Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* New York: Houghton Mifflin; 2002.

6

Quasi-Experiments: Interrupted Time-Series Designs

Time (tīm): [Middle English from Old English *tima*; see *da* in Indo-European Roots.] n. 1. Abbr. t., T. a. A nonspatial continuum in which events occur in apparently irreversible succession from the past through the present to the future. b. An interval separating two points on this continuum; a duration: *a long time since the last war; passed the time reading.*

Time (tīm): adj. 1. Of, relating to, or measuring time.

Se·ries (sîr´ēz): [Latin series, from *serere,* to join; see *ser-*[2] in Indo-European Roots.] n., pl. series. Abbr. ser. 1. A number of objects or events arranged or coming one after the other in succession.

IN JULY 1982, the Arizona state legislature mandated severe penalties for driving while intoxicated. A comparison of monthly results from January 1976 to June 1982 (the control condition) with monthly totals between July 1982 and May 1984 (the treatment condition) found a decrease in traffic fatalities after the new law was passed. A similar finding occurred in San Diego, California, after January 1982, when that city implemented a state law penalizing intoxicated drivers. In El Paso, Texas, which had no relevant change in its driving laws during this period, monthly fatality trends showed no comparable changes near the months of either January or July 1982. The changes in trends over time in San Diego and Arizona, compared with the absence of similar trends in El Paso, suggest that the new laws reduce fatalities (West, Hepworth, McCall, & Reich, 1989). This interrupted time-series design is one of the most effective and powerful of all quasi-experimental designs, especially when supplemented by some of the design elements discussed in previous chapters (see Table 5.2), as we illustrate in this chapter.

WHAT IS A TIME SERIES?

Time series refers to a large series of observations made on the same variable consecutively over time. The observations can be on the same units, as in cases in which medical or psychiatric symptoms in one individual are repeatedly observed. Or the observations can be on different but similar units, as in cases of traffic fatalities displayed for a particular state over many years, during which time the baseline population is constantly changing.

In this chapter we describe a special kind of time series that can be used to assess treatment impact, the *interrupted* time series. The key here is knowing the specific point in the series at which a treatment occurred, for example, the date on which a mandatory seat belt law took effect. If the treatment had an impact, the causal hypothesis is that the observations after treatment will have a different slope or level from those before treatment. That is, the series should show an "interruption" to the prior situation at the time at which treatment was delivered. Such designs have been widely used to assess the impact of interventions in areas as diverse as attorney advertising (Johnson, Yazdi, & Gelb, 1993), community interventions to improve child-rearing practices (Biglan, Metzler, & Ary, 1994), epidemiology (Catalano & Serxner, 1987; Tesoriero, Sorin, Burrows, & LaChance-McCullough, 1995), consumer product safety (Orwin, 1984), gun control (Carrington & Moyer, 1994; O'Carroll et al., 1991), history of marriage (Denton, 1994), human rights (W. Stanley, 1987), political participation (Seamon & Feiock, 1995), real estate values (Brunette, 1995; Murdoch, Singh, & Thayer, 1993), environmental **risk analysis** (Teague, Bernardo, & Mapp, 1995), spouse abuse (Tilden & Shepherd, 1987), surgery (Everitt, Sourmerai, Avorn, Klapholz, & Wessels, 1990), substance abuse (Velicer, 1994), tax policy (Bloom & Ladd, 1982), workplace safety (Feinauer & Havlovic, 1993), and the effects of television (Hennigan et al., 1982). The interrupted time series is a particularly strong quasi-experimental alternative to randomized designs when the latter are not feasible and when a time series can be found. Here, we begin with the simplest interrupted time-series design, then introduce variants on that design, and finally discuss practical problems in implementing time-series studies.

Describing Types of Effects

A posttreatment time series can differ from a pretreatment series in several ways. First, there may be a sharp discontinuity at the point of the intervention, the time at which we expect the overall series to be "interrupted." Consider a short series with 20 observations, 11 before the intervention and 9 after it. If the values for the pretreatment series were 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12 and the values for the posttreatment series were 10, 11, 12, 13, 14, 15, 16, 17, and 18, then we

would conclude that the values decreased following the intervention, for the 12th observation is a 10 and not the expected 13. The change from a value of 12 to 10 is called a change in *level* or *intercept,* because (1) the level of the series drops and (2) the pre- and posttreatment slopes would have different intercepts.

Second, there may be a change in the slope of the series at the point of interruption. Imagine that the pretreatment values are 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12 and the posttreatment values are 14, 16, 18, 20, 22, 24, 26, and 28. So before treatment the series shifted one score unit per time interval, but after it there were two score units of change per time interval. This is variously called a change in *drift, trend,* or *slope.*

Though changes in level and slope are the most common forms of change, they are not the only ones possible. For instance, posttreatment changes can occur in the variances around each mean if a treatment makes people more homogeneous or heterogeneous in the outcome measure compared with the preintervention time period. Or a cyclical pattern can be affected; for example, the introduction of air conditioning probably caused a new relationship between time of year and time spent indoors. Though one typically looks for changes in intercept and slope in interrupted time-series research, investigators should remain open to other kinds of effects.

Effects can be characterized along another dimension. A *continuous* effect does not decay over time. Hence, a shift in level of X units that is obtained immediately after an intervention is still noted in the later part of the total series. A *discontinuous* effect does not persist over time. Commonly, the initial effect drifts back over time toward the preintervention level or slope as the effect wears off. This is especially likely to occur if the intervention is introduced and then removed, but it can also occur if a treatment with transitory effects is left in place. Sometimes a discontinuous effect can take the opposite form, with an effect becoming larger over time, creating a sleeper effect (Cook et al., 1979). But in our experience this is rare.

Effects can also be *immediate* or *delayed* in their initial manifestation following the treatment. Immediate effects are usually simpler to interpret, for their onset can be matched exactly to the time of intervention. Delayed effects are more difficult to interpret unless there is a theoretical justification for how long a delay should elapse before an effect is expected (e.g., biology leads us to expect about a nine month delay between the introduction of a new birth control method and the first effects on birth rate). With delayed effects, the longer the time period between the treatment and the first visible signs of a possible impact, the larger the number of plausible alternative interpretations.

It clarifies an interrupted time-series effect to describe it on all three dimensions simultaneously, that is, the form of the effect (the level, slope, variance, and cyclicity), its permanence (continuous or discontinuous), and its immediacy (immediate or delayed). We provide examples of all these different kinds of effects in this chapter.

Brief Comments on Analysis

We focus almost entirely on design rather than analysis issues, although we occasionally mention some problems of statistical conclusion validity that particular examples highlight.[1] However, it is useful to know a few basic points about time-series statistics. Ordinary statistics cannot be used, for example, for comparing preintervention observations with postintervention observations using a *t*-test. Ordinary statistics assume observations are independent of each other. But time-series data are typically autocorrelated. That is, the value of one observation is usually related to the value of previous observations that may be one, two, three, or more lags away. Estimating this **autocorrelation** usually requires a large number of observations, typically 100, to facilitate correct model identification (Box, Jenkins, & Reinsel, 1994; Velicer & Harrop, 1983). However, the exact number of observations needed cannot be fully specified in advance. Statistically, one needs only enough to identify the model, and this identification depends on the amount of error in the data, any periodicity effects, the timing of the intervention within the series, and the number of lags to be modeled. Although having 100 observations is typically cited in the literature as desirable, some examples we use in this chapter have fewer observations.[2] Sometimes the reason is that the model can be identified adequately with fewer observations. Mostly, however, we present shorter time series to illustrate how even an abbreviated time series rules out many more threats to validity than is possible in cases in which there are only a few pretest or posttest time points. The interpretation of time series is not purely a statistical matter; it also depends on design features and the immediacy and magnitude of the effect.

1. Statistical methods for time-series analysis are classified into the time domain or the frequency domain (Shumway, 1988). Time domain approaches model time-series observations by predicting current values from past ones. The best known of these is the autoregressive integrated moving average (ARIMA) approach by Box and Jenkins (1970). Many econometric analyses are close to this tradition, sometimes labeled as structural regression models (Kim & Trivedi, 1994; Ostrom, 1990). Frequency domain approaches model time-series observations as a combination of periodic sine and cosine waves, often called spectral, harmonic, or Fourier analysis, an approach used most in the physical sciences and engineering. Shumway (1988) claims that both frequency and time approaches yield similar conclusions if the time series is long. We know of no strong consensus that one approach is preferable in general. Analysts are rapidly developing multivariate time series, methods for dealing with missing data, nonlinear time series, pooled time series that combine multiple time series on different units, and estimation of time series that are robust to outliers (Box, Jenkins, & Reinsel, 1994; Caines, 1988; Hannan & Deistler, 1988; Kendall & Ord, 1990; Sayrs, 1989; Tong, 1990; W. Wei, 1990). The *Journal of Time Series Analysis* offers cutting edge developments, and the *Journal of the American Statistical Association* includes pertinent articles on a regular basis. Many time series analysis books are available at both basic (Cromwell, Labys, & Terraza, 1994; Cromwell, Hannan, Labys, & Terraza, 1994; McDowall, McCleary, Meidinger, & Hay, 1980; Ostrom, 1990; Sayrs, 1989) and advanced levels (e.g., Box, Jenkins, & Reinsel, 1994; Fuller, 1995; Hamilton, 1994; Harvey, 1990; Judge, Hill, Griffiths, & Lee, 1985; Kendall & Ord, 1990; Reinsel, 1993; W. Wei, 1990). Some of them are accompanied by computer software programs (e.g., Brockwell & Davis, 1991; Shumway, 1988), and standard statistical packages have extensive time-series analysis capabilities (Harrop & Velicer, 1990a, 1990b; Newbold, Agiakloglou, & Miller, 1994; Kim & Trivedi, 1994).

2. Unless otherwise noted explicitly, none of the figures in this chapter aggregate raw data across time points; that is, all the time points available in the raw data are included in each figure.

SIMPLE INTERRUPTED TIME SERIES

A basic time-series design requires one treatment group with many observations before and after a treatment. A design with 10 observations might be diagrammed as:

O_1	O_2	O_3	O_4	O_5	X	O_6	O_7	O_8	O_9	O_{10}

A Change in Intercept

Our first example of a simple interrupted time series (McSweeny, 1978) has 180 observations and is nearly ideal for ruling out many threats to internal validity (Figure 6.1). In March 1974, Cincinnati Bell began charging 20 cents per call to local directory assistance. Figure 6.1 clearly shows an immediate and large drop in local directory assistance calls when this charge began. A little thought suggests very few plausible rival hypotheses. Regression to the mean is implausible because the very long preintervention time series shows that a high number of calls to directory assistance were occurring for many years, not just immediately prior to the intervention (see Maltz, Gordon, McDowall, & McCleary, 1980, for a time series with prominent regression artifacts). Selection is implausible if the population making calls to local directory assistance in Cincinnati had no atypical changes over adjacent months before and after the intervention, so that pre- and postintervention samples should be largely the same on any variables affecting outcome. Similarly, attrition is implausible because it seems unlikely that such a large number of customers would disconnect their phones in response to the charge, and this could be easily checked with phone company records. Further, no known, naturally occurring maturation process could cause such a dramatic drop in local directory

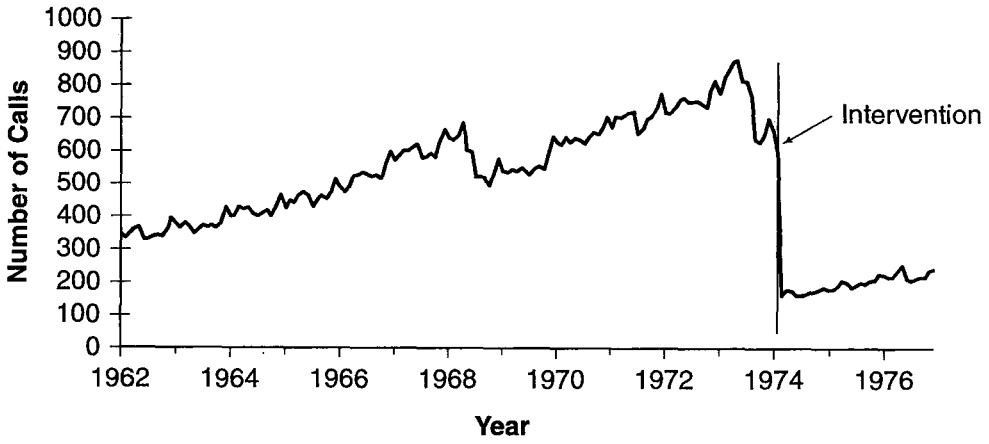

FIGURE 6.1 The effects of charging for directory assistance in Cincinnatii

assistance use. Testing effects are unlikely; here, the test is essentially the bill, and a testing effect would require that the phone company changed its billing to, say, highlight the number of directory assistance calls made prior to instituting the charge, so that customers changed their behavior based on that feedback rather than the charge. Whether they did so could be easily checked. History is plausible only if one could find another event that occurred simultaneously with the 20-cent charge and could have produced such a large effect, which seems unlikely. When effects are as immediate and dramatic as in this example, most threats to internal validity are usually implausible. Other examples of interrupted time series with immediate and dramatic effects include the effects of vaccinations on diseases such as tetanus (Veney, 1993) and of screening for phenylketonuria (PKU) on retardation (MacCready, 1974).

