15

Visual Data Analysis

In behavior analysis, as data are collected, information is graphed and analyzed on a continuous basis until the experiment is completed. Information from each session is plotted in a graphic display, and patterns in the data are studied to decide what the next step in the experiment will be. An example will help illustrate this process. After selecting the initial procedures for an experiment and deciding on a design tactic (e.g., an A-B-A-B design), data collection begins. After the first observational session, the data are scored, summarized, and charted in graphic form. The research team then visually inspects the data. After the first day, there are no trends or patterns in the data apart from the specific levels obtained for each variable. However, as this process is repeated, visually analyzing the data begins to yield more interesting patterns. For instance, the occurrence of problem behavior and on-task behavior may occur at consistent levels day after day. If this pattern is replicated within a participant, decisions need to be made about the introduction of an independent variable. If the pattern in the data makes this a logical decision, then the intervention could be introduced. Or, if the pattern in the data had downward or upward trends or was highly variable, the research team may choose to continue baseline until a stable pattern is obtained. As the study unfolds, decisions are made on a daily basis whether to continue, change, or end a particular set of procedures. These decisions are ongoing during a study and always made in reference to the data. This process is summarized by the expression, often heard among behavior analysts, "Follow your data."

This is a dynamic process and roughly analogous to a chess match. The first observation is made (you move a chess piece), data are collected (your opponent makes a move), the results analyzed (your next step is planned), and the next move executed, the data collected, and so on. In chess, when errors are made or unexpected moves from an opponent occur, the tactical plan is changed accordingly and as often as needed to be successful. In single-case research, plans are changed as patterns in the data unfold, with the goal of revealing some type of behavioral process in the form of a functional relation. The use of graphic displays to visualize quantitative information is central to this process. The data, in graphic format, can act as a road map for conducting a study, since the course cannot be predicted in an a priori manner (Latour, 1990).

This process can be contrasted with group comparison designs in which an experimental design (e.g., a pre/postcontrol group design; cf. Campbell & Stanley, 1966) is selected before the start of the study, data are then collected prior to and after intervention, the data are summarized, and the appropriate inferential statistics are used to test for an experimental effect (see Box 15.1). In such a case, the analysis occurs after the data are collected, and no change in the experimental design can be made without limiting internal validity. Unlike group comparison designs, single-case designs are highly dynamic, and often the most exciting analyses unfold over time as patterns in the data emerge and the experimental tactic is adjusted in reaction to the data (e.g., see Chapter 14).

The process of inspecting graphic data is a very powerful way of revealing functional relations (Hacking, 1983; Smith, Best, Stubbs, Archibald, & Roberson-Nay, 2002). In behavior analysis, the use of graphs to analyze data is as old as the field itself. In B. F. Skinner's seminal work, *The Behavior of Organisms* (1938), which led to the development of behavior analysis, visual displays of data were prominently featured as the primary means of data analysis. Figure 15.1 shows the first graph from Skinner (1938). In this figure, the cumulative number of responses by a rat is shown during the initial shaping of a lever press. The data show that three lever presses occurred during the first 120 minutes of the session, with large temporal gaps between responses. However, approximately 130 minutes into the conditioning session, the fourth response was emitted, and lever pressing began to occur at a rapid and regular rate for the remainder of the session. At this point, lever pressing had been brought under the control of a positive reinforcer. Figure 15.1 reveals this process in a manner that is easily accessed by other researchers. It is the most revealing way of analyzing the data and provides the most information to the viewer.

BOX 15.1 • *Use of Inferential Statistics in Single-Case Designs*

The use of inferential statistics in single-case research has an intellectually interesting, but fractious, history. Some researchers who use single-case designs have written eloquently about the potential that inferential statistics possess for this experimental methodology (e.g., Kratochwill, 1978; Thompson et al., 1999). Other behavior analysts have written equally eloquent expositions on why inferential statistics are not useful for researchers using single-case designs (Baer, 1977; Michael, 1974). However, these arguments—for and against—are largely moot points.

The practical problem with using inferential statistics in single-case designs is that the currently existing statistics either violate fundamental statistical assumptions or are intractable in the large majority of applied research (e.g., time-series analysis, randomization tests, or R_n test of ranks; see Hartmann et al., 1980; Kazdin, 1982, app. B). This former concern centers around issues such as normalcy of distributions and serial dependency. The latter is largely an issue of the statistical test placing severe constraints on an experimental design, either in terms of the number of data points needed or the requirement of randomizing conditions.

Therefore, the use of inferential statistics in single-case designs is largely an academic debate and not a practical issue for researchers looking for new analytical tools. If, in the future, inferential statistics can be developed that fit the design requirements of single-case research, then the issue will require renewed debate. Until that time arrives, however, single-case researchers will continue to use the visual analysis of data as a primary means of examining their data.

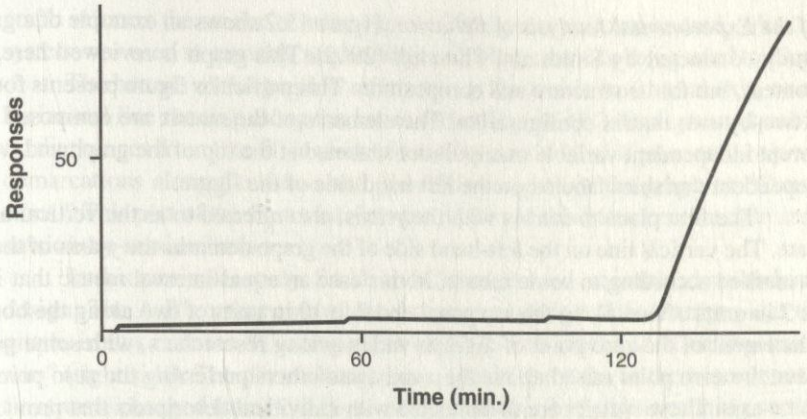

FIGURE 15.1 ***The first data set presented in B. F. Skinner's*** **The Behavior of Organisms** ***(see Catania, 1988).*** The *y*-axis displays the occurrence of individual responses as a cumulative function. The *x*-axis represents the passage of time.

Source: From B. F. Skinner (1938). *The behavior of organisms: An experimental analysis,* fig. 2, p. 67. Acton, MA: Copley. Copyright 1991 by the B. F. Skinner Society. Reprinted by permission.

The use of graphic displays by Skinner was no accident. He was using the tools of experimental biology to analyze psychological phenomena (see Chapter 2). Graphic displays—because of their flexibility, ease of use, and aid in visualizing functional relations—quickly became a mainstay in the experimental analysis of behavior, a practice that continues today. Not surprisingly, when researchers began to apply the behavioral processes discovered in the laboratory to natural settings, such as public schools, the use of graphs was continued, along with many other scientific practices. In fact, one of the most powerful interventions that allows educators to be more effective in teaching is the graphing of student performance and the visual analysis of the data on a regular basis, with appropriate adjustments to teaching procedures as indicated by the data (Alper & White, 1971; Lindsley, 1991).

This chapter explores how to use graphs in single-case research. First, the elements comprising a graph are reviewed. The discussion then explores how to visually inspect data so that a conclusion can be reached about patterns in the data and whether experimental control was demonstrated. Examples of how to use graphs to explore and analyze different facets of the data collected in an experiment follow. Finally, the topic of how to teach people to visually inspect data so that there is a high level of consistency among those inspecting data is examined.

Elements of a Graph

Although there are a multitude of graphic formats that can be used to visually display data, there are some elements that most graphs share in common (see Parsonson & Baer, 1978, for an extensive treatment of this topic). An excellent source for how to create graphs for publication purposes is contained in the January issues (2000 through 2003) of the *Journal*

of the Experimental Analysis of Behavior. Figure 15.2 shows an example of a graph from a study conducted by Smith and Churchill (2002). This graph is reviewed here, not for its content, but for its structure and composition. This particular figure presents four panels in a two-by-two matrix configuration. The elements of the matrix are composed of two different independent variable manipulations labeled at the top of the graph and two different dependent variables labeled on the left-hand side of the figure.

The first place to start is with the *y*-axis, also referred to as the vertical axis or ordinate. The vertical line on the left-hand side of the graph demarks the *y*-axis of the graph and is marked according to some metric, in this case an equal-interval metric that is labeled 0 to 3 in units of one along the top panel and 0 to 10 in units of two along the bottom panel. Placement of the zero point of a graph varies among researchers, with some preferring to have the zero point raised above the *x*-axis, and others preferring the zero point to rest on the *x*-axis. These metrics are then labeled with individual descriptors that provide information regarding the measurement unit and topography of behaviors. In this instance, the upper panel is labeled "Responses per Minute (SIB)," and the lower panel is labeled "Responses per Minute (precursors)." The information arrayed along the vertical axis of this

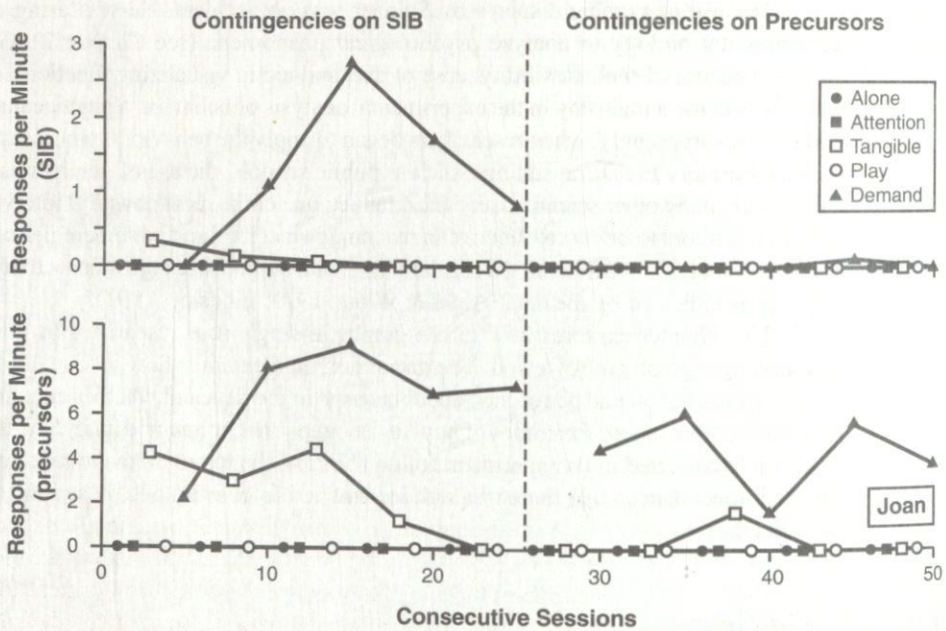

FIGURE 15.2 ***Example of a graph from a study used to illustrate the elements in a graph.***

Source: From R. G. Smith and R. M. Churchill, "Identification of Environmental Determinants of Behavior Disorders through Functional Analysis of Precursor Behaviors," *Journal of Applied Behavior Analysis, 2002, 35,* fig. 1, p. 130. Copyright 2002 by the Society for the Experimental Analysis of Behavior. Reprinted by permission.

graph informs the reader of what types of behaviors were measured, what measurement system was used to quantify responding, and what metric is being used to display the data.

The bottom of the graph is referred to as the *x*-axis, horizontal axis, or abscissa. The horizontal line running along the lower end of Figure 15.2 displays an equal-interval metric that ranges from 0 to 50 in units of ten. The metric label is presented below the numeric demarcations along the *x*-axis and identifies this part of the graph as "Consecutive Sessions." This part of the figure informs the reader, in this instance, that data are plotted on a session-by-session basis.

In this particular study, Smith and Churchill used a combined multielement and A-B design. The A-B components of the design are labeled with descriptors at the top of the graph, in this instance "Contingencies on SIB" (left side) and "Contingencies on Precursors" (right side). The phase change line—that is, the point in the experiment when conditions were changed from A to B—is designated as a dashed line running vertically through the graph at midpoint. Some researchers prefer to use solid lines, and others prefer to use broken lines to denote phase changes.

Within Figure 15.2 are the data. In this example, five different experimental conditions were analyzed: alone, attention, tangible, play, and demand. The experimental conditions are labeled in a legend contained within the graph and presented adjacent to the specific symbol (closed circle, closed square, open square, open circle, and closed triangle, respectively) that represents each condition. The quantitative outcomes from individual sessions are presented as individual data points connected by lines. Note that the data points are not connected across phase changes. Finally, the pseudonym or other identifier for the individual whose behavior is being analyzed in the study is placed in a box within the graph (in this case "Joan"). Some researchers use boxes for the names and legends in a graph to clearly set them apart from the data, but others do not.

One last component of a graph is the figure caption, which is placed below or next to the figure if the study is published. The figure caption provides a written description of the information contained in the actual figure. Ideally, the reader should be able to look at the graph and read the figure caption and understand what the data represent without having to refer to the Method section or other text in the published paper. This information can include a general description of what the graph represents, a description of the *y*- and *x*-axes, and the phases and experimental conditions used in the study.

