**Assessing Program Impact: Alternative Designs**

Required reading: RLF, Chapter 9.

The focus of this chapter is on how to design an evaluation when random assignment to groups is not possible but constructing a control group is possible. ***Quasi-experiments*** are defined as experiments that do not have random assignment but do involve manipulation of the independent variable. You learned in the last chapter about the power of random assignment as a technique for controlling all known and unknown extraneous variables by equating the groups at the start of an experiment. In quasi-experiments we must come up with other strategies for equating groups and ruling out alternative explanations that an observed relationship between the IV (independent variable) and DV (dependent variable) is due to one or more uncontrolled extraneous confounding variables. Quasi-experiments are not nearly as strong as randomized experiments for establishing firm evidence of cause and effect (e.g., evidence of program impact). If done well, however, quasi-experiments can provide moderately strong evidence of program impact. Don’t forget that once you determine impact, you must make an evaluative decision about the program. Remember the four steps in the logic of evaluation?

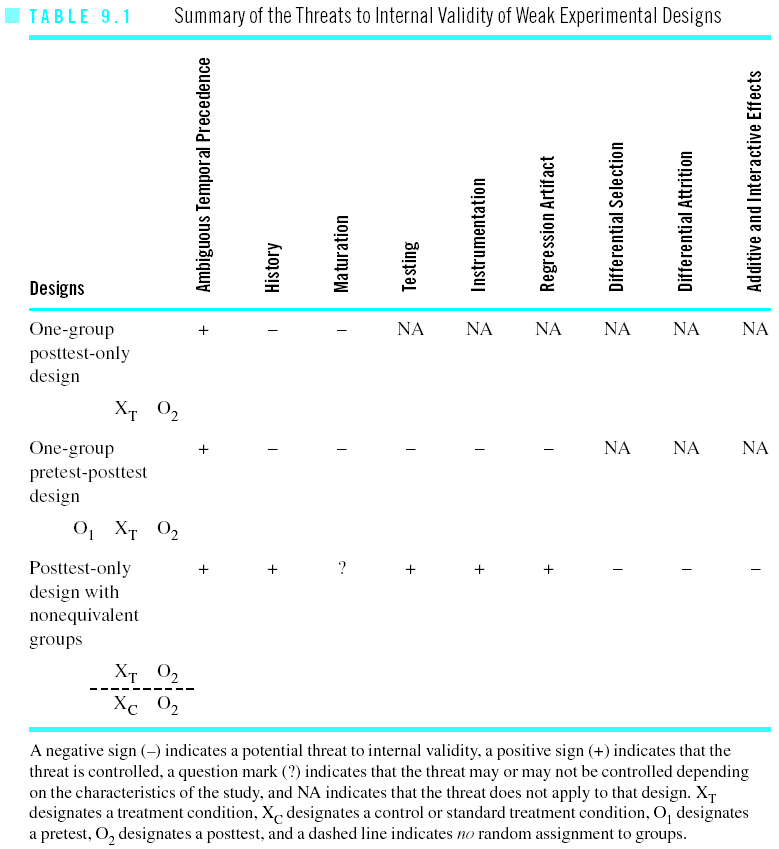
I will provide comments about each of the major sections of this chapter, as well as a section that I am adding to the chapter on additional alternative designs not discussed by the authors.

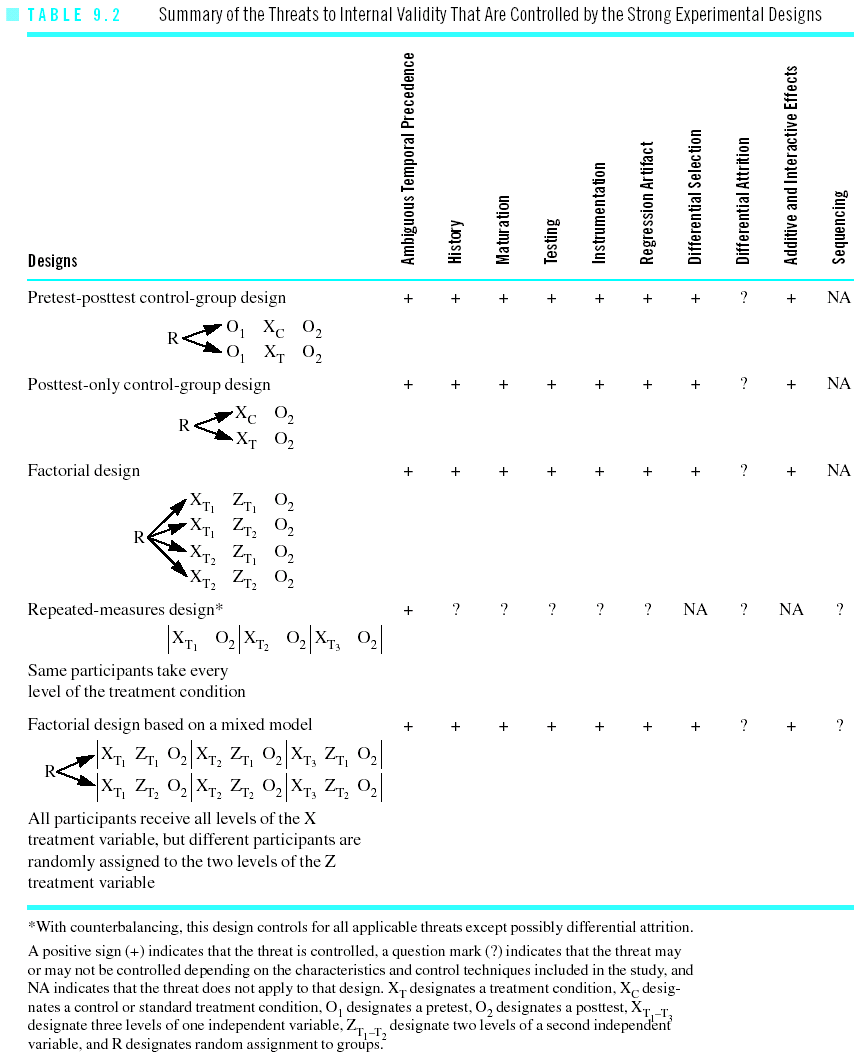
* Bias in Estimation of Program Effects
* Quasi-Experimental Impact Assessment.
* Additional Alternative Designs Not Discussed by RLF.
* Some Cautions About Using Quasi-Experiments for Impact Assessment.