A Change in Slope

Figure 6.2 shows an example of an interrupted time series with 96 observations in which the outcome is a change of slope rather than of intercept. In 1983, Canada reformed its Criminal Code pertaining to sexual assault. The old law had two categories: rape and indecent assault. The new law used three categories of increasing severity and other provisions to increase the number of victims who report the crimes to the police, and it was implemented with much national publicity. To assess the impact of this change, Roberts and Gebotys (1992) used monthly data from the Canadian Uniform Crime Reporting System on sexual assault reports from 1981 through 1988. The resultant time series shows a seasonality effect, with more sexual assaults reported in summer and fewer in winter. Taking that into account, the slope was close to zero before the law was passed. After the law was passed the slope of the time series increased, suggesting that the law did have the desired effects on reporting sexual assaults.

Given the pattern of results, most threats to validity seem implausible. For example, maturation is unlikely because the slope changed dramatically at the point of the intervention, a result that does not mimic a known maturation process. For history to be plausible, another event must exist and have an abrupt effect on reporting of sexual assaults. With the exception of the publicity that surrounded the implementation of the law (which is arguably part of the intervention), no such event could be identified. However, the authors did present data suggesting that the publicity may have influenced people's attitudes toward sexual assault, which may have contributed to the increase in reports. This raises a question of the construct validity of the treatment. Should the treatment be construed as reform legislation or reform legislation with great publicity? The authors argue for the latter interpretation.

The intervention changed reporting categories, so Roberts and Gebotys (1992) identified four possible instrumentation threats. First, the new law allowed wives to charge their husbands with sexual assault, and it included assaults against

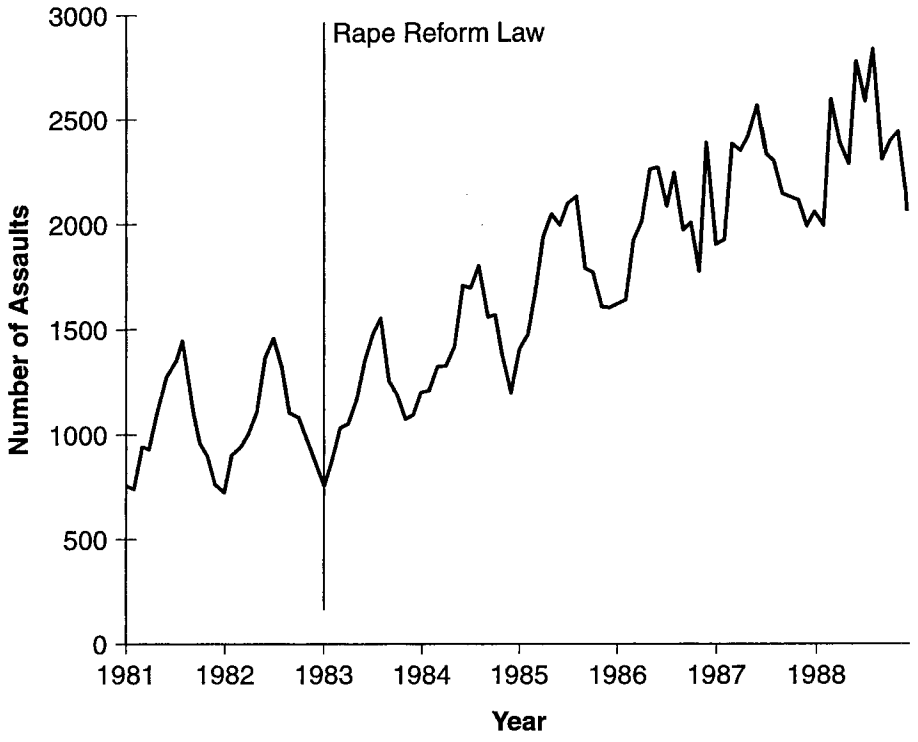

FIGURE 6.2 The effects of sexual assault law reform in Canada

From "Reforming rape laws: Effects of legislative change in Canada," by J. V. Roberts and R. J. Gebotys, 1992, *Law and Human Behavior, 16,* 555–573. Copyright 1992 by Kluwer Academic/Plenum Publishers.

both males and females. Countering this threat, the authors showed that the number of cases in which the suspect was either a woman or the husband increased only 5 percent after the law, insufficient to account for the far greater increase shown in Figure 6.2. Second, it could be possible that crimes that had been earlier reported as "other sexual assault" (and that therefore were not included in the preintervention section of Figure 6.2) were now being added to the sexual assault category and so were showing up only in the posttreatment series, which now for the first time included sexual exploitation, invitations to sexual touching, and bestiality. But analyses showed no change in reports of these highly specific crime categories over time, making it less likely that they caused the observed increase in slope. Third, perhaps these reports of other sexual assaults were rarely reported before reform but were reported more often afterward because of the publicity generated by the new law. The authors could find no data to directly rule out this threat, but they reported that indirect data suggested that it was unlikely. Fourth, some data suggested a temporal increase in the number of sexual assaults against children and juveniles. Was the increase shown in Figure 6.2 restricted to juvenile offenses? Unfortunately, the national statistics could not be disaggregated by age to test this possibility. But disaggregated crime statistics in cities such as Montreal showed too small an increase in the number of assaults against juveniles to account for the large increase in reporting sexual assaults shown in Figure 6.2.

Weak and Delayed Effects

Although interpretation of Figure 6.2 is not as clear as interpretation of Figure 6.1, few simple interrupted time-series designs are even as clear as Figure 6.2. Figure 6.3 presents a more common example. It includes 63 observations, and the effect seems both delayed in onset and weak in strength. This study (Hankin et al., 1993) examined the effects of an alcohol warning label law on the frequency with which pregnant women drank alcohol. Starting November 18, 1989, a federal law required a warning label on all alcohol containers. The label specifically warned that drinking alcohol during pregnancy can cause birth defects. The authors looked at how much alcohol was consumed by 12,026 pregnant African American women during the 2 weeks prior to their first prenatal visit to a Detroit clinic, both before and after the law took effect. The time series spanned September 1986 through September 1991 at monthly intervals. Visual inspection of the time series does not yield a clear-cut interpretation about whether the warning labels affected drinking.

The authors' statistical analysis, however, suggested a delayed rather than an immediate impact, beginning about 7 months after the law took effect. The reason to expect a delay is that the law affected newly produced containers, not those already on store shelves. Hence some time elapsed before bottles with warning labels could even appear on shelves in numbers sufficient to be noticed by consumers. Supporting this hypothesis, the authors asked women if they were aware of the labels, and they saw no increase in awareness until March 1990, 4 months after the implementation of the law. Hence a delayed effect would be expected, though not necessarily a 7-month delay. Furthermore, drinking among pregnant women appeared to have been decreasing before the intervention, and so maturation is a possible threat to validity given that the effect is a gradual change in slope, not in intercept. However, time series permits us to contrast the size of the pretreatment slope with the size thereafter. This contrast suggested that the (delayed) impact of the law was to speed up slightly the already occurring decrease in drinking among pregnant women.

We also see a seasonal trend in Figure 6.3. The rate of alcohol drinking goes up around the end of the year holidays and in the summer. However, seasonality is not a threat to validity in this case. Drinking actually increased very briefly immediately after the law took effect, coincidental with the approach of the Christmas season. Because the change in the law was supposed to decrease drinking, seasonality worked in the opposite direction to the hypothesis, and so cannot explain the obtained decrease. Indeed, Figure 6.3 shows that drinking during winter holidays and summer was visibly lower after the law's impact than before. However, if the law had been implemented in February, after the holidays, or in September, after summer ended, such seasonality effects could have been misinterpreted as a treatment effect. Good analytic practice requires modeling and removing seasonality effects from a time series before assessing treatment impact.

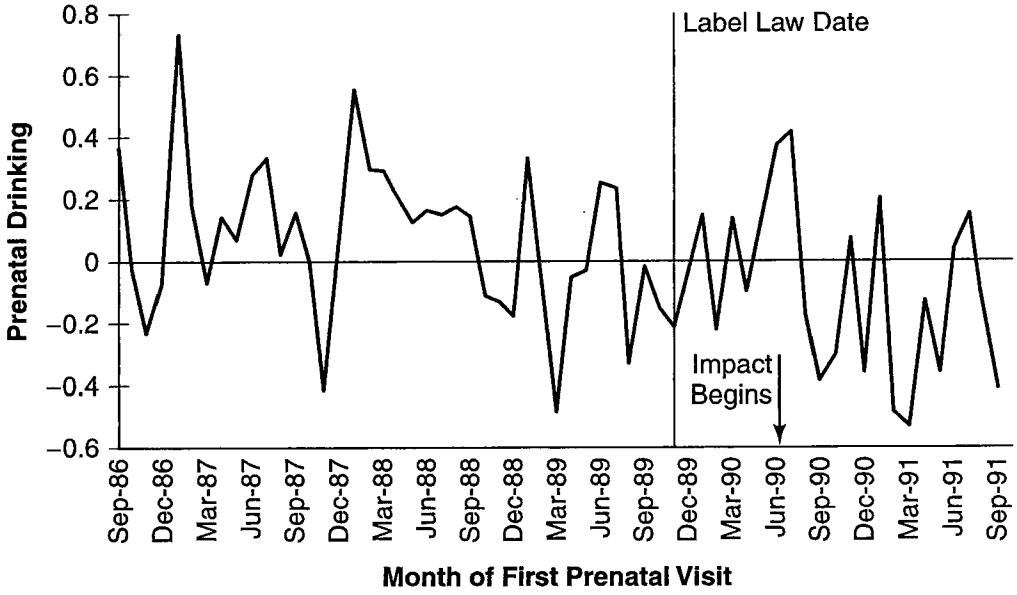

FIGURE 6.3 The effects of an alcohol warning label on prenatal drinking

From "A time series analysis of the impact of the alcohol warning label on antenatal drinking," by J. R. Hankin et al., 1993, *Alcoholism: Clinical and Experimental Research, 17,* pp. 284–289. Copyright 1993 by Lippincott, Williams & Wilkins.

The Usual Threats to Validity

With most simple interrupted time-series designs, the major threat to internal validity is history—the possibility that forces other than the treatment under investigation influenced the dependent variable at the same time at which the intervention was introduced. For example, in the Hankin et al. (1993) data, if the City of Detroit simultaneously passed laws restricting the sale of alcohol, this could have reduced drinking among pregnant women in a way that mimics the warning label effect. Several controls for history are possible, perhaps the best being to add a no-treatment time series from a control group, as we discuss shortly. But this is not always necessary. For instance, Hankin et al.'s measure of drinking was aggregated into monthly intervals, and the historical events that can explain an apparent treatment effect are fewer with monthly intervals than with yearly ones. Also, if a list is made of plausible effect-causing events that could influence respondents during a quasi-experiment, it should be possible using qualitative or quantitative means to ascertain whether most or all of them operated between the last pretest and the first posttest. If they did not, history is less plausible as a threat.

Another threat is instrumentation. For example, a change in administrative procedures sometimes leads to a change in how records are kept. Persons who want to make their performance look good can simply change bookkeeping procedures to redefine performance or satisfaction. Or persons with a mandate to change an organization may interpret this to include making changes in how

records are kept or how criteria of success and failure are defined. This seems to have happened, for instance, when Orlando Wilson took charge of the Chicago police. By redefining how crimes should be classified, he appeared to have caused an increase in crime. But the increase was spurious, reflecting record-keeping changes, not criminal behavior. Similarly, the Hankin et al. (1993) time series relied on women's self-reports of drinking. Self-reports are subject to demand characteristics, and publicity surrounding the law might have increased such demand characteristics. That is, women's self-reports of drinking might have decreased when the law took effect—even if their actual drinking did not change—as the women became aware that it was socially undesirable to drink while pregnant.