The components just reviewed comprise the basic elements of a graphic display. However, what specific form a graph takes depends on the nature of the data and what aspect of the experiment the investigators are trying to visualize (Tufte, 1997; Ware, 2000). One stylistic issue to be considered when constructing graphs is Tufte's (1983) concept of data ink and nondata ink. Data ink are those elements in a graph that are drawn to display information that is critical for the visual analysis of the data. Nondata ink are those parts of the graph that could be erased without removing any information critical to the visual analysis. In general, data ink should be maximized within a graphic display, and nondata ink should be eliminated whenever possible. An example of this concept is displayed in Figure 15.3 (page 196). In this figure, a histogram that might appear in a journal article is shown in the left-hand panel, which contains both data and nondata ink. In the center panel is are the nondata ink. In the right-hand panel is the remaining drawing, which presents only the data ink necessary for analyzing the figure. Although the right-hand panel may be

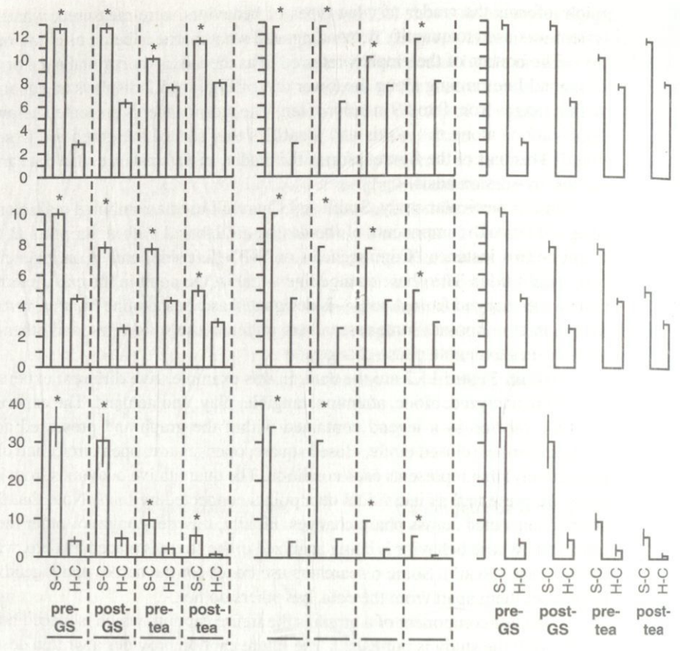

FIGURE 15.3 ***A histogram that might be seen in a scientific journal article.*** The histogram is shown in the left-hand panel, containing both data and nondata ink. In the center panel are only the nondata ink. In the right-hand panel is the remaining drawing that presents only the essential data ink necessary for analyzing the figure.

Source: From E. R. Tufte, *The visual display of quantitative information,* p. 102. Cheshire, CT: Graphics Press. Copyright 1983 by Graphics Press. Reprinted by permission.

considered an extreme case of minimizing nondata ink, it nicely illustrates the point that graphs should be kept as simple and uncluttered as possible so the eye is drawn to the data.

Visual Inspection of Graphs

The evaluation of quantitative information via visual inspection is accomplished by analyzing specific types of patterns in the data display. It may not seem obvious at first, but when researchers look at a graph, they look for a series of patterns that allow them to draw conclusions regarding what the data represent. These dimensions are familiar to someone who has been trained in their use, but for someone who is unfamiliar with them, their use is nonintuitive. In this section, we will review various dimensions of data that are visualized in graphs and used for analysis.

Within-Phase Patterns

The first dimension used in visual analysis is the level of the data. Level refers to the average of the data within a condition and is typically calculated as the mean or median. The left-hand panel in Figure 15.4 shows a baseline data set, with the level drawn over the data. There are six data points in the panel, with a mean of 4.7. Attending to the level of data within a phase allows for the estimation of the central tendency of the data during a particular part of an experiment. It also allows for comparison of patterns between phases (see below). Although the absolute level within a phase is important, it should be noted, particularly in applied research, that the last few data points contain the most essential information regarding the level of behavior before a phase change. The pattern of data shown in the right-hand panel of Figure 15.4 illustrates this point. Although the mean level of the data is 6.7, the last three data points deviate from this level enough to warrant special emphasis.

A second dimension used to visually inspect graphs is the trend of the data. Trend refers to the best-fit straight line that can be placed over the data within a phase. Trend has two distinct elements that must be simultaneously evaluated: slope and magnitude. Slope is the upward or downward slant or inclination of the data within a phase. Slopes are generally positive (upward), flat, or negative (downward). A positive slope is one in which the data points are increasing in value within a phase (see the upper left-hand panel of Figure 15.5, page 198). A negative slope is just the opposite, a downward pattern in the data within a phase (see the lower right-hand panel of Figure 15.5). The second element of a trend is magnitude, which is the size or extent of the slope. The magnitude of a trend is

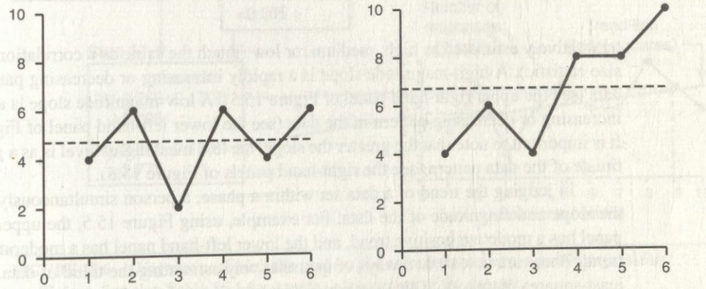

FIGURE 15.4 ***An example of level used to visually inspect data.*** The left-hand panel in the figure shows a baseline data set, with the level drawn over the data. There are six data points in the panel, with a mean of 4.7. The right-hand panel in the figure shows a data set in which the most essential information is within the last three data points, rather than the overall average.

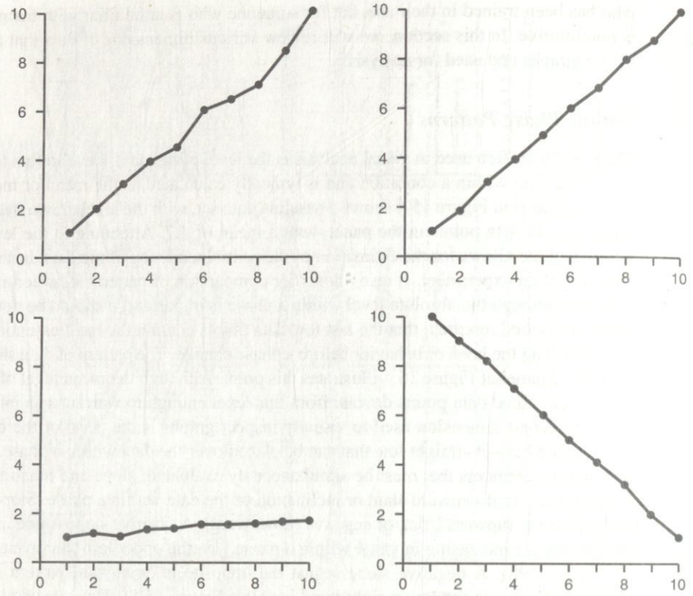

FIGURE 15.5 ***Examples of slope and magnitude used to estimate trend.*** Slope is the upward or downward slant or inclination of the data within a phase. Slopes are generally positive, flat, or negative. The magnitude of a trend is qualitatively estimated as high, medium, or low.

qualitatively estimated as high, medium, or low (much the same as a correlation or effect-size statistic). A high-magnitude slope is a rapidly increasing or decreasing pattern in the data (see the upper right-hand panel of Figure 15.5). A low-magnitude slope is a gradually increasing or decreasing pattern in the data (see the lower left-hand panel of Figure 15.5). It is important to note that the greater the slope, the less meaningful level is as a general estimate of the data pattern (see the right-hand panels of Figure 15.5).

In judging the trend of a data set within a phase, a person simultaneously estimates the slope and magnitude of the data. For example, using Figure 15.5, the upper left-hand panel has a moderate positive trend, and the lower left-hand panel has a moderate negative trend. There are at least two ways of quantitatively estimating the trend of data. The first, least-squares regression, fits a straight line to the slope of the data set by minimizing the sum of squared deviations of the observed data from the line. Figure 15.6 shows a diagram of how to calculate a least-squares regression line (from Parsonson & Baer, 1978, p. 131). Table 15.1 (page 200) describes how to calculate the same data shown in Figure 15.6. The

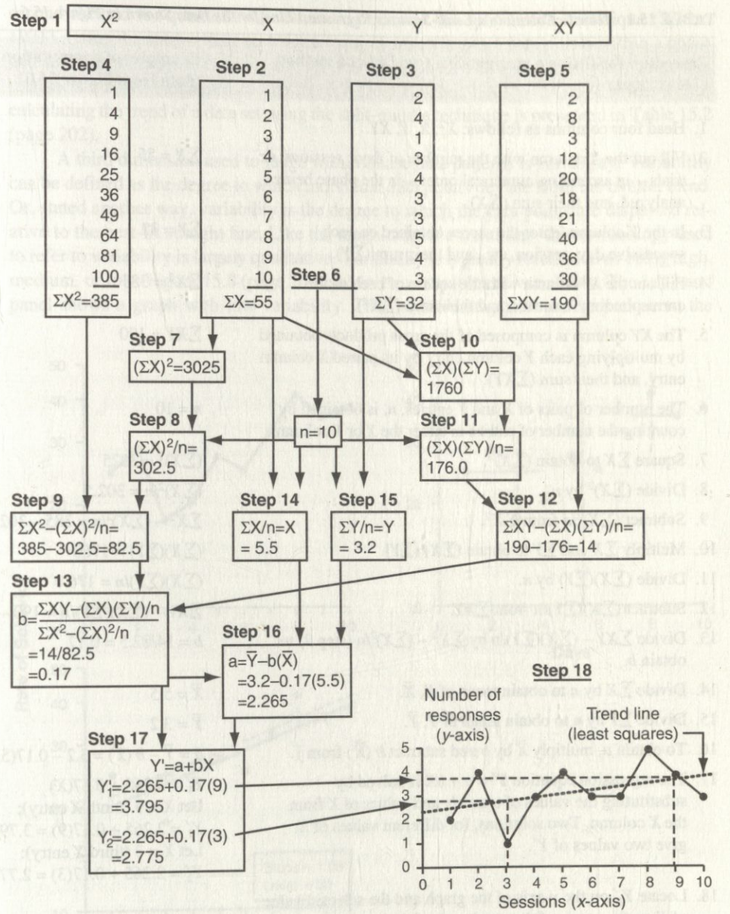

FIGURE 15.6 ***Diagram outlining the process for calculating a least-squares regression coefficient.*** See also Table 15.1 for a textual description of the process.

Source: From B. S. Parsonson and D. M. Baer, "The Analysis and Presentation of Graphic Data, 1978 (fig. 2.21, p. 131). In T. R. Kratochwill (Ed.), *Single subject research: Strategies for evaluating change* (pp. 101–166). New York: Academic Press. Copyright 1978 by Academic Press. Reprinted by permission.

TABLE 15.1 ***How to Calculate a Least-Squares Regression Line for the Data Shown in Figure 15.6***

Procedure for fitting a straight-line trend by the method of least squares	Examples of computation (data from Figure 15.6)
1. Head four columns as follows: X^2, X, Y, XY.	
2. Fill out the X column with the number of days, sessions, or trials—in ascending numerical order—in the phase being analyzed, and their sum ($\sum X$).	$\sum X = 55$
3. In the Y column, enter the scores obtained on each successive day, session, etc., and their sum ($\sum Y$).	$\sum Y = 32$
4. Fill out the X^2 column with the square of each of the corresponding X entries, and their sum ($\sum X^2$).	$\sum X^2 = 385$
5. The XY column is composed of the cross products obtained by multiplying each Y column entry by its paired X column entry, and their sum ($\sum XY$).	$\sum XY = 190$
6. The number of pairs of X and Y entries, n, is obtained by counting the number of entries in either the X or Y columns.	$n = 10$
7. Square $\sum X$ to obtain $(\sum X)^2$.	$(\sum X)^2 = 3025$
8. Divide $(\sum X)^2$ by n.	$(\sum X)^2/n = 302.5$
9. Subtract $(\sum X)^2/n$ from $\sum X^2$.	$\sum X^2 - (\sum X)^2/n = 385 - 302.5 = 82.5$
10. Multiply $\sum X$ and $\sum Y$ to obtain $(\sum X)(\sum Y)$.	$(\sum X)(\sum Y) = 1760$
11. Divide $(\sum X)(\sum Y)$ by n.	$(\sum X)(\sum Y)/n = 176.0$
12. Subtract $(\sum X)(\sum Y)/n$ from $\sum XY$.	$\sum XY - (\sum X)(\sum Y)/n = 190 - 176 = 14$
13. Divide $\sum XY - (\sum X)(\sum Y)/$n by $\sum X^2 - (\sum X)^2/n$ (step 9) to obtain b.	$b = 14/82.5 = 0.17$
14. Divide $\sum X$ by n to obtain mean of X, $\overline{X}$.	$\overline{X} = 5.5$
15. Divide $\sum Y$ by n to obtain mean of Y, $\overline{Y}$.	$\overline{Y} = 3.2$
16. To obtain a, multiply $\overline{X}$ by b and subtract b ($\overline{X}$) from $\overline{y}$.	$a = \overline{Y} - b\,(\overline{x}) = 3.2 - 0.17(5.5) = 2.265$
17. The regression equation $Y' = a + bX$ is solved by substituting the values of a and b, and values of X from the X column. Two solutions, for different values of X, give two values of Y'.	$Y_1' = 2.265 + 0.17(X)$ Let $X = 9$ (ninth X entry): $Y_1' = 2.265 + 0.17(9) = 3.795$ Let $X = 3$ (third X entry): $Y_2' = 2.265 + 0.17(3) = 2.775$
18. Locate Y_1' on the y-axis of the graph and the selected value of X on the x-axis of the graph and mark the point at which they intersect. Similarly, locate Y_2' on the y-axis and mark their point of intersection. A straight line drawn through the two points is the line of best fit and describes the trend in the data.	

Source: From B. S. Parsonson and D. M. Baer, "The Analysis and Presentation of Graphic Data, 1978, in T. R. Kratochwill (Ed.), *Single subject research: Strategies for evaluating change* (pp. 101–166). New York: Academic Press. Copyright 1978 by Academic Press. Reprinted by permission.

second method for quantitatively estimating trend is the split-middle technique (White, 1971). Using the split-middle technique requires seven or more data points within a phase and splits the data set in half, establishes a median for each half, and then plots a line that intersects the two medians (see Figure 15.7, from Kazdin, 1982, p. 313). A procedure for calculating the trend of a data set using the split-middle technique is presented in Table 15.2 (page 202).

