Psychology 205, Spring 2010
Research Methods in Psychology
Mid-Term

Name: _____

Short Answer (13 questions):

1. How can subject variables threaten external validity? Give examples of two different types of subject variables.

(4 points)

Variability among subjects threaten external validity by limiting generalizability. Some examples include: ability, practice, motivation, interest, gender, age, culture and personality.

2. What is falsifiability and why is it important in scientific research?

(4 points)

Falsifiability - the requirement that a hypotheses must be capable of being falsified - is the characterisitic that distinguishes science from other ways of seeking knowledge that do not rely on evidence (such as philosophical argument, personal experience, casual observation or religious insight).

3. What is counterbalancing? With which type of design is it used? Why is it used?

(4 points)

Counterbalancing is a procedure in which different participants receive the levels of the independent variable(s) in different orders. It's used in within-subject designs. It's used to avoid systematic order effects.

4. Distinguish between a Type I and Type II error.

(2 points)

Type I - rejecting null when true (claiming effect when none present). Type II - failing to reject null when false (overlooking effect when present).

5. Define reliability and compare it to validity.
(4 points)

Reliability is the consistency or dependability of a measurement technique. Validity is the extent to which a measurement procedure actually measures what it is intended to measure.

6. Define variance (in words) and explain how systematic and error variance differ.
(4 points)

Variance is the amount of observed variability (or dispersion). Systematic variance is due to the variables of interest in the study while error variance is due to all other unidentified factors.

7. State one advantage of using interviews relative to questionnaires.
(2 points)

Interviews are ideal for probing for elaboration about complex topics. Also better for illiterate/disabled populations.

8. What is power (in relation to data analysis)?
(2 points)

Power is the probabilty that a study will correctly reject the null hypothesis. Or… increases ability to detect an effect. Or 1-beta. Several students mentioned the relationship between power and effect size, which was worth 1 point.

9. What are the three methods that researchers use to estimate the reliability of their measures?
(3 points)

Test-retest reliability, interitem reliability and interrater reliability.

10. Draw a (rough) graph that illustrates an interaction effect between two independent variables.
(2 points)

11. Now draw a graph that illustrates a fan fold effect and explain why this type of graph suggests a potential measurement problem.
(4 points)

Recall the writing tests given to students from three schools in the Boston area.
Potential measurement error - namely those caused by ceiling or floor effects.

12. What are factorial designs?
(2 points)

Factorial designs are experiments that include two or more independent variables.

13. List (and briefly explain) three threats to internal validity.
(3 points)

Threats to internal validity are systematic differences in experimental conditions other than differing levels of IVs. Examples include biased assignment of participants to conditions, differential attrition, pretest sensitization, carryover effects, maturation, etc… NOT looking for a description of the different types of validity. Also, not looking for sources of error variance or threats to reliability.

Long Answer (8 questions):

14. (8 points) An education specialist has been asked to test the effectiveness of a lunchtime math program for middle school students. Based on the results of a district-wide math assessment, students scoring in the top 5% in each grade are invited to attend the program twice a week for one month. After the program has ended, a second math assessment is given to all students in the district. Based on the results of the second assessment, the school board informs the education specialist that the program will not be continued – not only were mean scores unchanged but several of the students who attended the program failed to remain in the top 5%. Is the basis for this decision reasonable? Why or why not?

    No – ceiling effect and/or regression to the mean (4 points possible).
    Best student answer: "No! This is the selection bias effect. Since only top students were selected in the study, not only might there be a ceiling effect but the top students were more likely to be in the top through random chance, and thus through random variability, they were more likely to show a decrease in performance.

    Describe a better way of assessing the program's effectiveness.

    Random assignment rather than only top 5% of performers in first assessment (4 points possible).
    Best student answer: "There could be students randomly selected for the program from every level of each grade, thus reducing the selection bias."

