Research design:

1. Consider a study with two factors of interest, and an experimental design in which these factors can be studied as the independent variables in a crossed ANOVA analysis. In such cases it is common to try to design a research study with equal sample sizes in every cell of the design.

   a) Give at least two reasons why equal sample sizes are an advantage.
   b) What complications can arise if sample sizes are not equal across cells?
   c) Are there modifications or adjustments that will compensate for the problems caused by unequal sample size?
   d) Are there analyses where unequal cell sizes cause serious problems that cannot be ignored? If so, why? If not, why do you think that is the case?

2. You have been taught that when you conduct a research study you have to be careful about confounding variables. Discuss the following questions: What is a confounding variable? Give an example of a situation that includes such a variable. Why do you need to avoid such variables, if you can? What problems do they cause for the researcher? If you cannot avoid having the variable in the study, what statistical methods can be used to reduce the inferential problems it raises?

3. The National Assessment of Educational Progress is a national large-scale survey of student achievement. It includes tests of cognitive abilities such as reading and mathematics, and gathers background data on students' demographics and educational experiences. The students are a true random sample of the grade populations. It is a one-occasion survey rather than an experiment or longitudinal study.

An interesting finding appeared in the 1992 NAEP reading assessment, concerning the relation between extra instruction in reading and children's reading levels. In that study, the amount of reading instruction fourth-graders received was correlated negatively with their performance on the reading tasks:

<table>
<thead>
<tr>
<th>Time Spent in Reading Instruction</th>
<th>30-45 Minutes</th>
<th>60 Minutes</th>
<th>90 Minutes or More</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Proficiency</td>
<td>220</td>
<td>219</td>
<td>216</td>
</tr>
</tbody>
</table>

There is thus a negative correlation ($r = -.1$) between reading performance and time spent in reading instruction. It appears on the surface that increasing reading instruction decreases reading performance.
a) Discuss whether this causal inference is justified, using concepts of internal and external threats to validity.

b) Describe an experimental study that you could carry out that would address the relationship between reading performance and time spent in reading instruction. Compare it with the NAEP study in terms of validity.

4. Consider five designs for estimating the effect of an educational treatment:
   
a) One group, randomly sampled from the population of interest. Pretest, Treatment, Posttest.
   
b) Two groups, randomly sampled from the population of interest, with randomized assignment to Treatment or Control conditions. Posttest only for both groups.
   
c) Two groups, randomly sampled from the population of interest, with randomized assignment to Treatment or Control conditions. Pretest, Treatment, Posttest in both groups.
   
d) Two groups, one with volunteers for the Treatment and the other who don't want the treatment, Posttest only for both groups.
   
e) Two groups, one with volunteers for the Treatment and the other who don't want the treatment, Pretest, Treatment, Posttest for both groups.

Which of these designs provides the weakest evidence for estimating the effect? What threats to validity does it suffer from? Which design is the strongest? Which of the weaknesses in the weakest design does it address, in what ways?
Measurement:

1. One can distinguish among reliability coefficients calculated by (1) using item-level responses on a single administration of a test, (2) by administering alternate forms of the same test, and (3) by administering a test twice to the same sample of students.

   a) What are these three kinds of reliabilities called?
   b) How are they calculated?
   c) Should the results be the same, other than sampling error, or different? Why?
   d) What are the interpretations of each?
   e) Give an example of test use for each type of reliability where that type is particularly relevant to the inferences that are of interest.

