Note: Please see Ashley during her office hours (or make an appointment) if you have any specific questions about the test!

**Psychology 205, Fall 2011**
**Research Methods in Psychology**
**Mid-Term**

**Name: _____**

**Short Answer (13 questions):**

1. What tests did Platt refer to as "touchstone(s) of strong inference"? (2 pts)

*("Ask what (experiment) could disprove your hypothesis" OR "Ask what hypothesis an experiment could disprove")*
2 points: something along the lines of "tests that pit two alternative hypotheses against each other"
1 point: on the right track but not quite there; e.g. mention of "induction and observation of data"
0 points: "ANOVAs," "t-tests," "statistics," etc. or if left blank

2. How do we formally represent the relationship between the data we collect and the data's uncertainty? List three sources of variability in data. How do researchers handle the uncertainty that these types of variability introduce? (4pts)

*((i) Data = Model + Residual; (ii) between person variability, within person variability across conditions, person x condition interactions; (iii) can use within-subjects designs to control for between-person variability)*
4 points: something like (i) "true score + error", (ii) "within subjects, between subjects, person x condition interaction", (iii) correct descriptions of ways in which to control for error variability in data
3 points: (i) no statement of relationship between error and data, but with (ii)/(iii) correct lists and descriptions of types of variability and control
2 points: (i) no statement of relationship between error and data, (ii)/(iii) partially incorrect lists of types and variability and control

3. List and define three measures of central tendency. What are some advantages and disadvantages of these measures relative to each other (give at least two). Why do we refer to some statistics as 'robust'? (4 pts)

*((i) Mean, median, mode (+ definitions). (ii) Mean is more sensitive to outliers than median or mode, but mean may not always be the best way to summarize the data (often because of its sensitivity). References to in-class example welcome here. (iii) Robust stats are those that don't change very much if you do weird things to the data (eg the median).*
4 points: (i) correct (above) list of central tendency measures (however, I didn't count off if you included "interquartile range"), (ii) correctly stated advantages and disadvantages (even if

different from the ones above), mostly correct statement regarding 'robustness' (again, didn't need to be identical to above)
3-2.5 points: (i) correct list of central tendency measures and (ii) statement of advantage/disadvantage that contained an error OR (i) list of central tendency measures with one error (I did count off for plain "range" here) and (ii) a correct statement of advantages and disadvantages. (iii) Again, if the 'robustness' statement was mostly right, you got credit for it. In several instances, however, I subtracted half a point for clearly incorrect statements about robustness.
2 points: in this case, the sum of errors across the response was 2 or more, whether these occurred in parts (i), (ii), or (iii).
0 points: blank

4. How do psychologists define 'Type I' and 'Type II' errors? What relevance these have for the analysis of experimental data? How are they related to each other?(4 pts)

4 points: (i) Type 1: rejecting the null hypothesis when it is true; Type 2: failing to reject the null hypothesis when it is false. (ii) Various correct statements about the utility of alpha and beta settings, power, etc. were accepted. (iii) Relationship: something along the lines of "inversely related" or "as you decrease the probability of making a Type 1 error, you increase the probability of making a Type 2 error and vice versa"
3 points: (i) correctly defined Type1 and Type 2, (ii) various responses accepted, (iii) partially incorrect statement regarding relationship
2 points: (i) correctly defined Type 1 and Type 2, (ii) various responses accepted, (iii) wholly incorrect statement regarding relationship OR failure to address the question regarding relationship
1 point: (i) incorrect definitions of Type 1 and Type 2, (ii) various responses accepted, (iii) incorrect statement regarding relationship

5. List three ways in which a researcher might improve the power of a study. Discuss the advantages or disadvantages of each option. (3 pts)

*((i) make alpha level less stringent, (ii) decrease error variance by increasing experimental control or homogeneity of sample, (iii) increase number of subjects)*

3 points: (i) anything like "make alpha bigger" or "make beta smaller" or "make 1-beta bigger" was fine; (ii) any correct statement about ways to increase control or sample homogeneity, (iii) "larger N" or "more participants." In addition, if you omitted one of the preceding and instead gave me more ways to control for error variability, that was okay, too (given rather vague question wording).
2 points: Here you could have given either (a) a correct list of only 2 items OR (b) a list of 3 items that included some incorrect statements or details.
1 point: A list of 2 items that included some incorrect statements or details
0 points: blank

6. How are reliability and validity related? List and define three types of validity that are relevant to the development of psychological tests; in addition, list and define two types of validity that are relevant to the design of psychological experiments. (4 pts)

