Alternative Designs

The typical experimental designs (between, within, mixed) with large numbers of subjects are not suitable for all situations. Next, we will examine three alternative (non-experimental) designs: *small-n designs*, *quasi-experimental*, *developmental*.

I. Small-N Designs

There are situations where very few individuals are available to serve as participants or where differences between individuals are minimal.

In Psychophysics, where we are interested in basic sensory processes, differences between individuals are relatively small and small numbers of subjects are used.

In clinical settings (medical, psychological), we often have a situation where the number of subjects is 1. That is, we'd like to know if a particular therapy or treatment is effective for a particular individual.

Before considering the small-n designs, we will (re)consider a design that does not work: the AB design. (We referred to this as a pretest-posttest design earlier, Chapter 8.)

A) The AB design.

The basic design is simple. Observe behavior (the dependent measure) during a baseline period. Then, institute the treatment and observe the behavior again. If the treatment is effective, the behavior should change.

Example: Operant conditioning to treat temper tantrums. A therapist who uses behavior modification techniques would approach this problem by trying to uncover the reinforcement contingencies that support the temper tantrums. That is, the tantrums are a learned response of the child that is reinforced by the parents (or others).

The normal course of therapy is to examine the situation to learn the contingencies. This is the baseline period. The parent is then trained to reinforce the child in situations where tantrums are not occurring and to ignore the tantrums (or shape them, reinforcing the milder ones, responding only during a lull in the tantrum, etc.). This is the treatment phase.

Suppose we observe that the tantrums disappear. The therapist and parents are satisfied since the original problem is gone, but a researcher would not accept, as valid, the conclusion that the therapy produced the change in behavior.

The problem is that there is no control over other factors that could have co-occurred with the treatment. History effects could be present. Over a long enough interval, maturation could alter behavior. Both within and between designs would have a control condition. This design lacks the control of the experimental designs.

B) The ABAB design (reversal or return to baseline design)

Is there any way around the lack of control? One approach, called the ABAB or return to baseline, is to do the baseline, then the treatment, then go back to baseline and finally the treatment again. If the behavior changes back and forth with the conditions, and it is unlikely that other factors are perfectly correlated with our baseline-treatment alternation, then the relation between treatment and behavior is likely to be causal.

Our example is from Hart et al. (1964) using behavior modification in a nursery school to treat the crying of a 4-year old boy. They had the teacher monitor crying for a 10 day baseline period, then use a new reinforcement system with the child for 10 days (treatment), then go back to the baseline situation for 10 days, then back to the treatment for 10 more days.

The data showed 5 - 10 crying episodes per day in the first baseline (A). After the first few days of treatment, crying was down to 0 or 1 time per day (B). In the second baseline (A), crying reverted to around 5 or more times per day. During the second treatment phase (B), it quickly dropped back to no crying.

Since the behavior is changing with the treatment and baseline conditions, we conclude that the operant conditioning used by the teacher was effective in controlling the crying behavior.

The are no inferential statistics in this type of design. Instead, good control of the independent variable and multiple measurements of the dependent variable are sufficient.

Why can't history and maturation explain the results?

C) Extensions to the Return to Baseline.

Often, we want to examine more than two levels of the independent variable. This can be done in the same way as the ABAB design.

Rose (1978) studied the effects of artificial colorings (a food additive) on hyperactivity in children. Feingold (1975) had reported that a strict diet that eliminates artificial preservatives, flavors and colorings from food, plus eliminating foods with natural salicylates reduced hyperactivity. Feingold's conclusions were based on an AB design.

Rose studied two hyperactive 8-yr old girls. There were three conditions, the K-P diet which the girls were normally on (control or baseline condition), the diet plus occasional oatmeal cookies with no colorants (placebo condition) and the diet plus occasional oatmeal cookies with artificial yellow dye (experimental condition).

