

1. Dr. I. P. Freeley was interested in replicating the Middlemist et al. study (effect of invasion of personal space on time to urinate), but using a repeated measures design for greater power. For several days, he had a rotating cadre of confederates (so that the students wouldn't think that someone was stalking them) follow a set of students enrolled in an introductory psychology class. Whenever one of these students would enter a restroom to urinate, a confederate would check to ensure that no one else was using a urinal. If the participant were alone at one of the urinals, the confederate would either: 1) go to the urinal immediately next to the student (Near Stall); 2) go to a urinal one urinal away from the student (Distant Stall); or would simply go to the mirror and comb his hair (Alone). The dependent variable, as in the Middlemist study, was the time (in minutes) between when the unwitting participant unzipped his pants and when he began to urinate (micturate). Complete the source table below and interpret these data as completely as you can. [10 pts]

ANOVA Table for Distance

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	1.228	.136				
Category for Distance	2	2.282	1.141	111.217	<.0001	222.433	1.000
Category for Distance * Subject	18	.185	.010				

Means Table for Distance

Effect: Category for Distance

	Count	Mean	Std. Dev.	Std. Err.
Alone	10	.550	.227	.072
Distant Stall	10	.560	.196	.062
Near Stall	10	1.140	.259	.082

There is a significant effect of the Presence/Distance of another person, $F(2,18) = 111.217$, $MSE = .01$, $p < .001$. Post hoc analysis was conducted using Tukey's HSD:

$$HSD = 3.61 \sqrt{\frac{.01}{10}} = .11$$

Men with another man at the near stall took significantly longer to begin urination ($M = 1.14$) than men with another man at a distant stall ($M = .56$) or men who were alone in the restroom ($M = .55$).

2. Mook argues that external validity is not always the purpose behind psychological research. For each of the studies below, indicate why the study is not externally valid, then why it's not a concern, given the intentions of the researcher(s). [10 pts]

Use Mook article to respond to this question.

Study	Why not externally valid	Why lack of EV is not a concern
Argyle (glasses and intelligence)		
Harlow (infant monkeys and drive reduction theory)		

Hecht (dark adaptation)		
Brown & Hanlon (parental role in grammar acquisition)		

3. Correlational designs do not allow you to make casual claims. Why not? Be very explicit about the difficulty of claiming that changes in one of the two variables in a correlational study *causes* the related changes observed in the second variable. We also discussed the shortcomings of using non-manipulated characteristics of the participants as “independent variables” in an experiment. How is this class of variable related to the notion of correlational designs? [10 pts]

In class, we discussed the three possibilities for an observed significant correlation between two variables (X and Y). One possibility is that X does cause Y . Another possibility is that Y causes X . And, finally, there’s the possibility of a third variable (Z) that might cause both X and Y . And of course, we’re not talking about a host of potential third variables ($Z_1, Z_2, Z_3\dots$).

Each person carries around a whole host of characteristics (resulting from nature and nurture). Thus, when a researcher looks at such characteristics as an “independent variable,” he or she runs the risk that the characteristic is not actually having a causal impact on the dependent variable. Instead, some “third variable” that is related to the nominal characteristic may be actually causing the behavior observed in the dependent variable. For example, suppose that a researcher were interested in studying the impact of IQ on performance in a public speaking task. It may well be the case that low IQ causes people to perform more poorly on such a task and high IQ causes people to perform better on the task. However, it may be that other variables related to IQ may be the causal agents. For instance, people with low IQ may have lower self-esteem, which may lead them to be more uncomfortable when speaking in public. Or it may be that people with high IQ are asked to give their opinions more often, so they have more experience speaking in public.

In essence, then, studies using non-manipulated characteristics of participants as “independent variables” are not true experiments, but actually correlational, because the “independent variable” is not being manipulated. And then, of course, there’s the definite possibility of a third variable potentially causing any observed effects.