A third dimension used to judge within-phase data patterns is variability. Variability can be defined as the degree to which individual data points deviate from the overall trend. Or, stated another way, variability is the degree to which the data points are dispersed relative to the best-fit straight line. Like the magnitude of a trend line, the terminology used to refer to variability is largely qualitative. Variability is typically referred to as being high, medium, or low. Figure 15.8 (page 202) shows two examples of variability. The left-hand panel shows a graph with low variability. That is, the data points are very close to the

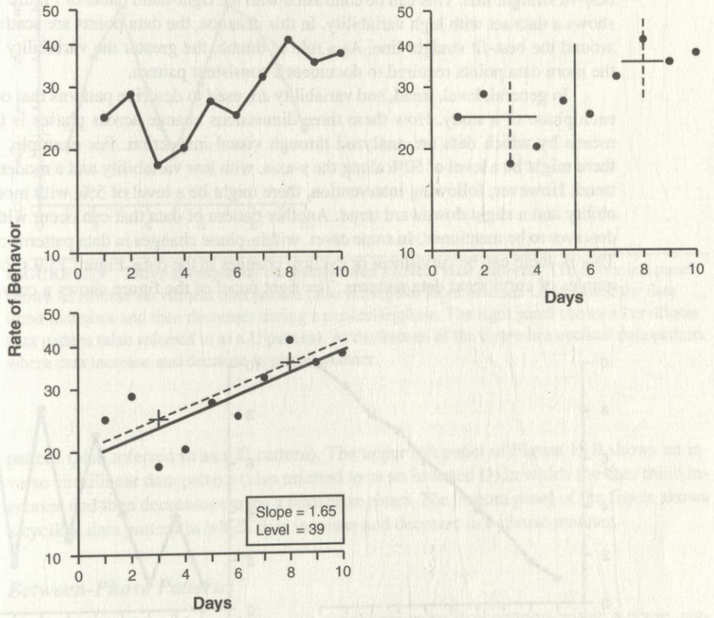

FIGURE 15.7 ***Series of graphs illustrating how to calculate trend using the split-middle technique described in Table 15.2.***

Source: From A. E. Kazdin, *Single-case research designs*. New York: Oxford University Press. Copyright 1982 by Oxford University Press. Reprinted by permission.

TABLE 15.2 ***How to Calculate a Split-Middle Trend Estimation Line***

1. Count the number of data points in the phase that is being used for trend estimation.
2. Draw a line on the graph at the median data point to divide the graph into two halves.
3. Divide each of the halves in half using the technique described in step 2.
4. Identify the median level of the data in each half of the split graph.
5. Mark the point at which the median number of sessions (*x*-axis) and median data level (*y*-axis) intersect for each half of the split graph.
6. Plot a straight line that intersects the two marks made in step 5.
7. Adjust the straight line plotted in step 6 so that 50% of the data points are above and below the line, making sure not to alter the slope of the line.

Note: Example uses data from Figure 15.7.

best-fit straight line. This can be contrasted with the right-hand panel of Figure 15.8, which shows a data set with high variability. In this instance, the data points are scattered widely around the best-fit straight line. As a rule of thumb, the greater the variability in the data, the more data points required to document a consistent pattern.

In general, level, trend, and variability are used to describe patterns that occur within each phase of a study. How these three dimensions change across phases is the primary means by which data are analyzed through visual inspection. For example, in baseline there might be a level of 50% along the *y*-axis, with low variability and a moderate upward trend. However, following intervention, there might be a level of 5%, with moderate variability and a slight downward trend. Another pattern of data that can occur within a phase deserves to be mentioned. In some cases, within-phase changes in data patterns can emerge. That is, there can be curvilinear or cyclical changes in the data. Figure 15.9 show three examples of curvilinear data patterns. The right panel of the figure shows a curvilinear data

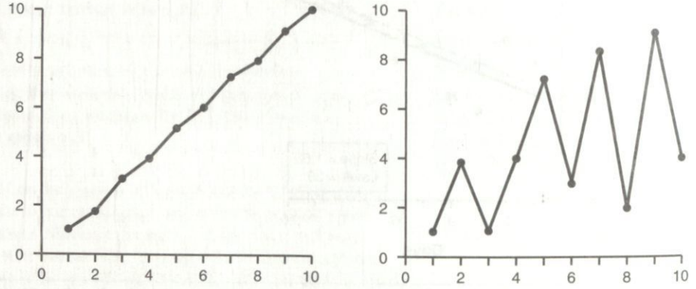

FIGURE 15.8 ***Examples of two different degrees of variability in data.*** The left-hand panel shows a graph with low variability. This can be contrasted with the right-hand panel, which shows a data set with high variability.

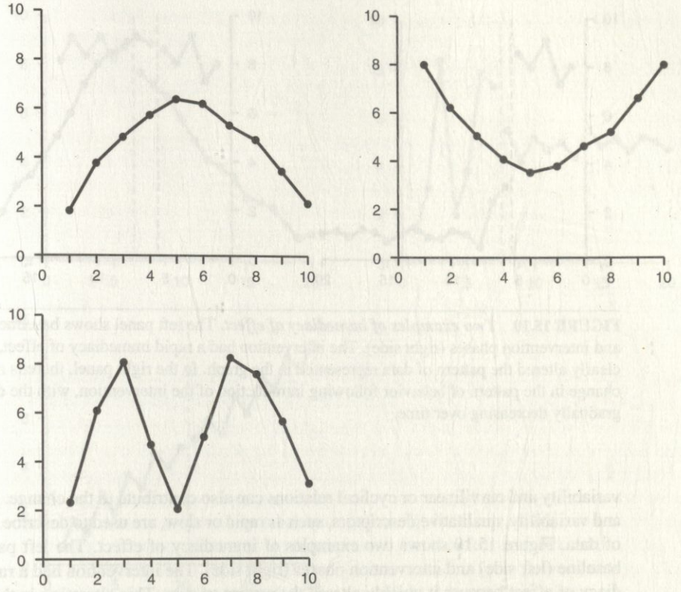

FIGURE 15.9 ***Three examples of curvilinear and cyclical data patterns.*** The upper left panel shows an inverse curvilinear data pattern (also referred to as an inverted U) in which the data trend increases and then decreases during a particular phase. The right panel shows a curvilinear data pattern (also referred to as a U pattern). At the bottom of the figure is a cyclical data pattern, where data increase and decrease in phasic manner.

pattern (also referred to as a U pattern). The upper left panel of Figure 15.9 shows an inverse curvilinear data pattern (also referred to as an inverted U) in which the data trend increases and then decreases during a particular phase. The bottom panel of the figure shows a cyclical data pattern in which data increase and decrease in a phasic manner.

Between-Phase Patterns

Along with level, trend, variability, and curvilinear or cyclical patterns within a phase, patterns occurring between phases are also used to visually inspect data. The first such pattern is referred to as immediacy of effect (or rapidity of change). This dimension of data display can be defined as how quickly a change in the data pattern is produced after the phase change. This is typically expressed as changes in the level and trend of the data, although

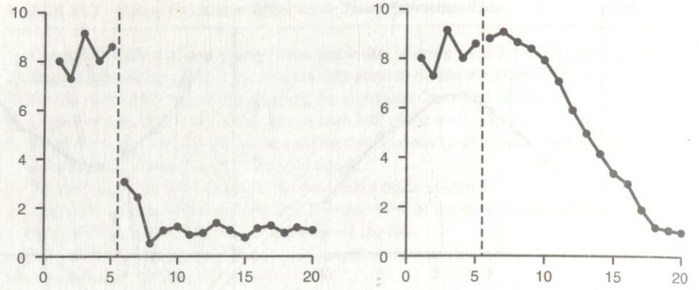

FIGURE 15.10 ***Two examples of immediacy of effect.*** The left panel shows baseline (left side) and intervention phases (right side). The intervention had a rapid immediacy of effect, because it clearly altered the pattern of data represented in the graph. In the right panel, there is no initial change in the pattern of behavior following introduction of the intervention, with the data then gradually decreasing over time.

variability and curvilinear or cyclical relations can also contribute to the change. Like slope and variability, qualitative descriptors, such as rapid or slow, are used to describe this aspect of data. Figure 15.10 shows two examples of immediacy of effect. The left panel shows baseline (left side) and intervention phases (right side). The intervention had a rapid immediacy of effect because it quickly altered the pattern of data. This alteration in the data pattern across phases can be contrasted with the data in the right panel of Figure 15.10. In this data set, there is no initial change in the pattern of behavior following introduction of the intervention, but then the data gradually decrease over time. Such a pattern would be referred to as having a slow immediacy of effect. In general, the greater the immediacy of effect, the briefer a phase can be and the more convincing is the functional relation.

A second between-phase pattern is referred to as overlap. Overlap can be defined as the percentage or degree to which data in adjacent phases share similar quantitative values. Figure 15.11 shows three distinct patterns of overlap between baseline (left side) and intervention phases (right side). In the upper left panel of the figure, there is no overlap (i.e., 0%) between baseline and intervention phases. This is because there are no overlapping data values between the two phases. The right panel of Figure 15.11 shows an example of complete (i.e., 100%) overlap between baseline and intervention phases. In this case, the intervention data completely overlap with the baseline data (although the converse is not true, so specifying the directionality of overlap is important). The final data set in Figure 15.11 (bottom panel) presents a common data pattern that leads to misinterpretation. Although there is no overlap between the data in adjacent phases, there is a trend that is continuous across phases. In such cases, trend overrides the importance of overlap in evaluating whether a functional relation has been established.

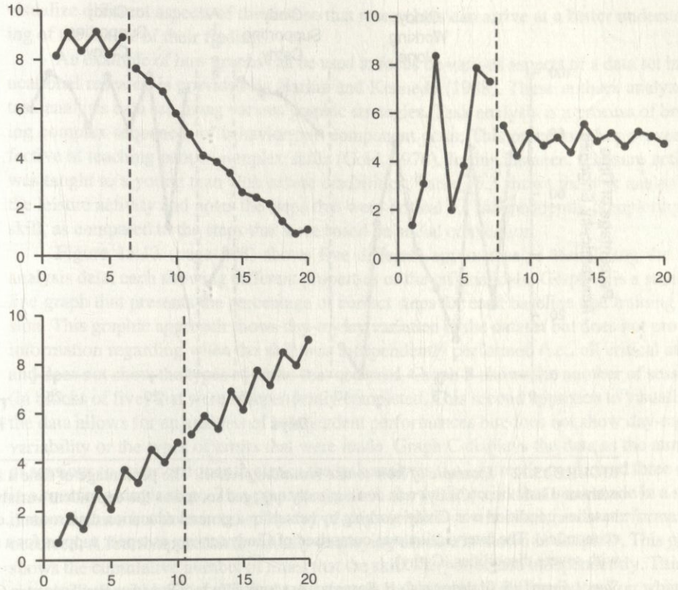

FIGURE 15.11 ***Three examples of overlap in data between phases.*** In the upper left panel, there is no overlap (i.e., 0%) between baseline and intervention phases. The right panel shows an example of complete (i.e., 100%) overlap between baseline and intervention phases. The final data set (bottom panel) presents a common data pattern that often leads to misinterpretation.

An Example

An illustration of how to use visual inspection to characterize a data set and arrive at a judgment regarding whether a functional relation has been established can be done using Figure 15.12 (page 206). In this study by Cushing and Kennedy (1997) the percentage of time a student without disabilities (Cindy) was academically engaged served as the dependent variable. The baseline condition was Cindy working by herself in a general education classroom. During the initial baseline, there was a high degree of variability, with a mean of 42% (range, 0 to 76%) and a moderate downward trend. The intervention was comprised of Cindy serving as a peer support for a student with severe disabilities (Cathy). Following intervention, there was an immediate increase in the level of the dependent variable ($M = 91\%$), with a small upward trend, little variability, and no overlap with the previous baseline. The

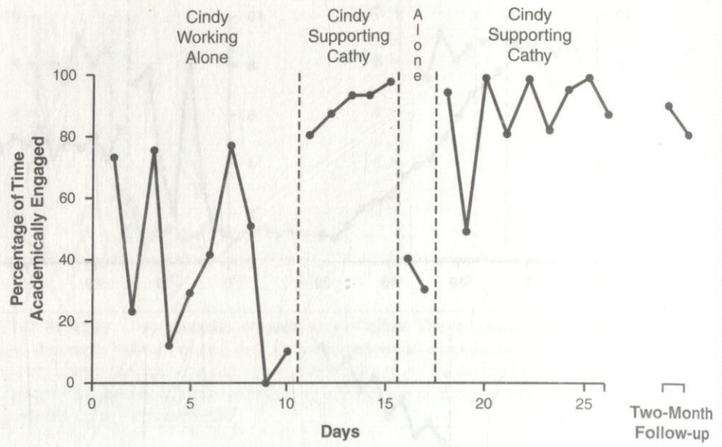

FIGURE 15.12 ***Example of how to use visual inspection.*** The percentage of time a student without disabilities (Cindy) was academically engaged served as the dependent variable. The baseline condition was Cindy working by herself in a general education classroom (i.e., home economics). The intervention was comprised of Cindy serving as a peer support for a student with severe disabilities (Cathy).

Source: From L. S. Cushing and C. H. Kennedy, "Academic Effects of Providing Peer Support in General Education Classrooms on Students without Disabilities," *Journal of Applied Behavior Analysis,* 1997, *30,* fig. 1, p. 145. Copyright 1997 by the Society for the Experimental Analysis of Behavior. Reproduced by permission.

withdrawal of the intervention coincided with a reversal to baseline levels of performance (M = 38%). Reintroducing the intervention resulted in an increase in academic engagement (M = 85%), which after several sessions returned to levels similar to the previous intervention phase. Using the A-B-A-B withdrawal design, Cushing and Kennedy were able to demonstrate that Cindy was more academically engaged when she assisted Cathy than when she worked alone.