15. (4 points) A consumer research company wants to compare the "coverage" of two competing cell phone networks throughout Illinois. To do so fairly, they have decided that they will only compare survey data taken from customers who are all using the same cell phone model - one that is functional on both networks and has been newly released in the last 3 months. They also plan to focus their analysis only on actual phone calls (not text messaging or other functionality). Assuming that all technological variables are controlled for, will this provide an accurate comparison of network coverage for the state? Explain.

    Customers who are using a newly released cell phone may not be representative of the state's population of cell phone users – may be more likely to live in (or outside of) urban areas, may be early adapters who have usage patterns dramatically different from the general population, may include a higher percentage of users who have recently switched networks based on coverage, etc.

    In addition, the limited focus on phone calls is a large confound as many users rely on their phone for texts and other non-voice functions exclusively. Missing data about coverage for these services is likely quite important.

    Best student answer: "No this will not provide an accurate comparison of network coverage. For one, both networks offer coverage to more than 1 type of phone. By only testing 1 phone they are limiting the generalizability of the results. Secondly, this new phone was newly released and probably more expensive. Since it was newly released, fewer people may have it and if it is expensive only certain people in certain areas will have this new phone. Again this could eliminate a large area and proportion of coverage in IL. Finally, only testing coverage for actual phone calls again presents a problem and leads to findings not being generalizable to all the functions of "coverage" that these networks offer.

16. (6 points) A researcher is interested in decision-making and judgments under uncertainty. He suggests that in the absence of enough knowledge, people tend to use irrelevant information as anchoring points for their answers. He decides to test his idea by running an experiment with 400 students from his Intro to Psychology class. He presents each student with a sheet of paper where they are instructed to write down the last four digits of their SSN and to answer the question 'How many countries are members of UN?' Then he collects the papers, finds the median of the SSNs and separates the papers into two piles – one higher than the median and one lower than the median. When he computes the average answer for the two piles he finds that students whose SSNs were above the median had significantly higher answers on the question about UN members. He concludes that the students' answers were influenced by the irrelevant SSN question and accepts this as evidence that people use the information at hand while making decisions, regardless of whether or not it's useful.

    What kind of study is this? (3 points)

    Correlational. Quasi-experimental is also acceptable as this frequently overlaps with correlation.

    Is the conclusion sound? Justify your answer. (3 points)

    Conclusion is sound – no obvious confounds. However, some students mentioned that he should have used a control group, which was worth a couple of points.

    Best student answer: "Correlational study. This is a sound conclusion assuming that the last four digits of a SSN are random. Although this is only a correlational study (the IV of SSN in not manipulated but only observed), it is impossible for the number of members guessed to affect SSN and it is unlikely that there is a variable that would affect both. However, much data is lost by separating the numbers into above median and below median groups. A better correlation might be found if a regression line was used to consider the data."

17. (8 points) An investigator is interested in the usefulness of GRE preparation courses. To put a specific course to the test, he runs a study at the university where he works. He invites students to participate in his abbreviated prep course via an ad in the school newspaper. Fifty students respond to the ad and, subsequently, take part in the two-week long course. The investigator enlists the help of a seasoned KAPLAN instructor to teach basic verbal, quantitative, and analytic skills for two hours a day for the duration of the course. At the end of the two weeks, the participants take the GRE.

    Very enthusiastic about his project, the investigator decides to take it a step further. He randomly selects another 20 universities from across the country. For each of those universities, he completed an identical procedure. Once all of the participants had been run nationwide, the investigator averages all of their test scores and finds it to be a 600. Comparing this to the national average of 500, the investigator deems the prep courses useful in improving GRE scores.

