

1. Briefly define the term *confound*. Then, using a *very explicit* example of practice effects (maybe even with numbers?), illustrate why conducting a repeated measures experiment without counterbalancing makes the study confounded. How does counterbalancing eliminate the confound? [10 pts]

A confound is a threat to internal validity in which some systematic but unplanned factor may well explain our results. In general, a confound emerges when we treat our conditions differently in more than one way. That is, the conditions may differ because of the IV, but they also may differ because of another factor.

Let's model a practice effect as a simple additive effect (+5). That means that the second time that a person is tested, that person's score will improve by 5 simply because of being tested a second time. Below you'll see the data from 6 participants. Note that they have different initial scores (because of individual differences). However, if we do nothing to them, but simply test them twice, their scores will improve (due to the practice effect) as seen below:

Participant	Time 1	Time 2
1	4	9
2	8	13
3	3	8
4	7	12
5	5	10
6	6	11

So, that's what our data would look like if all that happened was practice. But suppose that we have two conditions (Control and Experimental). Let's model the treatment effect as a simple additive constant (+2), so that the Experimental Condition would go up by 2 and the Control Condition would go up by 0 (no treatment). Adding that information to the data above would yield a data set that looks like this:

Participant	Control (Time 1)	Experimental (Time 2)
1	4	11
2	8	15
3	3	10
4	7	14
5	5	12
6	6	13

There would now be a "large" difference (7) between the two conditions (5.5 for Control and 12.5 for Experimental). However, that difference emerged not because of the treatment difference alone, but also because of the practice effect (and no counterbalancing).

If we were to appropriately counterbalance this design (e.g., have Participants 1, 2, and 3 get the Control Condition first and the Experimental Condition second; and have

Participants 4, 5, and 6 get the Experimental Condition first and the Control Condition second), the data would look very different. For example, consider Participant 4. Participant 4 would have a Treatment Condition score of 9 (initial state of 7 + 2 for the treatment + 0 for practice effect) and a Control Condition score of 12 (initial state of 7 + 0 for the treatment + 5 for practice effect).

Participant	Control (Time 1)	Experimental (Time 2)
1	4	11
2	8	15
3	3	10
4	12	9
5	10	7
6	11	8

Note that now the difference between the Control Condition and the Experimental Condition is more modest. In fact, the difference (2) reflects only the treatment effect. The practice effect is now distributed equally over the two conditions. Note that one side effect of counterbalancing is to increase the variability within each condition (which will serve to increase the error term).

2. OK Jeff, here's an interesting study. 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. [15 pts]

ANOVA Table for Attractiveness

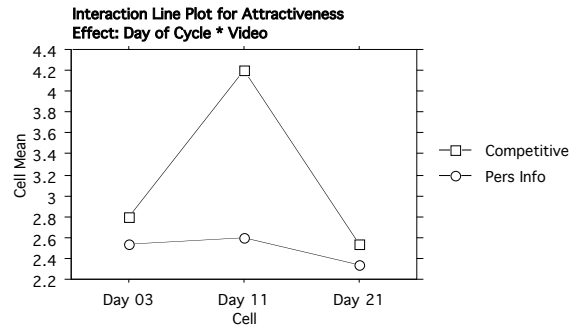
	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Day of Cycle	2	15.267	7.633	10.477	<.0001	20.954	.992
Video	1	10.678	10.678	14.656	.0002	14.656	.979
Day of Cycle * Video	2	9.356	4.678	6.420	.0025	12.841	.906
Residual	84	61.200	.729				

Means Table for Attractiveness

Effect: Day of Cycle * Video

	Count	Mean	Std. Dev.	Std. Err.
Day 03, Competitive	15	2.800	.862	.223
Day 03, Pers Info	15	2.533	.743	.192
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

The interaction is significant, $F(2,84) = 6.42$, $MSE = .729$, $p = .003$. Thus, the first step would be to create a graph of the means as seen below. It appears that there is little difference between Type of Video at Day 3 and at Day 21. However, at Day 11 it appears that the Competitive Video yields a higher rating than the Personal Information video.