Using Graphs to Analyze Data

Following this review of the elements of line graphs and the basic issues in visual analysis, the focus will now shift to how to use graphs to explore various aspects of a data set. Data sets are often multifaceted and lend themselves to a variety of analyses. Depending on the approach taken, different aspects of the nature of the data will be revealed. Therefore, visual analysis of data is much more than simply putting data into a graphic template and describing the information. Instead, visual analysis is a process of using graphs to explore and

visualize different aspects of the data so that researchers can arrive at a better understanding of the nature of their findings.

An example of how graphs can be used to focus on various aspects of a data set in educational research is provided by Haring and Kennedy (1988). These authors analyzed a task analysis data set, using various graphic strategies. Task analysis is a process of breaking complex sequences of behavior into component parts. This procedure has proven effective at teaching people complex skills (Gold, 1976). In this instance, a leisure activity was taught to a young man with severe disabilities. Table 15.3 shows the task analysis of the leisure activity and notes the steps that were critical for independently completing the skill, as compared to the steps that were based on social convention.

Figure 15.13 (page 208) shows five different approaches to visualizing the task analysis data, each showing different properties of the original data. Graph A is a standard line graph that presents the percentage of correct steps for each baseline and training session. This graphic approach shows day-to-day variation in the dataset but does not provide information regarding when the skill was independently performed (i.e., all critical steps) and does not show the types of errors that occurred. Graph B shows the number of sessions (in blocks of five) that were independently completed. This second approach to visualizing the data allows for an analysis of independent performances but does not show day-to-day variability or the types of errors that were made. Graph C displays the data as the number of sessions to criterion for each step in the task analysis (i.e., correctly performed three days in a row for a single step). This graphic approach shows the errors that were made in a summative fashion but does not display day-to-day variation or when competent performances occurred. A fourth approach to visually displaying the data is shown in Graph D. This graph shows the cumulative number of times that the skill was performed independently. This approach allows analysis of whether a session was independently performed and on what day (preserving day-to-day variation at a certain level) but provides no information regarding what types of errors were made. Graph E combines elements of Graphs A and D to display the percentage of steps correct and the occurrence of independent performances. By doing this, the graph shows all of the characteristics of the data previously described, with the ex-

TABLE 15.3 ***Task Analysis of a Leisure Skill***

Steps in Task Analysis
1. Gets radio and magazine[a]
2. Sits down in leisure area[a]
3. Turns radio on[a]
4. Selects radio station[a]
5. Puts headphones on appropriately
6. Looks at magazine[a]
7. Stops activity when signaled that break is finished[a]
8. Takes headphones off[a]
9. Turns radio off[a]
10. Puts magazine and radio away

[a]Critical steps.

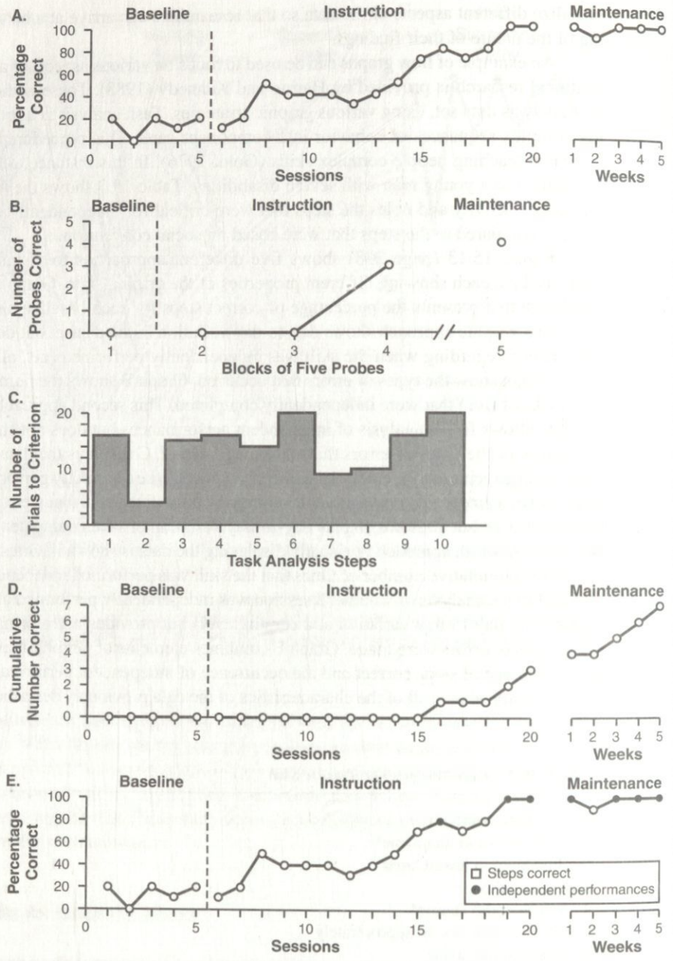

FIGURE 15.13 ***Example of how graphs can be used to focus on various aspects of a data set in educational research.*** Each graph reveals a different aspect of the original data.

Source: From T. G. Haring and C. H. Kennedy, "Units of Analysis in Task-Analytic Research," *Journal of Applied Behavior Analysis,* 1988, *21,* fig. 1, p. 209. Copyright 1988 by the Society for the Experimental Analysis of Behavior. Reproduced by permission.

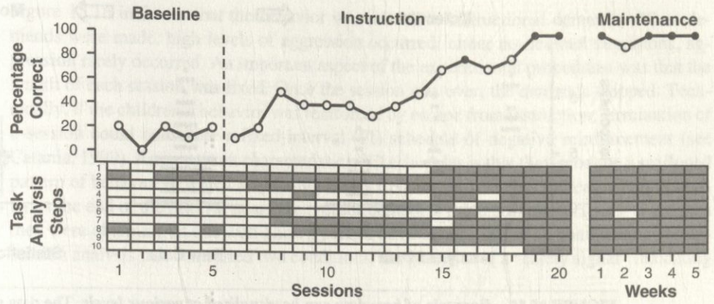

FIGURE 15.14 ***This graph allows for a comprehensive visual analysis of the original data displayed in various ways in Figure 15.13.***

Source: From T. G. Haring and C. H. Kennedy, "Units of Analysis in Task-Analytic Research," *Journal of Applied Behavior Analysis,* 1988, *21,* fig. 2, p. 213. Copyright 1988 by the Society for the Experimental Analysis of Behavior. Reproduced by permission.

ception of an error pattern analysis. A final graph was constructed by Haring and Kennedy that met each of the criteria previously discussed (see Figure 15.14). This final graph allows for a comprehensive visual analysis of the original data.

The previous examples illustrate how using various approaches to graphically display data can allow for the visualization of different aspects of a data set. The variety of means for visualizing data is enormous, and researchers need to select those graphic approaches that best represent salient aspects of their data. An issue that needs to be addressed when graphing data is the level at which the information will be summarized. In some instances, it may be more desirable to summarize data as averages (e.g., percentage of intervals), while at other times it may be best to graph data at a more fine-grained level (e.g., cumulative occurrences in real time). The former approach is sometimes referred as a molar approach, and the latter as a molecular approach.

Figure 15.15 (page 210) illustrates how the same data set can be visualized at various levels (Iversen, 1988). The data are from a laboratory experiment that analzyed the number of times a response was emitted following the delivery of a response-independent positive reinforcer. The *y*-axis displays this as the number of responses that occurred within thirty seconds of each reinforcing event. In the center panel of the graph are the raw data. It shows that on one occurrence, twelve responses were emitted; on four occurrences, eleven responses were emitted; on one occurrence ten responses were emitted; and so on. If the data were summarized at a molar level, such as the mean and standard deviation of the raw data, it would look like the graph in the right-hand panel of Figure 15.15. If the data were analyzed at a more molecular level, one approach would be to adopt the strategy shown in the left-hand panel. In this instance, the data were further analyzed in terms of the contiguity between reinforcer delivery and the last response prior to the reinforcer delivery. The data in

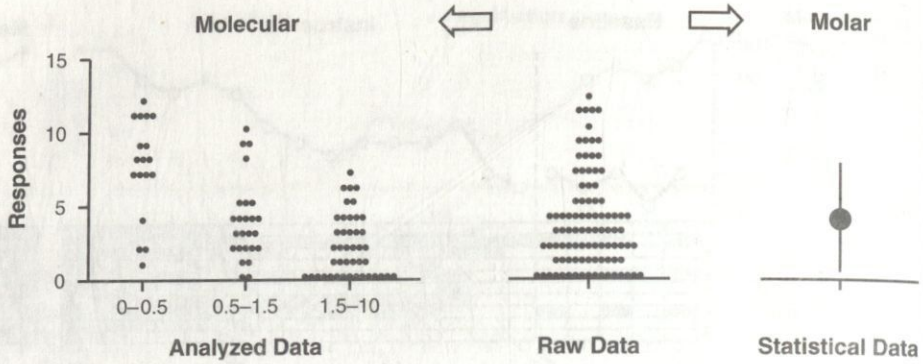

FIGURE 15.15 ***Example of how data can be visualized at various levels.*** The data are from a laboratory experiment that measured the number of times a response was emitted following the delivery of a response-independent positive reinforcer. The vertical axis displays this as the number of responses that occurred within thirty seconds of each reinforcing event. Each of the panels aggregates the data at a different level of refinement.

Source: From I. H. Iversen, "Tactics of Graphic Design: A Review of Tufte's *The Visual Display of Quantitative Information*," *Journal of the Experimental Analysis of Behavior,* 1988, *49,* fig. 6, p. 179. Copyright 1988 by the Society for the Experimental Analysis of Behavior. Reproduced by permission.

the left-hand panel clearly show that the greater the contiguity (e.g., 0 to 0.5 s) between prior responses and reinforcer delivery, the greater the number of postreinforcement responses.

The data displays in Figure 15.15 show how information can be summarized at different levels of aggregation and disaggregation. The raw data show the general distribution of events, the statistical data show central tendency and dispersion at the most coarse level, and the disaggregated data show specific patterns in the data. The information shown in the left-hand panel also illustrates the power of graphic displays to more completely analyze data. With the addition of a second variable along the horizontal axis, the author was able to account for variability in the behavior unexplained in the other panels. In this sense, graphs can be used to explore, visualize, and explain variability in behavior.

Another approach to visualizing information that is becoming increasingly common is to conduct within-session analyses. The information in Figures 15.2 through 15.15 show data-summarizing events occurring during an entire session. That is, the data represent the average level of events within the experimental session. However, in some instances, information regarding the pattern of events within a session can aid in the understanding of variables influencing behavior. One assumption that is implicit in summarizing whole-session data as individual data points is that the pattern of behavior observed in an experimental session was uniform throughout that session. For instances in which there are changes in responding during a session, within-session data analysis can reveal potentially important patterns that otherwise might not be discovered.

One of the first applied examples of within-session analysis was provided by Carr, Newson, and Binkoff (1980). Carr and colleagues were interested in understanding why aggression occurred in two boys with developmental disabilities. The analysis shown in

Figure 15.16 indicates that the behavior was related to instructional demands. When demands were made, high levels of aggression occurred; under no-demand conditions, aggression rarely occurred. An important aspect of the experimental procedures was that the length of each session was fixed. Once the session was over, the demands stopped. Technically, if the children's behavior was reinforced by escape from instruction, termination of a session could constitute a fixed-interval (FI) schedule of negative reinforcement (see Catania, 1999). An important characteristic of FI schedules is that they produce a scalloped pattern of behavior in which responding is less frequent early on but increases in probability as the end of the interval nears (i.e., reinforcement is more proximal). To see if the data they were obtaining fit this well-known pattern of behavior, Carr et al. conducted a within-session analysis that contrasted two conditions: the presence of a "safety signal" indicating

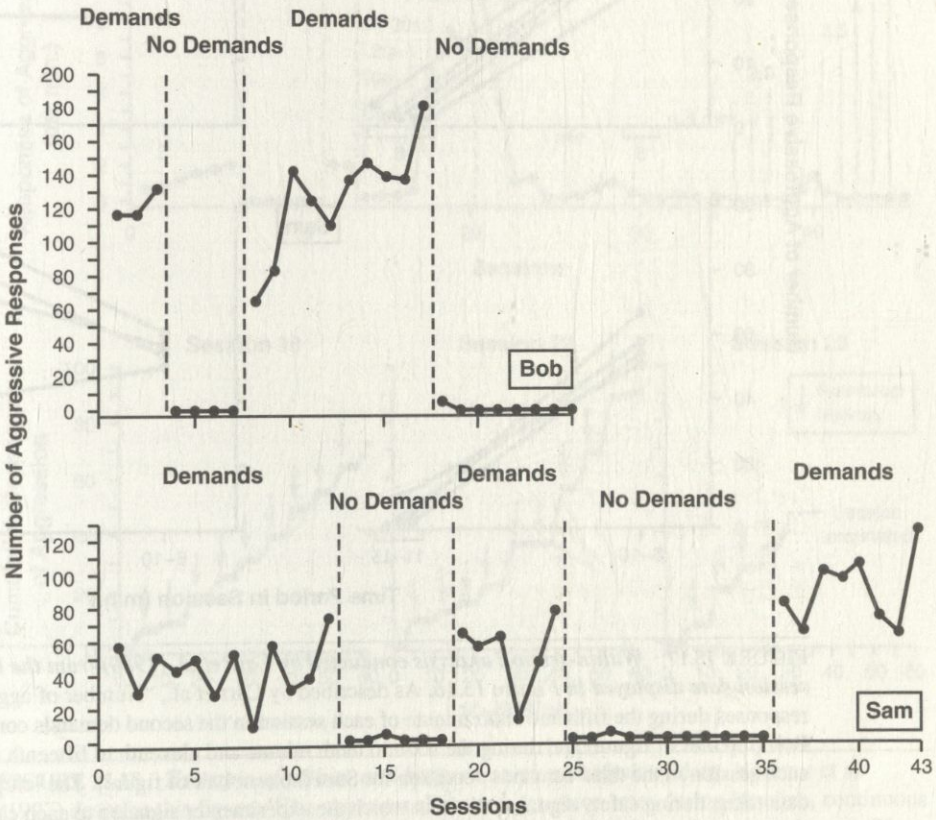

FIGURE 15.16 ***Initial analysis conducted by Carr et al. (1980).*** The number of aggressive responses for Sam and Bob (two boys with developmental disabilities) served as the dependent variable. The independent variable was the presence or absence of instructional demands.