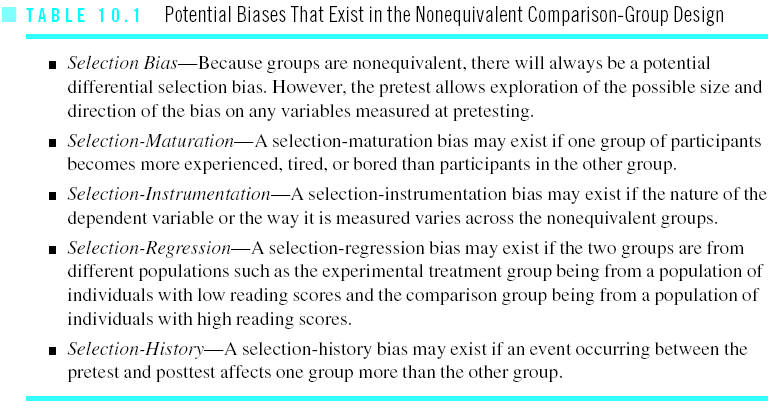
**Bias in Estimation of Program Effects**

In this section RLF discuss some of the threats to validity that can pop up when you are estimating program effects. For a review of the concepts of internal and external validity read my noted from my book (link: <http://www.southalabama.edu/coe/bset/johnson/dr_johnson/lectures/lec8.htm> and study the tables that connect designs to the threats to internal validity (Table 9.1, and 9.2 here <http://www.southalabama.edu/coe/bset/johnson/dr_johnson/2oh_masters.htm> ). You also should examine Table 10.1 (not online) that shows the threats to the nonequivalent comparison-group design (on page 303 of Johnson and Christensen).

Here they are for your convenience:







Here are the “threats” mentioned by RLF:

* *Selection bias* (this includes any factor that the groups differ on, often as a result of their initial composition, but also including them dropping out in unique ways which is called differential attrition)
* *Secular Trends* (this is not typically mentioned in research methods books but should be, especially for studies that take place over an extended period of time as in interrupted time-series designs and on-going field experiments of any type)
* *Interfering events* is basically what we call history effects in the above links.
* *Maturation* is the same as mentioned in the above links.

**Quasi-Experimental Impact Assessment**

If you can’t randomly assign participants to groups, what do you think you should do? The typical quasi-experimental answer is to try to construct a control group through matching and/or statistical control. You can use both matching and statistical control in the same study. For example, you may construct a control group by finding participants who match the experimental group participants on relevant characteristics (e.g., age, gender, education) and you may also measure all of the participants on variables that you couldn’t match them on and then use the statistical control method to further equate the groups on these measured variables.

Now let’s see how the impact assessment formula for quasi-experimental impact assessment compares to the impact assessment formula for randomized designs discussed in the last chapter. Here is the impact assessment formula for randomized designs (our strongest experimental research designs) that was discussed in the last chapter:

**Net = Gross outcome - Gross outcome ± chance differences**

**Effect for an for a comparable**

**intervention control group**

**group**

Here is the impact assessment formula for quasi-experiments that include a control group:

**Net = Gross - Gross ± Uncontrolled ±chance differences**

**Effect outcome outcome difference**

**for an for a between**

**intervention constructed intervention**

**group control and**

**group control**

**groups**

Do you see the difference between the two impact assessment formulas?

Unlike the impact assessment formula for randomized designs, the new formula has an additional component for uncontrolled differences between the intervention and control groups. This new threat could be labeled “effects of other processes (i.e., extraneous confounding variables).”

Again, because of the lack of random assignment, we are worried that our the groups will differ on ***extraneous confounding variables***—more specifically, we are worried about the threats to internal validity from my chapter as well as secular trends and any other factor on which the groups may differ. We are worried that our groups will differ because of the operation of one of these threats. Any factor that the treatment and constructed control groups differ on, other than the treatment variable, are potentially confounding variables that threaten our ability to attribute group differences after the intervention to a treatment effect.

**Quasi-Experimental Impact Assessment**

The most commonly used quasi-experimental design is the ***nonequivalent comparison group design*** and it is shown here to give you a visual image of the design:

O1 Xtreatment group O2

-----------------------------------

O1 Xcontrol group O2

Pretest Independent Posttest

Measure Variable Measure

(Dependent (Dependent

variable) variable)

Notice that there are treatment and control groups, pretests, and posttests, but there is no random assignment. The design is appropriately labeled the nonequivalent control group design because our big worry when we do not have random assignment is that the groups will be nonequivalent.