Selection can be another threat if the composition of the experimental group changes abruptly at the time of the intervention. This might occur if the treatment causes (or even requires) attrition from the measurement framework. If so, the interruption in the time series might have been due to different persons being in the pretreatment versus the posttreatment series. One can sometimes check this by restricting a data analysis to the subset of units that were measured at all time periods, but this is not always possible (e.g., when the third-grade achievement scores from a single school over 20 years are involved; the third-grade children are mostly different each year). Alternatively, the characteristics of units can sometimes be analyzed to see whether a sharp discontinuity occurs in the profile of units at the same time the treatment is introduced.

The typical statistical conclusion validity threats apply as much to time series as to any other design, such as low power, violated test assumptions, and unreliability of measurement. But the Detroit study of labels against drinking highlights a particular problem. The time-series analyst must specify at what point the intervention began and must have some theory or data that maps its diffusion through the units exposed to treatment. But in the Detroit example, the intervention diffused slowly, with the exact diffusion rate and pattern being unknown. Hence the researcher has discretion over specifying at what time the intervention begins and so can inadvertently capitalize on chance by picking a start time to match the time of apparent maximum effect. Because a number of changes may occur by chance alone in a long time series, a poorly specified intervention point can seriously weaken the causal logic of the design. (If diffusion rates are known, a diffusion curve could be used to model the intervention rather than an abrupt start. We discuss this option later in this chapter.)

Those who use time series must also guard against all the generic issues of construct validity, such as inadequate explication of constructs or confounding of constructs. However, time series raises some special construct validity issues. Many time series use archived data such as driving records or school grades. In these cases, many reactivity-related threats are less relevant because it is often harder for respondents to affect the outcome measures. Indeed, respondents often do not know they are part of a study. Reactivity is more likely in clinical time series, especially when the time interval between observations is short and respon-

dents can remember their past responses. So each time-series experiment has to be carefully examined on its own merits to determine whether the observed results might be due to evaluation apprehension, demand characteristics, or some similar threat to construct validity.

Also regarding construct validity, time-series work often uses only one outcome measure. The reason is partly that concerns about expense or old-fashioned definitional operationalist thinking led the persons who set up archives to measure, say, academic achievement, burglary, or unemployment in just one way. Compounding the problem, the researcher is often forced to use the outcome measures available, even if they are not fully relevant to the treatment being tested. This makes the available measures less sensitive for detecting outcome than they are when researchers collect their own dependent variable data, can tailor their measures to the treatment theory, and can add more items to increase reliability and validity. Of course, in time-series work in which several measures of an effect are available and each is reasonably reliable, changes in the time series can be separately examined for each measure. Furthermore, treatments are often events that respondents see as naturally occurring, such as changes in laws; and the outcomes are often (but not always) less obtrusively collected because respondents are used to government and corporations collecting data on them, at least more often than is the case with other kinds of research. So reactivity threats to construct validity of both treatment and outcome may be less.

Regarding external validity, it is sometimes possible to investigate it more actively by using available data on the background characteristics of units to stratify them into, say, males and females, or into different age groups, to see if the effect holds over such variability. There is no need to restrict such exploration to person variables. For example, setting variables can also be used to explore the range of an effect; and time variables can be used to see if an effect holds at different times of day (arrests during the day versus at night). This disaggregation has to be accomplished with caution, for statistical power can be reduced by creating subgroups. Moreover, in archival research one has a restricted flexibility in creating subgroups—necessary variables and cutoff points must be in the record. Thus, if the last age category is "over 65," and one is interested in studying the so-called old old (over 75), one cannot do so.

ADDING OTHER DESIGN FEATURES TO THE BASIC INTERRUPTED TIME SERIES

In previous chapters, we showed how to build stronger causal inferences by adding carefully chosen design features to basic quasi-experimental designs. The same principle applies to the interrupted time series, as well, as the following examples illustrate.

Adding a Nonequivalent No-Treatment Control Group Time Series

Consider the addition of a control group time series to the simple interrupted time-series design. The resulting design is diagrammed below:

O_1	O_2	O_3	O_4	O_5	X	O_6	O_7	O_8	O_9	O_{10}
O_1	O_2	O_3	O_4	O_5		O_6	O_7	O_8	O_9	O_{10}

An example of this design is given in Figure 6.4. In June 1992, the Spanish city of Barcelona passed legislation requiring riders of small motorcycles to wear helmets. Helmet use for large motorcycles had already been mandatory for several years. Ballart and Riba (1995) used a time series with 153 observations to examine the impact of that legislation. The outcome was the number of motorcycle accident victims with serious injuries or death. Both outcomes should decrease once the law took effect, but only for accident victims on small motorcycles. Victims riding large motorcycles, for whom the law had already been in effect for years, should show no such decrease, and so served as the control group. Figure 6.4 shows that the law had this hypothesized effect, and statistical analysis supported that interpretation.

Because the experimental and control groups rode their motorcycles and had accidents over the same time period, it is unlikely that a treatment-correlated historical event caused the decrease in serious injuries in the small-motorcycle group. Such an event should have decreased control group accidents, as well. The ability to test for the threat of history is the major strength of the control group time-series design. However, local history can be problematic if one group experiences a set of unique events that the other does not. Even then, local history can only threaten internal validity if the confounding event occurs *at the time of the intervention and would have an effect in the same direction as the intervention.* Such an event is unlikely with groups as comparable as those Ballart and Riba used. But with less comparable groups the probability of local history increases.

The untreated control series also allows tests of the other threats to internal validity that operate on the single time series. In the Ballart and Riba case, for example, the measurement instrument was the same between the treatment and control groups and before and after the law went into effect. Each group seemed to be changing at similar rates before the intervention (i.e., maturation). The intervention did not occur right after a particularly extreme observation, so statistical regression seems implausible. However, because the control group time series was formed nonrandomly, selection bias is a potential problem, though not often plausible. In the Ballart and Riba case, for example, perhaps the motorcycle riders most concerned with safety tended to ride large cycles until the law was passed because the law

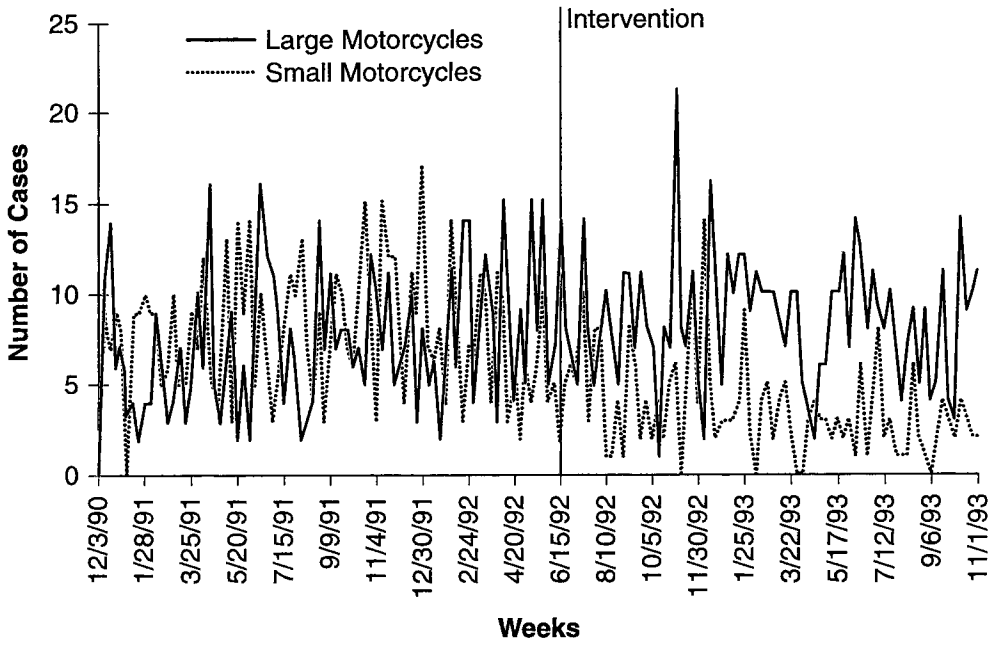

FIGURE 6.4 The effects of legislation requiring helmets on serious motorcycle injuries

From "Impact of legislation requiring moped and motorbike riders to wear helmets," by X. Ballart and C. Riba, 1995, *Evaluation and Program Planning, 18,* pp. 311–320. Copyright 1995 by Elsevier Science Ltd.

forced them to wear helmets. Then, when the law for smaller cycles changed, they felt free to go to such cycles. But it is possible that these safety-conscious individuals could have chosen to wear a helmet prior to the law about small cycles. We could add a corollary—perhaps that they would have been reluctant to wear helmets earlier because of peer pressure from other small cycle riders. How plausible this convoluted selection threat is for the Barcelona circumstance we cannot say.

Another interrupted time series (using 111 observations) with a control (Figure 6.5) shows the effects of a 1985 change in Iowa liquor laws on wine sales (Mulford, Ledolter, & Fitzgerald, 1992). On July 1, 1985, Iowa ended its public monopoly on liquor sales. Prior to then, only about 200 state-owned stores could sell liquor. After then, private sector liquor stores were licensed, and about 1,200 such private stores were rapidly established. Some people feared that increased availability of alcohol would lead to increased alcohol consumption with negative effects. Indeed, an early time-series analysis examined alcohol sales for 2.5 years after the law took effect (until December 1987) and found that wine consumption increased by 93% (Wagenaar & Holder, 1991). Mulford et al. (1992) investigated this matter further by adding a control series and extending the data to 1990. Figure 6.5 presents their data from 1981 through early 1990, nearly 5 years after the law took effect. As a control, they used national wine sales data for the same years. The Iowa wine sales time series shows an increase in wine sales when the law took effect. In fact, consistent with Wagenaar and Holder (1991), the increase persisted until early 1988, 6 months after Wagenaar and Holder stopped their data collection. After that time, Iowa wine sales returned to their preintervention levels. Mulford et al. (1992) were able to show that the sales

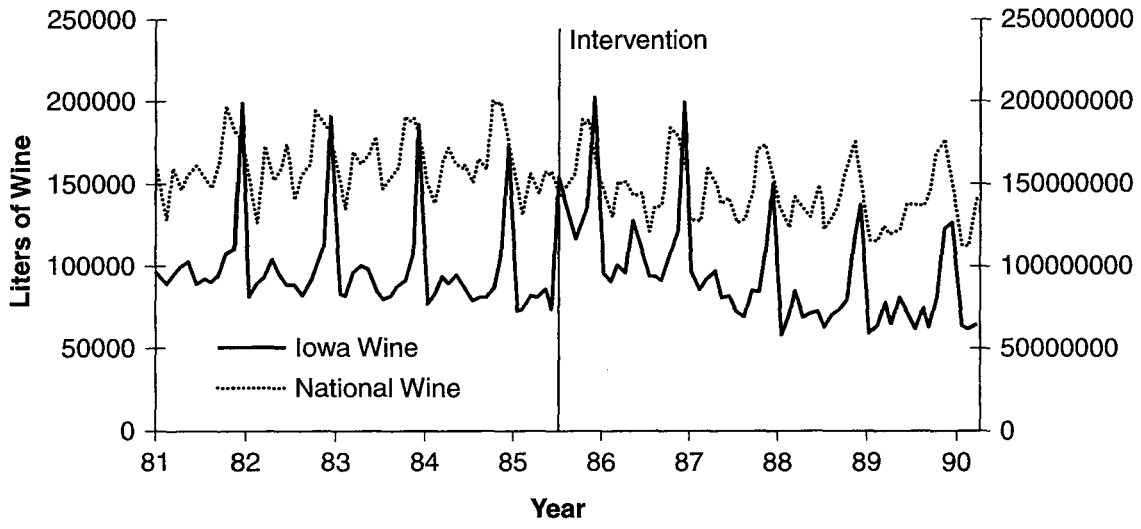

FIGURE 6.5 The effects of legislation in Iowa allowing private sector liquor stores on wine sales, using national data as a control

From "Alcohol Availability and Consumption: Iowa Sales Data Revisited," by H. A. Mulford, J. Ledolter, and J. L. Fitzgerald, 1992, *Journal of Studies on Alcohol, 53,* pp. 487–494. Copyright 1992 by Alcohol Research Documentation, Inc., Rutgers Center of Alcohol Studies, Piscataway NJ 08855.

increase was temporary and due partly to the 1,200 new liquor stores stocking their shelves rather than to increased consumption by retail consumers. Once these new stores fully stocked their shelves, sales returned to normal. A lesson from this example is that a long-duration time series may be needed to detect the temporal persistence of an effect. The addition of the national control in the Mulford et al. (1992) data helped to rule out history as an alternative explanation and made it easier to assess the temporal persistence of treatment effects.