Source: From E. G. Carr, C. D. Newsom, and J. A. Binkoff, "Escape as a Factor in the Aggressive Behavior of Two Retarded Children," *Journal of Applied Behavior Analysis,* 1980, *13,* fig. 1, p. 105. Copyright 1980 by the Society for the Experimental Analysis of Behavior. Reproduced by permission.

no more demands would be made versus a "no-safety-signal" condition in which demands were continued (see Figure 15.17). In the absence of the safety signal, aggression increased as the sessions continued; when the safety signal was presented, aggression decreased. This within-session analysis helped provide additional evidence that the aggression was

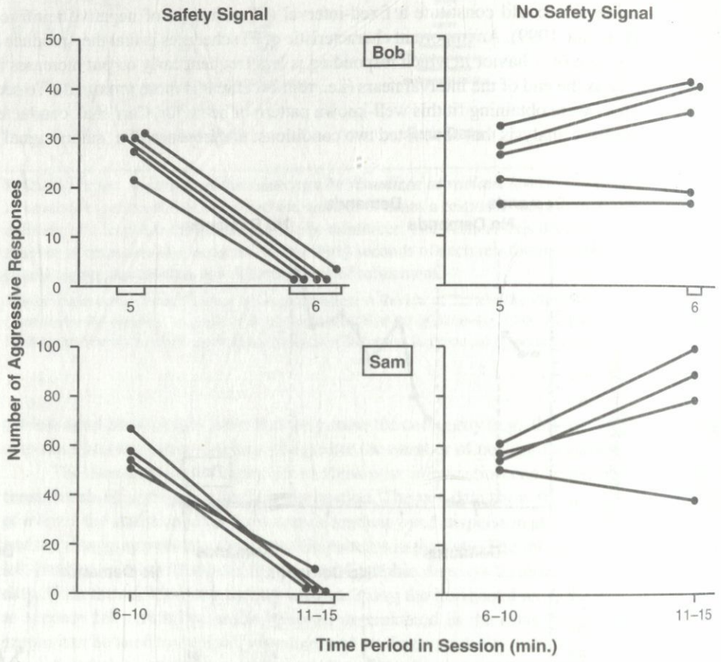

FIGURE 15.17 ***Within-session analysis conducted by Carr et al. (1980) from the between-session data displayed in Figure 15.16.*** As described by Carr et al., "Number of aggressive responses during the fifth and sixth minute of each session in the second demands condition for Bob (top half of figure) and during the sixth to tenth minute and eleventh to fifteenth minute of each session in the third demands condition for Sam (bottom half of figure). The left panels show data taken during safety signal sessions in which the experimenter signaled to each child that demands were no longer forthcoming. The signal was given to Bob at the start of the sixth minute and to Sam at the start of the eleventh minute. The right panels show data taken during no-safety-signal sessions in which neither child received any cue that demands had ended. Some of the data points have been slightly displaced horizontally to enhance presentation clarity" (p. 106).

Source: From E. G. Carr, C. D. Newsom, and J. A. Binkoff, "Escape as a Factor in the Aggressive Behavior of Two Retarded Children," *Journal of Applied Behavior Analysis,* 1980, *13,* fig. 2, p. 106. Copyright 1980 by the Society for the Experimental Analysis of Behavior. Reproduced by permission.

negatively reinforced (see Carr, 1977) and followed principles of behavior that were well established in basic research (see Sidman, 1960b).

Another example of within-session analysis is provided by Vollmer, Ringdahl, Roane, and Marcus (1997) in an analysis of adventitious positive reinforcement using non-contingent reinforcement (NCR) procedures to treat behavior problems. The top panel of Figure 15.18 shows an analysis of NCR to treat the behavior problem of an adolescent with developmental disabilities. During the second NCR phase (sessions 14 through 20),

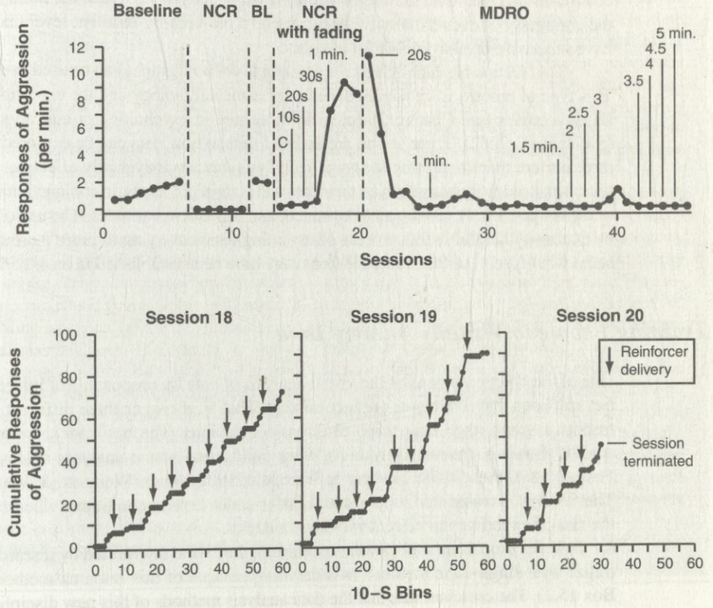

FIGURE 15.18 ***Example of within- and between-session analysis of data.*** Vollmer et al. (1997) describe this figure: "The upper panel shows aggression rates during baseline, continuous NCR, NCR with fading, and MDRO conditions (values for MDRO are shown in minutes). The fading steps are indicated by the lines and reinforcer delivery schedule values. The lower panel shows cumulative records of aggression during sessions 18 through 20. Arrows indicate when reinforcers were presented" (p. 163).

Source: From T. R. Vollmer, J. E. Ringdahl, H. S. Roane, and B. A. Marcus, "Negative Side-Effects of Noncontingent Reinforcement," *Journal of Applied Behavior Analysis,* 1997, *30,* fig. 1, p. 163. Copyright 1997 by the Society for the Experimental Analysis of Behavior. Reproduced by permission.

problem behavior began to rapidly increase in frequency. Vollmer et al. conducted within-session analyses of these sessions, which are displayed in the lower panels of the figure. Using the NCR procedure, a reinforcer was delivered on a response-independent fixed-time (FT) schedule. The pattern of responding and stimulus delivery on the NCR FT schedule showed a pattern of responding very similar to that previously discussed for FI schedules of reinforcement. That is, the authors had inadvertently (i.e., adventitiously) established a positive reinforcement contingency for the problem behavior, thus making it increase over time. Given the results of the within-session analysis, Vollmer et al. were able to arrange a momentary-differential-reinforcement-of-other (MDRO) schedule of reinforcement to counteract the adventitious positive reinforcement schedule established via the NCR procedure. In this example, the use of within-session analysis provided the ability to analyze the contiguity of events that were occurring on a moment-by-moment level that would not have otherwise been available for analysis.

This section has highlighted a few examples of how graphs can be used to explore data. This type of process is frequently done in experimental biology and the experimental analysis of behavior (see Chapter 2) but is far more rare in psychology or educational research (Smith et al., 2002). There are no prescribed limits to how data can be explored and visualized, nor are there templates for graphic analysis that are universally effective. Instead, researchers need to look carefully at their data and at multiple levels, in conjunction with a range of events potentially serving as independent and dependent variables. The use of graphs can be extremely helpful in this process of exploring data and trying to more thoroughly understand what type(s) of functional relations may have been established in an experiment.

Training People to Visually Analyze Data

One of the chief criticisms of the visual analysis of data by proponents of inferential statistics has been that judgments are inconsistent. That is, if two or three different researchers inspect a graph, they may reach different conclusions. The basis for this concern were studies showing interrater variability when conducting visual analyses of data (e.g., DeProspero & Cohen, 1979; Furlong & Wampold, 1982; Jones, Weinrott, & Vaught, 1978). The findings showed that judges could differ under certain conditions when categorizing the size of experimental effects via visual analysis.

At the time these studies were conducted, they were interpreted by researchers not familiar with single-case methods as a damning critique of this research methodology (see Box 15.2). The criticism was that the data analysis methods of this new discipline were invalid, or at least inconsistent, rendering the methodology inadequate (Fisch, 2001; Shavelson & Towne, 2002). The logic was that if researchers varied in their visual interpretation of graphs, the data analysis methods used in behavior analysis were flawed. On the surface this criticism seems to have merit. However, the logic seems inconsistent with certain aspects of the scientific process. First, the judgment of raters differed primarily under conditions of small level changes between conditions and high variability within phases. Few, if any, single-case researchers would attempt to claim functional relations under such circumstances. If they did, their claims would likely be challenged by other single-case researchers on the bases that (1) any assertion regarding a functional relation needs to be highly qualified and (2) the finding would require direct replication with additional participants before it was publishable. Hence, the criticism that visual inspection may lead to increased Type I errors (i.e., claims of an effect when none exists) ignores the larger social context of how data are evaluated in science (see Smith et al., 2002). Rarely, in single-case research, do claims of an experimental effect rest on a single ambiguous A-B analysis. Such a claim would not survive the peer review process, which serves as the primary quality control system for introducing new findings into the research literature (see Chapter 1). Given that multiple individuals (i.e., researchers, reviewers, and editors) visually analyze each

BOX 15.2 • *Applied Psychology's Disfavor with Visual Data Analysis*

Considerable attention was given to the emergence of applied behavior analysis in the 1960s and 1970s. Part of that attention was from applied psychologists, who criticized the nascent field's primary approach to data analysis—graphic displays. This approach to data analysis contradicted the predominant approach in psychology, which was the use of inferential statistics. Applied psychologists have criticized researchers using single-case designs for not following proper scientific method (Shavelson & Towne, 2002).

One of the underlying issues of this debate was two fundamentally different approaches to designing experiments and analyzing data. Using group comparison methods, subjects are randomly assigned to prescribed conditions; after the data are collected, the results are analyzed to estimate the degree to which a particular finding is due to chance. If that probability is low enough, the results are accepted as due to experimental effects and not extraneous variables. In behavior analysis, experiments are designed to reveal how behavioral processes operate, and data analysis is focused on demonstrating functional relations via repeated experimental manipulations. The designs are inductive and changed as the data require. The two approaches have distinct epistemological assumptions, meaning that they are not reconcilable (Catania, 1973). Rather, they exist as distinct approaches to conducting experiments. Criticizing one approach by using the experimental criteria of the other is logically unsatisfactory.

There is a long-standing truism among researchers that the appropriate analytical technique is not tied to a specific research methodology but to the success of researchers in answering their experimental questions. Whether single-case designs are an adequate approach to answering experimental questions should not be judged by a particular person's aesthetic sense of what a research method should be but by how useful that method is in answering the experimental questions being posed. Given that research in the behavior-analytic tradition has thrived for most of the twentieth century, from laboratory to school settings, and is only showing signs of increasing in prevalence across a range of academic disciplines, this approach to experimental design must be producing useful results. Otherwise, it would have ceased to be used as a tool for understanding human nature long ago.

Perhaps what is fundamentally at issue with applied psychology's disfavor of various methods used in single-case research is not the inadequacy of those methods but the critics' lack of familiarity with single-case methods and their underlying assumptions. To dismiss a functional relation derived from a single-case design as not adequate because inferential statistics were not used is no more substantive than behavior analysts complaining that group comparison designs hide variability or are inefficient because they require too many participants to demonstrate experimental control. Ultimately, the adequacy of any particular experimental method will rest on researchers' ability to produce findings and solve problems by using that particular method.

data set prior to publication, the odds of every person making a Type I error following this type of training is extremely small.

A second flaw in criticisms leveled against visual inspection is that they ignore the self-corrective nature of the research process. If researchers were to claim a functional relation from a limited data set with small effects, those findings would still need to be independently replicated. If researchers did make a Type I error by overinterpreting their data, they and others would be unable to replicate the original findings. The result would be that the original finding would be viewed as an anomaly and disregarded as a valid claim. It is important to remember that at the heart of the scientific method is a "quality control system" called replication (see Chapter 4).

A final limitation of critiques of visual analysis is that this approach to data analysis is demonstrably effective as a scientific technique. Ignoring for the moment the findings of experimental biology and medicine over the last two centuries (which often rely on visual analysis of data; see Latour & Woolgar, 1986), the single-case literature in educational research has repeatedly produced important findings over the last forty years. Those findings, as noted in Chapter 2, have been repeatedly replicated, extended, and refined over time and have led to important insights into behavioral processes and innovative new teaching techniques. If the basis for data analysis in single-case research were fundamentally flawed, it is unclear how a continuous stream of important discoveries could be made (and repeatedly replicated).

A more adequate solution to concerns about interrater variability using visual analysis is one very familiar to educators and behavior analysts: teach people how to analyze graphed data. In every research team, such a process is done informally (or at least without an explicit curriculum) using visual analysis methods and is a standard component of university courses in behavior analysis. Recent data have shown that untrained observers can interpret graphed data (as indexed by a consensus of experts) with approximately 55% accuracy. However, following training, those same raters improve their performances to approximately 95% (Fisher et al., 2003).

Fisher et al.'s (2003) recently validated approach to training visual analysis skills will be used as an example. First, a set of A-B graphs are developed that show a range of effects relating to trend, level, variability, immediacy of effect, and overlap. The graphs are then categorized into varying degrees of change between conditions as an index against which trainees' judgments can be indexed. Trainees are then exposed to the concepts of visual inspection (reviewed in the previous sections of this chapter). The A-B graphs are then presented individually to the trainees, who are asked to record whether an effect is present or not, and they are then provided with feedback regarding the accuracy of their performances. This relatively simple method can be used to train consistent visual inspectors in less than an hour, according to Fisher et al. (2003).