*((i) Reliability is necessary but not sufficient for validity (ii) Face/content, construct, criterion; (iii) internal and external)*
4 points: (i) correct definitions of reliability and validity accepted here; the 'necessary but not sufficient' condition was sufficient but not necessary for you to get credit for your response. (ii) face/criterion/construct. (iii) internal/external
3 points: (i) same as above, (ii) 2/3 of the face/criterion/construct mentioned plus one other form of validity (some of these, like convergent/divergent, were okay), (iii) convergent/divergent (these are validities relevant to tests, not experiments)
2 points: (i) same as above, (ii) test validity list with at least two mistakes (e.g. an unrelated form of validity, a substitution of reliability for validity, etc.), (iii) partially or otherwise incorrectly addressed part 3 OR correct lists for parts (ii) and (iii) with no definitions
1 point: largely incomplete; 1-2 sections of question absent
0 points: blank

7. Describe three ways of assigning participants to conditions in a psychological experiment. What are the advantages associated with each method of assignment? (3 pts)

3 points: (i) random assignment, (ii) matched assignment, (iii) within subjects + definitions of all three of these (note: if you had the definition right but the name a bit wrong, I still gave you full credit; 'wrong' answers involve very clear mistakes, very unclear wording, or omissions)
2.5 points: Mostly correct list of three assignment strategies, but with less clear definitions than in full-credit responses
2 points: 2/3 assignment strategies right
1 point: 1/3 assignment strategies right

8. What is effect size? Why do researchers care about effect sizes? (2 pts)

2 points: (i) full credit for answers similar to "effect size is the amount of change in the dependent variable seen after the manipulation of the dependent variable." (ii) A number of different answers were acceptable here, again due to the vague wording! You could say things like "Effect size tells researchers more about the effect of the IV on the DV than conventional significance levels do, since it shows the 'size' of that effect as opposed to simply telling us that we can be fairly confident that the effect isn't zero." As long as you said something about effect size here that wasn't completely wrong, you got points.
1 point: Good attempt at both parts of question, but with one significant error
0.5 points: More than one significant error, but appears to have some understanding
0 points: Most statements incorrect/blank

9. In general, if we draw a very large number of samples from a population, how will the mean and variance of the resulting distribution compare to the mean and variance of the parent population? (2 pts)

2 points: By the central limit theorem, the (i) sampling distribution mean = mean of parent distribution and (ii) variance of sampling distribution is less than variance of parent distribution. (Note: "sigma-squared/N" was also acceptable for sampling distribution variance)
1.5 points: If you said "the mean and variance (of the sampling distribution) are the same (as for the parent distribution)" and mentioned the central limit theorem (or 1/N).
1 point: "mean and variance are the same"
0.5 points: "mean and variance are highly correlated" or an answer that seemed to reflect a misinterpretation of the question (like "why do researchers prefer that their studies have larger sample sizes?" instead of the question actually asked)
0 points: Mostly incorrect statements

10. What is a Latin Square design, and why is it often preferred to complete counterbalancing?(2 pts)

2 points: (i) something like "In a Latin Square design, every condition appears in every position with each condition preceding/following (one or other okay for full points; however, the answer is only strictly correct if you have both) every other condition." (ii) preferred because it is simpler, more efficient, etc.
1.5 points: (i) "way of counterbalancing" plus some reasonable attempt at showing how the Latin Square design works; (ii) same as above.
0.5-1 points: (i) similar to 1.5 point answer, but with at least one clear mistake; and/or (ii) left blank
0 points: blank or mostly incorrect responses for both parts

11. Describe the process of null hypothesis testing. What have been some criticisms of this form of hypothesis testing?(4 pts)

4 points: Given if you correctly stated (i) what a null hypothesis is, (ii) what p value (typically) is, (iii) made some statement about a statistical decision, and (iv) made some correct statements regarding drawbacks (e.g. Type 1 and 2 errors, significance levels, impossibility of proof, etc.)
3.5 points: Above elements are present and seem correct, but answers are less clear and thorough than for a 4 point response.
3 points: 3/4 of the above present and correct
2.5 points: 3/4 of the above present and seem correct, but answers are less clear and thorough than for a 3 point response
2 points: 3/4 of the above present, but with 1-2 clear errors
1.5 points: 2/4 of the above present; somewhat vague

12. List (and briefly explain) three ways in which researchers' and participants beliefs about an experiment can have an effect on the experiment's validity. (4 pts)