Adding Nonequivalent Dependent Variables

The plausibility of many threats to internal validity in time series can be examined and the construct validity of the effect enhanced by collecting time-series data for a dependent variable that a treatment should affect and for a nonequivalent dependent variable that the treatment should not affect but that would respond in the same way as the primary dependent variable to a pertinent validity threat. The two dependent variables must be conceptually related. The design is diagrammed as:

O_{A1}	O_{A2}	O_{A3}	O_{A4}	O_{A5} X O_{A6}	O_{A7}	O_{A8}	O_{A9}	O_{A10}
O_{B1}	O_{B2}	O_{B3}	O_{B4}	O_{B5} X O_{B6}	O_{B7}	O_{B8}	O_{B9}	O_{B10}

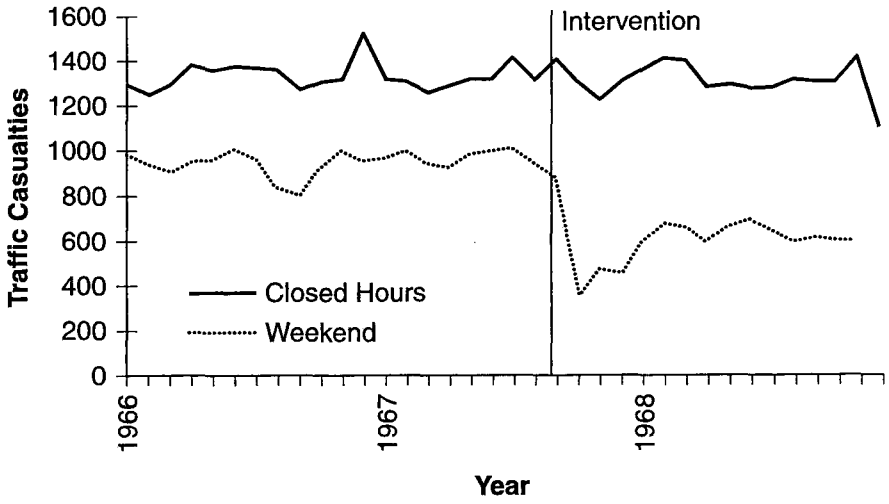

FIGURE 6.6 The effects of the British Breathalyzer crackdown on traffic casualties during weekend nights when pubs are open, compared with times when pubs were closed

From "Determining the social effects of a legal reform: The British 'breathalyser' crackdown of 1967," by H. L. Ross, D. T. Campbell, and G. V. Glass, 1970, *American Behavioral Scientist, 13,* pp. 493–509. Copyright 1970 by Sage Publications.

In this diagram, the A observation series represents the dependent variable of interest, and the B observation series represents the nonequivalent dependent variable.

McSweeny (1978) used a nonequivalent dependent variable in the Cincinnati directory assistance example, though it is not graphed in Figure 6.1. The new directory assistance charge was for local directory assistance, not long-distance directory assistance. If the effect were the result of the charge, only local calls should change; if the effect were the result of some other history event affecting all kinds of directory assistance calls, then long-distance directory assistance calls should change as well and at the same time that the local directory assistance time series changed. McSweeny plotted both time series and found that only local directory assistance calls changed; long-distance calls were unchanged.

Another example of this design comes from a study of the effectiveness of the British Breathalyzer crackdown (Ross, Campbell, & Glass, 1970; Figure 6.6). This time series has only 35 observations, but its effects are dramatic. The Breathalyzer was used to curb drunken driving and hence reduce serious traffic accidents. Under British drinking laws at that time, pubs could be open only during a limited time of day. If a large proportion of traffic casualties are due to drinking that takes place in pubs rather than at home, the Breathalyzer should decrease serious traffic accidents during the daytime hours or during weekend nights, when drinking was heaviest at pubs, and accidents should be less affected during commuting hours, when pubs were closed. Indeed, Figure 6.6 shows a marked drop in the accident rate on weekends (the outcome of interest) at the time of the intervention but little or no drop when pubs were closed (the nonequivalent dependent variable). Statistical analysis corroborated this decrease.

The distinction between accidents that occurred when pubs were either open or closed is important because most history threats to a decrease in serious accidents should affect *all* serious accidents *irrespective of the time of day.* This would be true for weather changes, the introduction of safer cars, a police crackdown on speeding, contemporaneous newspaper reports of high accident rates or particularly gory accidents, and so forth. So it is harder to find fault with either the internal or statistical conclusion validity of these data.

However, questions can be raised about external validity. For example, would the same results be obtained in the United States? Another concerns whether the effects are stronger with some kinds of drivers than with others. Another relates to possible unanticipated side effects of the Breathalyzer. How did it affect accident insurance rates, sales of liquor, public confidence in the role of technological innovations for solving social problems, the sale of technical gadgetry to the police, or the way the courts handled drunken driving cases? Such issues are examined by Ross (1973). Figure 6.6 reveals that some of the initial decrease in serious accidents on weekends is lost over time. The accident rate drops at first but then drifts part of the way back up toward the level in the control time series. Thus the effect is really only a *temporary and partial* reduction in serious accidents.

Ross also noted that the Breathalyzer was introduced with much nationwide publicity. Did the publicity make the public more aware of the desirability of not driving after drinking? Or did it make the police more vigilant in controlling the speed of traffic, especially during and immediately after pub hours? Did it reduce the overall number of hours driven? Did it cut down on drinking? Did it make drunken drivers drive more carefully? Ross very ingeniously ruled out some of these explanations. He took the regular surveys of miles driven done by the British Road Research Laboratory and showed that the introduction of the Breathalyzer was still associated with a decrease in accidents when the estimate of accidents per mile driven was used. This makes it less plausible to say that the Breathalyzer's effect is due to a reduction in the number of miles driven. Ross examined the sale of beer and spirits before and after the introduction of the Breathalyzer and found no evidence of a discontinuity in sales when the Breathalyzer was introduced, ruling out the interpretation that the Breathalyzer had reduced all drinking. He was able to show, for 10 months after the Breathalyzer was introduced, that more persons reported walking home after drinking than had been the case in the 10 months preceding the use of the Breathalyzer. Finally, he showed that fewer post-Breathalyzer traffic fatalities had high alcohol levels in their blood than had the corpses of pre-Breathalyzer fatalities. These analyses suggest that the explanation was a reduction in the number of heavy drinkers who drove, rather than a significant reduction in either aggregate drinking or driving. Ross's use of data to rule out alternative explanations of the causal construct highlights the importance of doing this, the difficulty and expense sometimes encountered in doing so, and the number of irrelevancies associated with the introduction of new practices into society.

Finally, Ross was faced with the problem of explaining why the effects of the Breathalyzer were not more permanent. His analysis suggested that the British

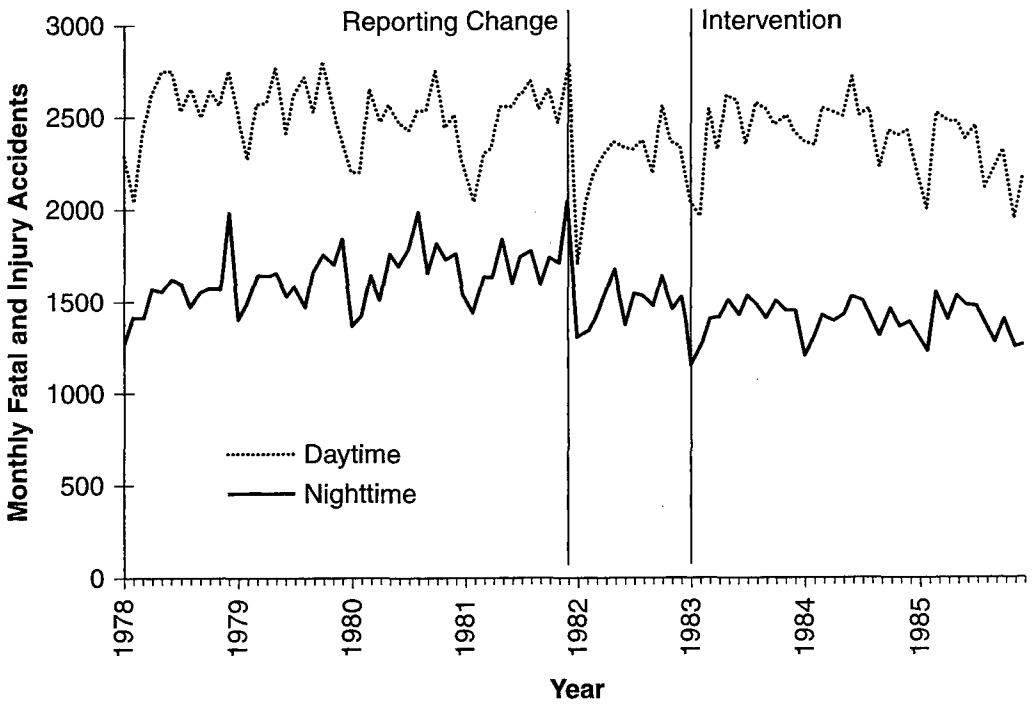

FIGURE 6.7 The effects of drunk driving legislation preceded by a change in accident reporting practices

From "The impact of drunk driving legislation in Louisiana," by M. W. Neustrom and W. M. Norton, 1993, *Journal of Safety Research, 24,* pp. 107–121. Copyright 1993 by Elsevier Science.

courts increasingly failed to punish drinkers detected by the Breathalyzer so that it lost its deterrent power. Thus Ross's final inference took a highly useful form: A Breathalyzer will help reduce serious traffic accidents when it is used to restrict drunken driving, but it will have this effect only if the courts enforce the law about drinking and driving.

Neustrom and Norton (1993) provide a related time series with 96 observations (Figure 6.7) on the effects of drunk-driving legislation introduced in Louisiana in 1983. Neustrom and Norton (1993) hypothesized that the effects of this legislation would be stronger at night than in the day because past research had shown that alcohol-related accidents are more serious and frequent at night. Problematically, however, a temporary shortage of police resources had occurred earlier and had ended at the same time the drunk-driving law took effect. This event might have resulted in more police officers writing up more of the very accident reports that the new law was meant to decrease, thus setting up two countervailing forces that might cancel each other out. As Figure 6.7 shows, the control series helped clarify the effects of this confound. In the daytime accident time series, accident reports rose after the new law passed, but the nighttime series dropped very slightly. Taking into account the increase in accident reports caused by the end of the personnel shortage, Neustrom and Norton (1993) estimated that

the new law caused a reduction of 312 accidents in the nighttime series and a reduction of 124 accidents in the daytime series. So the effects of a reporting change can sometimes be estimated if its onset is known and if the design has a nonequivalent dependent variable—in this case, day versus night, where the effect is expected to be stronger at night.

Removing the Treatment at a Known Time

The influence of treatment can sometimes be demonstrated by showing not only that the effect occurs with the treatment but also that it stops when treatment is removed. The removed-treatment design is diagrammed here, with X indicating the treatment and $\not{X}$ its removal.

O_1	O_2	O_3	O_4	X	O_5	O_6	O_7	O_8	O_9 $\not{X}$	O_{10}	O_{11}	O_{12}	O_{13}

The design is akin to having two consecutive simple interrupted time series. The first, from O_1 to O_9 here, assesses the effects of adding treatment; and the second, from O_5 to O_{13}, tests the effects of removing an existing treatment. The most interpretable pattern of effects occurs if the intercept or slope changes in one direction between O_4 and O_5 and then changes in the opposite direction between O_9 and O_{10}.

Figure 6.8 shows an example with 36 observations by Reding and Raphelson (1995). In October 1989, a psychiatrist was added to a mobile crisis intervention team to provide immediate, on-the-spot psychiatric treatment in hopes of preventing subsequent state hospital admissions. This effect occurred. Six months later, several factors led to the termination of the psychiatrist's services, which led to a rebound in state hospital admissions. Although not pictured in Figure 6.8, the authors strengthened their causal inference even further by comparing these results with data on admissions to a local private psychiatric hospital at the same time, which showed no changes in admissions. Here three design features facilitated causal inference: the interrupted time series, the treatment removal, and the control private hospitals.