This type of formal training scenario, combined with regular reading and discussion of published research, seems to be the best approach to teaching people how to visually analyze data. Combine these approaches with the daily and weekly decisions that are required when engaging in either research or practice, and a rigorous training regimen is established. Twenty-five years after the debate over whether visual data analysis is an acceptable method, the focus has changed from whether the approach works to how to effectively teach people to use this demonstrably effective data analysis technique. In other words, with the vantage gained by hindsight, the issues regarding the use of visual data analysis are not if, but how.

Conclusion

The visual inspection of graphed data has proven to be one of the most powerful analytical techniques in science. By visualizing different aspects of a data set, the results of a study can be explored and described in a variety of ways. Such a process allows researchers to delve into various aspects and patterns of their data to gain a deeper understanding of the nature of their findings. The use of visual analysis techniques match the inductive nature of single-case designs by facilitating data exploration and the provision of readily developed and updated data analysis formats. That is, as an experiment unfolds, the data can be tracked on a daily basis for decision-making purposes and the experimental procedures adjusted accordingly.

Because of these properties, graphed data have been the central means of data analysis for the experimental analysis of behavior since its inception in the 1930s. As basic laboratory findings were extended to socially relevant issues, the data analysis techniques of basic researchers were adopted by applied investigators. Although the use of graphs instead of inferential statistics as a primary means of data analysis is contrary to traditional approaches to psychology, this technique has proven useful in the biological and medical sciences. This chapter has provided an introduction to the visual analysis of data as well as a few examples of how graphic displays can be used to better understand the nature of functional relations that result from single-case research. However, like all aspects of research methodology, there is no single "correct way" of graphing data. Instead, researchers need to adapt the techniques available to them to the analytical task at hand and use, or develop, the most appropriate means to reveal what patterns exist in their data.

16

Social Validity

If this book were strictly about behavioral processes analyzed in laboratory settings, then this final chapter would not be necessary. However, this book is not about basic research. Instead, it has explicitly focused on how to analyze behavioral processes in educational settings. Such a focus requires researchers to work directly with people, many of whom are in some type of need and require help from others. This makes applied behavior analysis different from the experimental analysis of behavior. Although both disciplines seek to understand the underlying nature of human behavior through the analysis of behavioral processes, applied research does so within a highly public context.

Typically, in educational settings, researchers are working to ameliorate some type of problematic situation, whether it be increasing the acquisition of reading skills, improving a child's articulation, or intervening to reduce aggressive behavior. In addition, because schools are public settings, other students, teachers, paraprofessionals, school administrators, and related services personnel are likely to be involved. Most directly affected are the students and their families. Such a research context, by definition, occurs within a social milieu in which multiple individuals may be affected by the behavior change intervention, even if that intervention is focused on a single student.

Because of the applied nature of educational research, additional analytical activities are necessary to evaluate the effects of interventions on a range of consumers. If researchers want to understand the impact of their intervention on a classroom, more may be needed than to graph a functional relation between the intervention and improvements in math performances. Additional information may be gained from the following questions: Did the teacher find the intervention easy or hard to implement? How did the students react to the new procedure? Did the teacher use the intervention after the study was completed? Were there any positive or negative side-effects associated with the intervention? Did the recipients of the intervention perform at levels typical of other students their age? Did the principal view the results as worth replicating in other classrooms? Such information helps investigators understand the larger context of effects their intervention may produce. Such an understanding can help in interpreting functional relations within the social contexts in which they occur and potentially can improve the effectiveness, or at least the acceptability, of educational interventions. In order to understand the social context within which

single-case research is conducted, researchers have developed a concept referred to as social validity.

Social Validity

Social validity is the estimation of the importance, effectiveness, appropriateness, and/or satisfaction various people experience in relation to a particular intervention. If, after fifteen chapters focusing on the scientific virtues of being precise, objective, and analytical, this seems somewhat subjective, that is because it is subjective. And this is the reason why social validity is so integral to understanding the effects produced in applied settings. Because educational research occurs in applied contexts, knowing how people in those settings react to an intervention is an important component in understanding the effects of a behavioral intervention.

The concept of social validity was introduced to the field of applied behavior analysis by Kazdin (1977) and Wolf (1978). However, there were antecedents to this concept in other disciplines. In the 1930s, the business sector became interested in whether the employees making products and the consumers using those products were satisfied (e.g., Rothlisberger & Dickson, 1939). In psychotherapy, psychologists and psychiatrists were interested in the expectations their clients had toward what they would experience and whether they believed they benefited from therapy (e.g., Rogers, 1942). Finally, in medicine, researchers and clinicians became interested in measuring whether patients were satisfied with the medical treatments they received (e.g., Makeover, 1950). Each of these lines of inquiry focused on establishing what people expected, experienced, and perceived were the effects of a particular endeavor.

When Kazdin (1977) and Wolf (1978) developed the concept of social validity, it was during the initial, rapid growth of applied behavior analysis. As was noted in Chapter 2, applied behavior analysis had emerged from earlier laboratory research and by the early 1970s was a well-established discipline, but one that was controversial (Kazdin, 1978). Much of the controversy was due to public concerns about researchers "controlling" the behavior of other people. These concerns, in part, were due to the effectiveness of behavioral interventions and their explicit, operationally defined procedures focusing on the consequences of responding (Goldiamond, 1976). Whatever the actual basis for concern, there was a great deal of public debate about whether "behavior modification" was ethical or desirable.

Unfortunately, when behavior analysts tried to address these public concerns, there were scant data from their own studies to buttress their arguments about the acceptability of their work. Because behavior analysts tended to focus on carefully defined behaviors that are directly relevant to a particular experimental question, there were little available data about how people "felt" about a particular experiment. This left behavior analysts in the uncomfortable situation of having very little data on which to argue in favor of their interventions, other than the changes in behavior they typically documented. This historical context set the occasion for Kazdin (1977) and Wolf (1978) to suggest measuring the social impact of behavioral interventions using the construct of social validity.

In Wolf's (1978) original description of social validity, he focused on the use of subjective judgments regarding the adequacy and desirability of behavioral interventions. He

suggested that by understanding the subjective nature of interventions, applied behavior analysts could gain a better understanding of the social importance of their work. Specifically, he outlined three general domains for subjective analysis: goals, procedures, and outcomes. Goals refer to the targets of an intervention, including individuals, settings, and specific behaviors. Procedures are the techniques used in a study to change behavior—that is, what the experimenter did to increase or decrease the probability of specific behaviors. Outcomes are the behavioral changes produced by an intervention, both direct and indirect.

These three domains were a framework to begin the study of social validity in applied behavior analysis. Such a system allowed for the systematic study of subjective data. However, these suggestions ran contrary to decades of research in behavior analysis in which only objectively defined variables were permitted into a behavioral analysis. In a sense, the empirical approach championed in behavior analysis required that these researchers incorporate subjective data into their studies in order to better understand the effects being produced. This was, and still is, somewhat ironic, but it was a necessary condition for understanding the effects of behavioral interventions in applied settings (see Box 16.1).

This historical context led Wolf (1978) to write the following:

> Earlier in our history, Watson and Skinner argued forcefully against subjective measurement because they were concerned about the inappropriate causal roles that hypothetical internal variables, subjectively reported, were playing in social science. As a result, many of us concluded that all subjective measurement was inappropriate. A new consensus seems to be developing. It seems that if we aspire to social importance, then we must develop systems that allow our consumers to provide us feedback about how our applications relate to their values, to their reinforcers. This is not a rejection of our heritage. Our use of subjective measures does not relate to internal causal variables. Instead, it is an attempt to assess the dimensions of complex reinforcers in socially acceptable and practical ways. It is an evolutionary event that is occurring as a function of the contingencies of the applied research environment; contingencies that our founders would probably say they appreciate, if we had the nerve to ask them for such subjective feedback on our behavior. (p. 213)

BOX 16.1 • *Applying Social Validity*

Given the value of understanding the social impact of an experiment, should all studies in applied behavior analysis use social validity assessments? One could argue that any study seeking to change a person's behavior in an applied context should have the social validity of that endeavor assessed. However, the appropriateness of using social validity assessments depends on what one means by "applied context." In many respects, the "applied" versus "basic" distinction is a false dichotomy. Rather than being a binary distinction, these concepts actually span a continuum from basic to applied research. That is, some researchers may use humans and even study a behavior of clear social importance (e.g., self-injurious behavior) but be focused on how basic behavioral processes produce these behaviors.

Such studies have been referred to as "bridge studies" because they fall between applied and basic research (Mace, 1994). In such instances, the use of social validity data may not meaningfully contribute to the interpretation of the experimental results. However, the rationale for such investigations is gaining a better understanding of behavioral processes impacting socially important situations so that more effective interventions can be developed (Lerman, 2003). If this is the case, and bridge studies are successful at identifying new behavior-environment mechanisms, then such findings necessarily need to be translated into practical interventions.

Because such a translation has a clear therapeutic intent, those studies would clearly need to assess the social validity of their goals, procedures, and/or outcomes. However, at this point in time, given the complex range of research occurring within the field of behavior analysis, the use of social validity assessments should probably be reserved for studies in which some type of intervention effect is being studied.

Approaches to Social Validity

Over the last twenty-five years, three approaches have been introduced to estimate social validity (see Box 16.2). Each approach focuses on a different aspect of the construct of social importance. As one might deduce, each has its strengths and limitations, and no single approach to assessing social validity can be referred as "the gold standard." Therefore, this section reviews the different approaches to social validity estimation, explains the purpose of each approach, provides examples of their use, and critiques the strategies.

Subjective Evaluation

The original conceptualization of social validity focused around what Kazdin (1977) and Wolf (1978) referred to as subjective evaluation. This approach is used to gather information regarding people's perceptions of some dimension of the goals, procedures, and/or outcomes of an experiment. The purpose is to estimate how people view some dimension of the experimental situation. Which aspect of the experimental situation is assessed is largely

BOX 16.2 • *Language Use and Social Validity*

Unlike most aspects of behavior-analytic research, which deal with explicit events, social validity presents researchers with a different type of analytical situation. Behavior analyses, by definition, focus on physical events that can be operationalized and directly measured. This rigorous approach to experimental methodology has been one of the key aspects of the success of behavior analysis over the past century. With the introduction of social validity, however, this situation was altered in some respects.

The essence of the argument for social validity, particularly in the use of subjective evaluation, is to allow verbal constructs such as "like," "acceptable," and "inappropriate" into experimental analyses. While this is entirely appropriate within the framework of studying the social validity of behavior analyses, it has led to some confusion in the language used to describe social validity.

Because social validity is a social construction, that is it is based on social conventions and poorly defined concepts, it is not a thing. Therefore, its use as a noun, as in "We need to show that this intervention has social validity," is inaccurate and misleading. Rather, social validity is an adjective that describes some characteristics of the goals, procedures, and/or outcomes of an experiment in light of some defined social context. Because of this aspect of language use (see Hineline, 1990), it is more accurate to refer to estimating or assessing social validity.

a function of the experimental question and what the researcher wants to learn about. For example, if the investigators are working on a novel applied problem, then they might gather social validity data on whether this topic is viewed as important and whether their goals for reducing these behaviors are desirable. However, if the investigators are focusing on the use of a novel intervention, then they might want to assess whether people view their new technique as acceptable. Or if the experimenters want to demonstrate the desirable qualitative outcomes of their techniques, they might have people subjectively evaluate the behavior of interest before and after intervention. Depending on what the experimenters want to learn, any or all of these approaches to subjective evaluation can be used.

Conducting Subjective Evaluations. The first step in using subjective evaluation is to identify whether the aim is to receive feedback about the goals, procedures, or outcomes or some combination of these. Once this has been decided, researchers need to identify who they will solicit information from. Schwartz and Baer (1991) identified four types of consumers: (1) direct consumers, (2) indirect consumers, (3) members of the immediate community, and (4) members of the extended community. Direct consumers are the immediate recipients of the intervention—for example, the student whose spelling is being improved or the teacher who is receiving technical assistance to improve recommended research practices. Indirect consumers are people involved in the situation being studied. These individuals can include the parents of a child who is receiving the spelling intervention or the principal in charge of supervising the teacher who is learning to use new instructional techniques. Members of the immediate community are those who are indirectly impacted by the study but who have some type of contact with the direct and indirect consumers. These individuals can include other children and their parents, other teachers in the school, or school board members. Members of the extended community are individuals who do not have direct contact with consumers but who may be interested in the potential beneficial or detrimental effects of a study. Examples could include taxpayers, legislators, media reports, content experts, or anyone else who might be interested in the researchers' efforts (see Kennedy, 2002a).

Again, which group(s) is the focus on the social validity assessment is a function of the question being posed. At one of end of the continuum, a researcher may want to understand how children and teachers react to a particular type of educational intervention. For example, researchers might compare lecture-based instruction with cooperative learning groups and ask the direct consumers which approach they prefer and why. At the other end of the continuum, researchers might be interested in polling a regionally representative group of home owners (i.e., people who pay the property taxes that finance local school systems) about whether they view school violence as an important enough issue that they would endorse cuts in other school programs (e.g., extramural sports) to increase services to reduce violence.

Once the consumer group(s) is identified, researchers need to select the assessment strategy to be used. In general, there are four approaches to collecting subjective evaluation information: (1) questionnaires, (2) forced-choice procedures, (3) structured interviews, and (4) open-ended interviews. Questionnaires are the most frequently used method (Kennedy, 1992). Questionnaires typically present a series of questions to which a particular person responds in writing or some other medium. The questions focus on some aspect

of the investigation the experimenter wants to learn about. For example, the questionnaire might ask a series of questions regarding the acceptability of the intervention procedures given certain circumstances (see Kazdin, 1980). Forced-choice procedures require informants to make choices among possible goals, procedures, and/or outcomes. For example, individuals might be asked to sort in order of acceptability a variety of interventions used to reduce behavior problems. In some instances, these choices are abstractions sampling a person's opinion; in other instances, individuals may be asked to actually choose among possible goals or procedures they will be the recipients of (see Schwartz & Baer, 1991).