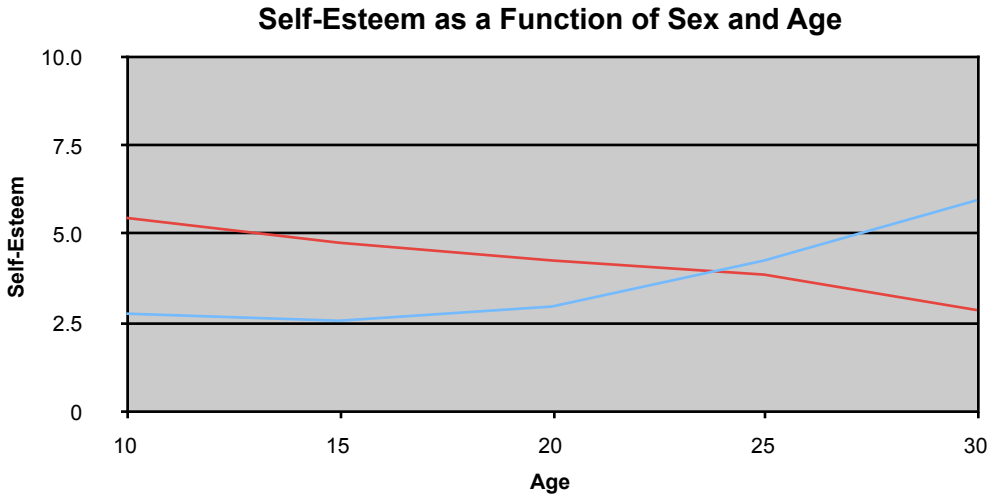
    Are his conclusions correct? Why or why not?

    Selection bias.
    Students could also mention that information is lost by averaging across the three subtests, but that issue is peripheral.

    If you think his study is flawed, how could you improve it?

    Randomly assign folks to course or no course conditions or pre- and post- measures could be given. Also, though he randomly selected the 20 universities, he may need to compare the scores with students from those same schools as variation in the quality of schools would present a confound.

18. (8 points) If an experimenter were to run a study and receive the data displayed below, how would she interpret them? (Note: The self-esteem scale ranges from 1-7 with higher scores indicating higher self-esteem).



*Sex x age **interaction** such that self-esteem increases with age for women but decreases in age for men. (I generally required students to mention the interaction effect explicitly for full credit of 2 points on this part.)*

Describe the design of a study which might produce these results. (3 points)

Many options (e.g., cross-sectional or longitudinal studies)

Most students answered this with a longitudinal study.

Are there any limitations to the study you proposed?

Students who suggest longitudinal should mention the length of time, large cost and high risk of attrition. Students who describe a cross-sectional design should mention something about generational effects.

19. (8 points) Some social psychologists were interested in investigating tipping behavior as a function of group size. They had the idea that the "diffusion of responsibility" theory developed in the helping behavior literature might be relevant to tipping behavior in groups. Diffusion of responsibility means that the more individuals witness an event in which they might offer some help, the less responsible each individual feels to actually help out. Thus, they hypothesized that individuals in groups would tip a smaller percentage of the total bill overall than individuals dining alone, because they would feel less responsibility for the tip.

    The results these researchers obtained confirmed their expectations: while most tips were around 15%, single diners tipped more (around 19%) than did individuals dining in groups (13.5%).

    Due to these findings, the researchers concluded that wait staff should happily allow for separate checks from groups, because they will likely get tipped more overall. What do you think about this conclusion?

    Different tip levels may reflect much more than diffusion of responsibility – it may also be that quality of service per individual served is lower for large groups than it is for small groups (whether real or perceived). If this is true, or some other confound is responsible for the lower tip amounts given by customers in large groups, then splitting the bill will have no impact.

    Give an alternative explanation for the tendency of individuals in a group to tip less. In addition, give another (different) explanation for why individuals dining alone might tip more.

    As mentioned above, it may be that service is worse for groups than it is for individuals, which leads people in groups to tip less. Additionally, the absolute value of the tip is often very large for big groups so that the percentage tip becomes less meaningful.

    Also, individuals dining alone may feel more of a personal connection to the waiter and tip more on that basis, which is entirely unrelated to diffusion of responsibility.

    Finally, it's reasonable to speculate that people dining in groups are different from people who eat out alone in ways that are not taken into account by the diffusion of responsibility theory (age, economic status, etc). These differences could play a part in differential tipping patterns.