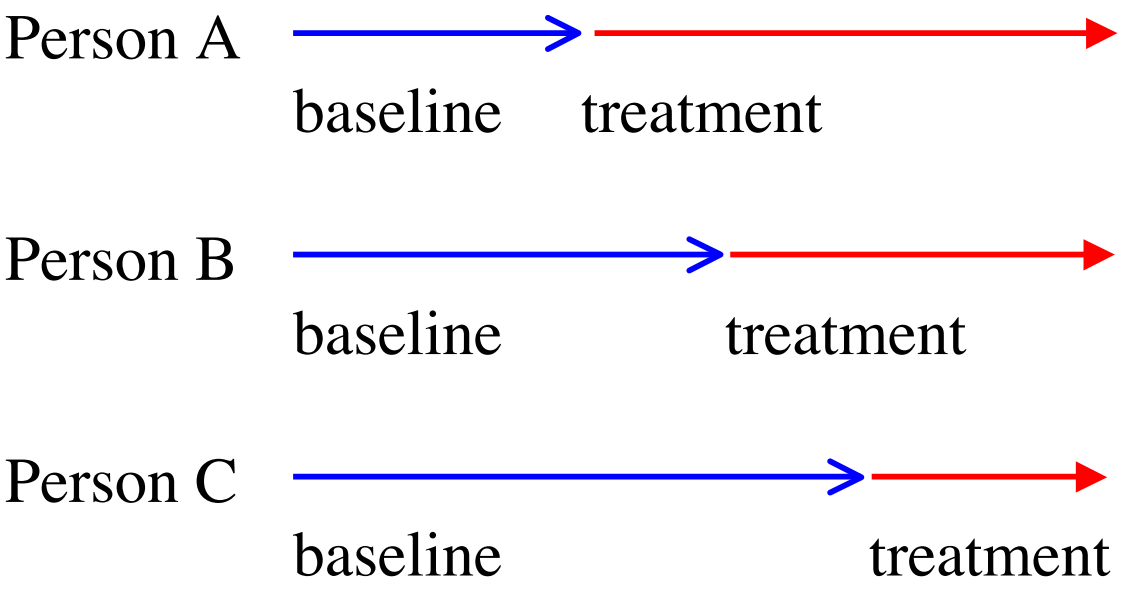
The girls got an ACABCBCB design. (A is the control, B is placebo, C is experimental).

What this does is repeat the different conditions, over time, with various other conditions before and after each (counterbalancing the order).

In the case of the two girls, the dependent measures showed that they were much more active after eating the oatmeal cookie with yellow dye, while their activity was roughly the same in the baseline and placebo conditions. So, artificial colors in foods can lead to an increase in hyperactivity (at least for these two participants).

D) Multiple baseline.

There are situations where the treatment has permanent carry-over effects, so we can not use the return to baseline type of design. For these cases, we have a small-n equivalent of the between subject design: the multiple baseline.



In this design, everyone gets the same order of conditions: baseline then treatment. The baseline conditions all start simultaneously and behavior is measured "continuously" during the study. The treatment conditions for each individual start at different times, so that it is unlikely that a variable correlated with the change from baseline to treatment for one individual would also be present for each of the other individuals *when they are moved from baseline to treatment*.

Our example is from work by Schreibman et al. (1983) in the treatment of autistic children. Older siblings (mean age 10) were trained to use behavior modification with their younger (mean age 7) autistic siblings. The older children were to reinforce (teach) their autistic siblings behaviors such as counting, learning about money, letter names, etc.

Each of the three pairs of normal and autistic children (siblings) was started at the same time. The baseline went for 4 sessions with one pair, 6 with the second and 8 sessions with the third. The study lasted 16 sessions.

The results showed that the older siblings were adept at learning the behavioral techniques and that their autistic siblings made progress in learning to count, letter identification, etc. relative to the baseline. In addition, Schreibman et al. observed the behavior of the siblings outside training sessions (unobtrusively) and found that the behavior in the treatment setting generalized.

II. Quasi-Experimental Designs

Designs in which the researcher does not directly manipulate at least one of the variables are called quasi-experimental.

For example, if we examined the role of gender (male, female) and mnemonic strategy on the recall of maps, then we have a quasi-experimental design. The mnemonic strategy is a true independent variable, but gender is a subject characteristic that can be measured and assigned to groups, but not manipulated. Since we can not form equivalent groups with regard to a subject characteristic, we need to be cautious in interpreting the results.

Another type of quasi-experiment is one involving natural treatments where the effect of some event in the environment is observed. Again, since the occurrence of the event and assignment of subjects to conditions is not under the control of the experimenter, we must be cautious in interpretation.