4 points: (i) placebo effect, (ii) experimenter expectancy effect, (iii) demand characteristics + correct definitions/explanations of all three (note: if you gave the correct definition but were a little off on the name, you still got credit)
3.5 points: (i) placebo effect plus (ii-iii) two different kinds of experimenter expectancy effect
3 points: (i) placebo effect (or some other correct response) plus (ii-iii) one or two forms of participant response bias, such as 'nay-saying' (these are for psychological tests, not experiments).
2 points: Gave a list of three items, but with 2 or more clear errors unrelated to question interpretation

13. Describe the distinction between descriptive and inferential statistics. (2 pts)

2 points: (i) Descriptive: these "summarize" or "describe" the data; reduce data to more manageable chunks. Examples here are fine. (ii) Inferential: help us "draw conclusions about" or "extrapolate from" or "make predictions about" the data; allow researchers to "generalize" about reliability, validity, etc. Examples here are fine.
1.5 points: (i) see above; (ii) Statements more vague than above, or involve over-specific language like "help us determine significance". Alternatively, (i) could be incorrect and (ii) correct.
1 point: At least one very clear error; statements brief and/or vague.

**Long Answer (7 questions):**

14. An experimenter is interested in whether memory can be improved if people use visual imagery. Participants (all female) are placed in one of two groups—some are trained in imagery techniques, others are trained to use rote repetition. The imagery group is given a list of concrete nouns (easier to form images than abstract nouns) to study and the other group is given 20 abstract words (ones that are especially easy to pronounce so repetition will be easy), matched with the concrete words for frequency of general use. To match the method of presentation with the method of study, subjects in the imagery group are shown the words visually (on a computer screen). To control for any "techno-phobia," rote subjects also sit at the computer terminal, but for them the computer is programmed to read the lists to them. After hearing their respective word lists, subjects have 60 seconds to recall as many words as they can, in any order that occurs to them.

    (4 pts) What are the constructs of interest in this study? How were they operationalized?

    4 points: (i) Visual imagery and memory; (ii) training in imagery vs. rote techniques, number of words recalled in 60 second recall task (respectively)
    3.5 points: (i) see above; (ii) one operationalization correct, the other incorrect
    3 points: Like 3.5 point responses, but with additional incorrect constructs mentioned in part (i)

2.5 points: 1/2 of the above correct (for instance, only mentioned memory and its operationalization), with second construct and its operationalization omitted or incorrect.

(4 pts) Is anything wrong with this study? Elaborate on any problems you see.

4 points: Confounds are (i) noun type and (ii) presentation method. All correct elaborations accepted.
3.5 points: 1/2 of the above correct plus statements to indicate a partial understanding of what the other correct answer probably was.
3 points: 1/2 of above correct; no strong indication of understanding regarding other correct confound
2.5 points: Neither of the above confounds identified; however, some effort put into coming up with other problems the study might have had

15. A clinical psychologist interested in treating flying phobia administered a pretest of anxiety to 50 phobic participants. The pretest involved a simulated flight aboard a functional aircraft. To increase the effect of the manipulation, the researcher chose not to inform participants that the flight was simulated. While 'flying,' participants filled out an anxiety questionnaire, but five participants became too anxious to complete the measure. Afterward, individuals with completed measures were matched on anxiety and randomly assigned to either an "Implosion Therapy" group or a "Systematic Desensitization" group. Four of the participants who panicked during the simulation were offered spots in the implosion group; the fifth was offered a spot in the desensitization group. All five declined and withdrew from the study.

Treatment groups met regularly for four weeks. Three participants dropped out of the implosion group before treatment was complete. Following treatment, remaining participants took part in a flight simulation identical to the first and completed the anxiety measure again. Anxiety levels for both groups had declined, but declines were greatest for the implosion group. The psychologist concluded that implosion therapy is better for treating individuals with flying phobia.

The pretest and posttest means for the two groups are shown in the table below:

| Group | Pretest | Posttest | Total |
|---|---|---|---|
| Desensitization (n=24) | 43.70 | 33.54 | 38.63 |
| Implosion (n=18) | 34.17 | 18.44 | 26.31 |
| Totals | 39.62 | 27.07 | |

(4 pts) What can you conclude about the researcher's findings on the basis of these data? Can you think of at least two alternative explanations for the results obtained?

4 points: Any two good reasons for the above results were accepted, as long as they were stated correctly (e.g., attrition, familiarity/practice, state changes in anxiety, placebo effect, etc.)
3.5 points: Good attrition discussion, but different aspects of attrition problem treated as two separate explanations for results.

(5 pts) Leaving aside obvious ethical considerations, how would you make this study better?

