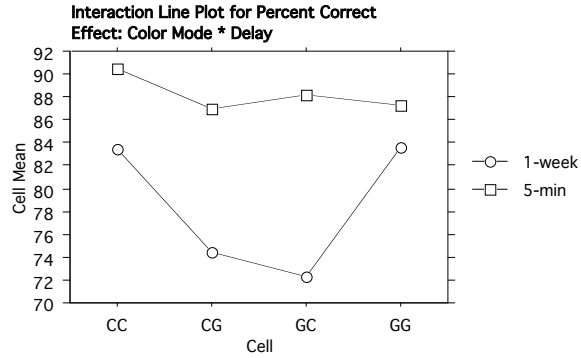
2. In an encoding specificity study, Suzuki and Takahashi (1997) were interested in whether pictures were better remembered if the test pictures were similar to the acquisition pictures in terms of color. Let's treat their study as a completely independent groups design. That is, participants studied a set of pictures that were either all in color or were all in gray-scale. Then, at test half of the participants who got the pictures in color saw a test set of pictures in which the pictures were all in color (half old and some new). The other half of the participants who got the pictures in color saw a test set of pictures in which the pictures were all in gray-scale (half old and half new). The same procedure was used for participants who saw the original pictures in gray-scale (that is half got color test pictures and half got gray-scale test pictures). Thus, we can conceive of one factor in their study as the four types of Color Mode (CC, CG, GC, GG; C = Color and G = Gray, with the first letter acquisition and the second letter test). In addition, they were interested in the impact of delay. Let's suppose that half of the participants were tested after a 5-min delay and half of the participants were tested after a 1-week delay, which defines the Delay factor. The DV is the percent of the old test items correctly identified as old. Thus, this is a 2x4 independent groups design as I've portrayed it here. Complete the source table below and analyze these data as completely as you can. How would you discuss your results (in a Discussion section)? [20 pts.]

**ANOVA Table for Percent Correct**

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Color Mode	3	804.865	268.288	15.211	<.0001	45.634	1.000
Delay	1	2271.760	2271.760	128.804	<.0001	128.804	1.000
Color Mode * Delay	3	540.948	180.316	10.224	<.0001	30.671	.999
Residual	88	1552.083	17.637				

**Means Table for Percent Correct**  
Effect: Color Mode \* Delay

	Count	Mean	Std. Dev.	Std. Err.
CC, 1-week	12	83.417	4.078	1.177
CC, 5-min	12	90.417	4.358	1.258
CG, 1-week	12	74.500	2.611	.754
CG, 5-min	12	87.000	3.814	1.101
GC, 1-week	12	72.333	7.215	2.083
GC, 5-min	12	88.167	2.980	.860
GG, 1-week	12	83.667	3.420	.987
GG, 5-min	12	87.250	3.388	.978



$$HSD = 4.4 \sqrt{\frac{17.637}{12}} = 5.33$$

There is a significant main effect of color mode,  $F(3,88) = 15.211$ ,  $MSE = 17.637$ ,  $p < .001$ . There is also a significant main effect of delay,  $F(1,88) = 128.804$ ,  $p < .001$ . There is also a significant interaction between color mode and delay,  $F(3,88) = 10.224$ ,  $p < .001$ . Post hoc analyses using Tukey's HSD indicated that the interaction was due to the fact that at five minutes, the effects of color mode were not significant (each group performed fairly well). However, after a 1-week delay, people who saw the stimuli in color on both occasions (CC) ( $M = 83.417$ ) and people who saw the stimuli in gray on both occasions ( $M = 83.667$ ) correctly identified significantly more items those who first saw the stimuli in color and then in gray ( $M = 74.5$ ) and those who first saw the stimuli in gray and then in color ( $M = 72.333$ ). No other differences at one week were significant. Thus, there is some basic support for the encoding specificity hypothesis at a one-week delay, but not at a five-minute delay.

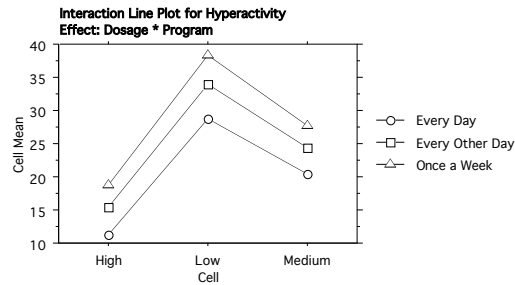
3. In a study of hyperactivity among elementary school boys, nine groups of participants were randomly selected from a school population of ADHD, 7-year-old boys. (ADHD is Attention Deficits with Hyperactivity, and left untreated, it can prevent a child from attending to incoming learning stimuli and may also create major disruptions in the classroom.) The researcher wanted to study the classroom effects on the activity levels of the participants. Both the drug Ritalin as well as a behavior modification program served as factors. The drug was varied from a Low dosage, to a Moderate dosage, to a High dosage of Ritalin. The behavior modification program consisted of giving the child ten tokens to start the day and then taking away a token for each hyperactive infraction. The tokens that were saved could then be exchanged for some valued prize. The behavior mod program was varied from using the program Every Day, to the program using the program Every Other Day, to using program Once a Week. After 4 weeks, all the children were evaluated for hyperactivity and were assigned scale scores ranging from a possible low of 0 (no indication of hyperactivity) to a high of 40 (extreme hyperactivity). Complete the source table below and interpret the data from this study as completely as you can. [20 points]