Natural treatment designs in which we make our measurements only after the natural event are *ex post facto* designs. They can be used to study phenomena, such as the effects of natural disasters on people, that can not be studied any other way. They can also be used as a means of assessing the effects of government programs, television violence, and many other aspects of an individual's cultural environment on behavior.

A) Nonequivalent control group pretest-posttest design

This design is similar to the pretest-posttest design. Here, we make our dependent measure before the natural event that constitutes the treatment and again afterward.

Many of these "treatments", whether the occurrence of a natural disaster, the introduction of a new curriculum for teaching reading or the introduction of fluoride in a water supply, have long-term carry-over effects. We can not use a return to baseline design.

There are two basic classes of potential confounding effects that we must pay particular attention to: maturation and history.

1. Maturation effects are biological and physiological changes that occur as a function of age.
2. History effects are events (other than the one we are interested in) that co-occur with the "treatment" we are studying.

One method for dealing with these problems is the use of a *non-equivalent control group*. This is a group that is as similar to the group that received the treatment as possible, but did not receive the treatment. This is not a true control group, because we did not use random assignment. The control subjects, like the experimental group, were selected after the fact. They are a control for effects of maturation and history. If the control group data are similar to the experimental group, then the treatment did not produce any observable effects.

B) Deviant Case Analysis.

Case studies have no controls for history, maturation and other factors.

One approach to dealing with the case study is the deviant-case analysis. Here, we identify an individual as similar to the case as possible, except for the "treatment".

Example: P.Z. is an elderly, world famous scientist who suffers from sever memory loss. P.Z. has a history of alcohol abuse. A "matched" case would be a world famous scientist, similar background, etc. except for no history of alcohol abuse. The two individuals would be compared on memory tasks.

C) Interrupted Time Series Designs

In this approach, we make our observations over an extended period of time, making multiple observations before and after the "treatment".

For example, to examine whether the number of serious automobile accidents is influenced by speed limits, we could look at the monthly number of serious accidents (defined as any auto accident involving bodily injury that required medical treatment) for 6 months before and 6 months after a change (reduction or increase) in the highway speed limit.

This design does not permit a causal inference any more than the other non-experimental designs we have discussed. Basically, we have a relational (correlational) design. The relationship we observe could be caused by other factors that co-occurred with the change in speed limit (e.g. increased enforcement of speed limits, a crack-down on drunk driving, a spike is gas prices that led to a reduction in driving).

Finding a non-equivalent control group(s) that is as similar as possible to our treatment group would increase our confidence that the effect we were observing was related to the treatment. This is the control series and this design is called a *control series design*.

A further step is to make measurements of potential third variables and try to find a non-equivalent control group that is matched on these third variables. If the effects of the treatment are still present in only the treatment group, then this increases our confidence in the results.

Remember, though, that we can't match on everything.

An example of use of the interrupted-time-series design is the research of Phillips (1983) on possible causes of homicides. The basic idea was that violence on TV that appeared to be justified, exciting, real and rewarded might model this behavior and lead to an increase in homicides. The homicide rate was examined before and after heavyweight boxing matches that were televised.

A 12.4 percent increase was found following matches. The homicide rate following other sporting events (which produce gambling, etc.) was not higher than before the events. So, some careful "detective work" here shows that it is plausible that some homicides are the result of violent behavior being modeled.

III. Program Evaluation

Program evaluation is more than simply assessing whether a program is effective using outcome measures.

Rossi et al. (1999) identified five steps in program evaluation:

1. Needs Assessment
2. Program Theory Assessment
3. Process Evaluation
4. Outcome Evaluation
5. Efficiency Assessment (cost-benefit)

A) Needs Assessment

Are there problems that need to be addressed? Research here may use surveys, interviews or statistical data from various governmental agencies and foundations. Once a need is identified, we can proceed to design a program to address it.

B) Program Theory Assessment

In designing a program, it must be based on valid assumptions about the causes of the problem. The program must address specific needs that have been identified and have a focused, specific set of goals.

This may require further research to provide details about needs and causes of the problems.

C) Process Evaluation

Once a program is designed, research is needed to assess whether it is reaching the target population and whether the services are being adequately delivered. This phase evaluates the *implementation* of a program.

Questionnaires and interviews, observational studies of the program staff and ex post facto analysis of the records kept by the program are often used in this step.