20. (6 points) One of the tests for ADHD (Attention-Deficit Hyperactivity Disorder) in children involves bringing the child into a room and measuring their "attention span" by noting the number of toys they play with inside of a ten minute period, and the average amount of time spent playing with each toy. The rooms are stocked with toys and games for this purpose.

    If a child plays with a large number of toys in the ten-minute period, he or she is classified as having a short attention span. This conclusion then contributes ultimately to a diagnosis of ADHD.

    What is wrong with how "attention span" is measured in this study? In other words, is there reason to think that the measure of attention span might be confounded with other characteristics?

    Possible confounding factors include intelligence, arousal, gender, age, anxiety about the situation, prior depravity, distractability…

    Best student answer: "Yes. Attention span is very poorly operationalized in this situation. It is probably confounded with interest, positive affect, curiousity, etc."

    Can you think of a better way to measure attention span?

    Best student answer: "It may be better to, first of all, not measure attention span in relation to emotionally loaded items such as toys, where interest/mood/positive affect would clearly be confounds. It is probably better to give the child a set of tasks to perform that are fairly emotionally neutral. However it should also be made clear to the child that their goal is to actually finish these tasks so that some motivation is in play (an incentive may be necessary, although this may present a confound). Attention span should also be operationalized different. The amounts of time that the child can focus and work at the question before turning his attention to other things (as indicated by clear distraction from another stimuli) should be encoded and averaged."

21. (12 points) It so happened that of 100 young people (ages 16-20) in a certain town, 50 were male and 50 were female. Half of the males and half of the females were smokers, and half of each were nonsmokers. An industrious investigator tracked these 100 people for 20 years, administering a physical-endurance test to all of them once every 5 years. The values tabled, below represent the mean scores on the test as a function of age, sex, and whether the subjects were smokers or nonsmokers. For purposes of this exercise, assume that none of the subjects died or were lost for any other reasons, and assume further that all of the smokers remained smokers throughout the 20 years and that all nonsmokers remained nonsmokers throughout the period. In the table, the higher the score the greater the endurance. The maximum possible score on the test was 75, and a difference of 10 between means represents a statistically significant difference. Do not do any statistical tests, but just consider cell means.

	age in years				
	16-20	21-25	26-30	31-35	36-40
Males					
Smokers	65	55	35	20	5
Non-smokers	70	60	50	40	30
Females					
Smokers	55	50	40	32	20
Non-smokers	60	55	50	45	40

a) What are the three independent variables? (2 points)

Age, sex, smoking.

b) State verbally the influence of each of the three independent variables on endurance scores. (2 points)

Age decreased endurance. Smoking decreased endurance. Influence of gender varied depending on age (though females had higher endurance scores on average).

c) Determine if there is an interaction in the scores produced by age and sex. If so, state the nature of the interaction. What statistical test should you apply to test for this interaction? (2 points)

It appears there is an interaction – endurance decreases more rapidly for men than it does for women as they age. Use ANOVA to test for this interaction effect.

d) Determine whether the differences between smokers and non-smokers were greater for the males than for the females. (Just report means, do not do a statistical test.) State the nature of this effect verbally. (3 points)

The average difference scores between smokers and non-smokers was greater for males (14) than it was for females (10.6). In other words, the impact of smoking on endurance was, on average, more deleterious for males than it was for females.

Students had to state the difference of the means for full credit. Correct description verbally was worth only 2 points.

e) Determine whether there was a difference in the effect of the age variable depending upon whether the subjects were smokers or non-smokers. To do this, get a single score for each age for smokers and for nonsmokers by summing across sex. Plot the results on a graph below. Do these two variables interact? (3 points)

There was a difference. Even though the lines do not intersect on the graph, students should still point out that there WAS an interaction because the effects of smoking on endurance differed across age. In this case, we have a fan-fold interaction however so we do have to consider the potential for errors caused by scale…

Graph with one error worth on point. Correct graph worth two points. Correctly stating that variables interact worth 1 point.