Because this above design does not have random assignment to the groups, it needs to be improved by using the following processes when feasible:

* matching (individual or group matching)
* statistical controls (e.g., adding control variables, modeling the determinants of outcome and/or selection)

Next, RLF make a distinction between *ex ante* and *ex post* quasi-experiments. For clarity of language (for non-Latin experts such as myself), I will call these *prospective* and *retrospective* quasi-experiments.

Note that usually the above nonequivalent comparison group design is done prospectively (it might also be called an ex ante nonequivalent comparison group design). Here the design is planned and started before delivery of the program services to the intervention group. When using the prospective nonequivalent comparison group design the evaluator constructs a control group, administers the pretest measures, implements the treatment, and then administers the posttest measures (this is shown visually in the above picture of the nonequivalent control group design).

Sometimes you will be brought in during a program or after the fact. In this case you might have to use some for of a retrospective design. These (*ex post*) or retrospective impact designs are undertaken after the delivery of program services to the intervention group. Here the evaluator comes in after the program has been started; the program may even be ended. To obtain data, the evaluator often uses retrospective survey questions or searches for retrospective documents to construct a control group and to obtain the pretest measures. In short, if you look at the picture of the nonequivalent control group design above, you will see that it will be very hard to get all of the pieces after a program is in operation or after it is completed. As a result, the retrospective nonequivalent control group design is weaker than the prospective nonequivalent control group design, and, more generally, retrospective designs are weaker than prospective designs for estimating program impact.

Now, we will discuss in more depth the process of locating a counterfactual comparison group, which often involves constructing comparison groups to use for comparison with your group receiving the program to be evaluated.

**I. One major strategy (that has been mentioned briefly) is called *matching***. Don’t forget that matching can be a useful technique (except when you select opposite extremes from your groups, such as high scoring at risk students with low scoring not at risk students). Matching can also be used in combination with statistical controls.

When using the technique of matching, you need to first decide on what ***matching variables*** (i.e., extraneous variables on which you match groups) that you intend to use. What you hope to do is to equate the experimental and control groups on the matching variables so that those potentially confounding variables can be ruled out as alternative explanations of the observed relationship between the treatment variable and the outcome variable. If we want to conclude that variations in the IV cause changes in the DV then we must rule out any variables that may be related to both the IV and the DV—these are the kinds of variables we match on.

When deciding on your matching variables, consider these points:

* You want matching variables that will be feasible to use in your evaluation study.
* You want matching variables that you are pretty sure your groups will be different on if you don’t match them on these variables.
* You want matching variables that are related to the outcome variable (i.e., the DV).
* Match participants on their matching variable status that exists before the intervention is administered if possible.

Here is a list of participant characteristics often leading to selection bias that you may consider for matching (from Johnson and Christensen).

Characteristics on Which Research Participants Can Differ

-------------------------------------------------------------------------------

Ability to do well on tests Home environment Reading ability

Age Intelligence Religious beliefs

Anxiety level Language ability Self-esteem

Attitudes toward research Learning style Socioeconomic status

Coordination Maturity Spelling ability

Curiosity Motivation to learn Stress level

Ethnicity Personality type Time spent on homework

Gender Political beliefs Vocabulary

Hearing ability Quality of eyesight

--------------------------------------------------------------------------------

Here are a few more factors on which to consider matching:

* characteristics of families or households in your study, including life-cycle stage, number of children, number of members, SES, housing type, ethnicity
* characteristics of the organized units in your study such as schools, classes or organizations, including class size, levels of authority, number of subunits/students, growth rate.
* characteristics of communities from which the different groups might come (geographically based groups), including types of industry, population size, growth rate, population density, location to metropolitan area.

Note that there are two general approaches to matching. The first is ***individual matching*** which means that for each individual in your treatment group, you attempt to find a comparable individual to place in the control group. For example if you are studying gifted high school students and you want to match on IQ, you might have a group of 25 gifted students for the experimental intervention and you then try to find a matching individual (on IQ) for each experimental group member to place in your control group. The key is that you are directly matching individuals (and, as a result, the groups will also be matched).

In ***aggregate matching or group matching***, your focus is on making sure that the experimental and control group averages or proportions on the matching variable(s) are similar. For example, if you are studying young adults and the average age of your experimental group is 23, you will attempt to include in your control group people in the age range of 18 to 28 or so (which is the age range of young adults) and you will specifically make sure that the average age of all the people in the control group is 23 years old so that it will have the same average age as the intervention group. The key is that you are directly matching the groups rather than particular individuals.

Individual matching is preferable to aggregate matching on scientific grounds; however, if group matching is more feasible, it should be used because it is much better than using no matching at all!

Before concluding I want to point out a case where matching can fail. Let’s say that you are designing an evaluation for a local head start type program. Program administrators include an income requirement (e.g., no more than 175% of poverty level) for participation and they invite all children in a local area who meet this requirement to participate. They measure these potential participants on an achievement test, and decide to find (match) control group children from a neighboring area who have similar achievement test scores. They ignore the fact that the neighboring area has a higher overall income level. Because the control group and experimental and control groups have the same average score on the achievement test the program administrator is not worried, thinking that all the kids are at the same level of intelligence.

The program is conducted, but you get what appear to be poor results—in fact, your head start kids actually became a little worse on the posttest achievement measure and the control group kids actually got a little better on the achievement. It looks like the program HARMS the children in the program.