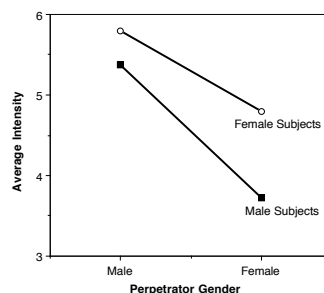
The next step would be to compute Tukey's HSD as a post hoc test.

$$HSD = 4.14 \sqrt{\frac{.73}{15}} = .91$$

I can now say that the attractiveness ratings at Day 3 and Day 21 did not differ significantly. However, at Day 11 (a high fertility day) women rated the male in the Competitive video as more attractive ($M = 4.2$) than the male in the Personal Information video ($M = 2.6$).

3. In a study by Baron, Burgess, and Kao (1991), male and female participants read accounts of stories that included a description of a sexist act perpetrated by either a male or a female against a female. The 193 participants described the perpetrator in a way that could be scored for intensity of sexist behavior. The displayed sexist behavior was rated 1 for *slightly displayed* to 7 for *extremely displayed*. Part of their *Results* section reads:

Perpetrator gender and participant gender main effects were both significant. Female participants, compared with male participants, gave more intense ratings to both male and female perpetrators...: $F(1,189) = 5.06$, $p < .03$...Furthermore, male perpetrators were seen as displaying more intense gender bias than female perpetrators: $F(1,189) = 15.97$, $p < .001$. The interaction between participant gender and perpetrator gender was nonsignificant... $p < .34$...These results can be seen in the figure below:



Briefly interpret the meaning of these results, as you would in a discussion section. [10 pts]

Given the non-significant interaction, I'd focus on the two main effects, which can both be interpreted without any post hoc test because each has only two levels. Thus, Male perpetrators received higher sexist scores than Female perpetrators. Female participants rated the acts as more sexist than Male participants.

Thus people (male and female) tend to think that when a male perpetrator engages in a sexist act against a female it's worse (rated more sexist) than when the same act is performed by a female perpetrator. This effect may reflect a reaction against males treating females in a sexist fashion. Interestingly, female perpetrators can get away with the same act and not be viewed as behaving in quite the same sexist fashion.

Female participants also seem less tolerant of sexist behaviors directed against a woman (whether by a male or a female) than are male participants. That is, regardless of the gender of the perpetrator, females rate the sexist act as more sexist.

4. 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." So, how well do you think that you'll do on this exam? 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. [15 pts]

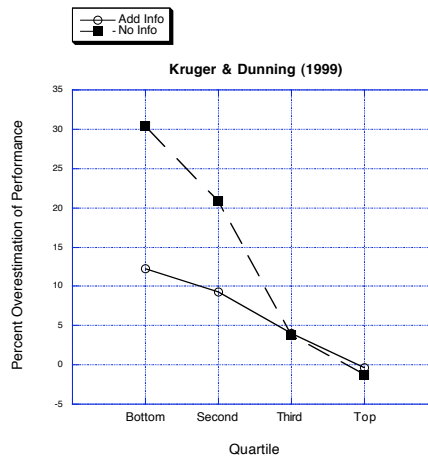
ANOVA Table for Estimate

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Quartile	3	6184.6	2061.5	420.7	<.0001	1251.531	1.000
Add Info	1	1008.2	1008.2	205.8	<.0001	204.020	1.000
Quartile * Add Info	3	1308.1	436.0	89.0	<.0001	264.708	1.000
Residual	72	355.8	4.9				

Means Table for Estimate
Effect: Quartile * Add Info

	Count	Mean	Std. Dev.	Std. Err.
Bottom, Add Info	10	12.300	3.164	1.001
Bottom, No Info	10	30.400	4.300	1.360
Second, Add Info	10	9.300	1.418	.448
Second, No Info	10	20.900	2.079	.657
Third, Add Info	10	4.000	1.155	.365
Third, No Info	10	3.700	1.337	.423
Top, Add Info	10	-.300	.823	.260
Top, No Info	10	-1.300	.949	.300

Once again, because the interaction is significant, that's where you would focus your attention. First, I would graph the data, as seen below:



It appears that there is no difference between **Additional Information** and **No Information** for the people in the top two quartiles. However, for people in the lower two quartiles, it appears that the **No Information** groups produce higher overestimation of performance. To determine what's actually going on, however, I'd need to compute a post hoc test, as seen below:

$$HSD = 4.4 \sqrt{\frac{4.9}{10}} = 3.1$$

With the post hoc test, our visual inspection of the data is confirmed. That is, the people who did worst on the test (**Bottom and Second Quartile**) actually tended to overestimate their performance more when they were given no additional information and to overestimate their performance less then they were given additional information. However, people who did better on the exam (**Third and Top Quartile**) tended to overestimate their performance less and did not differ whether given additional information or not.

5. Suppose that you are interested in studying the impact of a drug on maze learning in rats. Because you are unsure of the level at which the drug might be most effective, you decide to use 4 different levels of the drug. First of all, tell me (in very general terms) how you would determine the 4 levels that you would use in your experiment. You want to avoid carry-over effects of the drug, so each rat will be exposed to only one level of drug. Because you think that the drug may lead to better performance on some mazes than on other mazes, you want to run each of your rats through three different mazes (Easy, Moderate, and Difficult). Thus, this experiment would be a 4x3 mixed design. In *very explicit fashion*, tell me how you would run this experiment, including the minimum number of rats you'd need for your study and how many you'd actually use, the procedure you'd use, etc. [20 pts]

I would choose four levels of drug that would include a placebo control (No Drug). The other three levels (Low, Medium, High) would be determined based on some pre-testing. I would want to choose the drug levels such that they would enhance the power of the study, but would not be outrageous (e.g., High level is ridiculously high).

I would construct three different mazes that were equal in length (from start to finish) but differed in difficulty (maybe use more blind alleys in the more difficult mazes). The three mazes would be Easy, Moderate, and Difficult. As a repeated factor, all rats would do each of the three mazes, but in a different order (EMD, EDM, MED, MDE, DME, DEM). If I felt that I could run the rats through the maze sufficiently quickly (and leave them motivated to run the maze), then I'd administer the drug, wait X minutes (determined by pre-testing) for the drug to take effect, then run them through the three mazes. If I didn't feel that I could run them through all 3 mazes at one time, I'd conduct the study over the course of 3 days for each rat (administer drug, wait, test in maze).

For power considerations, I would probably go with 30 rats in each of the 4 drug conditions, for a total of 120 rats. Thus, I would use each of the 6 orders 5 times each.

6. 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. [10 pts]

ANOVA Table for Score

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Major	3	647.344	215.781	21.159	<.0001	63.478	1.000
Gender	1	3.781	3.781	.371	.5483	.371	.088
Major * Gender	3	37.344	12.448	1.221	.3238	3.662	.278
Residual	24	244.750	10.198				

Means Table for Score

Effect: Major * Gender

	Count	Mean	Std. Dev.	Std. Err.
English, Female	4	37.750	1.708	.854
English, Male	4	37.750	2.630	1.315
Math, Female	4	43.000	5.598	2.799
Math, Male	4	39.000	5.292	2.646
Phys-Ed, Female	4	35.500	2.082	1.041
Phys-Ed, Male	4	34.750	.957	.479
Voc-Ed, Female	4	27.750	1.708	.854
Voc-Ed, Male	4	29.750	2.062	1.031

Because the interaction is not significant, I would focus my attention on the main effects. There is no significant effect of Gender, but there is a significant effect of Major, $F(3,24) = 21.16$, $MSE = 10.20$, $p < .001$. The appropriate means would be:

English	Math	Phys-Ed	Voc-Ed
37.75	41	35.1	28.75

With 4 levels of the factor, I would need to compute a post hoc test:

$$HSD = 3.9 \sqrt{\frac{10.2}{8}} = 4.4$$

For these data, people who majored in Math, English, and Phys-Ed scored higher than those in Voc-Ed. People who majored in Math did better than people who majored in Phys-Ed. No other differences were significant. Note, also, that neither of these factors was manipulated, so you'd be unable to make causal claims.