4. Two researchers were interested in studying the effects of reward magnitude on performance. Both researchers used introductory psychology students as participants, the same total number of participants (21), the same type of reward and reward magnitudes (\$1, \$5, \$20), the same apparatus, the same task, and the same performance measure (DV). One researcher used an independent groups design and, on the basis of the results, cannot reject the null hypothesis (that reward has no effect on performance). The other researcher used a repeated measures design and found a statistically significant effect of reward magnitude—larger rewards lead to better performance. Assume that neither study has a major flaw (e.g., repeated measures design is properly counterbalanced, random assignment to conditions). There are two fundamental reasons why the two researchers might have reached different conclusions. One reason concerns the sensitivity of the test of the null hypothesis. The other reason concerns the nature of the participant’s experience in the two studies. Provide me with a clear explanation of the two reasons for the different results that the two researchers obtained. Would you trust the results of one study more than the other? Why? Finally, complete the source tables for the two experimenters seen below. [10 pts]

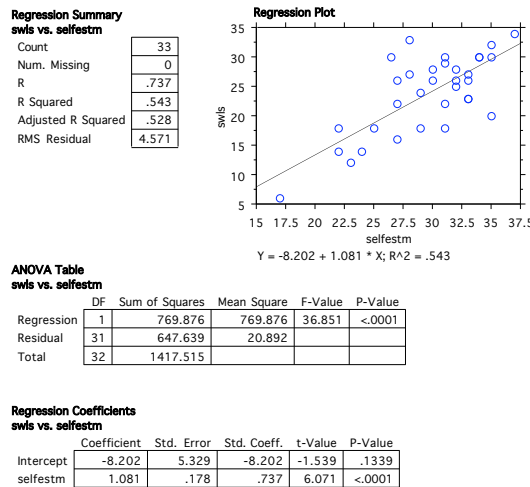
Independent Groups Design ($F_{crit} = 3.55$):

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Treatment	2	28	14	3.5
Error	18	72	4	
Total	20	100		

Repeated Measures Design ($F_{crit} = 3.23$):

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Subject	20	100	5	
Treatment	2	20	10	5
Error (Subj x Treat)	40	80	2	
Total	62	200		

5. In your first lab, there were a number of different personality measures. One was the Rosenberg Self-Esteem Scale (*selfestm*) and another was the Satisfaction with Life Scale (*swls*). Had you correlated those two measures, you would have seen an output like the one below. Interpret the output below as completely as you can. If a person had a self-esteem score of 30, what would you predict that person's SWLS score to be? What proportion of variance do these two measures share? If you were to talk about this result in a Discussion, what might you say about the relationship? [10 pts]



There is a significant positive correlation between the Satisfaction with Life Scale and the Rosenberg Self-Esteem Scale, $r(31) = .737, p < .001$. Using the regression equation, a self-esteem score of 30 would lead you to predict a SWLS score of 24.2. The two variables share 54% of their variability ($r^2 = .543$).

Of course, in a discussion, I'd be very careful to avoid any language that sounded the least bit causal. I could say that people with lower self-esteem also experienced lower satisfaction with life and that people with higher self-esteem experienced higher satisfaction with life. I might propose a mechanism by which the relationship might emerge (directly causal or indirectly through some other variable or variables), but would do so tentatively and possibly as a means of discussing future research that might clarify the relationship.

6a. First of all, imagine a repeated measures design with seven levels. Can you tell me *why* you'd need to counterbalance such a design, what kind of counterbalancing you'd use, and how many participants you'd need? What is the impact of counterbalancing on order and carry-over effects? [3 pts]

Of course, in the typical repeated measures design you'd need to counterbalance to ensure that order effects or carry-over effects did not contaminate your results. That is, because your participants may become fatigued by the later stages of the experiment or because they may become increasing better at the task over time (practice effects), if you did not counterbalance, you would be confounding treatment level with position. Similarly, if exposure to one treatment level has an impact on the results of exposure to a subsequent treatment level, then you'd want to ensure that there wasn't a consistent order of the treatment levels.

With seven levels, I'd use incomplete counterbalancing, which would require some multiple of 14 participants (e.g., 14, 28, 42, 56).

Keep in mind that counterbalancing doesn't eliminate order or carry-over effects, which are often present in repeated measures designs (and in life). Instead, the role of counterbalancing is to distribute those effects equally over the conditions of your study, thereby eliminating a potential confound. For instance, because each condition will occur equally often in the first and last position, you can be sure that any impact of position will fall equally on all your conditions.