There is, however, an alternative explanation for this finding: the result could be a regression artifact (i.e., a threat to internal validity where people with extreme scores tend to regress toward the mean of their group over time because of measurement unreliability). In this case, when constructing the control group through matching, perhaps you selected kids who were relatively advantaged but simply had very low scores on the test you gave—upon a second administration of this test many of these kids may do better simply because they are regressing to the mean. Further, perhaps the kids included in the treatment group were generally quite disadvantaged but happened score relatively high on the test compared to other kids in their disadvantaged group—upon a second administration of the test, many of these kids may do worse simply because they are regressing to their group mean. Putting these two regression artifacts together, and it may even appear that a program is detrimental, when in fact the problem is the threat called regression artifacts.

**II. Another strategy for equating groups is called *statistical control***. (More generally, this is simply a statistical approach used to control for potentially confounding variables in any kind of research.) In order to statistically control for a variable, you measure all participants (e.g., the experimentals and controls) on the control variable and then (using a statistical package such as SPSS) use a statistical technique such as multiple regression to control for the variable. The most common control variable in the nonequivalent control group design is the pretest measure of the dependent variable; that is, you use statistical control (ANCOVA) to see if the experimental and control groups posttest scores on the dependent variable significantly differ, on average, after controlling for differences on the pretest measure. By the way, you can include multiple control variables when using statistical control, although a pretest measure of the outcome variable tends to be the single best control variable (Mohr, 1995).

The logic of statistical control is similar to matching logic. When deciding on your ***control variables***(i.e., the variables you will statistically control for) consider these points:

* You want control variables that will be feasible to use in your evaluation study (i.e., variables that you can obtain data on).
* Use control variables that you are pretty sure your groups will differ on.
* Use control variables that are related to the outcome variable (i.e., the DV).
* Measure the participants on the control variable(s) before the intervention is administered.

You may consider the same variables for statistical control that you consider for matching--just look again at the potential matching variables I listed above. A good understanding of relevant theory, however, is the best way to identify control variables.

The bottom line is that you want to control for variables that may represent plausible rival explanations for an observed correlation between your independent and dependent variables. You do not want another researcher or program expert to come up to you after your evaluation study (where you concluded that the treatment variable was related to increased learning) and then that other person says “but why didn’t you control for group differences on IQ and motivation?” If you didn’t control for these variables, all you will be able to say is “well…I dunno…you know you have a good a pretty good point there…I’ll be sure to do that next time” Remember to think about the variables you need to control and then to do it. This will make your study more credible and more defensible.

Here are a couple of examples of statistical control from Johnson and Christensen that I hope will make the concept of statistical control a little clearer for you. I kind of like the ANCOVA example (where a woman tells her husband that she thinks that there is gender discrimination in wages in the U.S. and her husband say “No honey, It ain’t noting but education. Men got more education than ladies and that’s why they got more money.” In the example, the wife decides to control for education to check out her husband’s “rival explanation” of the relationship between gender and income).

One “special case” of the general linear model is called **partial correlation**, which is used to examine the relationship between two quantitative variables “controlling for” one or more quantitative extraneous variables (Cohen, 1968, 1983). It is called a partial correlation because the effect of the third variable is “partialled out” or removed from the original relationship. Perhaps you want to examine the correlation between intelligence and test results controlling for amount of time spent studying to see if effort or IQ seem to be more important for your test. Typically, all of the variables used in partial correlation analysis must be quantitative rather than categorical.

Here is a relatively easy way to think about partial correlation. If you determine the regular correlation between your independent variable and your dependent variable at each of the levels of your extraneous variable, you will have a lot of correlations (e.g., if your extraneous variable had 10 levels then you would have 10 correlations; if your extraneous variable had 100 levels then you would have 100 correlations). Next, if you take the weighted average of those correlations, you will have a single number, and that number is called the partial correlation coefficient (Pedhazur, 1997). The range of a partial correlation coefficient is the same as a regular correlation coefficient (i.e., -1.00 to +1.00, with zero signifying no relationship at all). As a general rule, if a researcher uses a regular correlation coefficient (i.e., the correlation between two variables) rather than a partial correlation coefficient (i.e., the correlation between two variables controlling for one or more additional variables), then you can be pretty sure that he or she was not thinking about extraneous variables. On the other hand, if a researcher uses a partial correlation coefficient (or another control technique), you can be pretty sure that he or she was thinking about controlling for extraneous variables. As a general rule, you should upgrade your evaluations of research articles where the authors controlled for extraneous variables.

Another “special case” of the general linear model is called **Analysis of Covariance** (which was discussed in earlier chapters). ANCOVA is used to determine the relationship between one categorical independent variable and one quantitative dependent variable controlling for one or more quantitative extraneous variables (Pedhazur & Schmelkin, 1991). Here is an example. There is a relationship between gender (which is a categorical variable) and income (which is a quantitative variable) in the United States. Men earn more money, on average, than women. You might decide, however, that you want to “control for” education; that is, you want to make sure that the difference is not due to education. You could eliminate education as a rival explanation (i.e., you could control for it) by comparing the average income levels of males and females at each of the levels of education in your data. You could also have the computer analyze your data using the ANCOVA technique, and it will tell you if gender and income are still related after controlling for education. If gender and income are still related then the researcher can conclude that education has been eliminated as a rival hypothesis. The details of ANCOVA and partial correlation are beyond the scope of this book. However, the key point is that ANCOVA shows the relationship between a categorical independent variable (e.g., gender) and a quantitative dependent variable (e.g., income level) controlling for a quantitative extraneous variable (e.g., level of education).