*(Note: Probably a better way to word this question would have been "Describe three ways in which you would make this study better, **other than** by performing a less drastic and deceptive pretest manipulation.")*

5 points: (i) an untreated control group is needed, (ii) further reasonable discussion of how to solve problems with attrition, (iii) at least one other valid improvement (e.g. how to fix clumsy assignment of the five participants who panicked)

4 points: (i) no mention of need for control group, but (ii-iii) at least 3 other good solutions advanced

3.5 points: (i) no mention of control group, (ii) attrition discussion, (iii) at least one other valid improvement

3 points: (i) no mention of control group, (ii) no further discussion of how to control attrition, (iii) at least one other valid improvement

2.5 points: (i-ii) no mention of attrition or control group, (iii) improvement suggested is somewhat flawed

16. Three psychologists hypothesized that (a) depressed people are more likely to perceive time as passing slowly and (b) both depressed and nondepressed people will feel that time passes more slowly during unpleasant experiences than during pleasant ones. In a test of these hypotheses, 60 students completed a measure of depression (possible scores 0-100, higher scores indicating more severe depression). The median score was 20.63. All who scored above the median comprised the 'depressed' group; all who scored below comprised the 'nondepressed' group. Participants individually underwent three mood manipulations. First, to obtain a baseline 'time perception' measure, both groups completed a "neutral" task. They spent five minutes writing descriptions of their bedrooms, then estimated the amount of time they had spent writing. Next, half the participants in each group were randomly assigned to perform a "pleasant" task followed by an "unpleasant" task; other participants underwent conditions in the opposite order. Each task involved writing five-minute essays (now about pleasant or unpleasant experiences), then estimating time spent writing. Participants were not allowed access to timepieces during the experiment. Unfortunately, the researchers found support for neither hypothesis. Disappointed, they resolved to replicate the study using a larger sample size.

(4 pts) What are the constructs of interest in this study? How would you improve the study's design?

*Note: Here we have a median split using a depression score that's much lower than the halfway point on the test; we're not sure that there were actually any 'depressed' students in the sample. Other problems include poor experimental design (everyone gets 'neutral' first, but then half of each group gets U-P while other half gets P-U) and dubious assumptions about manipulations and their criterion measures (is writing going to be at ALL pleasant for some participants? Is describing your living room always going to be a neutral experience?).*

4 points: (i) depression, time perception, pleasantness of experience; (ii) at least one valid improvement (best if regarding the suspect median depression score and subsequent data 'splitting').

3.5 points: All the above except a direct statement that 'pleasantness' is a construct in this study; however, an indirect statement regarding pleasantness is present in response.
3 points: All the above except any statement, direct or indirect, about 'pleasantness'
2.5 points: (i) depression and time perception; (ii) flawed suggestion regarding improvement

(4 pts) Why do you think the researchers believe that they need to replicate with a larger sample size?

4 points: Larger sample size increases the power of the study
3.5 points: Getting at idea of power and its relationship to sample size, but 'power' never explicitly mentioned
3 points: Mentioned power but crossed it out and wrote in something different
2 points: No mention of power or relationship to sample size; explanations less clear and more 'off-base' than explanations given 3-3.5 points.

17. (10 points) In class, we discussed Madsen and McGaugh (1961).

a) What were the theoretical variables of interest?
b) How were they operationalized?
c) What specific hypotheses were being tested?
d) How did Madsen and McGaugh extend the work of Duncan (1949)?
e) Why was this particular form of extension chosen?

Note: If you did parts a-e right but got the wrong study, you got 4.5 points.
Also, if you left out some parts and got the wrong study, you got 4 points.
If you left this question blank, you got 0 points.
Finally, everyone who got the right study got at least 5 points.
So, total credit (10 points) would have been given for the following set of answers:

(a) Memory/amnesia and fear
(b) Memory = not stepping down (after receiving a shock) or
Amnesia = stepping down (after receiving a shock followed by ECS) and
Fear = not stepping down (after receiving a shock followed by ECS)
(c) Rats that receive a shock followed by ECS will step down from the platform (also okay to say something like "ECS does not simply induce fear; instead, it impairs memory consolidation")
(d) Duncan operationalized memory and amnesia in the following way:
Memory = run to lit room (after receiving a shock in unlit room)
Amnesia = stay in unlit room (after receiving a shock in unlit room followed by ECS)
Madsen and McGaugh introduced the possibility that fear was a confound, because it might be that
Fear = stay in unlit room (after receiving a shock in unlit room followed by ECS)
(e) Madsen and McGaugh designed their experiment in such a way as to remove the confound between fear and amnesia.