D) Outcome Evaluation

Experimental or quasi-experimental research is utilized here to see if the program is achieving its intended outcome.

Basically, we want to know what would have happened without the program and what happens with the program and compare these two.

E) Efficiency Assessment

Weigh the costs of the program against its benefits. While the determination of whether a program is worth doing is a political decision, research has a role in determining the costs of the program and the degree and value of the benefits.

IV. Designs with Subject Variables.

There are many aspects of human behavior that we would like to study that differ across subject groups. We would also like to understand how differences between individuals and groups contribute to behavior. Here, we are forming our groups based on qualities that they (the participants) posses: subject (participant) characteristics. These designs *can not show* causation with respect to participant variables.

A) Third variables

Any subject variable in a design is essentially a dependent variable. That is, we can measure the subject characteristic and determine the degree of relationship with our dependent measure(s), but we can not determine causation because of problems of directionality and third variables.

Two particular areas of psychology are primarily devoted to the study of subject variables: clinical and developmental.

One approach to reducing the influence of third variables is to use matching.

For example, if we wished to study the influence of schizophrenia on language comprehension, attention or memory, we could examine three groups of individuals: normal, schizophrenic, and a non-schizophrenic disorder group. Each of these two "control" groups would be matched to the schizophrenic group on demographic variables, intelligence, etc.

The normal group is, of course, our baseline for the absence of any clinical disorder. Our non-schizophrenic group is our baseline for a disorder that is not schizophrenia.

To the extent that the two matched groups are "equivalent" to the experimental group, we have eliminated some third variables. However, since we can not match on everything, there are no guarantees that we have eliminated all third variables.

B) Regression to the mean.

If you select subjects for groups based on extreme scores, then re-measure them later, their scores will, on average, be closer to the population mean. This is regression to the mean. It occurs because *part* of the *original measurement* that was used to select subjects was *error variance* (error in measurement). That is, the original measurement was not a perfect measurement of every subject. The most likely error for someone that scores extremely high is that their real score is lower (closer to the mean). The most likely error for someone who scores very low is that their real score is higher (closer to the mean).

Repeated testing, with equivalent forms, can be used to get scores closer to the "true score" for each individual. This type of problem is a particular issue in educational research where we are trying to determine the effectiveness of some form of remedial education or alternative teaching method for children who have low scores.

V. Developmental Research: Age as a Subject Variable.

Things don't simply change as a function of age, but many interesting phenomena do change as we age. For example, early in life, there are substantial neurological and hormonal changes.

The two most common ways of studying changes in behavior that occur with aging are longitudinal and cross-sectional.

A) Longitudinal Design

In a longitudinal design, we follow a group of individuals over time.

Changes in participant behavior could be the result of maturation or history effects. The cultures and environment in which individuals live are not constant and many events take place, so sorting out what produces changes in behavior is difficult.

Also, subjects are likely to drop out of this design, over time. This is *differential mortality*. The subjects that dropped out may have shown different results, on the dependent measure(s), than the subjects that remain. Thus, our group at the end of the study may not be representative of the population or similar to the group at the start.

B) Cross-Sectional Design

In a cross-sectional design, we study multiple groups (different ages) at one point in time.

Differences in their behavior could be the result of "generational differences" (different cohorts). That is, the history and culture at the time of birth and in the early years were different for the different age groups (cohorts).

Differences include changes in the educational system, the political climate (conservative, liberal), child-rearing practices, and economic conditions.

(Note, the term *cohort* refers to a group of individuals of the same age.)

C) Comparing the Designs

Longitudinal designs are expensive and take a lot of time. However, they are the only design that can show an effect is likely to be the result of processes associated with aging. It also has the advantage of sensitivity: the data are examined for each individual over time. That is, each participant serves as their own baseline.

Longitudinal designs are particularly sensitive to differential dropout of participants.

Cross-sectional designs take less time and do not suffer as much from problems with differential dropout. Cross-sectional designs can not separate cohort effects from maturation and learning. That is, the effects may be based in the cohort and not developmental/age changes.

D) Sequential Design

Another design is the cross-sequential design. In this design, multiple cohorts are followed over time. That is, we start with a cross-sectional design and then follow it longitudinally. We can directly estimate the influence of both cohort and time of testing from the results.