An advantage of statistical control (compared to matching) is that researchers can base their research on samples of participants who are randomly selected from a population (Pedhazur and Schmelkin, 1991). (You don’t have to “throw out” cases from the data like you do in matching when you can’t find a match for an individual.) In order to statistically control for one or more extraneous variables, a researcher must collect data on the extraneous variables in addition to data on the independent and dependent variables (i.e., collect data on all of the important variables). In effect, the researcher incorporates the extraneous variables into the design of the research study. Then, after collecting the data, the researcher controls for the extraneous variables during data analysis (using ANCOVA, partial correlation, or another technique).

RLF show a good example of controlling for education and employment in Exhibit 9-E.

RLF also go into more ***advanced strategies for statistical control.*** I’ll explain the basic concepts and ideas.

**(Note**: In case you have not had statistics beyond IDE 510, I want to remind you about regression which is just a form of the GLM. I am going to write out some general linear models (GLM) again which are just like the multiple regression equations discussed in the last chapter. As a review, the main idea (in regression or GLM) is that Y is the outcome variable (i.e., the DV) that you want to explain or predict. You want to “explain variance” in the outcome variable. You include in your equation variables that you think will successfully explain or predict Y (i.e., your DV). Don’t get blown away by the notation. It should not be hard to note that the variable on the left hand side of the equals side is the variable being explained by the variables on the right hand side of the equals sign. The regression coefficients (the little “bs” in the equations) are important (especially the one for the grouping variable) because they tell you about the relationship between that IV and the DV controlling for the other IVS in the equation. This is exactly the same logic of controlling for extraneous variables that was used earlier in this lecture when I included some material on partial correlation and ANCOVA.) For a review of regression, analysis of variance, and analysis of covariance, see Chapters 15 and 16 in Johnson and Christensen: go here for the lectures which are available on-line:

<http://www.southalabama.edu/coe/bset/johnson/dr_johnson/2lectures.htm>

**One of the two advanced strategies is called *modeling outcomes***. The basic idea here is to construct a general linear model (or multiple regression equation) where Y is your outcome variable, your primary independent variable is your grouping variable (e.g., treatment group versus control group) and you also include in this equation other variables that are known to be related to your outcome variable. The strategy is simply to include in your equation variables that are known to be related to you outcome variable—that’s probably why they call it “modeling the determinants of outcomes.” An excellent example of using a multiple regression or General Linear Modeling approach is shown in Exhibit 9-F; so look at this example closely.

For another example, let’s say you want to evaluate a management training seminar on participatory management practices for a large International Corporation, and one desirable outcome is that participants’ attitudes toward participatory management will improve as a result of the seminar. Two hundred managers are planning on attending the seminar next month (i.e., these managers signed up for the seminar). In the meantime, you determine that there are 350 additional managers working at for the corporation, and you decide to collect the following data from all 550 managers shortly before the seminar as your pretest measures: attitudes toward participatory management practices (as your pretest measure of your DV or outcome variable), group status (this is your primary IV and marks whether participants are in the treatment or the control group), and you collect additional data on the following control variables: years with the corporation, age, gender, attitudes toward subordinates, and desire for upward mobility. You conduct the seminar and afterwards you measure all 550 managers on the outcome variable one more time (as your posttest measure of attitude toward participatory management practices). Finally, you would run the following general linear model in SPSS using the data you have collected.

Y = bo + b1X1 + b2X2 + b3X3 + b4X4 + b5X5 + b6X6 + b7X7 + e

where,

Y is the posttest measure of the DV

b0 is the Y intercept

b1 through b7 are partial regression coefficients

X1 is the pretest measure of the DV

X2 is the group status variable (i.e., your IV)

X3 is years with corporation

X4 is age

X5 is gender

X6 is attitudes toward subordinates

X7 is desire for upward mobility

e is the residual or unexplained part of the model

You primary interest is in the statistical significance of the regression coefficient (i.e., b2) for the grouping variable (X2). If this coefficient (i.e. b2) is statistically significant, it tells you that the seminar appears to have had a statistically significant net impact on managers’ attitudes toward participatory management controlling for years with the corporation, age, gender, attitudes toward subordinates, and desire for upward mobility. If you excluded any important confounding variables from your equation, you have what is called ***specification error*** (i.e., this is the problem of omitting an important variable independent or confounding variable) and the results can not be trusted because they will be biased.

The beauty of a GLM is that if you want to control for a variable, you just include it in your equation along with the other variables. The difficulty is knowing which variables to include as control variables. RLF discuss two related classes of variables to consider for inclusion in your statistical equation: 1) include variables that are known to be related to your outcome variable (e.g., prior training, years with corporation, age, gender, attitudes toward subordinates) and 2) include variables that you believe affected why the participants ended up in the groups they are in (e.g., motivation to participate in the study, desire for upward mobility, knowledge of the training seminar instructor). Ultimately, for both of these classes of control variables, the point is that you want to include variables on which the groups initially differ and are related to the outcome variable.

