CHAPTER 15

Analysis of Experimental and Quasi-Experimental Data

Pinpointing Explanations

RICHARD GONZALEZ
TIANYI YU
BRENDA VOLLING

A major function of data analysis is to facilitate inferences about psychological and behavioral processes. Analytical tools help us learn about underlying processes. But textbooks typically do not highlight this key function of data analysis, instead focusing on the multitude of problems one must address when analyzing data, such as correcting the p-value for multiple tests or using the right error term in an F-test or discussing what goodness-of-fit measures to report in a paper. Textbooks tend to ignore the nonstatistical issues that can interfere with making solid inferences from data, even when one computes a p-value correctly.

In this chapter we present data analysis in a different light. We highlight the role of data analysis in making solid inferences about psychological, behavioral, and developmental processes from experimental and quasi-experimental designs. As researchers, we conduct studies because we want to learn something we did not already know (though sometimes studies are conducted for other reasons, such as gathering information to make predictions). The goal is to ensure that the design and data analytical procedures we use do not interfere with our ability to learn about underlying developmental processes. The chapter highlights how design decisions and data analytical procedures work together in service of illuminating our understanding of development, which sometimes goes beyond the practice of computing p-values in a test of significance.

The classic monographs by Campbell and Stanley (1963), Cook and Campbell (1979), and Shadish, Cook, and Campbell (2002) are consistent with this general view. These influential works list problems that can interfere with a researcher's ability to make solid inferences from data. Their lists can be characterized as "top pitfalls to avoid in psychological research." Random assignment into treatment conditions bypasses many of these pitfalls,

but often research settings in developmental psychology do not allow randomizing units into conditions. One cannot randomize participants into levels of gender (male or female) or randomize participants into having or not having an older sibling. The list of pitfalls, or threats to internal and external validity as Cook and Campbell called them, has played an important role in strengthening developmental research by bringing to mind a standard set of alternative explanations.

Several generations of psychologists have been influenced by this work. Our focus in the present chapter is more on what needs to be done in developmental research rather than what needs to be avoided (we do discuss what to avoid as well). To make a nonresearch analogy: A novice chef will not get very far when learning how to make a soufflé if she or he is provided with a list of what not to do in the kitchen; she or he also needs a set of instructions about what to do in order to make a successful soufflé. This analogy holds in data analysis as well: Researchers need a set of instructions for conducting successful developmental research.

This chapter takes a contemporary perspective on a few of the points raised by Campbell and his colleagues, presents them in a constructive manner, and highlights a few of the newer analytical tools. This chapter serves as an introduction to the topic and, we hope, will motivate readers to pursue the more detailed, book-length treatment in Shadish and colleagues (2002).

Design for the Research Question

Allen Edwards, an early methodologist in psychology, pointed out that the real concern in psychological research should be not in finding the significance of hypotheses but rather in finding significant hypotheses to test. All too often researchers become focused on concerns of significance testing, such as whether there is sufficient statistical power to detect differences or in using a fancy new statistical procedure, but they do not give sufficient attention to developing research hypotheses that are worth testing. We suggest spending time finding good research questions. Several researchers have written advice pieces about how to generate research ideas (McGuire, 1973; Wicker, 1985).

Once you have the research question and have decided that it is an empirical question (i.e., a question that data, in principle, can answer), the next step is to conceive of a research design that can test that research question. You need to consider a design that can answer the research question in a manner that is as clean as possible from alternative explanations. The design should be able to test the research question. If the research question is to test two competing predictions, then the design should be one that can produce data to adjudicate between the two theories. What tasks will you give the research participants? How will you recruit them into the study? Who are the appropriate participants for the research question? What will you measure or observe? What will you manipulate or control or counterbalance? Will you be able to make sense of other patterns in the data besides the one that you predict may emerge from the study? This last question can salvage a study because, if it is considered in advance, you may be able to include the appropriate measure or control that will permit conclusions from an alternative data pattern.

The selection of research design turns out to be critically connected to the research question. Typically, one cannot choose from among a set of ready-made research designs but must tailor the research design to fit the research question. It takes much effort, which

is the reason that having an important or "significant" research hypothesis from the beginning is so important. The research question has to be worth the effort.

We include this point about selecting the research design in a chapter on data analysis because often the best data analysis advice is to think through one's design from the beginning so analyses can be straightforward and simple. Too often, complicated and fancy statistical techniques are necessary because there were problems with the research design. Had those problems been addressed up front, perhaps there would be no need for heroic, life-saving, complicated data analytical procedures. Our own bias is that simple data analysis, when warranted, is best, and careful planning in the early stages of the research process can sometimes lead to simpler analytical procedures. Of course, sometimes the research question entails using more complicated statistical methods.

Attributing Explanations to Data Patterns: Experiments

Some methodologists and researchers have treated the lab experiment as the gold standard in psychological research. The experiment is characterized by (1) a manipulation of some sort by which some participants are placed in one condition and other participants are placed in different conditions, such as a control group; and (2) random assignment to those different conditions. In the case of two conditions, both conditions should be identical in all respects except for the key component that is manipulated. The comparison of these two groups allows us to assess the role of the key component because the two intervention groups are treated in an identical manner except for the key component. Any difference between the two groups on the outcome measures can be attributable to the key component. For example, in a developmental study on emotion regulation, the manipulation, say, could be whether the mother is present or absent in the testing situation. The presence of the mother would be the key component that distinguishes the two conditions; the conditions are identical in all other respects.

If the two intervention, treatment, or manipulation groups differed on more features, then it would be difficult to attribute the reason for the outcome differences to the key component (unless the components are arranged in a factorial design that permits teasing apart their separate influences). If the researcher selected, say, left-handed children to receive one intervention and right-handed children to receive the other, the assignment would be based on handedness. This would introduce a confound, the potential for an alternative explanation, because the two groups of children could differ on the measured variable because of their handedness rather than because of the difference between the treatments they received. Thus the ability to attribute the difference between group averages to the key component would be compromised—the observed difference could be due to the key component, could be due to handedness, or could be due to some combination of handedness and the key component. Random assignment to conditions helps reduce such attributional ambiguity in explaining differences between groups.