2. A 3-parameter IRT analysis of test items gives a picture of the item characteristic curve.

   a) Draw two such curves—one of an easy item and one of a hard item. Indicate on your graph the difficulty (b), discrimination (a), and guessing parameters (c).

   b) Consider measuring a high ability person, with a θ around 2. For each of the following pairs of items, which item would give more information about this person, and why?

   i) Item 1: $a_1 = 2, b_1 = 2, c_1 = .2$. Item 2: $a_2 = 1, b_2 = 2, c_2 = .2$.
   ii) Item 1: $a_1 = 1, b_1 = 2, c_1 = .2$. Item 2: $a_2 = 1, b_2 = -2, c_2 = .2$.
   iii) Item 1: $a_1 = 1, b_1 = 2, c_1 = .3$. Item 2: $a_2 = 1, b_2 = 2, c_2 = .1$. 
Measurement, continued:

3. This question concerns classical test theory. There is a set of ten parallel test forms that yield scores $X_1, X_2, \ldots, X_{10}$. The measurement error variance $\sigma_e^2$ associated with any form is 1.00. In a given student population, true score variance $\sigma_t^2$ for the trait the tests measure is also 1.00.

   a) What is the formula for the reliability of a single form, expressed in terms of $\sigma_e^2$ and $\sigma_t^2$?

   b) What is the measurement error variance associated with the average of two test forms?

   c) What is the reliability of the average of two test forms, expressed in terms of $\sigma_e^2$ and $\sigma_t^2$?

   d) What is the measurement error variance associated with the average of $N$ test forms?

   e) What is the reliability of the average of $N$ test forms, expressed in terms of $\sigma_e^2$ and $\sigma_t^2$?

   f) Using your answers to (a) and (e), what is the generalized Spearman-Brown formula for the reliability of a test increased in length $N$ times?

4. Reliability and validity are important technical values in testing. Fairness is an important social value in the use of tests.

   a) Define reliability and validity.

   b) How are reliability and validity related? Include in your discussion the issue of whether there can be validity with reliability, and reliability without validity.

   c) How do you think reliability and validity are related to fairness in test use?

Statistics:

1. The variability of the subjects who are included in a research study can be quite critical to the results of the statistical analyses in that study. Consider the implications of including a sample of only subjects who, compared to the population of interest, do not differ very much on some key variables that are being used in a multiple regression analysis.

   What are reasons that a sample of subjects might be more homogeneous than the population of interest? What would be the effect on the magnitude of the statistical relationships that involve the affected variables? If the subjects were not homogeneous with respect to those variables, would the relationships be larger or smaller than you
observed? Are there any statistical analyses that would help reveal the true relationship in the target population? Address both independent variables and independent variables. Discuss all your answers.

2. Suppose that you calculate a correlation between variables A and B using one sample of data, and then calculate the correlation between A and B using a different sample of data from the same population. Suppose you predict variable A from variable B in both sets of data. Finally, suppose you predict variable A from variable B in the second sample, using the regression coefficients you obtained from the prediction in the first sample. What would you expect to see in these data analyses and why would you expect to see it? In particular, do you expect to get the same correlation in both sets of data? Do you expect to get the same regression weights in both regressions? Do you expect the regression weights from the first sample to provide predictions that are just as accurate in the second sample as they are when applied to the data from the first sample? Discuss all your answers.

3. Your colleague asks you to help analyze the data set shown below. She wants to know whether the mean for Treatment A is significantly different from the mean for Treatment B.

a) Your colleague asks which test would be appropriate, between a t-test for matched pairs and a two-sample independent-groups t-test with a pooled error term. What further information do you need to know in order to answer? Why?

b) Suppose the answer to a) indicates that a t-test for matched pairs is appropriate. Carry out the analysis. Is the difference significant for a two-tailed test at the .05 level?

<table>
<thead>
<tr>
<th>Observation #</th>
<th>Treatment A</th>
<th>Treatment B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>-13</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>-14</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>-5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>-2</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>-2</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>-1</td>
</tr>
</tbody>
</table>
4. A partially filled ANOVA table appears below for a one-observation-per-cell randomized block design.

a) Fill in the highlighted cells.

b) Under what conditions would one use a randomized block design? What are its advantages over a one-way ANOVA for analyzing the treatment alone? Are there any circumstances in which the one-way ANOVA would be preferred to the randomized block analysis, in terms of statistical criteria?

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>160</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blocks</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>410</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>