This design, however, has two disadvantages. One is that it is cumbersome, requiring both lots of subjects and lots of time to run. Second, it is *very* sensitive to differential mortality.

Its advantages are that it requires less time than the longitudinal design and (maybe) fewer participants than the full cross-sectional. It can study process over time and thus partly control for cohort effects. Because of the multiple cohorts, it does not take as long to run as a full longitudinal design.

VI. Examples

A) Developmental Design Example: The Time Lagged Design

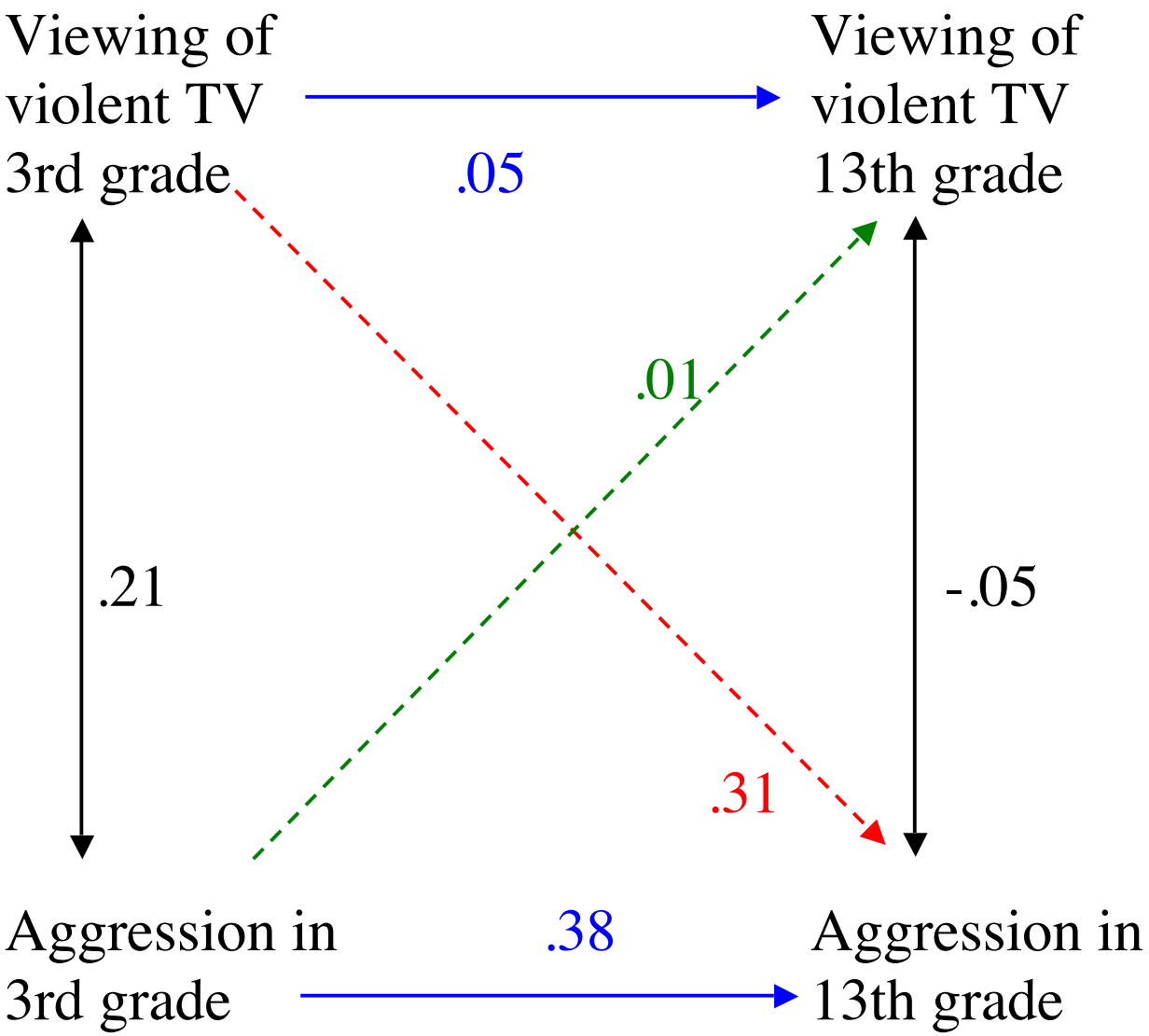
Obtain several measurements over time. On the basis of the size and direction of the correlations (+ or -) among the measures over time, the likely causal relations and the direction of any relationship between the variables can be determined.