Furthermore, the researcher needs to check that only the intended component was manipulated. The experimenter may intend both emotion regulation conditions to be identical in all respects except for the presence or absence of the mother, but in practice it may not turn out that way. For example, the research assistant may interact more with the child in the mother-absent condition, making it difficult to pinpoint the explanation for the difference observed on the dependent variable. Is the difference between conditions attributable

to the mother's absence, attributable to a change in the quality of the interaction between child and research assistant, or a combination of both? Differences between conditions can exist for many reasons, not only the intended manipulation, and sometimes those differences are not readily apparent to the researcher.

It is best to tackle these issues before the data collection phase. For example, children can be assigned to intervention groups through a random process. Under traditional statistical theory, random assignment to conditions allows one to rule out many alternative explanations when comparing differences across levels of a manipulation, such as two types of treatments, in which children who receive a new kind of instruction are compared with children who receive a traditional type of instruction. The experiment provides a level of control that permits clean explanations of data patterns.

However, in no way is random assignment sufficient for such inferences. Other details could cloud the ability to explain the findings, or "to make causal inferences," as some like to say. For example, an invalid measure for the outcome variable could damage an otherwise well-designed study that used random assignment, or the manipulation may involve more features than the researcher intended (such as in the emotion regulation example mentioned earlier in which the research assistant interacted differently with the participants in the different conditions). Random assignment is not sufficient, but it goes a long way toward eliminating many alternative explanations. Random assignment can eliminate alternative explanations involving comparison of different conditions, but it doesn't help much with other problems, such as issues with measurement.

There is an important lesson here. Sometimes a study is made stronger by careful thought and decision making before data are ever collected; this can sometimes make subsequent data analysis relatively easy. Too many researchers rely on complicated data analytical procedures to fix problems that sometimes could be addressed through relatively simple design decisions.

Attributing Explanations to Data Patterns: Quasi-Experiments

Random assignment to intervention, condition, or treatment group is not always possible. For example, a researcher studying the effect of divorce on children's academic performance cannot randomly assign families to the divorced or not divorced conditions. The researcher could, however, compare children whose parents divorced to children whose parents did not divorce. There is value in such a comparison, but the researcher must exercise care in interpreting group differences because the two conditions (divorced and nondivorced families) may differ on other attributes in addition to divorce status (e.g., Stewart, Copeland, Chester, Malley, & Barenbaum, 1997). Other variables commonly used in developmental research that cannot be subject to random assignment include gender and age.

Many research questions involve comparisons of naturally occurring groups. The researcher can compare boys with girls, divorced with nondivorced families, 3-year-olds with 5-year-olds, late adolescents with adults with older adults, and children who start first grade early with those children who delay first grade for 1 year. In these cases there is a natural "comparison" across levels, even though random assignment is not possible. But the researcher needs to deal with alternative explanations that are present in making such comparisons of naturally occurring groups. There are ways of dealing with such confounds,

which gave rise to quasi-experimental designs. These types of designs provide useful ways of minimizing alternative explanations, mostly through careful construction of comparison groups but sometimes through careful data analysis.

The validity of a research paradigm can be characterized on two major features: *internal validity* and *external validity*. Internal validity is the ability to make solid inferences about the process. Does the study design allow one to attribute an explanation to the correct key component without contamination from alternative explanations? External validity is the ability to generalize the findings to other subject populations, other types of research settings (e.g., from lab to classroom), or other types of stimulus materials (e.g., from printed page to computer screen or from a paradigm that uses stuffed animals with children to one that uses real pets). Methodologists have tended to view these two types of validity as a zero-sum game (you can have one only at the expense of the other). If you want to optimize internal validity, the common wisdom instructs, then you need well-controlled studies with clear manipulations and random assignment in order to attribute findings to relatively unambiguous explanations. If you want to optimize external validity, then you sacrifice internal validity in favor of generalizability, not only in terms of subject population but also in other senses, such as the robustness of findings across different settings (e.g., when classrooms naturally vary on many dimensions). We do not discuss other types of validity that have been studied in the literature (e.g., construct validity, predictive validity, face validity, discriminant validity).

New designs can be developed that yield both high internal validity and high external validity. There is a sense in which the Cook and Campbell tradition attempted to do this by focusing on quasi-experiments with their list of pitfalls, but there was still an element of sacrificing internal validity with the modifier *quasi* in front of the term *experimental*. There is room for much more innovation and creativity in developing new research designs that yield both high internal validity and high external validity. Shadish and colleagues (2002) offer some preliminary solutions using multiple methods across studies.

There has been an interesting counterproposal in the recent methodology literature that challenges the gold standard status of the experiment—that is, a challenge to the standard belief that internal validity has a higher priority than external validity. This gold standard status has not typically been given to the experiment in other social sciences such as economics and sociology. Indeed, some, such as the Nobel laureate econometrician James Heckman (2005a, 2005b), have argued that social science researchers should make the field experiment the gold standard because it maximizes external validity and because one can do a decent job at achieving internal validity if appropriate statistical techniques are used. A theme of the general argument is that sometimes threats to internal validity can be addressed through statistical means much easier than can threats to external validity. (An example is provided later in the chapter.) So, to paraphrase, his recommendation is to design studies to maximize external validity (which involves design decisions made before data are collected) and let statistical techniques deal with issues of internal validity (Heckman, 2005a, 2005b). It is still too early to tell whether this approach will be fruitful for developmental researchers, but it is worthwhile at least to think about the rarely challenged assumption in developmental psychology that the true experiment should be the gold standard. At minimum, we should give more attention to the general concerns of external validity in our research and to the creation of new designs that permit solid causal inferences when random assignment is not possible.

