CHAPTER 2

Exploratory Data Analysis

JOHN T. BEHRENS AND CHONG-HO YU

Quantitative research methods in the twentieth century were marked by the explosive growth of small-sample statistics and the expansion of breadth and complexity of models for statistical hypothesis testing. The close of the century, however, was marked primarily with a frustration over the limitations of common statistical methods and frustration with their inappropriate or ineffective use (Cohen, 1994). Responding to the confusion that emerged in the psychological community, the American Psychological Association convened a task force on statistical inference that published a report (Wilkinson & Task Force, 1999) recommending best practices in the area of method, results, and discussion. Among the recommendations in the area of conducting and reporting results, the task force suggested researchers undertake a cluster of activities to supplement common statistical test procedures with the aim of developing a detailed knowledge of the data, an intimacy with the many layers of patterns that occur, and a knowledge of the implications of these patterns for subsequent testing.

Unbeknownst to many psychological researchers, the general goals recommended by the task force, as well as specific graphical techniques and conceptual frameworks mentioned in the report, are rooted in the quantitative tradition of exploratory data analysis (EDA). Exploratory data analysis is a well-established tradition based primarily on the philosophical and methodological work of John Tukey. Although Tukey is clearly recognized as the father of EDA in statistical circles, most psychologists are familiar only with small aspects of his work, such as that in the area of multiple-comparison procedures. Although Tukey worked in mathematical statistics throughout his life, the middle of his career brought dissatisfaction with the value of many statistical tools for understanding the complexities of real-world data. Moreover, Tukey fought what he perceived as an imbalance in efforts aimed at understanding data from a hypothesis-testing or confirmatory data analysis (CDA) mode while neglecting techniques that would aid in understanding of data more broadly. To fill this gap and promote service to the scientific community, as well as balance to the statistical community, Tukey developed and implemented the processes and philosophy of exploratory data analysis to be discussed shortly. To introduce the reader to this tradition, the chapter is divided into four parts. First, the background, rationale, and philosophy of EDA are presented. Second, a brief tour of the EDA toolbox is presented. The third section discusses computer software and future directions for EDA. The chapter ends with a summary and conclusion.

This work was completed while Dr. Behrens was on leave from Arizona State University, Division of Psychology in Education. He would like to thank the staff and administration of the department for their support.

HISTORY, RATIONALE, AND PHILOSOPHY OF EDA

John Tukey and the Origins of EDA

The tradition of EDA was begun and nurtured by John Tukey and his students through his many years at Princeton University and Bell Laboratories. As a young academic, Tukey was a prodigious author and formidable mathematical statistician. He received his PhD in mathematics from Princeton at the age of 25 and at 35 reached the rank of full professor at the same institution (Brillinger, Fernholz, & Morgenthaler, 1997). A sense of Tukey's breadth and impact can be gleaned from examination of the eight volumes of his collected works. Volumes 1 and 2 (Brillinger, 1984, 1985) highlight his contributions to time-series analysis (especially through spectral decomposition). Volumes 3 (Jones, 1986a) and 4 (Jones, 1986b) address *Philosophy and Principles of Data Analysis,* and volume 5 is devoted to graphics (Cleveland, 1988). Volume 6 (Mallows, 1990) covers miscellaneous mathematical statistics, whereas volumes 7 (Cox, 1992) and 8 (Braun, 1994) cover factorial and analysis of variance (ANOVA) and multiple comparisons, respectively. More may appear at a future date because Tukey remained an active researcher and writer until his death in July of 2000.

In addition to the many papers in his collected works, Tukey authored and coauthored numerous books. In the EDA literature his central work is *Exploratory Data Analysis* (Tukey, 1977), whereas *Data Analysis and Regression: A Second Course* (Mosteller & Tukey, 1977) is equally compelling. Three volumes edited by Hoaglin, Mosteller, and Tukey (1983, 1985, 1991) complete the foundational corpus of EDA. Brillinger, Fernholz, and Morgenthaler (1997) provide a Festschrift for Tukey based on writings of his students at the time of his 80th birthday in 1995.

As Tukey became increasingly involved in the application of statistics to solve real-world problems, he developed his own tradition of values and themes that emphasized flexibility, exploration, and a deep connection to scientific goals and methods. He referred to his work as *data analysis* rather than statistics because he believed the appropriate scientific work associated with data was often much broader than the work that was followed by the traditional statistical community. Tukey did not seek to supplant statistics; rather, he sought to supplement traditional statistics by restoring balance to what he considered an extreme emphasis on hypothesis testing at the expense of the use of a broader set of tools and conceptualizations.