Using structured interviews requires the development of a series of questions that are read to the respondent followed by the opportunity for the individual to answer. Typically, these questions have a fixed number of response options to choose from. For example, the interviewer may ask the informant a series of questions, to which they respond "yes," "no," or "maybe." Open-ended interviews pose predetermined questions to a respondent that allow the individual to provide an extended and unstructured answer. For example, an experimenter may ask questions such as "What did you think of the cooperative learning intervention?" or "How did the students respond to whole-class instruction?" The answers are then recorded using some medium for later summarization.

The final step in conducting a subjective evaluation assessment is data analysis. This step in the process is the least clearly defined in the research literature. If the data are based on discrete, quantifiable elements (e.g., Likert-type scales or yes/no responses), then the use of descriptive statistics specifying the average and variation in responses would be appropriate. For example, in response to the statement "This treatment is one that I would agree to use with my child," parental reports could be summarized as a mean of 4.3 (range, 2 to 5) on a five-point Likert-type scale (with 1 being "strongly disagree" and 5 being "strongly agree"). Another option is summarizing the number of response options that were selected for each question. Table 16.1 (page 224) shows the results of a treatment acceptability analysis for students with severe disabilities and behavior problems, with respondents being special educators (Kennedy, 1994).

If the data are qualitative in nature (i.e., verbal responses), being most likely derived from structured or open-ended questions, then a qualitative analysis of the data may be required. A review of the research literature using social validity assessments suggests that content analysis is the most frequently published form of qualitative data analysis conducted on this type of data (see Miles & Huberman, 1984). In content analysis, the responses to each question are copied onto response cards or some other medium. Members of the research team then individually read and thematically sort them into self-constructed categories. Once two or more members have done this, then the research team meets and discusses their categorization schemes. Then the group revises the categorization scheme, as appropriate, and agrees as a group on how to sort each response item to a particular question. A similar process is undertaken for each question asked of respondents. These data can be summarized in at least two ways. A study by Cox and Kennedy (2003) will be used to illustrate both types of data summarization. The data reflect parental responses to open-ended questions about the hospitalization and subsequent recovery of their child who had a multiple disability. Table 16.2 (page 225) shows a content analysis that specified general response categories to each question and summarized the percentage of answers from respondents included in each category. Another technique for presenting these same data

TABLE 16.1 ***Results from the Treatment Evaluation Inventory Assessment***

1. How acceptable do you find this treatment to be for the student's problem behavior?				1	2	5
not at all acceptable			moderately acceptable			very acceptable
2. How willing would you be to carry out this procedure yourself if you had to change the student's problem behavior?					3	5
not at all willing			moderately willing			very willing
3. How cruel or unfair do you find the treatment?						8
very cruel			moderately cruel			not cruel at all
4. To what extent does this procedure treat the student humanely?						8
does not treat humanely at all			treats them moderately humanely			treats them very humanely
5. How much do you like the procedures used in this treatment?					1	7
do not like them at all			moderately like them			like them very much
6. How likely is this treatment to make permanent improvements in the student?					5	3
unlikely			moderately			very likely
7. To what extent are undesirable side effects likely to result from this treatment?				1	6	1
many undesirable side effects likely			some undesirable side effects likely			no undesirable side effects likely
8. Overall, what is your general reaction to this form of treatment?						8
very negative			ambivalent			very positive

Note: $N = 8$

Source: From C. H. Kennedy, "Manipulating Antecedent Conditions to Alter the Stimulus Control of Problem Behavior," *Journal of Applied Behavior Analysis,* 1994, *27,* 161–170. Copyright 1994 by the Society for the Experimental Analysis of Behavior. Reproduced by permission.

would be to present subcategories of answers to questions and exemplars of the actual responses that were received. Table 16.3 (page 226) shows the data from question 1 of Table 16.2 in this more detailed format. As with the visual analysis of data, what is most important in analyzing this type of data is that the process and presentation reveal the character of the data that are obtained in as clear and concise a manner as possible.

TABLE 16.2 ***Parent Responses to Open-Ended Questions***

What did your child's school do that was helpful?
- Nothing (14.3%)
- Offered support (85.7%)

What could the school have done to be more helpful?
- More communication and coordination (35%)
- Nothing (50%)
- School not responsible (15%)

What did the hospital do that was helpful?
- Nothing (20%)
- Provided health services to child (24%)
- Provided support services to parents (8%)
- Supportive staff (48%)

What could the hospital have done to be more helpful?
- Improve support for parents (13.3%)
- Improved care (26.7%)
- Increase continuity and collaboration (13.3%)
- More staff education (26.7%)
- Nothing (20%)

How successful was home-school-hospital communication?
- Communication not an issue (10.5%)
- No communication among entities (47.4%)
- Parent assumed lead (26.3%)
- Satisfactory (15.8%)

What could have been done to improve home-school-hospital communication?
- Improved communication (26%)
- Not sure what to recommend (8.7%)
- School and hospital separate issues (65.3%)

What was the effect of the hospitalization on your child's education?
- Improved performance (15.8%)
- No or little effect (42.1%)
- Small negative effect (21%)
- Substantial negative effect (21.1%)

Source: From J. A. Cox and C. H. Kennedy, "Transitions between School and Hospital for Students with Multiple Disabilities: A Survey of Causes, Educational Continuity, and Parental Perceptions," *Research and Practice for People with Severe Disabilities* (formerly *JASH*), 2003, *28*, 1–6. Copyright 2003 by TASH. Reproduced by permission.

TABLE 16.3 ***Examples of Responses to Types of Support Provided in Question 1 of Table 16.2***

Provided general support
"PT and teachers called and came to hospital."
"Teacher called and brought homework."
Offered general support
"Called once they heard child was in the hospital."
"School offered homebound services but we declined."
Homebound services
"Homebound teacher was already part of IEP. The teacher just came to the hospital."
"Homebound teacher provided laptop at the hospital."

Source: From J. A. Cox and C. H. Kennedy, "Transitions between School and Hospital for Students with Multiple Disabilities: A Survey of Causes, Educational Continuity, and Parental Perceptions," *Research and Practice for People with Severe Disabilities* (formerly *JASH*), 2003, *28,* 1–6. Copyright 2003 by TASH. Reproduced by permission.

Strengths and Limitations. There are several strengths and limitations to using subjective evaluation as a technique for estimating social validity. An important strength, and one championed by Kazdin (1977) and Wolf (1978), is that subjective evaluation allows qualitative information to be added to data gathered through an experimental analysis of behavior. A second strength of subjective evaluation is that its use broadens the range of dependent variables used in a study. Both of these strengths are based on including people's perceptions and opinions into the interpretation of what was done and what resulted from an experiment designed to have beneficial outcomes to particular individuals. An important limitation of subjective evaluation is that the questions posed are often biased toward receiving a positive outcome. That is, researchers often develop questions or present them in ways in which the situation predisposes respondents toward favorable answers. A second limitation of this technique is that people's perceptions of situations may not meaningfully reflect changes in a participant's behavior. A third limitation is that most instruments developed for subjective evaluation studies have unknown psychometric properties. That is, the reliability and validity of the instruments are typically unknown (see Sax, 1996). Overall, the use of subjective evaluation can be an important tool if a particular experimental question is developed in which this information would be useful.

Normative Comparison

A second approach to estimating social validity was developed largely in response to concerns about the highly qualitative nature of the data derived from subjective evaluations. The approach, referred to as normative comparison, was outlined by Van Houten (1979) shortly after the Kazdin (1977) and Wolf (1978) papers were published. In normative comparison, a particular behavior(s) engaged by a participant is compared to some reference sample of individuals. Typically, the reference group is chosen because it can serve as an exemplar of desirable levels or topographies of the behavior(s) of interest. The focus is to

reference the behavior change goals and outcomes for the participants in a study against some normative group whose behavior is considered typical or desirable.

An early example of the use of normative comparisons is provided by Walker and Hops (1976). These authors focused on improving the behavior of students considered to have conduct problems in general education classrooms. As a reference regarding the treatment goals and as an index of intervention outcomes, Walker and Hops sampled levels of appropriate and inappropriate behaviors among classroom peers identified by teachers as behaving appropriately. The children with conduct problems were then given an intervention in a separate setting until their behavior approximated that of their peers in the general education classrooms (see Figure 16.1). The children with conduct problems were then reintroduced into the general education classroom and maintained similar behavioral levels to those of their peers. Such a demonstration shows quantitatively that the appropriate behavior of the students who originally had conduct problems was similar to that of their peers without behavior problems following intervention.

Conducting Normative Comparisons. To conduct normative comparisons, researchers need to begin by identifying the behaviors of interest in the group of students whose behavior will receive intervention. Then, a decision is made whether to base the goals, outcomes, or both aspects of intervention on a normative sample. If goals alone are chosen for comparison, then researchers will have an intervention target to reach but no information

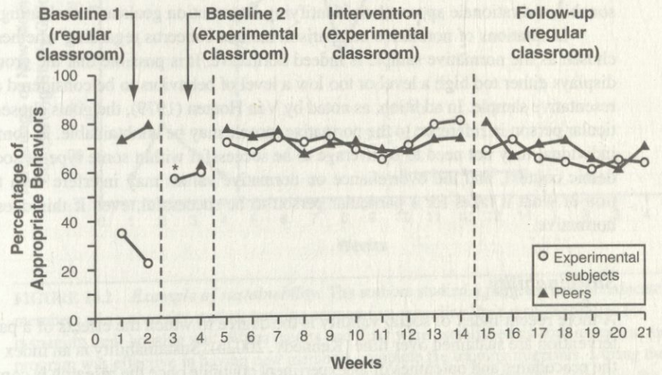

FIGURE 16.1 ***Example of the use of normative comparison.*** The data show the percentage of appropriate behaviors along the vertical axis and weeks of school along the horizontal axis. The variables were the behavior of children with conduct problems and the behavior of peers in the classroom deemed to behave appropriately by teachers.

Source: From H. M. Walker and H. Hops, "Use of Normative Peer Data as a Standard for Evaluating Classroom Treatment Effects," *Journal of Applied Behavior Analysis,* 1976, *9,* fig. 1, p. 164. Copyright 1976 by the Society for the Experimental Analysis of Behavior. Reproduced by permission.

on normative outcomes. If only outcomes are the focus of comparison, then researchers will have a metric of the normative outcomes of their intervention effects but no quantitative goals to guide their efforts. For these reasons, it is probably best to use normative comparison both to identify the goals to be achieved and then further demonstrate that after those goals have been met, they are still normative values.

Once these decisions are made, researchers will need to identify an appropriate reference group to sample data from. Often, the choice is made to sample from a group of individuals who already show desirable levels of the behavior of interest. Then, levels of those behaviors need to be measured in the environments in which they would naturally occur. These data, when summarized, provide the basis for comparing the goals and outcomes of the intervention that will then be experimentally analyzed.

Strengths and Limitations. Benefits of this approach to social validity assessment are primarily that there is a clear and defendable basis for setting the goals of an intervention. It is likely that in educational research when a child is identified as deviant in some aspect, the subsequent expectations for change in the child's behavior are higher than for their peers, who are not viewed as needing intervention. Collecting normative data may help with this concern. An additional strength of this approach is that it provides a reference, after interventions have been implemented, regarding whether the person's behavior is within the range of the peers, whose behavior has been deemed acceptable. A final strength of this approach is the logical foundation for basing treatment gains, which gives this strategy a high degree of face validity. That is, at face value most experts would say this is a reasonable and rationale approach to identifying intervention goals and evaluating outcomes.

Limitations of normative comparison include concerns regarding whether the group chosen as the normative sample is indeed normative. It is possible that the group sampled displays either too high a level or too low a level of behaviors to be considered a truly representative sample. In addition, as noted by Van Houten (1979), the goals chosen for a particular person in reference to the normative sample may be unobtainable. In some cases, an individual may not need to be average to be successful within some type of social or academic context, and the overreliance on normative values may interfere with the evaluation of what it takes for a particular person to be successful, even if this does not mean normative.

Sustainability

A more recent index of social validity is the degree to which the effects of a particular intervention are sustained over time (Kennedy, 2002b). Sustainability is an index of whether the procedures and outcomes of an experiment continue once the research is completed and the researchers are no longer involved. If the consumers present in a particular context consider the procedures being used and the behavioral changes that resulted from them as desirable, then they are likely to work at maintaining the program. The use of sustainability as an index of social validity comes from the observation that "if an intervention is socially invalid, it can hardly be effective, even if it changes its target behaviors thoroughly and with an otherwise excellent cost-benefit ratio; social validity is not sufficient for effectiveness but is necessary to effectiveness" (p. 323; Baer, Wolf, & Risley, 1987). Therefore, if an intervention is sustained over time, it must have some qualities that are consistent with what is meant by social validity.

An example of sustainability is provided by Altus, Welsh, Miller, and Merrill (1993). Altus et al. studied a program used to educate members of a university student housing cooperative regarding their responsibilities for managing their housing unit. As is shown in Figure 16.2, when credits and fines were established as contingencies, the program was effective in having new members complete the training materials. During the first fourteen weeks of the intervention, researchers managed the contingencies. After this, the research team turned over the program to the housing cooperative members. Nine years later, when the researcher again sampled the behavior of new members regarding training activities, the program had maintained at levels similar to the earlier analysis. Such a result suggests that this intervention was useful and acceptable to those using it.