The treatment removal adds many strengths to a time-series design. One is that the threat of history is reduced because the only relevant historical threats are either those that operate in different directions at different times or those that involve two different historical forces operating in different directions at different times that happen to coincide with the treatment introduction and removal. Selection and attrition are less of a threat unless different kinds of people enter or leave at different time points. Instrumentation is less likely, though problems can be created by a ceiling effect or floor effect, that is, by participants reaching either the lowest or highest possible score on the scale so that further changes in that di-

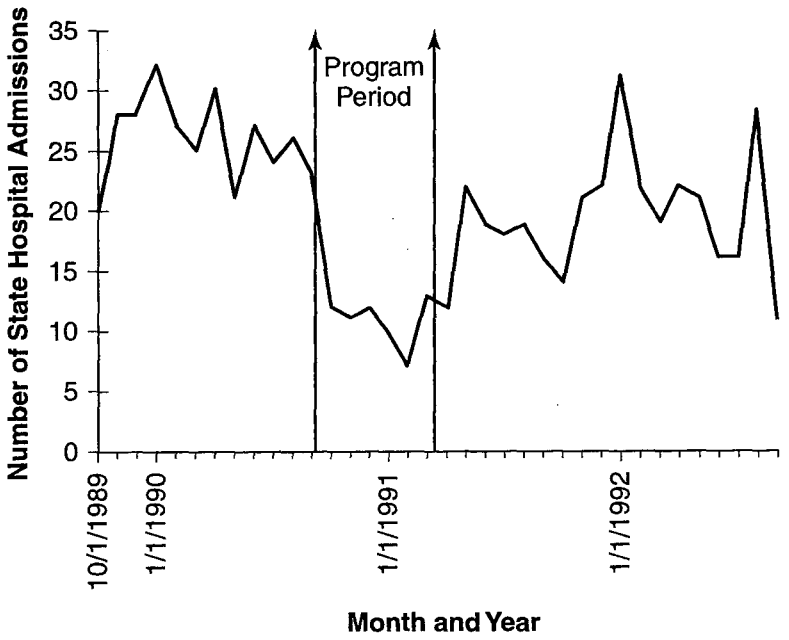

FIGURE 6.8 The effects of psychiatric crisis intervention on hospitalization

From "Around-the-clock mobile psychiatric crisis intervention: Another effective alternative to psychiatric hospitalization," by G. R. Reding and M. Raphelson, 1995, *Community Mental Health Journal, 31,* pp. 179–187. Copyright 1995 by Kluwer Academic Publishers.

rection are not possible, if the ceiling or floor is reached at the same point at which the treatment was removed. Other instrumentation effects may be implausible, insofar as the same instrumentation effect would have to account for both an increase and a decrease at different times.

The private hospital control series helped rule out history, the possibility that another event affected hospital admissions in general; and it helped rule out maturational trends such as cyclical patterns of admission to hospitals as the seasons change. The former threat is easily assessed using qualitative interviews with knowledgeable informants such as hospital directors. The latter threat is more plausible, as it is well known that some patients arrange to have themselves discharged from hospitals in cold climates during winter months, only to return during the summer to the cooler northerly climes.

Given this design and the effects in opposite directions at X and $\not{X}$, it is sometimes plausible to argue that the disappearance of the original effect is not due to removing the treatment but is due instead to resentful demoralization at having the treatment removed. In the Reding and Raphelson (1995) case, the higher level of hospital admission after the psychiatrist was removed from the team might have been due to demoralization among remaining team members that such a successful intervention was discontinued. If so, then removed-treatment designs are probably more interpretable the less conspicuous treatment is to participants. However,

such demoralization would not threaten the effect of introducing the treatment. When the removed-treatment design produces results in different directions, one usually needs two *different* alternative interpretations to invalidate a causal inference. Finally, the withdrawal design works well only if it is ethical to remove treatment and when the effects of treatment are temporary and end when treatment is removed. These conditions eliminate the design from consideration in the many cases in which treatment effects are presumed to last long.

Adding Multiple Replications

This is an extension of the previous design in which it is possible to introduce a treatment, remove it, reintroduce it, remove it again, and so on, according to a planned schedule. This design is diagrammed as:

$$O_1 \quad O_2 X O_3 \quad O_4 \not{X} O_5 \quad O_6 X O_7 \quad O_8 \not{X} O_9 \quad O_{10} X O_{11} \quad O_{12} \not{X} O_{13} \quad O_{14}$$

A treatment effect is suggested if the dependent variable responds similarly each time the treatment is introduced and removed, with the direction of responses being different for the introductions compared with the removals. The design has often been used to assess the effects of psychological (Marascuilo & Busk, 1988; Wampold & Worsham, 1986) or medical (Weiss et al., 1980) treatments with individual patients. For example, McLeod, Taylor, Cohen, and Cullen (1986) compared the effects of a drug treatment with those of a placebo for a patient with inflammation of her continent ileostomy reservoir. Drug and placebo were assigned randomly (and with double-blind, although blinding is frequently difficult with this design) to 10 treatment periods of 14 days each (although data were gathered daily, they are only available aggregated to the 10 treatment periods). The patient reported outcomes on well-being, nausea, abdominal pain, abdominal gas, stool volume, watery stool, and foul odor. Figure 6.9 shows results for well-being and pain, both suggesting that treatment was effective because well-being increased and nausea decreased with treatment, though the effects were far stronger for pain than for well-being.

An issue with this design is the scheduling of treatment and removal. Although scheduling is usually done systematically or in response to the state of patient symptoms (Barlow & Hersen, 1984), many advantages ensue from random scheduling of treatment, though perhaps in a way that preserves an alternation of X and $\not{X}$ (Edgington, 1987, 1992). Random scheduling rules out the threat of cyclical maturation, the possibility that the series would have exhibited a regular cycle of ups and downs even in the absence of a treatment. Other modifications increase the range of this design's application. For example, it can be used to compare two different treatments of theoretical interest, with X_1 being substituted for X and X_2

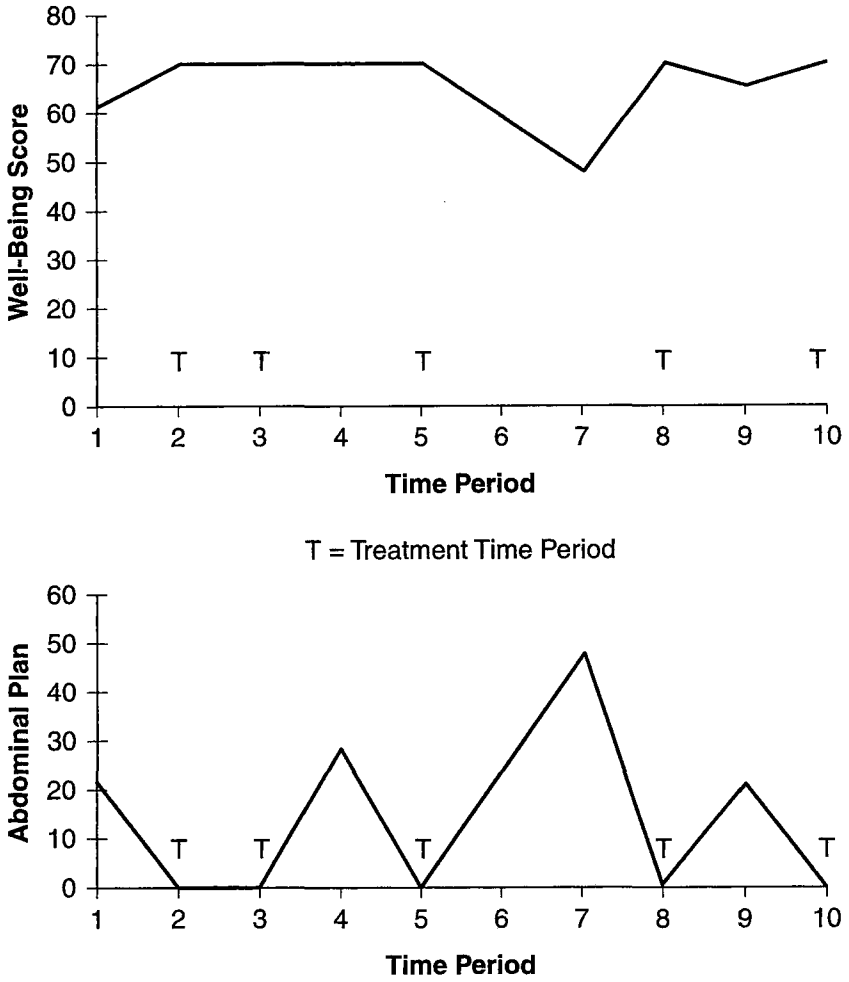

FIGURE 6.9 The effects of treatment for inflammation of continent ileostomy. In the graphs, the letter *T* indicates the time period during which treatment occurred

From "Single patient randomized clinical trial: Its use in determining optimal treatment for patient with inflammation of a Kock continent ileostomy reservoir," by R. S. McLeod et al., 1986, *Lancet, 1,* pp. 726–728. Copyright 1986 by The Lancet Publishing Group.

for ~~X~~. It is also possible to use a joint treatment factor ($X_1 + X_2$), as O'Leary, Becker, Evans, and Saudargas (1969) did in examining how the disruptive behavior of seven children in a classroom responded to first giving rules, then adding educational structure plus rules, then using rules, structure, and praise for good behavior or ignoring bad behavior, and then adding a token economy to all this. Then, to demonstrate control over the total phenomena, all treatments were removed and reinstated. In another design variation, the treatment can be implemented at different (usually increasing) strengths to examine a dose-response relationship; for example, Hartmann and Hall (1976) examined the effects of increasingly high penalties for an increase in the number of cigarettes smoked each day.

The major limitations of this design are practical. First, as with the removed-treatment design, it can be implemented only when the effect of the treatment is expected to dissipate rapidly. Also, the design normally requires a degree of experimental control that can rarely be achieved outside of laboratory settings, certain **single-case design** treatment contexts, or enclosed institutional settings such as schools and prisons. When the design is feasible, however, Barlow and Hersen (1984) provide a thorough discussion of various design options.

Adding Switching Replications

Imagine two (or more) nonequivalent groups, each receiving treatment at different times and in an alternating sequence such that (1) when one group receives the treatment, the other serves as a control, and (2) when the control group later receives the treatment, the original treatment group then serves as a continued-treatment control. The design can be diagrammed as:

O_1	O_2	O_3	X	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}
O_1	O_2	O_3		O_4	O_5	O_6	O_7	O_8	X O_9	O_{10}	O_{11}

The design controls most threats to internal validity, and it enhances external and construct validity. External validity is enhanced because an effect can be demonstrated with two populations at two different moments in history, sometimes in two settings. There may be different irrelevancies associated with the application of each treatment and, if measures are unobtrusive, there need be no fear of the treatment's interacting with testing.

An example is presented in Figure 6.10, which plots annual crime rates (logged and standardized) in 34 cities in which television was introduced in 1951 and in 34 other cities in which television was introduced in 1955 (Hennigan et al., 1982; McCleary, 2000), the gap being due to an FAA freeze on television licenses between 1951 and 1955. In both series, the introduction of television led to an increase in property crimes in the subsequent year. This replication of the effect at two time points 5 years apart makes history implausible. For example, the Korean War began in 1951, and the exodus of men to the army may have left more unguarded homes that were then robbed; but this could not explain why the effect occurred only in the earlier cities, for the war should also have caused the same effect in the later cities. Similarly, a recession that began around 1955 may have caused an increase in crime, but again this should have affected both series. Regression artifacts could possibly explain the increase in earlier cities, but that hardly seems likely for the later cities. More generally, it is hard for any single threat to explain both in-

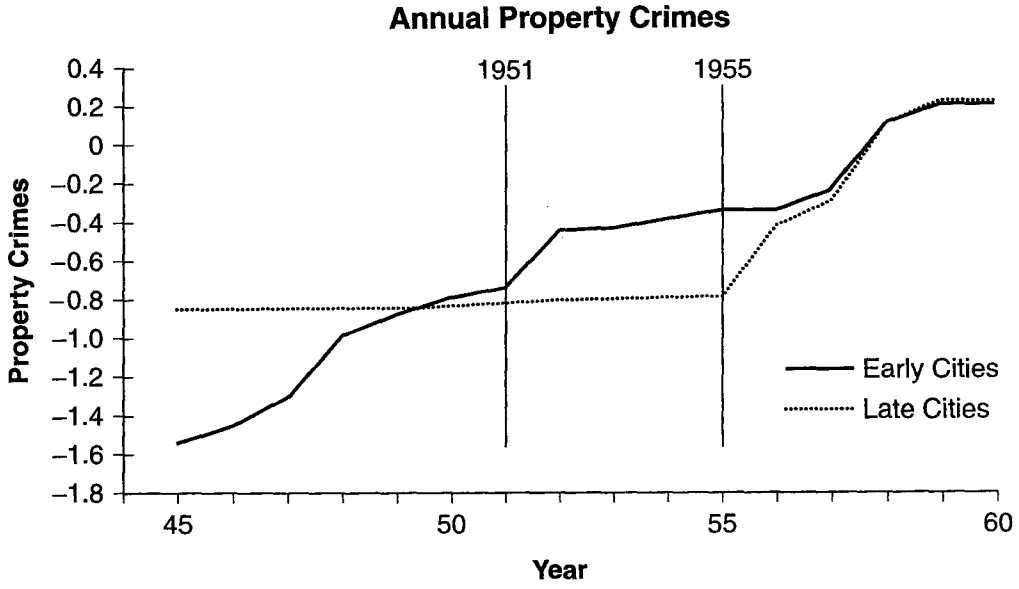

FIGURE 6.10 The effects of the introduction of television on property crime rates in cities in which television was introduced in 1951 versus 1955

From "The evolution of the time series experiment," by R. D. McCleary, 2000, *Research design: Donald Campbell's legacy, Vol. 2,* edited by L. Bickman, Thousand Oaks, CA: Sage. Copyright 2000 by Sage Publications.

creases. Threats such as selection-history or selection-instrumentation are possible alternative explanations—for example, if a critic can find two different historical events, one that occurred in most of the earlier cities (but not the later ones) and one that occurred in most of the later cities (but not the earlier ones), both of which could increase crime. But this possibility, too, seems implausible.