**The other advanced strategy mentioned by RLF for statistical control is called *selection* *modeling***. The last strategy was straightforward: just identify variables you need to control for and then include them in your statistical equation (your GLM). Modeling selection, however, involves a two stage process. Here is what you do:

Stage I. Identify variables that you believe predict or explain why participants are in the groups that they are in. For example, let’s say that you believe that the most important group selection variables for your training program are motivation to participate in the study, desire for upward mobility, and knowledge of the training seminar instructor. What you now do is to treat your grouping variable (treatment versus control group) as your dependent variable for the moment in the following equation:

Y = bo + b1X1 + b2X2 + b3X3 + e

where,

Y is the grouping variable (i.e., it’s your IV)

b0 is the Y intercept

b1 through b3 are partial regression coefficients

X1 is motivation to participate

X2 is desire for upward mobility

X3 is knowledge of trainer

e is the residual or unexplained part of the model

You run this equation using SPSS and tell SPSS to save the part of your grouping variable’s variance that is explained by the three selection variables. In particular, tell SPSS to save “predicted Y.” Then let’s rename this variable “selection” or whatever name you want. Now go to step two.

Stage II. Use the selection variable you obtained in the previous step in a new GLM where Y is your outcome variable (posttest scores on the DV), and also include you grouping variable (your original IV, treatment versus control) and you include any additional control variables you want. Here is the Stage II GLM you could run using SPSS:

Y = bo + b1X1 + b2X2 + b3X3 + b4X4 + b5X5 +

b6X6 + b7X7 + e

where,

Y is the posttest measure of the DV

b0 is the Y intercept

b1 through b7 are partial regression coefficients

X1 is the pretest measure of the DV

X2 is the group status variable (i.e., your IV)

X3 is the **selection variable** from stage one

X4 is years with corporation

X5 is age

X6 is gender

X7 is attitudes toward subordinates

e is the residual or unexplained part of the model

The main thing you want to check for in the above equation is the statistical significance of the coefficient (b2) for the grouping variable (X2). Believe it or not, this tells you whether the net impact (the difference between the groups) is statistically significant after controlling for selection differences (as measured by the selection variable X3) and controlling for the additional control variables in the equation. As always, you also want to know the amount of variance explained by your grouping variable if it is statistically significant.

In case you are worried, focus on the logic of these two advanced statistical approaches for your exam.

**III. The next major strategy for equating groups is called the *regression-discontinuity design***. This strategy is a specific research design (like the nonequivalent control group design was a specific research design). RLF point out that a better name for the regression-discontinuity design might be the **cutting point design** because the design works by measuring a group of potential participants on an ***assignment variable*** (also called the selection variable by some authors), setting a cutting point on that variable based on practical or theoretical considerations, and then putting the people falling above the cutting point in one group and the people falling below the cutting point in the other group. The researcher must be able to rigidly allocate all participants to their respective groups according to the cutoff rule. The participants can have absolutely no choice in which group they end up in. There can be no exceptions to the rule.

The **regression-discontinuity design** can be depicted visually as follows:

OA C Xtreatment ODV

OA C Xcontrol ODV

Measure Assignment Independent Posttest

of the by cutoff (C) Variable Measure

assign- score method of DV

ment

variable

As you can see from the above visual depiction, the regression-discontinuity design works like this:

1. Measure participants on the variable you are using for group assignment.
2. Assign participants to the two groups based on the cutoff score on the assignment variable.
3. Conduct the experiment (i.e., administer the treatment to one group and not to the other group).
4. Measure the participants on the dependent variable after the treatment has been administered.
5. Analyze the data.
6. Interpret the results. Answer the primary question, “Was there a net program impact?”

The ***assignment variable***is a variable chosen by the researcher and used to allocate participants into two groups (i.e., those scoring above the criterion and those falling below the criterion). As you can see, the selection process is not left up to chance; instead, the researcher fully controls the allocation process, and he or she understands exactly how the process worked. Then, during data analysis the researcher simply examines the relationship between the IV (treatment versus no treatment) and the DV (the outcome variable) statistically controlling for the assignment variable.

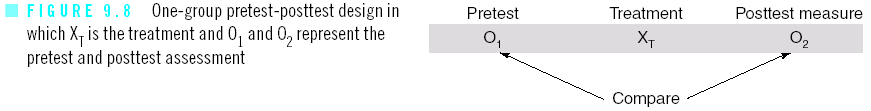
Assignment variables are often based on need or merit of some sort and are used to determine whether a participant qualifies for the intervention. In fact, this is one of the reasons that this design may be the design of choice: you can allocate those in greatest need to one group and the others to the other group. Here are a few possible examples of assignment variables: standardized achievement tests, number of days absent from school during the past academic year, number of nonadopted children, annual family income of children’s parents, mother’s IQ, and scores on a reading test.

For an example, let’s say that you are interested in determining if a new approach to increasing student’s academic self-efficacy works. You decide to use the 100 students in a large high school study hall as your participants (obviously after obtaining the appropriate consents). You measure these students on IQ (from student records) as your assignment variable. You choose as your cutoff point an IQ of 108, since that’s the median for your group of students (i.e., which conveniently gives you 50 students in each group), and you assign the groups of participants to the two groups. Next, you conduct the study (giving the treatment condition to one group and no treatment to the other group). Finally, you measure the students on their level of academic self-efficacy. You would run a GLM and see if the grouping variable is statistically significant, controlling for the assignment variable (IQ). If the treatment variable is statistically significant, you conclude that the program does have an effect.

In conclusion, the regression-discontinuity design is a very strong quasi-experimental design because it rules out selection differences, and once selection is ruled out, one can obtain an excellent estimate of net program impact.