The most common mistake was not providing an adequate description of Duncan's experiment in part (d); in general, if you only missed .5 or 1 points, then this is why (more thorough answers got 1.5 points; less thorough ones got 1 point).

18. An evolutionary psychologist collected a sample of eight rhesus monkeys for the purpose of studying the relationship between emotionality and the development of ulcers. Previous research suggested that monkeys placed in a restraining chair and allowed to press a level in order to escape from or avoid shock nearly always developed ulcers; however, it was unclear whether this was due to shock or to the stress of being in an avoidance conditioning procedure (ACP). To remove the confound, the researcher used two monkeys in each condition: one that had a lever to press and one that received a shock whenever the other failed to press the lever. To ensure that lever-pressing monkeys were up to their task, the researcher subjected all eight to a pretest ACP. The animals that learned the avoidance response most quickly were placed in the 'executive' condition (lever); the others were placed in the control condition (no lever). In all four pairs, 'executive' monkeys developed ulcers, but control monkeys did not.

    (4 pts) What were the constructs of interest in this study, and how were they operationalized?

    4 points: (i) stress/emotionality, control; (ii) whether monkeys developed ulcers or not, executive vs. control conditions
    3.5 points: mentioned stress/emotionality and correctly operationalized it; no mention of control as a construct
    3 points: mentioned stress but incorrectly operationalized it; no mention of control
    2.5 points: mentioned stress but incorrectly operationalized it; made additional incorrect statements

    (4 pts) What's a big problem with this study? Can you think of an alternative explanation for the results?

    4 points: no random assignment; therefore, there's an artifact: reactivity to threat confounded with 'executive' condition
    3 points: mentioned that there was no random assignment, but didn't mention the artifact; however, did come up with a creative, valid alternative
    2.5 points: No mention of artifact; no alternative given

19. A developmental psychologist suspected that parents' divorce has a greater impact on sons' aggressive behavior than on daughters'. To test this, 45 boys (ages 13-15) were recruited from several middle class junior high schools. Participants' divorced parents had been living apart for between one and five years.
Each boy was seated alone in an experimental room and given four mazes to complete. Experimenters told the boys that the first two mazes (which were moderately difficult) would be for practice; however, the boys would win a small cash prize if their performance on the remaining mazes was above average. Unbeknownst to participants, the second two mazes were insoluble. While boys worked on each of the mazes, the experimenter rated their behavior for signs of aggression. After working with the mazes (but before debriefing), boys were asked to rate how angry they had felt while attempting to solve each maze.
Although pre- and posttest self-ratings of anger did not differ significantly, boys' observer-rated behavior was significantly more aggressive in the posttest than in the pretest condition. Based on these findings, the psychologist concluded that anger is a distinguishing characteristic of sons of divorced parents.

(4 pts) What constructs did this study investigate? How were they operationalized?

4 points: (i) aggression, anger; (ii) observer ratings of aggressive behavior, self-ratings of recalled anger (okay to mention frustration, too. In addition, I didn't count off if you said 'divorce,' though it's not technically a construct under investigation here).
3.5 points: (i) correct; (ii) mentioned only one out of the two forms of operationalization
2.5 points: (i) mentioned only one of the two; (ii) mentioned only one of the two
1.5 points: (i) mentioned only one of the two; (ii) made an incorrect statement about how aforementioned construct was operationalized

(5 pts) Do you think that, given the hypothesis being tested, the researcher's choice of design was a good one? If not, what would you have done to make it better?

5 points: The study needs three more control groups before the researchers can justify their claims; specifically, groups needed are sons of nondivorced parents and daughters of both divorced and nondivorced parents. (Other valid responses were acceptable if given in addition to these)
4 points: no mention of the need for one (not both) of the following: control groups with daughters or control groups with children of nondivorced parents.

20. A freelance journalist's monograph on urban legends includes a chapter devoted to the infamous "Sports Illustrated Cover Jinx." The chapter describes, in breathless detail, over forty-five instances of the strange phenomenon. The essence of the legend is that teams and individuals who appear on the cover of Sports Illustrated subsequently experience losses and career upsets at a rate that seems by far to surpass that which one would expect by chance. "Perhaps," the author surmises in the chapter's closing paragraph, "this evidence is sufficient to confirm the adage that 'pride comes before a fall,' and celebrity athletes would do better to refuse the excessive publicity that contributes in no small measure to their fatal hubris."

(8 pts) How would you explain the Sports Illustrated Cover Jinx?

8 points: It's regression to the mean! (Plus a valid explanation of same)
6 points: Good, mostly valid effort, but no mention of regression to the mean.
4 points: More errors in response; no mention of regression to the mean.