Although most psychologists are unaware of the specific proposals Tukey made for EDA (but see Tukey, 1969; Behrens, 1997a, 2000), the work of EDA is slowly filtering into daily practice through software packages and through the impact of a generation of statisticians who have been trained under the influence of Tukey and his students. For example, although highly graphical data analysis was rare in the 1970s, the current reliance on computer display screens has led statistical graphics to hold a central role in data analysis as recommended in common software packages (e.g., Norusis, 2000; Wilkinson, 2001). Tukey's work inspired entire paradigms of statistical methods, including *regression graphics* (Cook & Weisberg, 1994), *robustness studies* (e.g. Wilcox, 1997, 2001), and computer graphics for statistical use (Scott, 1992; Wilkinson, 1999).

Despite these advances in the application of EDA-like technique, statistical training remains largely focused on specific techniques with less than optimal emphasis on philosophical and heuristic foundations (cf. Aiken, West, Sechrest, & Reno, 1990). To prepare the reader for appropriate application of the techniques discussed later, we first turn to a treatment of the logical and philosophical foundations of EDA.

Rationale and General Tenets

It's all about the World

Exploratory data analysis is an approach to learning from data (Tukey & Wilk, 1966/1986) aimed at understanding the world and aiding the scientific process. Although these may not be "fighting words" among psychologists and psychological methodologists, they were for Tukey as he first raised his concerns with the statistical community.

Tukey's emphasis on the scientific context of data analysis leads to a view of data analysis as a scientific endeavor using the tools of mathematics, rather than a mathematical endeavor that may have value for some real-world applications. A number of changes to standard statistical practice are implied in this view. First, the statistician cannot serve as an aloof high priest who swoops down to sanctify a set of procedures and decisions (Salsburg, 1985). Data analysts and scientists (not mutually exclusive categories) must work interactively in a cyclical process of *pattern extraction* (mathematics) and *pattern interpretation* (science). Neither can function without the other. This view has implications for the organization of academic departments and organization of graduate training.

Second, because the effort of data analysis is to understand data in all circumstances, the role of probability models relates primarily to confirmatory aspects of the scientific process. This leaves a wide swath of the scientific processes

for which researchers are left to use nonprobabilistic methods such as statistical graphics. This emphasis is based on the fact that in many stages of research the working questions are not probabilistic. When probabilistic methods are applied, there are layers of assumptions which themselves need to be assessed in nonprobabilistic ways to avoid an unending loop of assumptions. Contrasting classical statistics with data analysis, Tukey (1972/1986a) wrote, "I shall stick to 'data analysis' in part to indicate that we can take probability seriously, or leave it alone, as may from time to time be appropriate or necessary" (p. 755).

The probabilistic approaches taken in most confirmatory work may lead to different practices than the nonprobabilistic approaches that are more common to working in the exploratory mode. For example, a number of researchers have looked at the issue of *deleting outliers* from reaction time data. From a probabilistic view this problem is addressed by simulating distributions of numbers that approximate the shape commonly found in reaction time data. Next, extreme values are omitted using various rules, and observation is made of the impact of such adjustments on long-run decision making in the Monte Carlo setting. As one would expect in such simulations, estimates are often biased, leading the researcher to conclude that deleting outliers is inappropriate.

Working from the exploratory point of view, the data analyst would bring to bear the scientific knowledge he or she has about the empirical generating process of the data—for example, the psychological process of comparison. Using this as the primary guideline, outliers are considered observations whose value is such that it is likely they are the result of nonexperimental attributes. If common sense and previous data and theory suggest the reaction times should be less than 3 s, extreme values such as 6 or 10 s are most likely the result of other generating processes such as distraction or lack of compliance. If this is the case, then a failure to exclude extreme values is itself a form of biasing and is deemed inappropriate.

These divergent conclusions arise from approaching the problem with different assumptions. From a probabilistic view, the question is likely to be formulated as *If the underlying process has a distribution of X and I exclude data from it, is the result biased in the long run?* On the other hand, the exploratory question addressed is *Given that I do not know the underlying distribution is X, what do I know about the processes that may help me decide if extreme values are from the same process as the rest of the data?* In this way EDA emphasizes the importance of bringing relevant scientific knowledge to bear in the data-analytic situation rather than depending solely on probabilistic conceptualizations of the phenomenon under study. As with all techniques, EDA does not reject probabilistic approaches, but rather considers them within a larger context of the many tools and ideas that bear on scientific work.

A central idea in the EDA corpus is the goal of developing a detailed and accurate mental model that provides a sense of intimacy with the nuances of the data. Such an experience assists both in constructing scientific hypotheses and building mathematical representations that allow more formal confirmatory methods to be used later. All of these issues argue against the routine use of statistical procedures and the lockstep application of decision rules. Tukey (1969) saw the commonplace use of statistical tests to prove "truth" as a social process of "sanctification." In this approach, the codifying of specific actions to be undertaken on all data obscures the individual nature of the data, removes the analyst from the details of the world held in the data, and impedes the development of intimacy and the construction of a detailed mental model of the world.