**IV. Another strategy for equating groups is to use reflexive controls.**

Reflexive controls are measures of your outcome variable taken on the participating targets before the intervention and to use these measures as your countrol observations. For example, following is the one-group pretest-posttest design:



The pretest measure is your estimate of the counterfactual.

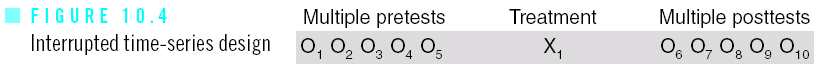
Reflexive controls also are used in “single case designs” that are adapted for use in program evaluations with groups such as the ABA (baseline, treatment, baseline) design, and the multiple baseline design (for more information on these see the Chapter 10 lecture here: <http://www.southalabama.edu/coe/bset/johnson/dr_johnson/lectures/lec10.htm>

Another design that uses reflexive controls is the interrupted time-series design. Here is a depiction of it in one of its forms and some discussion from my lecture notes from my research book:

**Interrupted Time-Series Design**

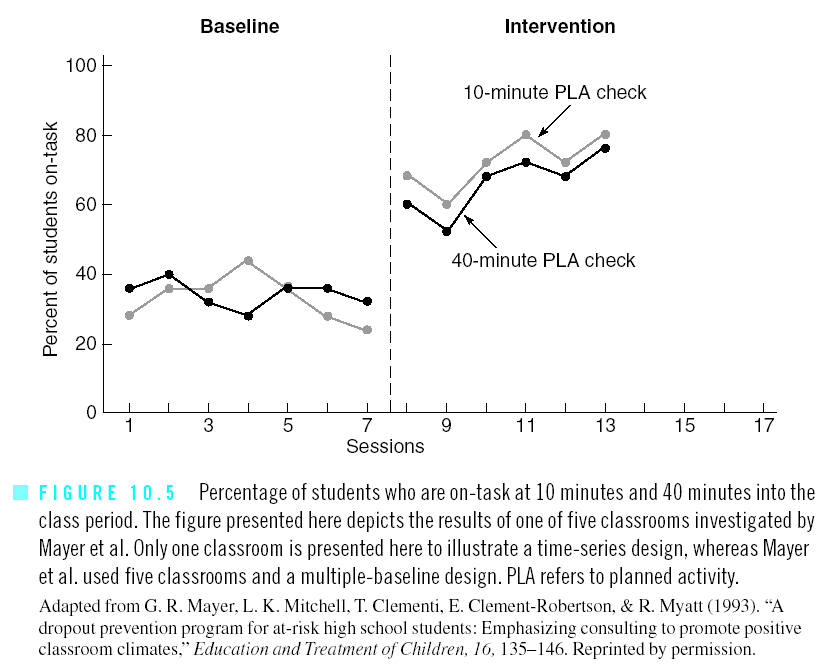
This is a design in which a treatment condition is accessed by comparing the pattern of pretest responses with the pattern of posttest responses obtained from a single group of participants. In other words, the participants are pretested a number of times and then posttested a number of times after or during exposure to the treatment condition.

Here is a depiction of the interrupted time-series design:



* The pretesting phase is called the baseline which refers to the observation of a behavior prior to the presentation of any treatment designed to alter the behavior of interest.
* A treatment effect is demonstrated only if the pattern of posttreatment responses differs from the pattern of pretreatment responses. That is, the treatment effect is demonstrated by a discontinuity in the pattern of pretreatment and posttreatment responses.
* For example, an effect is demonstrated when there is a change in the level and/or slope of the posttreatment responses as compared to the pretreatment responses.

* Here is an example where both the level and slope changed during the intervention:



        Many confounding variables are ruled out in the interrupted time-series design because they are present in both the pretreatment and posttreatment responses (i.e., the pretreatment and posttreatment responses will not differ on most confounding variables).

        However, the main potentially confounding variable that cannot be ruled out is a history effect. The history threat is a plausible rival explanation if some event other than the treatment co-occurs with the onset of the treatment.

**Some Cautions in Using Quasi-Experiments for Program Impact**

Here are some advantages:

* Constructed control groups might be the only feasible way to equate groups because, for example, random assignment may not be possible in a given situation for political or other reasons.
* Practicality and convenience may suggest a nonequivalent comparison group design (rather than the randomized design), and if you use this quasi-experimental design you can sometimes improve it by using matching or statistical controls or, preferably, by using both matching and statistical controls.
* Some other useful quasi-designs are the interrupted time-series design and the one group pretest-posttest design; be sure to note that the latter has more threats to its internal validity than does the former.
* Research shows that well done prospective nonequivalent comparison-group designs can approximate, under favorable conditions, the results obtained under a randomized design.
* The regression discontinuity design is a very strong quasi-experiment in which controls are constructed using a cutoff assignment rule. This design is the strongest of all the quasi-experimental designs.

Here are some disadvantages of using quasi-experimental designs:

* Although well done nonequivalent comparison-group experiments tend to approximate results of randomized designs, these results also tend to be more variable (in an unpredictable direction), which means that researchers will make more mistakes in estimating the net effect of a program.
* Retrospective designs are much more risky is estimating effects and are not as strong as prospective designs. For example, there are a host of problems in trying to approximate a nonequivalent comparison-group design retrospectively such as having to rely on recall (memory) and locating past records which are likely to be fallible.
* You should avoid sole reliance on generic controls as a method of constructing a control group comparison because it has so many problems. The main problem is that it is difficult to find a generic control group that is similar to your experimental group (e.g., in terms of demographics, motivation, location, etc.).
* The statistical methods we have discussed are based on good fundamental logic. The problem is that some of the important assumptions required to use these methods are violated in practice which mitigates their accuracy.
* The pervasive problem with matching and statistical control is that you can never identify all possible extraneous confounding variables on which the groups may differ. Therefore, you will always under-match or under-control to some degree, leaving your conclusion about program impact open to an attack.