The Right Comparison Group: Matching and Propensity Scores

The implicit assumption in experiments and quasi-experiments is that there are two or more groups that can be compared on the key components. Regardless of whether the study is an experiment or a quasi-experiment, the researcher should construct good comparison groups. Returning to the example of the effect of parental divorce on a child's academic performance, we could compare children whose parents have divorced with children whose parents have not divorced. But then we would need to be careful about equating all other differences. We could match families on demographic variables, such as age of parents and children, number of siblings, socioeconomic status, parents' race and ethnicity, parents' immigration status, parents' education, whether other family members such as grandparents live at home, and so forth. Hopefully, we could find pairs of families that are similar on all those dimensions, attributes, and variables except that one set of parents is divorced and the other is not. In effect, this is trying to mimic the result from a randomized experiment in which groups vary on key components and not much else. Such matching would be difficult to implement in practice, but it provides one option for creating natural comparison groups (see Shadish et al., 2002, for a nice review of matching). One example of a research paper using matching in the context of divorce is by Lansford and colleagues (2006), who used data analytical procedures that allowed sophisticated matching.

A new house on the methodological block is a set of techniques based on propensity scores. The logic of propensity scores is relatively simple (for more details, see Rubin, 1997, or the book-length treatment by Gao & Fraser, 2009). One takes a two-step approach to the data analysis problem. The first step is to use the available covariates to model the propensity to be in the treatment group. For example, although families cannot be randomly assigned to be divorced or not divorced or to have a second child or not to have one, it is possible to model the propensity for a family to be divorced or to have a second child. Thus a propensity to be divorced can be predicted from covariates such as the parents' employment and level of education, the length of the marriage, and so on. This first step resembles risk models seen in epidemiology and biostatistics in which the dependent variable, say, is a binary outcome (such as divorced–not divorced), and a set of covariates, or predictors, provides a basis for a logistic regression. The predicted scores from the logistic regression can be transformed into probabilities, and they become propensity scores.

With propensity scores in hand, the analyst can proceed to step 2. The analyst can create a subset of the data with suitable matches on propensity scores (e.g., two families with the same propensity score, with one divorced and the other not). There are several ways of accomplishing this matching on propensity (see Gao & Fraser, 2009). Another way to perform step 2 is to stratify the sample on the propensity score. For example, create five groups having propensity scores between 0 and 20%, 21 and 40%, 41 and 60%, 61 and 80%, and 81 and 100% (e.g., Rubin, 1997). Then, within each propensity group, compute mean differences between those families who are divorced and those who are not and perform tests within each propensity group. This general procedure works well and provides a diagnostic in that if there is an interaction between the five strata of propensity scores and the divorce status (say, in the context of a two-way ANOVA with treatment and propensity group as the two factors), then the first step may have been misspecified because there still is a relation between propensity and "treatment" (e.g., maybe the first step missed important predictors, or more complicated nonlinear or interaction terms are needed in the logistic regression in step 1). This approach also provides another diagnostic: It can identify whether there are a sufficient number of

participants in each group to make an appropriate comparison. For instance, if there are no participants in one group (e.g., at the lowest propensity level there are no divorced families), then one cannot estimate a treatment effect for that propensity subgroup.

The propensity score approach is one way to model self-selection into treatment conditions. This becomes important when modeling the primary dependent variable of interest, which could lead to biased results if propensity or self-selection effects are not properly controlled. This is one example of the kind of "statistical fix" to improve internal validity that Heckman had in mind when arguing for the field study as the gold standard, though Heckman's own techniques to deal with self-selection are different from propensity score techniques. Heckman (2005b) makes the important point that sometimes the theory that one is testing suggests different ways of handling covariates that may not be the same as that suggested by a general purpose statistical technique such as propensity scores.

A less desirable approach at step 2 would be to run an analysis of covariance (ANCOVA) using the propensity score as a covariate; this approach imposes strict assumptions on the relation of the propensity score and the "treatment" effect. ANCOVA is commonly used in developmental research for dealing with the problem of "equating" different groups when random assignment cannot be done. It adjusts statistical parameters by removing the linear and additive role of the confounding variables. In practice, most researchers merely enter the covariate as a linear predictor without also checking whether nonlinear or interaction terms are necessary. If the data suggest a nonlinear relation of the covariate but the analysis merely includes a linear term, then the analysis does not completely remove all variance associated with the covariate. Another problem with the ANCOVA approach to dealing with confounds is that one cannot easily assess whether the groups overlap on values of the covariate, which is an important condition for the logic of ANCOVA to apply. In general, ANCOVA methods depend on many assumptions, and the newer propensity score approaches may be preferred. ANCOVA is still a useful alternative to difference scores when modeling change, but that is a different use from the one we are describing here.

Indeed, much of the modern literature in statistics has not been kind to ANCOVA approaches to correcting for covariates and adjusting models. Psychologists have mostly ignored the important new developments in other social sciences, statistics, and computer science that have provided new methods for making causal inferences (e.g., Pearl, 2000), with propensity scores and selection models being two examples. Some psychologists have argued that ANCOVA tends to work relatively well in most cases and perhaps not much worse than the newer methods (e.g., Steiner, Cook, Shadish, & Clark, 2008). Still, it will be useful for developmental researchers to be aware of alternatives to ANCOVA, such as propensity scores.

The propensity score method provides one way of equating groups, because it can equate along multiple dimensions simultaneously. This is seen as a strength by some, given that many dimensions and attributes are addressed in the matching process, but it could be criticized by others as not being clear on the manner in which the units are matched. To illustrate, consider the many different ways a developmentalist could construct comparison groups to examine the effect of divorce on academic performance. One could compare single-parent homes of, say, widows and widowers to single-parent homes of divorced parents. In this case, the single-parent status is held constant but what varies is the reason for the single-parent status (divorce or death of parent). However, the death of a parent obviously has its own psychological consequences that may be different from the effects of a divorce, so this comparison could be criticized as not being a fair comparison, even when

holding all other aspects constant except divorce status. Another logical comparison could be to compare divorced couples who share equal custody of children with married couples who live in different cities due to work and whose children spend equal time in both households. The process of switching households is similar in the two types of families, but one family is divorced and the other family is not. Is this a fair comparison? It may be a good comparison if the goal is to equate the process of having children move around different households, but this may not be a fair comparison with which to examine the effects of divorce. Maybe couples who work in different cities have different financial struggles, so any difference between the two sets of families could be due to divorce status or could be due to differences in, say, financial difficulties.