Example: In a cross-sectional design, we find a correlation between aggressive behavior and the degree of violence in TV programs that children watch. The data on the relation between violence on TV and aggressive behavior can be interpreted in at least two different ways.

1. Watching violent programming causes aggressive behavior.

2. Aggressive children/adults show a preference for watching violent programming.

To deal with this, we use a longitudinal design with measurement of TV viewing and aggressive behavior in 3rd grade and again as college freshmen (13th grade, 10 years later).



The diagonal correlations reveal that there is a substantial correlation between watching violent TV in the third grade and exhibiting aggressiveness in the thirteenth grade (in red). No relation is found between 3rd grade aggression and later TV viewing (in green).

Since *cause precedes effect*, and the relation between aggressive behavior in the 3rd grade and viewing of TV violence ten years later is so low (green line), it appears that the *direction of any causation is from watching TV violence to aggressive behavior*.

Problems:

1) Internal validity is still partly suspect. The lack of control means that a third variable could be producing the observed relationship.

2) Time consuming.

B) Program Evaluation Example:
The Westinghouse-Ohio study of the effectiveness of the Head Start program.

Group 1 was randomly selected from children completing the Head Start program. Group 2 were children that had not done head start but were matched (to Group 1) for sex, race, ethnic group and kindergarten attendance. After the two groups were formed, additional measures of socio-economic status, demographics and attitudes were collected, with only small differences between the two groups being found.

This is a *post test only, non-equivalent control group* design.

No real difference was found in academic performance (after head start was completed) between the two groups. The study's authors drew the conclusion that Head Start was not effective in remediating the effects of poverty and social disadvantage.

However, because of the subject selection process, it is also possible that regression to the mean took place.

First, it is likely that the populations from which the two groups were drawn were *not equivalent*. Children were included in Head Start (Group 1) based on poverty and social disadvantage. Thus, as a group, they were likely to start out *lower* in scores of academic achievement than the population from which the control group (Group 2) was selected.

Second, the study authors tried to match on measures to form equivalent groups for their study. Thus, the control children were likely to have scores *below* their population mean.

Now, we have non-equivalent groups with the control group selected for "extreme" scores. On re-testing, what is likely to happen is regression to the means of their respective populations. That is, the control group would score higher.

This would "cause" the control group to appear to be doing as well, academically, as the Head Start group. In fact, much of the improvement in the control group is likely to be regression to the mean.

Does this mean that the Head Start group actually did improve more, since the two group scores were the same at the end of the study?

We *can not* draw this conclusion. Because we can not estimate the size (influence) of any regression to the mean, we can not draw any conclusions regarding Head Start from this study.

The Westinghouse-Ohio study is an example of a well intentioned study that resulted in bad science. Drawing invalid conclusions does not advance our understanding nor does it help to formulate good public policy. Rather, it causes controversy and leads to a general distrust, by the public, of both the issues and science in general.

While the self-correcting nature of science eventually reveals the problems, this is often "lost" on the public (or politicians) who just remember the first, erroneous, reports.

Answers for Chapter 10 Sample Questions

1) – a; 2) – d; 3) – a; 4) – d; 5) – a; 6) – c

Sample Exam Questions, Chapter 11

1) Subjects in a non-equivalent control group: a) are similar to the group that received the treatment, but do not receive the treatment b) are controls for maturation and history effects c) are randomly assigned to the control group d) a & b above

2) Which of the following are potential factors in determining the results in a longitudinal design? a) changes in the subjects' behavior may be the result of maturational changes b) changes in the subjects' behavior may be the result of history effects c) the group at the end of the study may not be representative of the population due to differential mortality d) all of the above

3) Which small-n design should be used when a treatment has permanent carry-over effects? a) ABAB design b) multiple baseline design c) return to baseline design d) none of the above

4) If we wanted to determine if hospital admissions for asthmatic attacks were related to air quality, which of the following would be an appropriate design? a) pretest-posttest b) between subjects with random assignment c) interrupted time series d) a & b above