**Additional Alternative Designs or Strategies That Are Not Discussed by RLF**

* One strategy is to compare your treatment group’s outcomes with *generic controls*.

A ***generic control***is an established measure of social processes, such as published test norms or local norms, that are used as comparisons with the outcomes of an intervention. In other words, the generic measure is used as your estimate of the counterfactual (i.e., what would have happened to the experimental group members had they not received the intervention). The problem is that there are rarely any high quality generic controls available in social science (unlike in the natural sciences). If you use a generic control, be very careful and supplement it with additional data and evidence.

* Another strategy is to use cross-sectional survey designs.

The idea here is to select a sample (hopefully a random sample) and compare the performance on the DV of those in the sample who were in the program with those in the sample who were not in the program. An obvious problem is that unless you are able to obtain a special sampling frame, there may be very few people in the sample who participated in the program. In making the group comparison, you also include control variables in your analysis so that you can statistically control for variables that you measured and believe may potentially be extraneous confounding variables.

* Another design is the **retrospective survey design** which I discussed in the chapter on the Kirkpatrick book in this class. Remember, many of the studies in the Kirkpatrick book used questionnaires after the fact to ask program participants rate how much they liked the program and to rate how effective they thought it was.

An advantage of the cross-sectional survey and the retrospective survey design is that it can be accomplished quickly and sometimes it is the only practical solution for evaluating a program. On the other hand, you must be very careful not to exclude any important variables from your statistical equation (i.e., watch out for specification error), and you must meet the statistical assumptions of the tests and you must realize that you are only looking at program participants perceptions rather than having more objective or hard outcome data measures.

* Another strategy to use with any of the designs if possible is to attempt to measure the amount of participation (such as no participation, attended one meeting, attended two meetings, attended three meetings, attended four meetings). When you can quantify your independent variable like this, you are able to document what is called the ***dose-response relationship***. That is, rather than comparing an experimental group to a control group, you can examine the quantitative relationship and functional form (e.g., is it linear or curvilinear?) of the relationship between the outcome variable and the quantitative independent variable.

Another category of strategies is to use some qualitative research approaches. I will group these under the term shadow controls...

**Shadow Controls**

A ***shadow control*** is an estimate of the impact of a program based on the opinions or judgments of experts, administrators, and/or program participants. In general, it can be difficult to obtain strong evidence of cause and effect using qualitative data. (Qualitative approaches are generally much better for studying need assessment, theory assessment, and implementation assessment, and to provide supplemental data for impact assessment.)

In some cases, a qualitative or mixed method approach (without the use of a rigorous experimental design) will be all that is needed. For example, an inexpensive approach will be called for when you are evaluating a one-shot program, a low cost program, an extraordinarily successful program, or a program that is relatively unimportant (e.g., a national study of the effect of national academic standards is probably more important than a study of a brief, one-time program affecting only a few people).

The primary approach that relies on shadow controls is sometimes called connoisseurial assessment.

***Connoisseurial assessment*** refers to judgments made by evaluators who claim to be experts. They hang out in the program, talk with and observe people, and make comparisons in their head (i.e., using shadow controls), and they ultimately decide whether they believe a program has an impact. A good connoisseurial evaluator comes into a situation, collects multiple sources of data (which may include quantitative data) and only then makes a judgment about program impact.

The accuracy or possibility of connoisseurial assessment depends on several factors:

1. Is the expert an expert in the subject that the program addresses?
2. Is the expert knowledgeable about the operation and outcomes of similar programs?
3. Is there relevant and relatively conclusive social science knowledge available that can be related to the program theory?
4. Does the expert, when collecting data, rely on multiple sources and search for corroborating results?

Further, the expert (i.e., you) should rely on as many of the following sources of evidence as possible:

* Administrative records and other extant data.
* Observations of program operation.
* Questionnaires and interviews with program participants.
* Interviews with program administrators.
* Interviews with stakeholders and informants, discussing, for example, contextual factors affecting the program.

There are problems with each of these data sources but it total the set of data and resulting evidence may suffice. Administrative records will be incomplete and will not directly provide evidence of program impact. Observations may be limited and perhaps biased because of the expert’s own subjectivities. Program participants can provide reactions and other useful data that can be used to think about program impact, but it is generally naïve to expect a participant to be able to provide a solid estimate of program impact. Program administrators tend to put a program in a positive light and they may be too involved with the program to provide a more distant or “objective” assessment of the impact of the program. Finally, an understanding of the contextual factors is essential but insufficient.

In sum, if an expert can answer yes to the above four questions and if he or she uses many of the data sources just listed, the commissioners of many programs (especially local and small programs) will be and should be completely satisfied with the outcome evidence provided by the expert evaluator. Often a precise estimate of net impact is not needed. A categorical judgment (the program is working “good enough” versus it is not working “at an acceptable level”) may be all that is needed. Remember, oftentimes in evaluation you have to rely on the “good enough rule?”