**ANOVA Table for Hyperactivity**

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Dosage	2	1549.852	774.926	1046.150	<.0001	2092.300	1.000
Program	2	296.963	148.481	200.450	<.0001	400.900	1.000
Dosage * Program	4	5.481	1.370	1.850	.1633	7.400	.446
Residual	18	13.333	.741				

Means Table for Hyperactivity  
Effect: Dosage \* Program

	Count	Mean	Std. Dev.	Std. Err.
High, Every Day	3	11.333	1.528	.882
High, Every Other Day	3	15.333	.577	.333
High, Once a Week	3	18.667	.577	.333
Low, Every Day	3	28.667	1.155	.667
Low, Every Other Day	3	34.000	1.000	.577
Low, Once a Week	3	38.333	.577	.333
Medium, Every Day	3	20.333	.577	.333
Medium, Every Other Day	3	24.333	.577	.333
Medium, Once a Week	3	27.667	.577	.333



$$HSD = 3.61 \sqrt{\frac{.741}{9}} = 1.04$$

	Hyper Score		Hyper Score
Low	33.67	Every Day	20.11
Medium	24.11	Every Other Day	24.55
High	15.11	Once a Week	28.22

There is a significant main effect of dosage,  $F(2,18) = 1046.15$ ,  $MSE = .741$ ,  $p < .001$ . There is also a significant main effect of program,  $F(2,18) = 200.45$ ,  $p < .001$ . However, there was no significant interaction,  $F(4,18) = 1.85$ ,  $p = .163$ . Post hoc tests using Tukey's HSD indicate that the main effect of Dosage was due to the fact that people who received Low dosage obtained significantly higher hyperactivity scores ( $M = 33.67$ ) than those on the Medium ( $M = 24.11$ ) or High dosage ( $M = 15.11$ ). People on the Medium dosage obtained significantly higher hyperactivity scores than those on the High dosage. The main effect of Program was due to the fact that people who received the drug once a week obtained significantly higher hyperactivity scores ( $M = 28.22$ ) than people who received the drug every other day ( $M = 24.55$ ) or every day ( $M = 20.11$ ). People who received the drug every other day had significantly higher hyperactivity scores than those who received the drug every day.

4. In independent groups ANOVAs, such as the one in Problem #3: [10 pts.]

a. If you were not provided the  $MS_{Error}$  (.741 in this case), you would still have been able to complete the source table, how could you have arrived at  $MS_{Error}$ ?

As long as you were provided the variances for each group (or the standard deviations, which you could square to turn them into variances), you could take the mean of the group variances. The mean of the group variances is the  $MS_{Error}$ . In this case, you'd square each standard deviation and add them together, then divide by 9 ( $1.528^2 + .577^2 + .577^2 + \dots + .577^2$ )/9 = .741.

b. What is the function of  $MS_{Error}$  in the analysis? That is, what population parameter is it intended to estimate?

$MS_{Error}$  estimates the population variance ( $\sigma^2$ ).

c. Suppose that you re-computed the ANOVA as a single-factor analysis on the Dosage factor. What would that source table look like?

SOURCE	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Dosage	2	1549.85	774.93	58.89
Error	24	315.78	13.16	
Total	26	1865.63		

5. Well, of course you expect to tell me about the impact of various designs on the number of participants needed. For this problem, assume that we want to have a minimum of 35 pieces of data in each cell/condition. [10 pts]

Design	# of participants	# of pieces of data
A 3x6 completely between (independent groups) design	$3 \times 6 \times 35 = 630$	<b>630</b>
A 3x7 completely within (repeated measures) design	$3 \times 7 \times 2 = 42$	<b>882</b>
A 3x7 mixed design, with the first factor between (independent groups) and the second factor within (repeated measures)	RM: $14 \times 3 = 42$ $3 \times 42 = 126$	<b>882</b>
A 3x7 mixed design, with the first factor within (repeated measures) and the second factor between (independent groups)	RM: $6 \times 6 = 36$ $7 \times 36 = 252$	<b>756</b>
A 4x8 mixed design, with the first factor between (independent groups) and the second factor within (repeated measures)	RM: $8 \times 5 = 40$ $4 \times 40 = 160$	<b>1280</b>

6. Suppose the data in Question 3 (Ritalin and Behavior Modification) had yielded the source table below. Tell me what you'd do next. [5 pts.]

**ANOVA Table for Hyperactivity**

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Dosage	2	27.185	13.593	1.266	.3060	2.531	.232
Program	2	19.185	9.593	.893	.4268	1.786	.175
Dosage * Program	4	104.593	26.148	2.434	.0849	9.738	.572
Residual	18	193.333	10.741				

**Nothing significant, so increase power. Increase the treatment effect of Dosage (consider adding a placebo—no Ritalin—control group). Increase the treatment effect of Program (consistent with no Ritalin, a group that never receives Ritalin; possibly a group that receives Ritalin more frequently, such as twice a day). Consider increasing the study to beyond four weeks. Decrease the individual differences (possibly use only males, or people with only very high ADHD). Decrease the random variability (clarify instructions, clarify the operational definition of hyperactivity). Increase the sample size.**