Overall, researchers should take seriously the problem of comparison and how to handle situations in which random assignment is not possible. Careful thought about design and construction of comparison groups (including matching, propensity score methodology, and ANCOVA approaches) will facilitate clear inferences from one's research.

Comparisons over Time versus Comparisons over Groups

There are other ways of addressing the comparison problem. Rather than focus on two different groups, it is possible to focus on change over time for a single group. For example, the researcher could collect data on the child's academic performance before and after the divorce. This is a different line of attack on the problem of addressing the role of divorce on academic performance. It compares the same child with him- or herself by examining the child before and after the divorce or by following the child multiple time periods after the divorce to estimate trajectories. The analyses of such data can be handled by repeated-measures and latent growth curve approaches.

This design approach, however, is not free from inferential problems. How far back in time before the divorce should the researcher go in the school records to collect data on academic performance? Maybe family life was affected by predivorce events that eventually led to the divorce, and those events may have affected the child's academic performance. Comparing predivorce to postdivorce academic performance may provide a biased estimate of the effect of divorce given that the predivorce academic performance score could have been influenced by emerging family dynamics. How far back in time does one go to assess the "pre"? How far into the future should the research collect "post" data? Does the process act relatively quickly within the same school year, or are the effects more distal, taking several years to show marked effects? Does the time course vary by individual?

There are clever hybrid designs that combine two types of comparison techniques—the comparison over two or more times and the comparison of two or more groups. For example, the researcher could compare families who have divorced to those families who have not in the context of a predivorce versus postdivorce temporal comparison. So, academic achievement could be tracked, say, 2 years prior to the divorce (e.g., by retrospective data collection) and 2 years after the divorce, and data could be collected over the same time frame for matched families who did not divorce. This type of comparison will allow, for instance, an examination of whether there were initial retrospective differences predivorce between the two types of families. It will also permit comparison of the trajectories, that is, changes over time, of these two types of families. If the children of divorced couples show a change in academic performance patterns over time that differs from those of the

nondivorced couples, especially if the two groups of students were relatively similar 2 years before the divorce, then we can be more confident about divorce as an explanation for the performance effect (e.g., Lansford et al., 2006). The analysis of such designs can proceed with repeated-measures or latent growth curve models, with the inclusion of a grouping variable to address the between-subjects comparison. The matching of divorced and nondivorced families provides an inferential boost when interpreting the effect of divorce on the academic performance trajectories.

Time: A Fundamental Developmental Variable

Developmental psychology is concerned with development, and development occurs over time. The issue of how best to capture the unfolding change over time in a research project turns out not to be as straightforward as one might think. Two simple research designs that may occur to the reader are to compare two groups that differ in age, such as comparing 3-year-olds with 5-year-olds, and to compare a single group of participants as they develop, such as observing a group of children at age 3 and then again 2 years later when they are 5 years old. The former design, making use of two groups of different ages, is called a *cross-sectional* design; the latter design, making use of a single group of participants over time, is called a *longitudinal* design. The analysis of these two basic designs is relatively simple. The independent sample *t*-test serves as a basis for the cross-sectional design (comparing two means from different groups of participants), and the paired *t*-test (comparing two means from the same group of participants) serves as a basis for the longitudinal design. When cross-sectional designs have more than two groups, then between-subjects analysis of variance (ANOVA) can be used to analyze data; when longitudinal designs include more than two testing times, then repeated-measures ANOVA can be used. Both types of ANOVA can be generalized to include covariates (though consider our previous discussion of ANCOVA with respect to covariates), to include more complicated random effects and error structures, especially in the case of longitudinal designs and latent growth curve models, and to include other types of dependent variables such as binary data. Both cross-sectional designs and longitudinal designs have their pros and cons, as we saw in the previous section with the example on divorce.

Consider a developmental question about emotion regulation. The researcher believes that there is a key developmental process that occurs roughly at age 3 or 4, so decides that the key comparison to make in a study is between children of age 2 and children of age 5. The researcher designs a laboratory task with behavioral observation to test the question and carefully administers the task to both age groups to avoid floor or ceiling effects (i.e., not using a task that is too difficult for the 2-year-olds or too easy for the 5-year-olds). There still is a choice to make between a cross-sectional and a longitudinal design. Should the researcher compare a group of 2-year-olds with a group of 5-year-olds (i.e., compare two different groups of participants) or measure a group of 2-year-olds, then wait 3 years until they are 5 years old to test them again (i.e., compare the same participants with themselves at different ages)? The answer rests mostly on whether one wants to study age differences or age changes, in addition to practical considerations such as whether waiting for the longitudinal data is feasible in the time frame of a dissertation, funding considerations, and so forth.

If the researcher hopes to collect data the same year, say 2012, then the cross-sectional design could be useful. Cross-sectional designs allow estimation of age differences. This

design also has drawbacks. Development is not directly assessed because the comparison is between different groups of participants, or cohorts. Maybe the two cohorts had different exposure to television shows (e.g., a new public television show for toddlers targeting emotion regulation aired in late 2011, such that many of the 2-year-olds were exposed to a type of information that was not available to the 5-year-olds when they were 2). Maybe the social interaction with the experimenter differed (the 5-year-olds interacted with the research assistant in a different way than the 2-year-olds), which may have affected the performance on the experimental task. In this case, the difference between the two age groups may not be attributable to the development of the emotion regulation construct but to, say, social interaction, so comparing the mean difference between groups of 2- and 5-year-olds may not lead to solid inferences. The researcher needs to be careful about making comparisons between different groups of participants when participants were not randomly assigned to those groups. Techniques such as matching and propensity scores may control for some alternative explanations, but a cross-sectional design is silent about age change.