Consider the story of the researcher who sought to analyze the ratings of university faculty of various ethnicities according to the ethnicities of the students providing the ratings. To make sure the job was done properly, the researcher contacted the statistical consulting center and spoke with the director. After a brief description of the problem, it was clear to the consultant that this situation required a series of two-way ANOVAs of rating value across levels of teacher ethnicity and student ethnicity. A graduate student was assigned to compute the ANOVAs using commonly available statistical software, and both the researcher and consultant were quite pleased with the resulting list of p values and binary significance decisions.

In this true story, as in many, it was unfortunate that the discussion focused primarily on choosing a statistical model (ANOVA) to fit the design, rather than being a balanced discussion of the need for a broader understanding of the data. When the researcher later sought help in answering a reviewer's question, a simple calculation of cell frequencies revealed that the scarcity of students and faculty in many minority groups led to a situation in which almost half of the cells in the analysis were empty. In addition, many cells that were not empty had remarkably few data points to estimate their means. In many ways the original conclusions from the analysis were more incorrect than correct.

The error in this situation occurred because a series of unspoken assumptions propagated throughout the data analysis processes. Both the researcher and the director were concerned primarily with the testing of hypotheses rather than with developing a rich understanding of the data. Because of this, the statistician failed to consider some basic assumptions (such as the availability of data) and focused too much

on an abstract conceptualization of the design. It was such lockstep application of general rules (factorial means between groups implies ANOVA) that Tukey sought to discourage.

Unfortunately, it is not unusual that authors of papers published in refereed journals neglect detailed examination of data. This leads to inferior mental models of the phenomenon and impedes the necessary assessment of whether the data conform to parametric test assumptions. For example, after reviewing more than 400 large data sets, Micceri (1989) found that the great majority of data collected in behavioral sciences do not follow univariate normal distributions. Breckler (1990) reviewed 72 articles in personality and social psychology journals and found that only 19% acknowledged the assumption of multivariate normality, and fewer than 10% considered whether this assumption had been met. Reviewing articles in 17 journals, Keselman et al. (1998) found that researchers rarely verified that statistical assumptions were satisfied and typically used analyses that were nonrobust to assumption violations. These authors noted that many of these types of problems could be detected by the application of techniques from the EDA tradition.

Detective Work and the Continuum of Method

In working to learn about the world, EDA holds several complementary goals: to find the unexpected, to avoid being fooled, and to develop rich descriptions. The primary analogy used by Tukey to communicate these goals is that of the data analyst as detective. Detective work is held up as a valuable analogy because the process is essentially exploratory and interactive; it involves an iterative process of generating hypotheses and looking for fit between the facts and the tentative theory or theories; and the process is messy and replicable only at the heuristic level. Detective work also provides a solid analogy for EDA because it is essentially a bottom-up process of hypothesis formulation and data collection.

Tukey (e.g., 1972/1986a, 1973/1986b) did not consider methodology as a bifurcation between EDA and CDA, but considered quantitative methods to be applied in stages of exploratory, rough confirmatory, and confirmatory data analyses. In this view EDA was aimed at the initial goals of hypothesis generation and pattern detection, following the detective analogy. Rough CDA is sometimes equated with null-hypothesis significance testing or the use of estimation procedures such as confidence intervals with the aim to answer the question, "With what accuracy are the appearances already found to be believed?" (Tukey, 1973/1986b, p. 794). With regard to strict confirmatory analysis Tukey notes, "When the results of the second stage is marginal, we need a third stage. . . . It is at this stage . . . that we require our best statistical techniques" (Tukey, 1973/1986b, p. 795). As a researcher moves through these stages, he or she moves from hypothesis generation to hypothesis testing and from pattern identification to pattern confirmation.

Whereas CDA is more ambitious in developing probabilistic assessments of theoretical claims, the flexibility and bottom-up nature of EDA allows for broader application. In many cases an appropriate model of parametric statistics may be unavailable for full CDA, while the simpler techniques of EDA may be of use. Under such a circumstance the maxim should be followed that "[f]ar better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise" (Tukey, 1962/1986c, p. 407).

Summarization and the Loss of Information

Behrens and Smith (1996) characterize the nature of data analysis as being oriented toward generating summary and extracting gist. When faced with numerous pieces of data, the goal of the analyst is to construct a terse yet rich mathematical description of the data. This is analogous to the summarization process that occurs in natural language processing. After reading a long book, one does not recall every individual word, but rather remembers major themes and prototypical events. In a similar way, the data analyst and research consumer want to come away with a useable and parsimonious description rather than a long list of data. An essential concept associated with summarization is that every summary represents a loss of information. When some aspects of data are brought to the foreground, other aspects are sent to the background.