Another example is given in Figure 6.11, from a study of newborns who were screened for phenylketonuria (PKU) to prevent later mental retardation (MacCready, 1974). This time series has 17 observations, the first 4 of which are aggregations of separate 3-year periods. PKU screening was introduced as a standard practice at birth into different U.S. states and Canadian provinces in different years—1962, 1963, 1964, and 1965—facilitating the replication feature of the design. The dependent variable was the number of annual admissions to state or provincial institutions with a diagnosis of retardation due to PKU. Figure 6.11 shows that such admissions dropped to zero and remained at zero the year after screening was implemented in each of the four time series. These results corroborated the hypothesis using archival measures, widely dispersed populations, different historical moments for introducing the treatment, different irrelevancies associated with how the treatment was introduced, and repeated measures to ascertain if an initial effect can be generalized over time.

But even the replicated time-series design can have problems of internal validity. Both the results in Figure 6.11 and anecdotal evidence suggest that some PKU screening occurred before it was officially introduced into some of the states or provinces. Medical professionals learned of the benefits of screening through

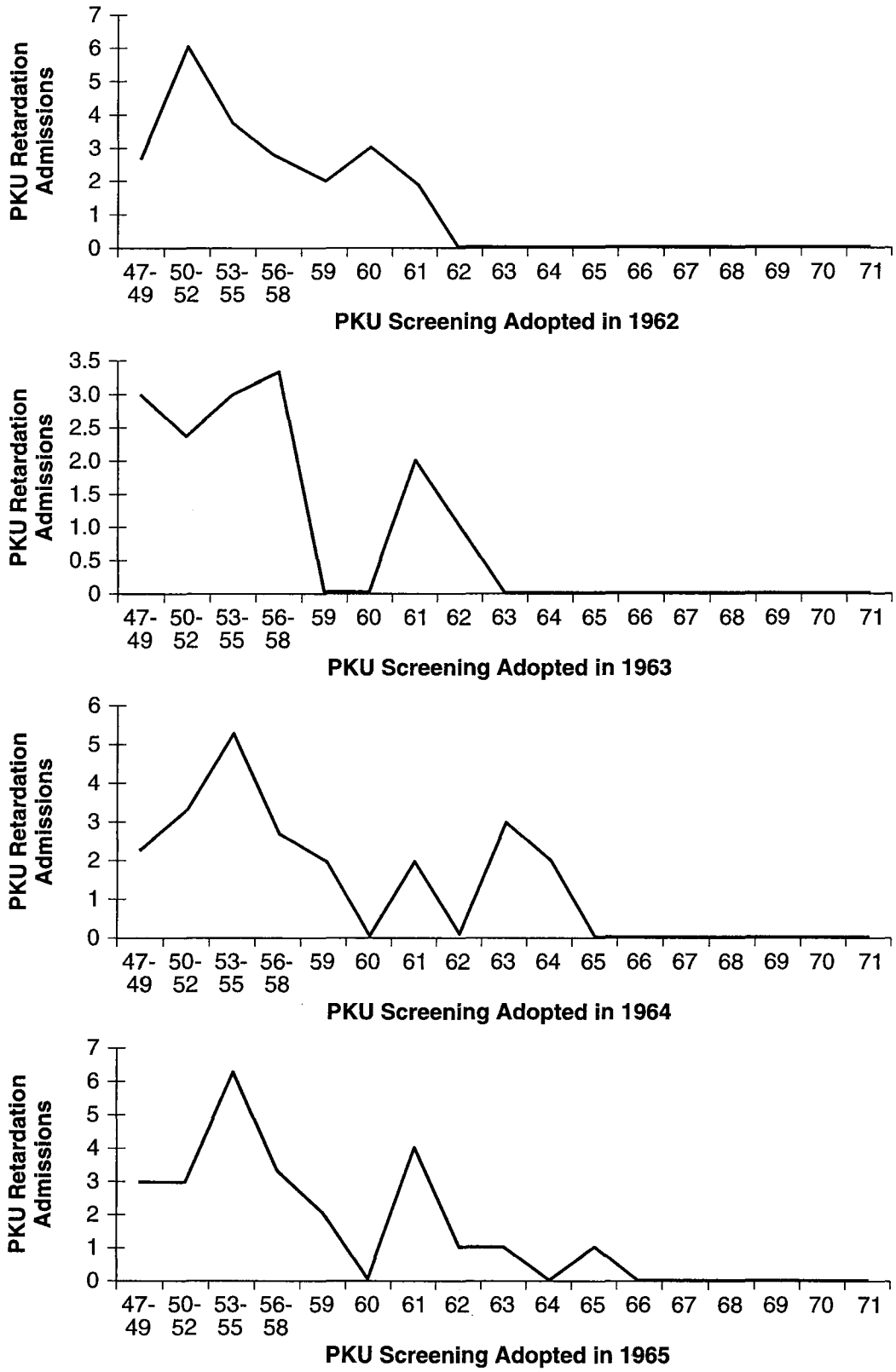

FIGURE 6.11 The effects of screening for phenylketonuria (PKU) on admissions for retardation due to PKU, with the implementation of screening staggered over 4 years in different locales

From "Admissions of phenylketonuric patients to residential institutions before and after screening programs of the newborn infant," by R. A. MacCready, 1974, *Journal of Pediatrics, 85,* pp. 383–385. Copyright 1974 by The National Medical Society.

journal articles, contact with newly trained physicians, and word of mouth at conferences and in discussions with colleagues. So a general trend toward reduced PKU admissions is observed even before the year in which screening was officially implemented—although admissions almost never dropped to zero before implementation. Perhaps such retardation was decreasing as mothers received better prenatal care, better nutrition, and better medical services at and right after birth. To investigate this possibility, MacCready (1974) gathered information on retardation from all causes from the 1950s through 1972 and reported finding no systematic upward or downward trend that could explain even part of the change in PKU admissions observed in Figure 6.11. Using nonequivalent dependent variables in this way renders this alternative explanation less plausible.

The switching-replications design can also help detect effects that have an unpredictable delay period. Assuming an equal delay of effect in each group, we would expect a discontinuity at an earlier date in one series than in the other. We would also expect the period between the discontinuities to be equal to the known period that elapsed between the implementations of the treatment with different groups. However, this will not always be plausible; for example, the effects of a treatment may be speeded up as new technologies for administering treatment are developed that make the treatment more effective. Therefore, it is more realistic to look for relative differences in the times at which an apparent treatment effect appears in each group. The switching-replication design is useful for probing delayed causal effects in cases in which there is no strong theory of the length of expected delay. However, it is most interpretable in the case in which the time difference between the groups receiving the treatment exactly matches the time period between effects appearing in each group—as is the case in Figure 6.11.

The replicated time-series design is practical wherever a time-series design with a no-treatment control group is feasible. Successful treatments will usually be of benefit to the groups or organizations that served as no-treatment controls. Representatives of these groups can be approached to see if they will agree to receive that treatment. For example, suppose Hankin et al. (1993; Figure 6.3) had been able to study the introduction of alcohol warning labels on bottled liquor introduced in different countries during different years. The causal inference would have been much clearer given their delayed, small effect.

SOME FREQUENT PROBLEMS WITH INTERRUPTED TIME-SERIES DESIGNS

As the previous examples begin to illustrate, a number of problems frequently arise in conducting interrupted time-series research:

- Many treatments are implemented slowly and diffuse through a population, so that the treatment is better modeled as a gradually diffusing process rather than as occurring all at once.

- Many effects occur with unpredictable time delays that may differ among populations and over time.
- Many data series are much shorter than the 100 observations recommended for statistical analysis.
- Many archivists are difficult to locate or are reluctant to release data.
- Archived data might involve time intervals between each data point that are longer than one needs; some data may be missing or look suspicious; and there may be undocumented definitional shifts.

We discuss all these in more detail below.

Gradual Rather Than Abrupt Interventions

Some interventions begin at a known point in time and quickly diffuse through the relevant population. The Cincinnati directory assistance intervention is an example; the financial charge took effect on a defined day and applied to all directory assistance calls. But other innovations diffuse quite gradually, so that the time span in which other events can occur that might affect the outcome increases, making history a more plausible threat to internal validity.

An example of a gradually implemented treatment is provided by Holder and Wagenaar (1994). They studied the effects on traffic crashes of mandated training for those who serve alcohol that was designed to reduce the intoxication levels and high-risk driving of alcohol drinkers. The State of Oregon introduced the training in December 1986, but not everyone could be trained at once. According to the Oregon Liquor Control Commission, about 20% of all servers were trained by the end of 1987, 40% by the end of 1988, and over 50% by the end of 1989. If the effect appears quickly for those individuals who receive treatment, then the immediate change in intercept should be small, but a gradual increase in slope should be apparent.

Knowledge of the form of the diffusion process is crucial to the analysis in such cases. For example, Holder and Wagenaar (1994) used the proportion of servers trained as their intervention variable rather than a dichotomous (1,0) dummy variable (sometimes called a **step function,** which assumes that all trainers received training on the day the law took effect). Treating a slower diffusion process as if it were a single step function can create serious problems. First, it can capitalize on chance and create false effects if the researchers assign the step function to a point in the time series at which the deviation in the outcome appears visually greatest. Second, one can overlook real but small early effects by assuming that the onset of the intervention (1986 for the Oregon law) occurred at the maximal point for potential impact (which was actually not reached until about 1988, and by then only 50% were trained in what is presumably a job with high turnover). Third, even when researchers carefully model treatment diffusion, they may look for patterns of effects that mirror that diffusion. However, to expect such patterns is often naïve, because causal thresholds can make the manifestation

of an effect dependent on reaching a certain level of the treatment. Holder and Wagenaar (1994) speculated that training only one server in a bar that employs many servers may not be very effective (for example, if the other servers exert peer pressure to continue serving as usual), so a majority of servers must be trained to have the effect. Without knowledge of the exact rate and form of diffusion, often the best one can do is to look for delayed effects sometime after the onset of the treatment. The difficulty is ruling out the historical effects that could operate between treatment onset and a change in the time series—for example, changes in enforcement of drunk-driving laws.

In this connection, it is worth considering why the Breathalyzer data (Figure 6.6) were so clear-cut. If a new law is not well publicized or is poorly enforced, we might expect only a gradual buildup in public reaction to it and would not expect the visually dramatic decrease in traffic accidents displayed in the Breathalyzer example. Fortunately, we know from background data that the British Breathalyzer was widely publicized, as was the date at which the police would begin to use it. This probably speeded up the usual process of informing the public about the use of the Breathalyzer and also contributed to policemen using it at frequent intervals right after its implementation date. Under these conditions the actual diffusion process better approximates a step function.

Delayed Causation

Not all effects are instantaneous. Delayed causation can occur even when a treatment is abruptly implemented, as with the delayed effects of smoking on lung cancer. The cancer does not develop until decades after regular smoking has begun. Even more delay may occur if the treatment implementation is diffuse. For example, Figure 6.3 shows a delayed impact due to the time it took for newly manufactured liquor bottles with warning labels to make their way to retail shelves and into consumer homes. There is little problem with delayed causation when strong background theory permits us to predict a specific lag, such as the 9-month lag between human conception and birth that helps predict the point at which alcohol warning labels could first affect births. Many times, however, no such theory exists, so the interpretation of a delayed effect is obscured by historical events between treatment onset and the possible delayed effect. In such cases, a switching-replications design permits the researcher to examine whether the replications show similar delay intervals between the treatment onset and the manifestation of an effect, reducing the threat of history. However, this procedure assumes that the treatment does not interact either with the different kinds of units that receive the treatment at different times or with the different historical moments at which each group experiences the treatment. For example, chronic inebriates may respond to a drunk-driving program more slowly than social drinkers; or media publicity given to an accident related to driving while intoxicated (DWI) in one city may heighten the effects of a new DWI law compared with some other city in which the law was implemented

without such publicity. Should there be such interactions, the delay between treatment onset and effect manifestation may differ over groups.