One way to view a cross-sectional design is that it has many missing data—2-year-olds are not observed at age 5, and the 5-year-olds were not observed at age 2. These missing data preclude direct examination of change over time and require additional assumptions in order to justify the usual comparison of means between a group of 2-year-olds and a group of 5-year-olds (e.g., Maris, 1998). It is difficult to make claims about individual growth or change patterns (*intra*individual comparisons) when one does not observe change directly but must infer change from cross-sectional data (*inter*individual comparison). If the goal is to make claims about a difference between two time points for a particular person but the researcher has cross-sectional data, then he or she can compare the means of two or more groups. This requires making assumptions, such as the similarity of change processes over the set of individuals, that justify the computation of a group mean. Such assumptions should be evaluated for plausibility.

The longitudinal version of this research design—observing a group of toddlers at age 2 and the same children again when they are 5 years old—more directly addresses change over time by focusing on comparisons within the same participants. This design offers a more direct assessment of development, the estimation of age changes, because data can be analyzed by comparing scores at age 2 and at age 5 for each child, something that cannot be done in the cross-sectional design. Although longitudinal designs have advantages, they also present interesting challenges, such as dealing with practice effects that can confound estimation of age change (Greenwald, 1976).

Obviously, for this longitudinal example data collection cannot happen completely in 2012. The researcher using a longitudinal design can collect data in 2012 when the toddlers are 2 years old and then again in 2015 when they are 5 years old. There may be practice or learning effects on repeated observation. For example, if a laboratory task is administered both at age 2 and at age 5, the 5-year-olds may remember aspects of the task, so that differences in performance may not be due solely to maturation but to other processes, such as memory, as well. There are also similar problems to those with cross-sectional designs. Events and experiences may have happened between observations, not all related to the particular maturation process under study, that may add confounds when attributing explanations to observed differences on the dependent variable.

Methodologists have offered hybrid designs that optimize the advantages of both cross-sectional and longitudinal designs. One such design begins with a cross-sectional design, such as collecting data in 2012 from a group of 2-year-olds and a group of 5-year-olds. Then the researcher assesses the same children in 2013 and again in 2014, so there is a longitudinal

design for a group of 2-year-olds that started in 2012 and a separate longitudinal design for a group of 5-year-olds that also started in 2012. This type of dual cross-sectional and longitudinal design, called cross-sequential design, permits the kinds of comparisons that characterize each of their component designs (Baltes, Reese, & Nesselroade, 1988). For example, within the same design one can compare age differences in 2012 by testing whether the mean on the dependent variable of interest differs between the 2-year-old group and the 5-year-old group. One can also examine longitudinal questions within the same design by comparing means on the same children over time (such as the group of 2-year-olds in 2012, then again in 2013 and in 2014). These hybrid designs also permit additional comparisons not possible with cross-sectional-only or longitudinal-only designs. For example, one can compare data from the 5-year-olds collected in 2012 with data from the 5-year-olds collected in 2015 to test whether 5-year-olds in 2012 differ from 5-year-olds in 2015. For a recent example of a quasi-experimental longitudinal design examining the effect of a school-based life skills intervention on alcohol use, see Spaeth, Weichold, Silbereisen, and Wiesner (2010).

The analysis of such hybrid designs turns out to be somewhat difficult. It appears on the surface that there are three factors: age of child, cohort, and time of measurement (see Masche & van Dulmen, 2004; Schaie, 1970). But it turns out that these three variables are related to each other. As Baltes (1968) showed, if you know two of the variables (such as age of child and cohort), you automatically know the third variable (time of measurement). One can conduct two-way ANOVA analyses on such designs (e.g., treating age of child as a within-subjects factor and cohort as a between-subjects factor). There are alternative analyses that permit comparisons of all three variables in a reduced design (e.g., sacrificing some of the higher-order interactions). These analyses allow the simultaneous study of age–cohort–period relations by making some simplifying assumptions, such as an equality assumption that two adjacent ages have the same effect (e.g., Mason, Mason, Winsborough, & Poole, 1973). Sometimes such simplifying assumptions may be reasonable (such as setting a constraint that the effect for 8-year-olds is the same as the effect for 10-year-olds in the context of a study that also includes 5-year-olds and 20-year-olds, given that 8- and 10-year-olds may not differ much on the dependent variable). But one must be careful that such equality constraints do not mask important data patterns (e.g., one may not want to set the effect for 3-year-olds to be the same as for 5-year-olds if one believes that relevant developmental processes have occurred during that period). Some advances have also been made in merging the analysis of longitudinal and cross-sectional data in the context of structural equation modeling (e.g., McArdle, Hamagami, Elias, & Robbins, 1991).

Describe What You Observed and What You Didn't

Data analysts spend much time on the details of the statistical test—choosing the model, checking its assumptions, fitting the model, learning the computer program, evaluating the fit, and reporting tests of significance. These are indeed important aspects of data analysis. Unfortunately, too little time is spent communicating the results in terms of what was observed. Too often researchers report the statistical model and its significance without also reporting the corresponding descriptive statistics that give substantive meaning to the results section. How often have we read sentences such as "A two-way ANOVA supported the predicted interaction; no main effects were observed" with little else to inform the reader about the observed data pattern? Report and discuss the descriptive statistics such as means, correlations, regression slopes, and relevant variance differences.

As with most use of descriptive language, one needs to be mindful that the words match the conclusions that are possible from the design. For example, in a cross-sectional study, one could correctly report means and standard errors, but it would be incorrect to discuss patterns in the means across cohorts as "age change" because change was not directly observed. Use the phrase "age difference" when discussing differences between means in a cross-sectional design, and use the phrase "age change" when discussing patterns of means in a longitudinal design.