6b. OK, now let's assume that there is a particular order effect—a practice effect. That means that scores on the DV will improve over time as a result of practice. What is the impact on your error term (MS_{Error}) of counterbalancing? [2 pts]

Unfortunately, one impact of counterbalancing is that when position effects are present (as in the practice effect proposed here), you'll actually inflate your error term (MS_{Error}) relative to not counterbalancing. Of course, you must counterbalance regardless. In advanced statistics courses, you'll learn how to circumvent this problem.

7. In Lab 2, you saw a set of photo-arrays. As you know, each participant in that study rated each of the faces on the extent to which that face was a match for the eyewitness description (1 = "Poor Match" to 7 = "Great Match"). If the photo-array were unbiased, then the ratings of the six faces would be similar. To the extent that some faces were rated as less similar to the eyewitness description, then they were not really fair alternatives. As you know, each participant rated each of the six faces (making it a repeated measures design). However, it would be possible to use an independent groups design. Suppose that we used $n = 64$ in an independent groups design. Thus, there would have been 64 people rating the extent to which Face 1 (F1) matched the eyewitness description. Another 64 people would rate the extent to which Face 2 (F2) matched the eyewitness description. Etc. Complete the analysis below and then interpret these results as completely as you can. [10 pts]

Means Table for Lineup1
Effect: Face

	Count	Mean	Std. Dev.	Std. Err.
F1	64	3.078	1.462	.183
F2	64	4.172	1.486	.186
F3	64	4.125	1.363	.170
F4	64	2.234	1.342	.168
F5	64	5.984	1.327	.166
F6	64	3.734	1.606	.201

ANOVA Table for Lineup1

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Face	5	508.513	101.703	49.434	<.0001	247.171	1.000
Residual	378	777.672	2.057				

$$HSD = q \sqrt{\frac{MS_{Error}}{n}} = 4.03 \sqrt{\frac{2.057}{64}} = .722$$

Table of Differences (Significant Differences with Asterisk*)

	Face1	Face2	Face3	Face4	Face5	Face6
Face1	-					
Face2	1.09*	-				
Face3	1.05*	.05	-			
Face4	.84*	1.94*	1.89*	-		
Face5	2.91*	1.81*	1.86*	3.75*	-	
Face6	.66	.44	.39	1.50*	2.25*	-

There was a significant effect of facial stimulus on ratings of match to an eyewitness description, $F(5,378) = 49.434$, $MSE = 2.057$, $p < .001$. Post hoc analyses using Tukey's HSD indicated that people viewed Face 5 as closest to the eyewitness description ($M = 5.984$), because the rating of that face was significantly higher than all other faces. Face 2 ($M = 4.172$), Face 3 ($M = 4.125$) and Face 6 ($M = 3.734$) did not differ, but both Face 2 and Face 3 were rated as more similar to the eyewitness description than Face 1 ($M = 3.078$) and Face 4 ($M = 2.234$). Face 6 was rated as more similar to the eyewitness description than Face 4, but didn't differ from Face 1. Face 1 was more similar to the eyewitness description than Face 4, making Face 4 the most discrepant from the eyewitness description.

8. Below you'll find a repeated measures analysis of the exact same data (as in Problem 7). Focusing solely on the source table, and comparing it with the source table found in the prior problem, what can you tell me about the nature of the two F -ratios? How and why do they differ? Would you have expected such a difference? [5 pts]

ANOVA Table for Lineup1

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	63	279.018	4.429				
Category for Lineup1	5	508.513	101.703	64.246	<.0001	321.228	1.000
Category for Lineup1 * Subject	315	498.654	1.583				

Means Table for Lineup1
Effect: Category for Lineup1

	Count	Mean	Std. Dev.	Std. Err.
L1.1	64	3.078	1.462	.183
L1.2	64	4.172	1.486	.186
L1.3	64	4.125	1.363	.170
L1.4	64	2.234	1.342	.168
L1.5	64	5.984	1.327	.166
L1.6	64	3.734	1.606	.201

Obviously, the same scores/means are being used in the analysis (as seen in Means Table and the fact that the treatment line in both source tables is identical). However, in this case the F is 64.246 instead of 49.434. The F -ratios differ because the MS_{Error} is smaller in the repeated measures ANOVA (1.583 compared to 2.057). I would certainly expect such a difference, given the greater power of a repeated measures analysis. There must have been substantial individual differences removed from the error term (into the Subject term) to reduce the error term in the repeated measures ANOVA.