When delayed effects occur along with slow diffusion of treatments (as in Figure 6.3), causal inference is particularly difficult. This is true because no knowledge exists as to where to place the onset of the expected or desired effect, and effects might be expected at any point after treatment is implemented. The longer the time after implementation, the more plausible it is to interpret a possible delayed treatment effect in terms of historical factors. In these cases, the addition of control groups, nonequivalent dependent variables, removed treatments, or switching replications are all invaluable aids to inference.

Short Time Series

Textbooks dealing with the statistical analysis of time-series data suggest different rules of thumb for the number of time points required for a competent analysis. Most suggest making about 100 observations in order to model trends, seasonality, and the structure of the correlated error in the series before testing for an intervention impact. In Figure 6.4, for example, it is extremely difficult to tell from visual inspection whether trend or seasonality patterns exist. Visual inspection is sometimes more profitable if the data are graphed after aggregating them over time intervals. For example, when the data in Figure 6.4 are aggregated at the quarterly rather than the weekly level, seasonality patterns are not observed, though there does appear to be the hypothesized steadily decreasing trend in the small-motorcycle series. However, aggregation shortens the time series, which reduces the ability to model other aspects of the data. So large numbers of data points are preferable, all other things being equal.

Situations frequently arise, however, in which many more observations are available than just a single pretest and posttest, but the number is not close to 100 observations. Such short time series can still be useful for causal inference even if statistical analysis by standard methods is inadvisable. There are four principal reasons for this usefulness. First, the addition of the extra pretests will help address some threats to internal validity compared with designs with just one or two pretests. Second, the extra posttest observations help specify the delay and the degree of persistence of the causal impact. Third, the use of control groups or control time series is often feasible and can greatly strengthen inferences in the short series. Fourth, some analyses of short series are possible, for example, making assumptions about the error structure rather than describing it directly. Economists do this regularly (Greene, 1999; Hsiao, 1986; Hsiao, Lahiri, Lee, & Pesaran, 1999).

The Usefulness of Multiple Pretests and Posttests

Some advantages of a short interrupted time series are apparent in Figure 6.12, which shows how attendance in a job training program during 1964 affected sub-

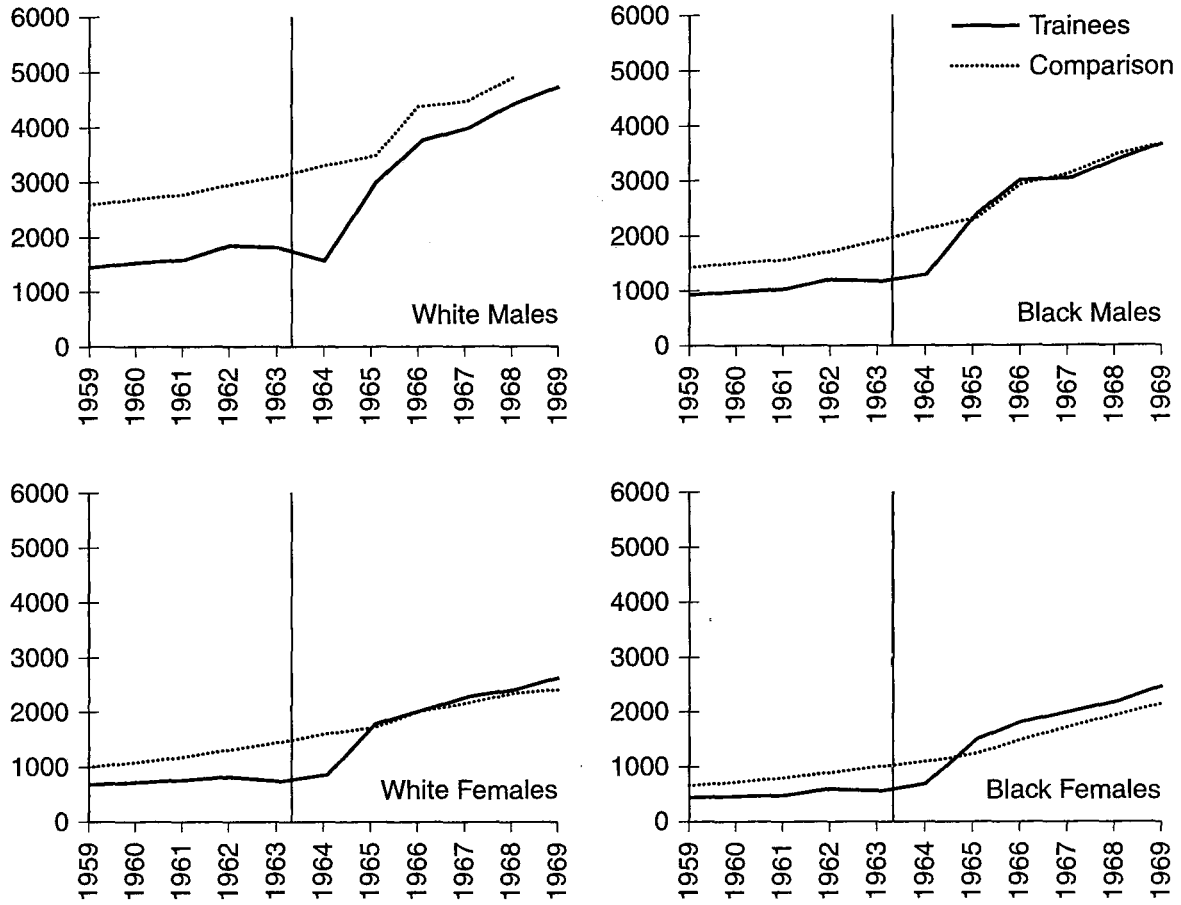

FIGURE 6.12 The effects of attendance in a job training program on subsequent earnings

From "Estimating the effects of training programs on earnings," by O. Ashenfelter, 1978, *Review of Economics and Statistics, 60,* pp. 47–57. Copyright 1978 by The MIT Press.

sequent earnings for groups of males and females who are Black or White (Ashenfelter, 1978). The treatment group comprised all those who began classroom training under the Manpower Development and Training Act in the first 3 months of 1964, a group that Ashenfelter noted were most likely to be successful. The control sample was constructed from the 0.1% Continuous Work History Sample of the Department of Labor, a random sample of earnings records on American workers. The outcome was earnings at eleven time points for each of the four groups.

Being in the work force, the controls had higher initial incomes (income data were taken from Social Security records). Figure 6.12 suggests a prompt causal impact in all four groups. Imagine now that only the 1963 and 1965 data had been available, the first full years before and after the program. An increase in earnings would still be suggested, but such a causal inference would be threatened by a number of alternative explanations. One is selection-maturation, the possibility that the training group was increasing its earnings faster than the control group but beginning at a lower starting point than the control group, even before 1963.

With the short pretest series, one can directly examine the plausibility of group differences in maturation.

Consider next the threat of regression. The persons eligible for job training programs in 1964 were those who were out of work in 1963. This fact could have depressed estimates of 1963 earnings for trainees relative to prior years if they had previously been employed and earning at the same level as the control group prior to 1963. If they were, and if their unemployment was just temporary, then the training group may have increased their posttreatment earnings in any case. Having 1963 as the sole pretest point does not allow us to estimate the plausibility of such regression, but having a few years of prior data does. In this case, regression may have added to the effect of training because a small decrease in earnings did occur in the treatment group between 1962 and 1963. But regression cannot account for all of the treatment effect because the average pretest earnings of the treatment group were much lower than those of the control group for many years, not just in 1963.

Without the later posttest years in Figure 6.12, one would not have been able to ascertain whether the effect of training was of any significant duration or quickly dissipated. Without the pretest series one might wonder, from consideration of only the years 1962 through 1965, whether the apparent change in earnings from 1962 to 1965 merely reflects a 4-year economic cycle within a general upward trend. The years from 1959 to 1962 help rule out this possibility, for they show no cyclical patterns. So the addition of multiple pretests and posttests greatly aids interpretation of quasi-experiments even when a full time series cannot be done (H. Bloom, 1984b, reanalyzed these data).

Strengthening Short Series with Design Features

The interpretability of short time series is enhanced by adding any of the design features discussed in this or previous chapters (e.g., Table 5.2), such as control groups, nonequivalent dependent variables, switching replications, treatment removals, and multiple replications (e.g., Barlow & Hersen, 1984; R. Franklin, Allison, & Gorman, 1997; Kratochwill & Levin, 1992; Sidman, 1960). For example, McKillip (1992) assessed the effects of a 1989 media campaign to reduce alcohol use during a student festival on a university campus. His primary dependent variable was a short time series (10 observations) on awareness of alcohol abuse in the targeted population. To strengthen inference in this short series, McKillip added two nonequivalent dependent variables (he calls them control constructs, to highlight their similarity to control groups) that were conceptually related to health and so would have shown changes if effects were due to general improvements in attitudes toward health. But these two variables (good nutrition, stress reduction) were not specifically targeted by the campaign, and so they should not have changed if the effect was due to the treatment. As Figure 6.13 shows, awareness of alcohol abuse clearly increased during the campaign, but

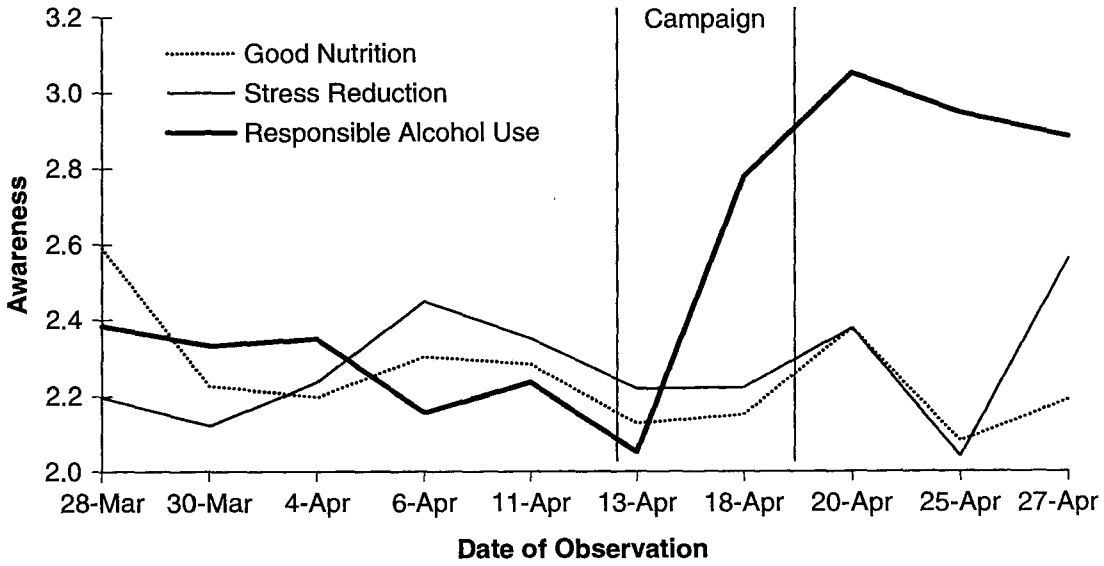

FIGURE 6.13 The effects of a media program to increase awareness of alcohol abuse

From "Research without control groups: A control construct design," by J. McKillip, 1992, *Methodological issues in applied psychology,* edited by F. B. Bryant, J. Edwards, R. S. Tindale, E. J. Posavac, L. Heath, & E. Henderson, New York: Plenum. Copyright 1992 by Plenum Press.

awareness of other health-related issues did not (see Fischer, 1994, for several similar applications that yielded more ambiguous results).

McClannahan, McGee, MacDuff, and Krantz (1990) added a switching-replications feature to their short time series (21 observations) that assessed the effects of providing regular feedback to married couples who supervised group homes for autistic children about the daily personal hygiene and appearance of the children in their home. The feedback was introduced after Session 6 in Home 1, Session 11 in Home 2, and Session 16 in Home 3. After each introduction, the personal appearance of the children in that home increased above baseline, and that improvement was maintained over time (Figure 6.14). However, both these examples illustrate a disadvantage of short time series, the difficulty in knowing how long the effect will last. Figure 6.13 shows this most clearly, with the beginnings of an apparent decrease in alcohol abuse awareness already apparent after the 2-week intervention.