Good descriptive reporting can include a table of measures of central tendency such as means and their corresponding standard errors. Or a table of correlations or regression slopes, along with their respective standard errors, can be reported if the association between variables represents the key measure. We support the recent trend to report confidence intervals for the key parameter estimates (e.g., Cumming & Finch, 2005). The combination of reporting parameter estimates and their confidence intervals highlights both the effect (such as differences between parameters) and its uncertainty (the confidence interval). A statistical test could have a p-value of 0.04 because the differences in parameters are small but the confidence interval is also small, or the same p-value of 0.04 could arise because the differences in parameters are relatively large but the confidence intervals are wide. But as statistical models become more complicated and dozens of parameters emerge, such as in a latent growth curve model, structural equation model, or mixture model (e.g., indicators of latent variables, error variances, covariances between factors and between errors, different parameter values for different latent classes, random effect variances, etc.), the suggestion of reporting parameters and confidence intervals can get unwieldy. Methodologists need to develop better visualization methods, especially for the case of multivariate longitudinal data that communicate the parameters, their variability, and the model fit.

We should also be mindful of dealing with what we did not observe. This is the point in the chapter at which we discuss the elephant in the developmental living room—missing data. Much of developmental research, especially in longitudinal studies, suffers from missing data. There could be many reasons, such as that the participant no longer wants to participate, or he or she can no longer be contacted, or the parent had to leave the session early so that the child could not finish the battery of tests, and so forth. It is important to report "missingness," such as percentages of missing data across conditions or data collection waves, much as any other descriptive statistic is reported. The handling of missing data is a major topic in data analysis and can get somewhat complicated if the missing data are related to variables or processes. Just because a statistical technique advertises that it can "handle" missing data does not mean that the user can proceed mindlessly with the analysis. The analyst should be mindful that proper inferences from such analytical tools depend on a set of important assumptions being satisfied and that those assumptions cannot always be checked. We refer the reader to the book-length treatment on missing data by Little and Rubin (2002). Learning how to handle missing data appropriately can go a long way to improve the inferential quality of the observations one has in hand.

Model the Statistical Process in Developmentally Meaningful Ways

We now turn to the important role of statistical models in data analysis. Even the simplest parameters or descriptive measures are computed in the context of a statistical model. When we compare the difference between the mean of 3-year-old boys and the mean of

3-year-old girls, we are making many assumptions. We are assuming that the mean is a good representation of the data. Does that single number for the boys provide an adequate summary of the observations? In order to attach a p-value to the difference between two means, we need to make assumptions about the nature of the variance (e.g., in the classical statistical test, that the population variance for the boys is assumed equal to the population variance for the girls). Maybe that is an unrealistic assumption for the particular data set. Maybe there are developmental or cohort reasons that the variances could differ across the two groups. In many classical statistical tests, the emphasis is on testing particular parameters such as means, with other parameters such as the variance playing supporting roles. There is no a priori reason that that should be the case. Maybe the difference in variance is the key research question that needs to be tested; then, rather than using a remedial measure such as transforming the data to eliminate the variance differences, one could test the variances themselves and let them serve as the key hypothesis test. We encourage developmental researchers to focus on parameters that are developmentally relevant and to consider statistical models that directly estimate and test those developmentally relevant parameters.

Variances also can play a more critical role in complicated models. In latent class models that are becoming popular in developmental psychology, it is common to make an equal variance assumption across classes. The typical assumption of equal variances across classes could be questioned. Why should variances be equal across classes? Of course, sometimes an equal variance assumption is necessary to identify the model, but this is an area to which methodologists should give much more attention. How can the equal variance assumption be relaxed in a way that minimizes convergence problems? How can substantive hypotheses about different variances across clusters or changes in variance across time be tested appropriately? General latent class and mixture models that relax many of the restrictive assumptions can be estimated using modern Bayesian statistical methods and their associated Monte Carlo sampling techniques. Psychologists have lagged behind other social scientists in the adoption of modern Bayesian statistical methods, which can sometimes solve problems that appear intractable or quite complicated from the perspective of the usual frequentist statistical theory that characterizes much of the statistics used in developmental psychology. For an introduction to Bayesian methods as applied to general regression methods, see Gelman and Hill (2006).

Heterogeneity: Random Effect Models and Latent Class Models

Developmental psychologists realize that the heterogeneity present in developmental studies can illuminate key developmental processes. For some research questions, capturing the heterogeneity present in a population is critical, whereas for other research questions one strives to minimize the heterogeneity present in the sample under investigation. The former is usually a concern in quasi-experimental designs and the latter in experimental designs. Capturing heterogeneity can be viewed as one step toward external validity (capturing the variability present in the population), whereas reducing heterogeneity can be viewed as one step toward internal validity (controlling the variability present in the population). Sometimes, though, heterogeneity can be modeled directly and can provide meaningful process information in both experimental and quasi-experimental designs.

The main point about heterogeneity is that every research participant can respond differently, and this variability can be quantified within a statistical model as a parameter

of interest. This is easiest to see in longitudinal designs, which produce trajectories over time for each participant. Suppose the researcher is testing the developmental trajectory of violent behavior in adolescent boys and assesses boys at ages 12, 14, 16, and 18 on a battery of aggressive measures. The researcher could compute the mean for each of the four times and draw a single curve that represents the average research participant. Does that average curve represent the majority of the data? Does it misrepresent important alternative data patterns? The mean curve may be fitted well by a line, meaning that a slope and intercept provide convenient summaries of the growth pattern, but there may be variability in the slopes and intercepts. Given the longitudinal nature of the design, in principle, it is possible to estimate separate regressions for each boy, thus computing a slope and intercept for each boy. If the study had 100 boys each measured four times, there would be 100 slopes and 100 intercepts. The slopes represent the linear (constant) change over time, and time can be scaled so that, for example, the intercept represents the starting level of aggression at age 12. We can conduct analyses to test which variables predict those boys who started relatively high versus relatively low (intercept differences) and which variables predict those boys who stayed relatively flat over time compared with those who increased (positive slope) or decreased (negative slope) over time. We can conduct analyses comparing different demographic groups, or different groups of boys, for different patterns in slope or intercept.

