

1. In the video at the beginning of the semester, two studies were described (the social psychological study on fear and affiliation and the perceptual development study on the effects of light deprivation). One of the studies would have benefited from a manipulation check and the other would not. Which one would, and why? Which one would not, and why not? Use either of these two studies to illustrate an operational definition by telling me how the researchers operationally defined one of their IV's or DV's. [5 pts]

Briefly, the social psychology study on fear and affiliation would have benefited from a manipulation check. It would have been really useful to know if the young women in the study were actually more fearful in the "high fear" condition because the level of fear is not directly observable. One way to assess that question would have been to conduct a manipulation check (at the end of the study) in which the participants were asked how afraid they were during the manipulation. To make the point of the questionnaire less obvious, the pertinent questions would have been buried among a host of irrelevant questions. Because the amount of light deprivation in the kitten study is not ambiguous, no manipulation check would be needed.

In the kitten study, "pacing" was the DV used to measure maze performance and was operationally defined as the number of times that the kittens reversed direction.

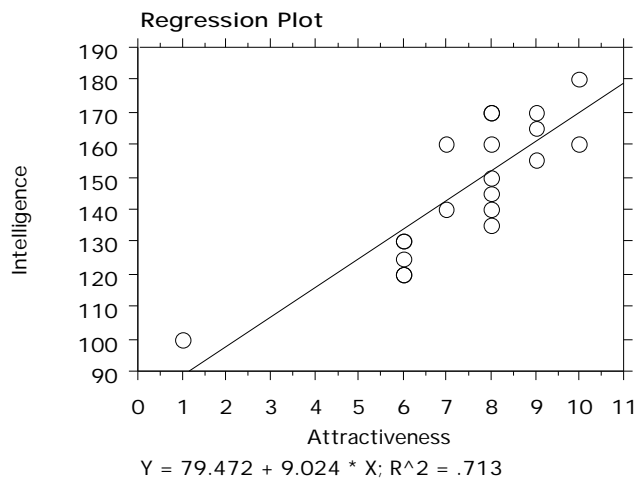
2. In the Mook article, he argues that external validity is not always a critical concern of the experimenter. How does Mook use the Hecht study on dark adaptation, the Argyle study on eyeglasses and intelligence, *and* the Milgram study on obedience to support his argument. What kind of study *would* require an experimenter to worry about external validity? [10 pts]

To answer this question well, a student would need to make good use of the studies described in the Mook article.

3. Many studies indicate that attractiveness is a big plus <darn!>. Thus, physical attractiveness is positively correlated with all kinds of good attributes. Dr. Luke N. Goode was interested in the relationship between perceived physical attractiveness and perceived intelligence. To investigate this relationship, Dr. Goode had 100 female participants use an 11-point scale to rate the attractiveness of 20 female faces (1 = very unattractive, 11 = very attractive). The participants also rated the intelligence of the women using the regular IQ scale. Dr. Goode took the mean of the ratings of the 100 participants for each of the 20 faces and analyzed the data as seen below. Interpret the results as completely as you can. If a woman's face had been given a mean attractiveness rating of 8, what would you expect that women would estimate her IQ to be? Do you have any comments on the design or outcome of this study? [10 pts]

Regression Summary		ANOVA Table					
Intelligence vs. Attractiveness		Intelligence vs. Attractiveness					
Count	20	Regression	1	6091.243	6091.243	44.706	<.0001
Num. Missing	0	Residual	18	2452.507	136.250		
R	.844	Total	19	8543.750			
R Squared	.713						
Adjusted R Squared	.697						
RMS Residual	11.673						

Regression Coefficients					
Intelligence vs. Attractiveness					
	Coefficient	Std. Error	Std. Coeff.	t-Value	P-Value
Intercept	79.472	10.323	79.472	7.699	<.0001
Attractiveness	9.024	1.350	.844	6.686	<.0001



These results indicate a significant positive linear relationship between attractiveness and intelligence ($p < .0001$). However, the relationship may be less strong than it might appear to be, because of the apparent outlier (with low attractiveness and low IQ rating). Using this regression line/equation, an attractiveness score of 8 would predict an IQ score of 151.7. Aside from noting the outlier, you should also question the use of only female participants and female faces (lack of generalizability).

4. What are the advantages of using a repeated measures (dependent groups) design compared to an independent groups design? Given those advantages, why would one ever use an independent groups design? [10 pts]

Briefly, the advantages noted should include efficiency and power. The independent groups design should be used whenever the manipulation is likely to leave a permanent effect (e.g., learning a language, cutting into a brain, etc.) or when deception is involved.

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

Complete the source table below and interpret the data as completely as you can. Note that the nature of this design does not lend itself to counterbalancing of the time of the GRE, because the time is critical (first must be first, etc.). [For extra credit, can you think of how the design would likely involve the counterbalancing of some aspect of the study?] As you can see in the source table, the SS_{Subject} is quite large. Using the data above to make your point, tell me where the SS_{Subject} comes from. What kind of error might you be making in your decision about H_0 and what is its probability? [20 pts]

ANOVA Table for Week

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	195980.000	21775.556				
Category for Week	2	3271.667	1635.833	7.370	.0046	14.741	.904
Category for Week * Subject	18	3995.000	221.944				

Means Table for Week

Effect: Category for Week

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

First, we would reject H_0 ($\mu_1 = \mu_2 = \mu_3$) because $p < .05$. Next, we would compute a post hoc test.

$$HSD = 3.61 \sqrt{\frac{221.944}{10}} = 17$$

Thus, Week 3 scores are significantly higher than Week 1 scores, but

no other differences are significant. Although you couldn't counterbalance the practice order, you should note that you'd need to use different tests each week. To ensure that one test is not easier or more difficult than other tests, you should counterbalance the order in which different participants get the different tests. With only 3 tests, that would imply 6 different orders of the 3 tests. To save space, I deleted the actual data, but you should recognize that the individual differences estimated by SS_{Subjects} are found in the variance of the mean scores across all participants. You could be making a Type I error in your original decision. However, when you interpret the post hoc tests and conclude that Weeks 1 and 2 or that Weeks 2 and 3 don't differ, those decisions could be Type II errors.

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

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

Complete the source table below and interpret the results as completely as you can. What might be some limitations of this study? [15 pts]

ANOVA Table for Response

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Gender	1	42.025	42.025	47.317	<.0001	47.317	1.000
Residual	38	33.750	.888				

Means Table for Response

Effect: Gender

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

No HSD is necessary because the study only involves two conditions. Thus, you could readily interpret the results to mean that women are more likely to get a positive response from men when asking them to go the concert ($M = 3.85$) compared to men asking women to the concert ($M = 1.8$), ($F(1,38) = 47.32, p < .0001$). You should note that the study is not a true experiment because gender of the person asking the questions is not a manipulated variable (nor is the gender of the participants). It’s also the case that the study is a bit confounded, in that women only ask men and men only ask women. Thus, you can’t know for sure if it’s the gender of the asker or the gender of the person being asked (or both) that’s producing the observed difference.

7. Power is a very important consideration in any experimental design. Describe at least three specific ways in which one might increase the power of a study. [5 pts]

1. Increase the sample size (n).

2. Increase the amount of the treatment (make the differences among the treatment levels more extreme).

3. Decrease the variability among the participants (make instructions clearer, use a more homogeneous group of participants, etc.).