Analysis of Short Time Series

When dealing with short time series, many researchers fail to do any data analysis. Some believe that visual analysis is adequate and that effects are not worth finding if they are so small that they require statistics to tease them out. But research suggests that visual inspection is fallible in the face of small or delayed effects (e.g., Furlong & Wampold, 1981; Ottenbacher, 1986; Wampold & Furlong,

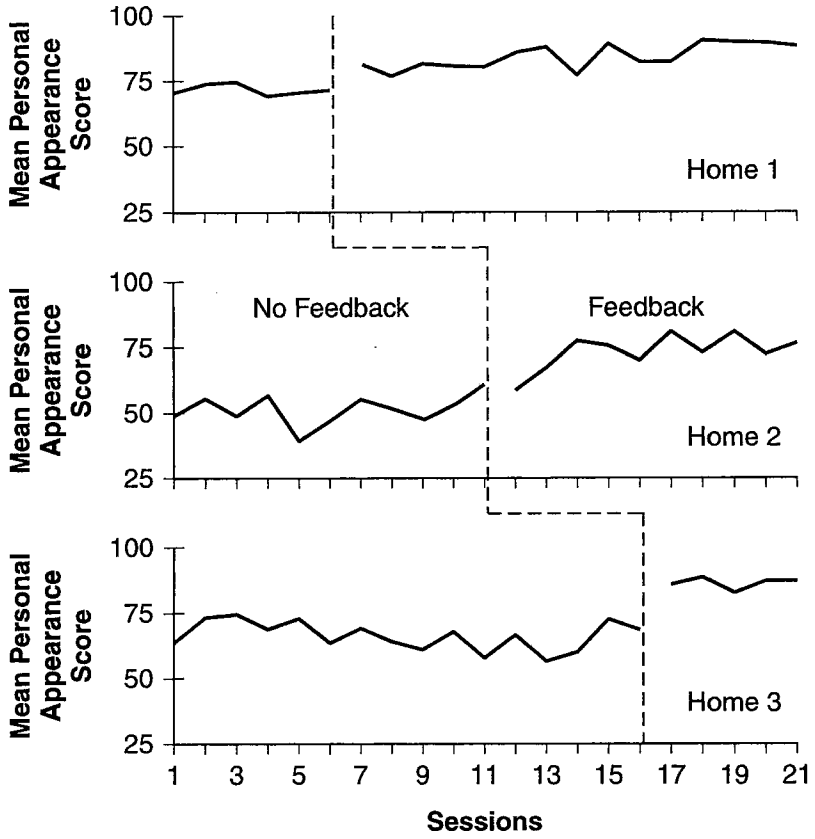

FIGURE 6.14 The effects of a parental intervention on the physical appearance of autistic children in three different homes

From "Assessing and improving child care: A personal appearance index for children with autism," by L. E. McClannahan et al., 1990, *Journal of Applied Behavior Analysis, 23,* 469–482. Copyright 1990 by The Society for the Experimental Analysis of Bahavior.

1981), and the belief that small effects are not important (e.g., Barlow & Hersen, 1984, p. 282) needs reconsideration given past demonstrations of important small effects (e.g., Rosenthal, 1994). Others have reviewed various analytic options but with mixed opinions about their worth (e.g., Franklin, Allison, & Gorman, 1997; B. Gorman & Allison, 1997; Matyas & Greenwood, 1997). Potentially accurate nonparametric options include randomization (exact) tests (Edgington, 1992; Gorman & Allison, 1997; Koehler & Levin, 1998) and bootstrapping methods (Efron & Tibshirani, 1993), although both often have low power for shorter time series. K. Jones (1991) and Crosbie (1993) suggest parametric alternatives, but critics point to significant problems with them (Reichardt, 1991). If the same people are measured over time, various kinds of repeated-measures ANOVAs can sometimes be used, as can growth curve models or event history analysis (e.g., Carbonari, Wirtz, Muenz, & Stout, 1994). Economists work with short time series by making assumptions about the error structure of the data to correct standard error estimates (Greene, 1999; Hsiao, 1986, 1999). If a short time series can

be gathered on many individual units, then pooling time series can sometimes be useful (Sayrs, 1989; West & Hepworth, 1991). Perhaps the best advice is to apply several statistical methods to short time series (see Allison & Gorman, 1997, for a summary of the options). If results of diverse analyses converge and are consistent with visual inspection, confidence in the findings may be warranted.

However, facilitating visual analysis of time-series data is important. Good visual presentation uses the power of graphics to engage the reader. The study of graphical techniques has become a specialty in itself (e.g., Tufte, 1983, 1990), and an increasingly diverse array of graphical techniques is being made more accessible by computer. Both data analysis and visual analysis are important to time series, as is knowing the limitations of each.

Limitations of Much Archival Data

Much time-series data come from archives kept by public and private institutions. However, guides to archives are often difficult to obtain. Gaining access to data from the private business sector or from local institutions (e.g., schools and city governments) can be difficult. Frequently, a researcher may be assured that some particular data were available and may travel to the archive, only to find that the data are not on hand. Recent years have seen some improvements in this situation. For example, Kiecolt and Nathan (1990) have described some of the many sophisticated archive services that have been established across the United States. Similarly, a database of over 560 time series is available for sale ("Time Series Database," 1992) that may be useful for some general time-series purposes. Various computerized networks such as the Internet are making time-series data sets more convenient to locate and retrieve.[3]

Just as important is the possibility of construct validity problems in archival data. Because most data that are collected and stored are used for social and economic monitoring, variables that have an "outcome" rather than a "process" flavor are stressed, and direct archival measures of psychological and small-group processes are rare. Consequently, current archives may not be useful for testing causal explanations or psychological constructs.

All archived data need close scrutiny. Operational definitions have to be critically examined, and the construct label applied to a particular measure may not necessarily be a good fit. Inquiries have to be made about shifts in definition over the time a record is kept. When possible, the nature of the shift should be documented, and it helps if overlapping data are examined for any periods during which data were collected using both the old and new definitions. Shifts in definition may be suggested as soon as the data are plotted, but it is more convenient

3. A comprehensive list of economic and social time series is at *http://www.economagic.com/*, and some United States government time series are at *http://www.fedstats.gov/* and *http://www.census.gov/*.

to document such shifts when they first occur. In many cases, the data for some times will be missing and, if the problem is not too serious, will have to be imputed. If the plotted data show suspicious regularities, as when values remain constant for a section of the series or the increase per time unit is constant, it may be that the data were not properly collected or that someone has interpolated values for the missing observations without documenting that fact. Conversely, the data may vary widely and resist any rescaling to reduce the variability, perhaps as a result of sloppy data collection procedures rather than inherently unstable phenomena.

The major difficulty with archival data is their inflexibility. The time-series analyst prefers data that can be disaggregated into more frequent time intervals, local regions, individual demographics, and fine topical breakdowns (Campbell, 1976). But archives rarely allow this flexibility. Often, the researcher would like weekly, monthly, or quarterly data because they provide longer series than annual data, are more sensitive in detecting immediate causal impacts, and are better fitted for ruling out historically based alternative interpretations. But if the data are collected and stored in annual form, disaggregation into finer time units is not possible. The researcher often wants to disaggregate data by various social variables, including demographic characteristics such as race, class, or gender. Such disaggregation permits examination of how an effect varies over groups to explore external validity or over groups that did or did not receive a treatment, creating a control series. For instance, in studying the effects of the Breathalyzer, it would be useful to break the data down into drinkers versus nondrinkers or into different religious groups that do or do not proscribe drinking. Often, it would aid the researcher if more dependent variables were available, for then he or she might be able to find a nonequivalent dependent variable. In all these cases, however, the researcher usually has to be content with what the archive contains.

We should not be overly pessimistic about the rigidity of archives or the quality of their data. As the examples in this chapter show, interrupted time-series designs that permit confident causal inferences can often be implemented. Moreover, with sufficient ingenuity and perseverance, the researcher may uncover data one would not expect to find in an archive. For example, Cook, Calder, and Wharton (1979) obtained 25-year series covering a wide range of variables that can be classified under the general heading of consumption, leisure, political behavior, workforce participation, local economic structure, public health, and crime. Though some of the data came from federal archives, the majority came from obscure state records that seem to be rarely consulted for research purposes. States differ in the quality of the records they maintain. To cover a substantive area adequately, the researcher may have to collect some variables in one state and others in another. However, a surprising amount of time-series data is available. If recent and future data are of better technical quality than was sometimes the case in the past, we can hope to see increased use of time-series analyses based on all these data sources.

A COMMENT ON CONCOMITANT TIME SERIES

The interrupted time-series designs we have discussed in this chapter differ substantially from another use of time series that is sometimes advocated for causal inference—the concomitant time series. The interrupted time series requires a treatment that is deliberately manipulated and implemented. Sometimes a potential causative agent is not implemented in this way but rather varies in intensity without experimental control over a period of time during which the outcome is also observed to vary. In a concomitant time series, the investigator correlates the time series from the presumed causal variable with a time series on the presumed outcome, both series being measured on the same units over the same time. The researcher then looks at how rises and falls in the causal series are related to rises and falls at later times in the effect series. This gets at the temporal precedence issue so central to conceptions of cause. McCleary and Welsh (1992) cite a previously published example of the correlation of citations to the Rorschach and the Minnesota Multiphasic Personality Inventory (MMPI) personality tests, the hypothesis being that as more clinicians started to use the MMPI, it displaced the Rorschach as the test of choice. The correlations reported by the original authors supported that interpretation, though the original analysis had significant statistical problems such as the failure to take autocorrelation into account or to allow for a time lag by computing lagged correlations.

Crucially, however, the putative cause in a concomitant time series is not manipulated experimentally but rather fluctuates in an uncontrolled manner. It may seem strange, therefore, that such uncontrolled correlations are cited as evidence for causation, given the many well-known reasons why correlation does not prove causation. Some advocates cite the notion of "Granger causality" to justify the approach, relying on the logic of Granger (1969). Specifically, this logic maintains that if the causal relationship is known to be unidirectional at a particular lag period and if the two variables meet an analytic condition for "white noise," then the (appropriately time-lagged) correlation yields an unbiased estimate of the causal relationship. Unfortunately, although the white noise condition can be tested, it is unlikely that the other conditions will be met in practice. The simple unidirectional causation condition is very unlikely in most real world applications (McCleary & Welsh, 1992); Shonkoff and Phillips (2000) note that the problem is commonly called simultaneity bias. Cromwell, Hannan, Labys, and Terraza (1994) conclude, "When we speak of 'Granger causality,' we are really testing if a particular variable precedes another and not causality in the sense of cause and effect" (p. 33; see also Holland, 1986; Menard, 1991; Reichardt, 1991).[4]

4. Similar problems exist with other proposed models of inferring causation from uncontrolled observations of the relationships between two variables (e.g., Wampold, 1992).

CONCLUSION

Interrupted time series constitute an important class of designs for gaining causal knowledge when circumstances permit collecting the required data across many time points. They gain their advantage from the pretreatment series that allows many potential threats to be examined, from exact knowledge of the time at which the treatment was made available that partially helps to deal with history, and from the posttreatment data that allow the form of the causal relationship to be described in terms of the speed of onset and the degree of persistence of the effect. We are great fans of interrupted time-series designs and would like to see more of them, whether with data from archives or collected first-hand by researchers themselves. We are fans even when the number of observations is fewer than required for traditional statistical analyses. The principle we endorse is that the more information one has about the temporal aspects of pre- and postintervention performance, the better the resulting study can decrease uncertainty about whether an association is causal.

The same advantage follows from increased knowledge about the intervention under study, including the time of its onset and the form of its diffusion through the population under study. The specificity that comes from knowing precisely at what point an intervention is supposed to influence an outcome is of the greatest utility, especially when the intervention occurs at a specific point along an interval scale with many levels. In the time-series case, time is that scale, and interruptions in the form of the response at the time of the intervention provide diagnostic clues about causation.

In the chapter that follows, many of these principles are seen again in similar form. The design in question, called the regression discontinuity design, also requires knowledge of the onset of the intervention at a point on a continuum (although the continuum is not time but the ordered points on a measured variable used to assign units to conditions), with the effect demonstrated by an abrupt change in the outcome variable at that point. The regression discontinuity design also shares some characteristics in common with the randomized experiment we describe later in this book, so it helps to bridge from the interrupted time series to the randomized experiment.