It turns out that we do not actually run separate regressions for each participant because that is inefficient and can introduce bias in some cases, such as in the presence of missing data. Instead, we use modern statistical models that estimate the regressions simultaneously. These models are referred to by many names, including random effect models, multilevel models, and latent growth models, but they are all essentially the same underlying model. The key idea is that there is variability around the slope and intercept. It is not the case that a researcher can simply estimate a single slope and intercept for the entire data set; these models estimate separate slopes and intercepts for each participant within a single, simultaneous regression model (these models also include "fixed effect" terms that capture the average patterns in the data). It does not matter whether we say that the slope is a random effect (a parameter with a distribution), that the slope is a dependent variable at a higher level of analysis (hierarchical modeling), or that there is a latent slope of which each participant represents an observation (growth curve modeling). In all those cases, we are allowing the slope to vary by person; the models lead to identical results given comparable identification constraints. The single, simultaneous model can also take into account differences in variance. For example, some participants may exhibit more variability, or volatility, than others—two participants may have the same slope and intercept but one person's data points are close to the regression line, whereas the other person's data points are more variable around the regression line. These models can also handle some types of missing data.

The standard repeated-measures ANOVA is a special case of the more general random effect model. The traditional repeated-measures ANOVA is a "random intercept model" in the sense that participants are treated as a random effect; it imposes a relatively restrictive set of assumptions on the covariances between the time points. More general latent growth models allow for more complicated random effects, such as both a random intercept and a random slope, as well as more general covariance patterns across time. One of the benefits of all this is that we can summarize individual subject trajectories to maintain the heterogeneity present in the data. We can then study those trajectories, such as figuring out what

predicts the slopes and intercepts or what the slopes and intercepts predict on the outcome variable. If age change trajectories are important in developmental psychology, then knowing what predicts them and what they predict could be valuable in both theory development and theory testing. The new technology of random effect terms allows us to maintain the heterogeneity in both our analyses and our theories.

We can see that the model specifics become important. For example, the model fitting linear regressions for each boy assumes that a straight line (slope and intercept) represents each boy's pattern well. What about the boy who exhibited low levels of aggression at ages 12 and 18 but high levels at ages 14 and 16? A straight line would not do justice to that boy's data pattern. We may need more complicated regression models to account for nonlinearities. In that case, the data are more complicated than mere slopes and intercepts. Blindly assuming linear relations in data may hold back theoretical advances if those linear relations are not present in the data. We believe developmental theory will progress more rapidly by tackling nonlinearity directly, finding ways to estimate it so that nonlinearity can be predicted, assessed, understood, and used to predict other variables.

We should not equate nonlinearity with polynomial regression (linear, quadratic, etc.). Nonlinearity is a more general concept, with polynomials as a special case. Sometimes it is possible to express nonlinear parameters directly in relation to theoretical parameters of interest. For example, in a study of adolescent drinking behavior, perhaps key parameters of interest are the age at which the adolescent begins to drink alcohol and, once the adolescent begins to drink, how quickly he or she accelerates his or her drinking behavior (if at all). Such processes can be modeled directly using nonlinear regression techniques by which a parameter can be assigned to the "liftoff," that is, the point at which the curve moves away from no drinking, and a second parameter assesses the degree of curvature (e.g., assessing the rate of increase). This type of direct assignment of a statistical parameter to a psychological parameter cannot always be accomplished in the context of polynomial regression, which is why more general nonlinear approaches are sometimes needed (Gonzalez, 2009b). The heterogeneity we discussed related to estimating separate parameters for each participant can be extended to nonlinear models as well.

An alternative method for dealing with heterogeneity is to define subsets of participants with similar parameters (i.e., similar trajectories). This latter method treats each subset as an equivalence class—participants in the same subset or equivalence class are treated as indistinguishable, but the subsets can be different from each other. There has been much progress in the past two decades on these kinds of models, which are also known by different names, such as latent class analysis and mixture models. Once the clusters are identified, then new research questions emerge about what predicts cluster membership (e.g., are there risk factors that can predict those boys who start low but increase their aggression over time, or are there any buffering processes that help maintain a low level of aggression in the group of boys who start low in aggression and remain low) or about what these clusters predict (e.g., does cluster membership predict future aggression, performance in school, or impulsivity?).

The general approach offered by latent growth mixture models holds promise in its usefulness for testing developmental theory. The ability to model trajectories for each participant, the ability to model hetereogeneity, and the ability to deal with missing data will likely lead to major new advances in testing developmental theory. It can also lead to new theory development as researchers ponder the developmental processes that yield heterogeneity.

Measurement Error

Developmental studies have the usual issue of measurement error that is typical in many psychological studies. This issue can be addressed through classic psychometric theory, modern structural equation models, or item response theory. Measurement error can have two key effects on the statistical analyses: It can reduce statistical power to detect differences, and it can introduce bias. The former effect is relatively innocuous and can be addressed by increasing the number of participants, reducing extraneous noise in one's procedure, or using statistical methods to create latent variables. The latter effect of measurement error, however, is more serious, especially in the context of multiple regression models. Measurement error in predictor variables can bias the coefficients (the "betas") of other variables in the model. The bias can be in either direction and can even change the sign of a regression coefficient. For instance, the population regression coefficient for a variable could actually be positive, but in a sample regression, the estimated coefficient could be negative due to measurement error on a different variable that is included in the model. The role of measurement error has mostly been underappreciated by developmental psychology, except in the context of estimating difference scores and change. But measurement error in predictor variables also needs attention and careful modeling.

The role of measurement error becomes important for developmental psychology in the context of change scores. At first blush, change seems so simple to estimate and describe—for example, take the difference of two time points and, *voila*, you have an estimate of change. Such a simple measure of change is at the heart of the standard paired *t*-test, which is typically used in a before–after design when one wants to compare the means between two time points. There is nothing wrong with the use of such difference scores as long as they are used appropriately (Rogosa, Brandt, & Zimowski, 1982). The paired *t*-test and the repeated-measures ANOVA, which can be conceptualized as using weighted difference scores in the form of contrasts over time (e.g., Gonzalez, 2009a; Maxwell & Delaney, 2004), tend to be legitimate uses of difference scores. Problems arise when one wants to use the difference score in other ways, for instance, as a measure of discrepancy (see Griffin, Murray, & Gonzalez, 1999). One needs to be careful about measurement error when interpreting difference scores directly or using them as predictors of other variables. For example, if one wants to use change in parental depression as a predictor of a child's behavior problems, then change is a measured variable with error. There are exciting new approaches that model change as a latent variable and that thus can take advantage of standard ways latent variables deal with measurement error (McArdle, 2009). These newer models still require more development and understanding (e.g., identification issues), but they hold promise for testing theory in developmental psychology.