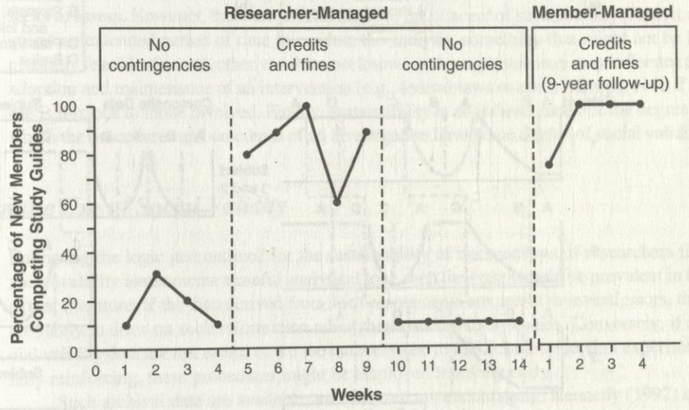

FIGURE 16.2 ***Example of sustainability.*** The authors studied a program used to educate members of a university student housing cooperative regarding their responsibilities for managing their housing unit. When credits and fines were established as contingencies, the program was effective in having new members complete the training materials. During the first fourteen weeks of the intervention, researchers managed the contingencies. After this, the research team turned over the program to the housing cooperative members. Nine years later, when the researcher again sampled the behavior of new members regarding training activities, the program had maintained at levels similar to the earlier analysis.

Source: From D. E. Altus, T. M. Welsh, L. K. Miller, and M. H. Merrill, "Efficacy and Maintenance of an Education Program for a Consumer Cooperative," *Journal of Applied Behavior Analysis,* 1993, *26,* fig. 1, p. 404. Copyright 1993 by the Society for the Experimental Analysis of Behavior. Reproduced by permission.

The study of maintenance was first proposed by Rusch and Kazdin (1981) within the context of experimentally analyzing the factors involved in treatment success over time. These authors suggested that the use of withdrawal designs over extended time periods can be used to analyze the maintenance of interventions and their effects on behavior. Figure 16.3 shows several hypothetical examples of withdrawal designs proposed to experimen-

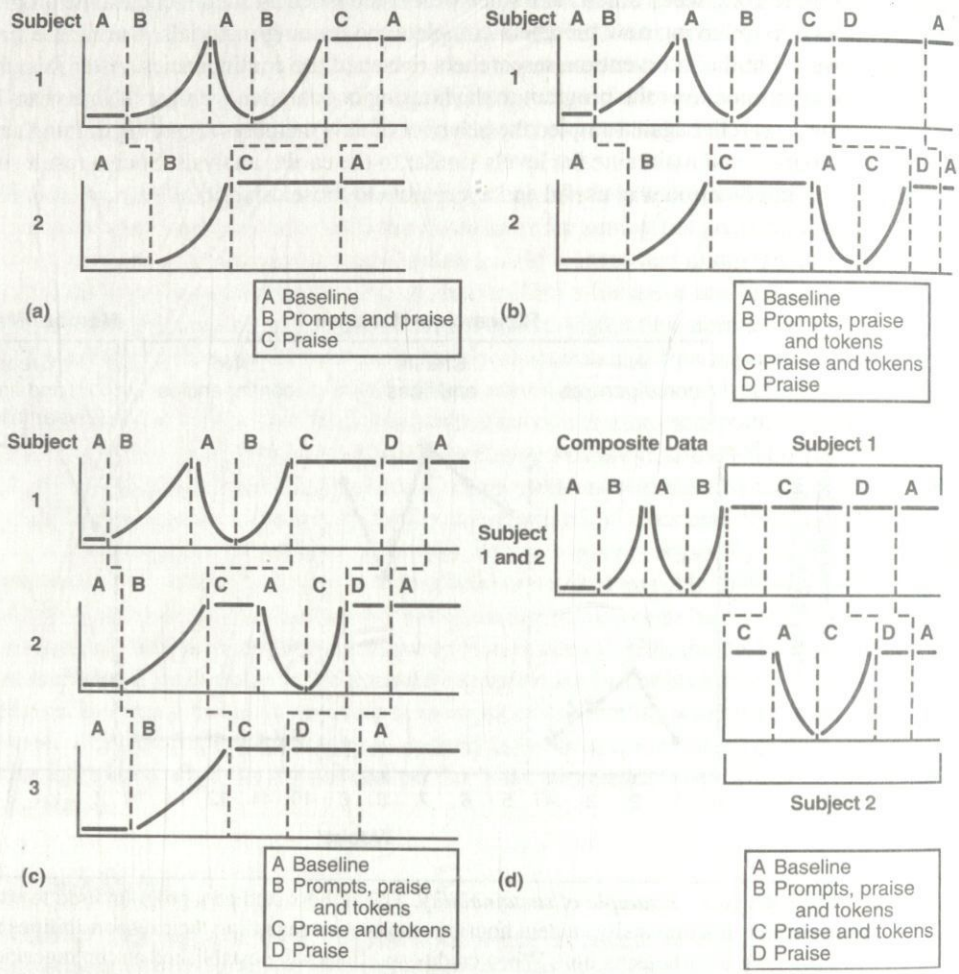

FIGURE 16.3 ***Hypothetical examples of withdrawal designs.*** A systematic withdrawal of a two-component treatment across two subjects is represented in the upper left graph (a). Withdrawal of a three-component treatment across two subjects is indicated in the upper right graph (b). A systematic withdrawal of a three-component treatment across three subjects is shown in the lower left portion of the figure (c). Finally, withdrawal of a three-component treatment across two subjects within an A-B-A-B reversal is depicted in the lower right portion (d).

Source: From F. R. Rusch and A. E. Kazdin, "Toward a Methodology of Withdrawal Designs for the Assessment of Response Maintenance," *Journal of Applied Behavior Analysis,* 1981, *14,* fig. 2, p. 137. Copyright 1981 by the Society for the Experimental Analysis of Behavior. Reproduced by permission.

tally study maintenance. If we put Rusch and Kazdin's suggestions within the framework put forward in Chapter 5 regarding experimental questions, such an analytical scheme sets the occasion for conducting component and parametric analyses (Kennedy, 2002). Component analyses allow investigators to remove one or more aspects of an independent variable, assess its effects on behavior, and then return to the previous conditions. Such analyses could be used to identify those components of an intervention that are necessary for it to be sustained by consumers. Conducting parametric analyses would permit a cost-benefit analysis of differing levels of intervention, the effects on behavior, and whether consumers choose to sustain the intervention. Such experiments could produce important information, not only of how interventions are sustained over time but why they are, or are not, sustained over time.

Strengths and Limitations. The primary strength of sustainability as an index of social validity is its face validity. If a group of consumers maintain an intervention over extended periods of time, there must be something about the intervention and its effects that are reinforcing to those consumers. This, perhaps, is an empirical test of what subjective evaluation seeks to assess. However, there are several limits to the concept of sustainability. First, it requires an extended period of time to conduct the analysis, something that might not be logistically feasible. Second, other variables not known to the experimenters may influence the adoption and maintenance of an intervention (e.g., federal laws or court rulings), even if its use is noxious to those involved. Finally, sustainability is an indirect index of the degree to which the procedures and outcomes of an investigation have some degree of social validity.

Trends in the Use of Social Validity

Following the logic just outlined for the sustainability of interventions, if researchers find social validity assessments a useful analytical tool, then their use should be prevalent in the extant literature. If the data derived from such assessments are useful to investigators, they are likely to draw on such information when they conduct their studies. Conversely, if social validity data are not useful or are too cumbersome to make their collection experimentally reinforcing, these procedures might be employed less frequently.

Such archival data are available, and they are not encouraging. Kennedy (1992) and Carr, Austin, Britton, Kellum, and Bailey (1999) have documented the degree to which social validity assessments are incorporated into applied behavior-analytic research. The results of the Carr et al. analysis are presented in Figure 16.4 (page 232). Arrayed along the vertical axis is the percentage of research articles published in the *Journal of Applied Behavior Analysis* that report data related to social validity. The horizontal axis represents the year of publication. The top panel shows data for outcomes, the center panel for procedures, and the bottom panel for either or both types of social validity assessment. Not surprisingly, few studies used social validity assessments prior to the Kazdin (1977) and Wolf (1978) articles, but an increase occurred following their publication. A gradual decline in the use of social validity assessments occurred during the 1980s, with their use stabilizing at approximately 20% in the subsequent decade. Kennedy noted a similar pattern of reporting social validity assessments for papers in a second journal, *Behavior Modification*. In addition, Kennedy noted that less than 5% of published studies used normative comparison methods.

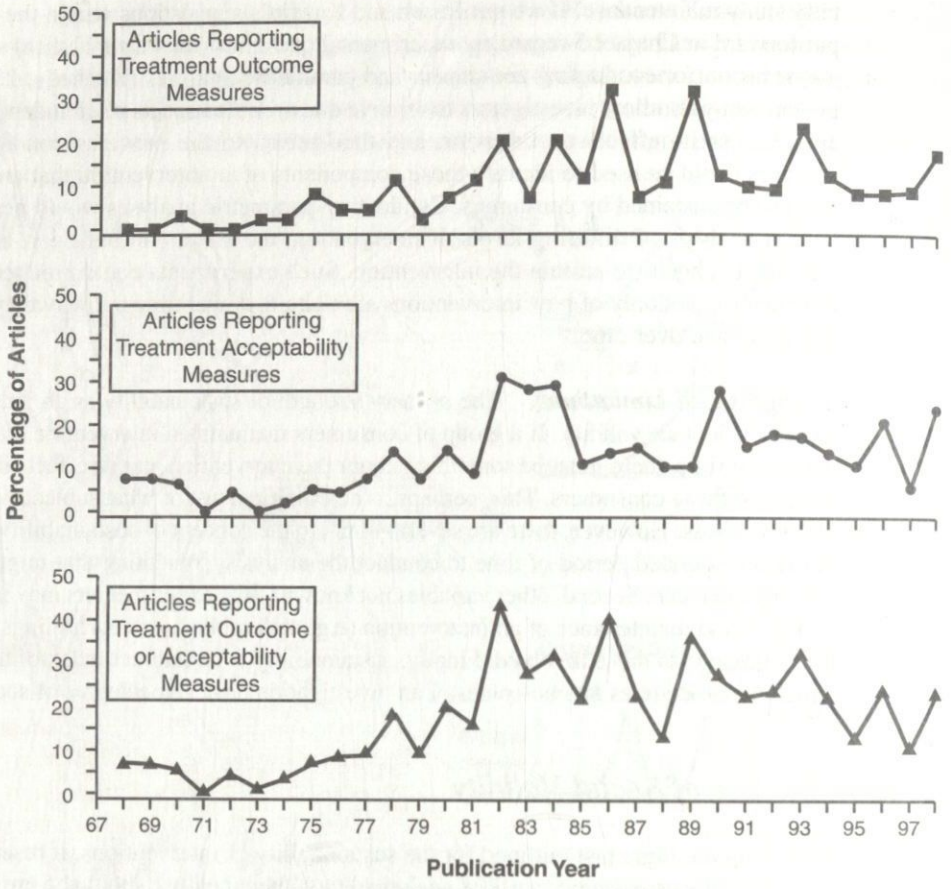

FIGURE 16.4 ***Example of use of social validity assessments in applied behavior analysis.*** Arrayed along the *y*-axis is the percentage of research articles published in the *Journal of Applied Behavior Analysis* reporting data relating to social validity. The *x*-axis represents the year of publication. The top panel shows data for outcomes, the center panel for procedures, and the bottom panel for either or both types of social validity assessment.

Source: From J. E. Carr, J. L. Austin, L. N. Britton, K. K. Kellum, and J. S. Bailey, "An Assessment of Social Validity Trends in Applied Behavior Analysis," *Behavioral Interventions,* 1999, *14,* fig. 1, p. 227. Copyright 1999 by John Wiley and Sons. Reproduced by permission.

These data seem to suggest that researchers are only occasionally incorporating social validity assessments into their experimental methods. Two possible explanations have been put forward for this pattern. First, as noted in Box 16.1, not all studies in applied behavior analysis are directly focused on improving a person's quality of life via intervention. Many experiments, referred to as bridge studies, focus on analyzing the underlying mechanisms of behavior and fall between basic and applied research (Mace, 1994). This could account for some percentage of applied studies not using social validity assessments. How-

ever, a perusal of the literature over the past several decades reveals that a number of studies that are explicitly focused on interventions to improve behavior do not incorporate social validity assessments into their data collection protocols.

A second possible explanation for the underutilization of social validity assessments is that the procedures may not be yielding data that researchers find useful in interpreting the outcomes of their research. It might be that the nature of the data collected in many social validity assessments is not as beneficial as the cost of collecting the data. Part of this might relate to the rigor used in gathering social validity data. For example, Fawcett (1991) has suggested increasing the psychometric rigor of social validity assessments so that the reliability and validity of assessments are established prior to their being used in a research study. Schwartz and Baer (1991) have argued that the incorporation of consumer input should be central to social validity analyses, rather than something that is conducted as a secondary experimental effort. Finally, Hawkins (1991) has suggested that social validity should not only be assessed but should be subjected to experimental analyses that yield functional relations relating to the social importance of research findings rather than descriptive data. Each of these suggestions, if incorporated into studies of social validity, would improve the quality of the information obtained and potentially increase researchers' efforts to incorporate this important construct into their applied studies.

Conclusion

An interest in studying social validity emerged during the 1970s in applied behavior analysis. The impetus for this development was public concern that behavior-analytic methods might be effective but might not be socially acceptable. In order to better understand whether these concerns were warranted, Kazdin (1977) and Wolf (1978) introduced the construct of social validity. These methods have come to incorporate subjective evaluation, normative comparison, and sustainability as procedures for collecting social validity information.

If used as intended, social validity assessments allow the study of how behavioral interventions impact a range of individuals directly and indirectly involved in the investigation. Subjective evaluations allow the collection of data relating to personal perceptions of the appropriateness of the goals, procedures, and outcomes of a study. Normative comparisons provide a method for indexing the goals and outcomes of a study against some social standard or reference. Finally, sustainability permits an assessment of the extent to which the procedures and outcomes of an experiment are adopted and maintained by a group of individuals. Each approach to social validity assessment can provide important information about the effects of an intervention above and beyond what is typically reported in terms of dependent variables in applied research.