Future Directions and Conclusions

Throughout this chapter, we viewed the problem of data analysis in experimental and quasi-experimental designs as one of reducing the number of alternative explanations and confounds to observed differences on the dependent variable. For us, the primary problem of data analysis is setting up a design and using analytical tools that facilitate clean inferences about explanatory mechanisms.

This chapter highlighted opportunities for innovation in methodology as related to developmental research. These include new designs that provide strength in both internal

and external validity, as well as new statistical procedures for the measurement and modeling of change. Other open problems are currently receiving attention from methodologically minded researchers. For example, it is not always possible to use the same dependent variable across time or with different age groups (e.g., there may be floor or ceiling effects). This problem also arises when pooling multiple studies that used different measures. The change in dependent variable introduces a confound, making it difficult to explain differences in age groups or changes over time. Some interesting attempts to tackle this problem have recently been studied (e.g., Curran & Hussong, 2009; McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009). We hope that methodologists and developmental psychologists continue working together to create new tools that facilitate the testing of developmental theory and advance our understanding of developmental processes.

The punch line of this chapter is that there is no substitute for good design. Statistics cannot salvage a poorly designed study. It is possible for a well-designed study to provide useful information when the data analysis is simple or even weak. Researchers should strive for research that is based on solid design principles. Good data analysis is necessary but not sufficient for making solid inferences from one's data. This is one of the major lessons we have learned from scholarship in quasi-experimental designs.

Acknowledgments

We thank Lindsay Bell, Jonathan Lane, and Julie Maslowsky for helpful suggestions on earlier versions of this chapter.

References

Baltes, P. (1968). Longitudinal and cross-sectional sequences in the study of age and generation effects. *Human Development, 11*, 145–171.

Baltes, P., Reese, H., & Nesselroade, J. (1988). *Life-span developmental psychology: Introduction to research methods*. Hillsdale, NJ: Erlbaum.

Campbell, D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Collins, L., & Horn, J. (1991). *Best methods for the analysis of change*. Washington, DC: American Psychological Association.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist, 60*, 170–180.

Curran, P., & Hussong, A. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14*, 81–100.

Gao, S., & Fraser, M. (2009). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gonzalez, R. (2009a). *Data analysis for experimental design*. New York: Guilford Press.

Gonzalez, R. (2009b). Transactions and statistical modeling: Developmental theory wagging the statistical tail. In A. Sameroff (Ed.), *Transactional processes in development: How children and contexts shape each other* (pp. 223–245). Washington, DC: American Psychological Association.

Greenwald, A. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin, 83*, 314–320.

Griffin, D., Murray, S., & Gonzalez, R. (1999). Difference score correlations in relationship research: A conceptual primer. *Personal Relationships, 6*, 505–518.

Heckman, J. (2005a). Rejoinder: Response to Sobel. *Sociological Methodology, 35*, 135–162.

Heckman, J. (2005b). The scientific model of causality. *Sociological Methodology, 35*, 1–97.

Lansford, J., Malone, P., Castellino, D., Dodge, K., Pettit, G., & Bates, J. (2006). Trajectories of internalizing, externalizing, and grades for children who have and have not experienced their parents' divorce or separation. *Journal of Family Psychology, 20*, 292–301.

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Maris, E. (1998). Covariance adjustment versus gain scores: Revisited. *Psychological Methods, 3*, 309–327.

Masche, J., & van Dulmen, M. (2004). Advances in disentangling age, cohort, and time effects: No quadrature of the circle, but a help. *Developmental Review, 24*, 322–332.

Mason, K., Mason, W., Winsborough, H., & Poole, W. (1973). Some methodological issues in cohort analysis of archival data. *American Sociological Review, 38*, 242–258.

Maxwell, S., & Delaney, H. (2004). *Designing experiments and analyzing data* (2nd ed.). Mahwah, NJ: Erlbaum.

McArdle, J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577–605.

McArdle, J., Grimm, K., Hamagami, F., Bowles, R., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Measurement, 14*, 126–149.

McArdle, J., Hamagami, F., Elias, M., & Robbins, M. (1991). Structural modeling of mixed longitudinal and cross-sectional data. *Experimental Aging Research, 17*, 29–52.

McGuire, W. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology, 26*, 446–456.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*, 726–748.

Rubin, D. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*, 757–763.

Schaie, K. W. (1970). A reinterpretation of age related changes in cognitive structure and functioning. In L. R. Goulet & P. B. Baltes (Eds.), *Life-span developmental psychology: Research and theory* (pp. 485–507). New York: Academic Press.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Spaeth, M., Weichold, K., Silbereisen, R., & Wiesner, M. (2010). Examining the differential effectiveness of a life skills program (IPSY) on alcohol use trajectories in early adolescence. *Journal of Consulting and Clinical Psychology, 78*, 334–348.

Steiner, P., Cook, T., Shadish, W., & Clark, M. (2008). The importance of covariate selection in controlling for selection bias in observational studies. Available at *www.northwestern.edu/ipr/events/workshops/QEPPTs/CovariateSelection08.pdf*.

Stewart, A., Copeland, A. P., Chester, N. L., Malley, J. E., & Barenbaum, N. B. (1997). *Separating together: How divorce transforms families*. New York: Guilford Press.

Wicker, A. (1985). Getting out of our conceptual ruts: Strategies for expanding conceptual frameworks. *American Psychologist, 40*, 1094–1103.