Algebra Lies, So You Need Graphics

Anscombe (1973) described a data set of numbers, each measured on the scales of x and y. He described the data as having a mean of 9 and standard deviation of 3.3 in x and a mean of 7.5 and standard deviation of 2.03 in y. The data were fit by ordinary least squares (OLS) criteria to have a slope of .5, an intercept of 3, and a correlation of .83. This allows the terse and easily interpretable summary for the data in the form $y = 3 + .5(x) + \text{error}$. As a thought experiment, we encourage the reader to try to visualize a scatter plot that depicts such data.

If you imagined a scatter plot similar to that shown in Figure 2.1, then you are quite correct, because this represents the data Anscombe provided that met the descriptive statistics we described previously. This, however, is only a small

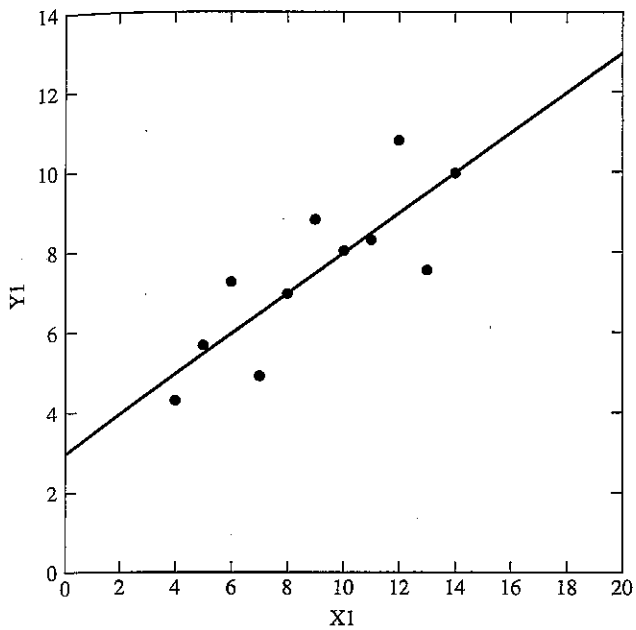

Figure 2.1 Plot of bivariate normal version of Anscombe data.

part of the story, for if you imagined the data to have the shape shown in panel A of Figure 2.2 then you are also correct. If you imagined the pattern in panels B or C of Figure 2.2, you are also correct because all the patterns shown in Figures 2.1 and 2.2 conform to the same algebraic summary statistics given by Anscombe. Although this example speaks to the weakness of overdependence on algebraic representations alone, it points to the larger issue that all summarization leads to a loss of information.

Graphics Lie, So You Need Algebra

Although graphics are a mainstay of EDA, graphics are not immune from this general principle. Consider the following data set: 1,1,2,2,3,3,4,4,5,5,5,5,6,6,6,6,6,6,7,7,7,7,8,8, 9,9,10,10,11,11. Entering this data into a standard statistics package produces the display presented in Figure 2.3. As the reader can see, a slight skew is evident that may not be detected in the listing of numbers themselves. It is important to consider the computational model that underlies this graphic. It consists of a mapping of bar height to frequency and bar width to bin width in the frequency table. *Bin width* refers to the size of the interval in which numbers are aggregated when determining frequencies. (The term *bandwidth* is similarly used in many domains, including nonparametric smoothing; cf. Härdle, 1991). Different bin widths and starting points for bins will lead to different tables, and hence, different graphics. Using the same data with different combinations of bin starting point and bin widths produces the displays seen in Figure 2.4.

In sum, all data analysis is a process of summarization. This process leads to a focus on some aspects of data while taking focus off of other aspects. Conscious of these issues, the exploratory analyst always seeks multiple representations of the data and always holds a position of skepticism toward any single characterization of data.

The Importance of Models

Apart from the cognitive aspects of statistical information just discussed, there is an additional epistemic layer of meaning that must be dealt with. As George Box (1979) wrote: "All models are wrong but some are useful" (p. 202). An important aspect of EDA is the process of model specification and testing with a focus on the value and pattern of misfit or residuals. Although some psychologists are familiar with this view from their experience with regression graphics or diagnostics, many individuals fail to see their statistical work as model building. In the EDA view, all statistical work can be considered model building. The simple *t* test is a model of mean difference as a function of group membership considered in terms of sampling fluctuation. Regression analyses attempt to model criteria values as a function of predictor variables, whereas analysis of variance (ANOVA) models means and variances of dependent variables as a function of categorical variables.

Unaware of the options, many individuals fail to consider the wide range of model variants that are available. In regression, for example, the "continuous" dependent variable may be highly continuous or marginally continuous. If the dependent variable is binary, then a logistic regression is appropriate and a multilevel categorical dependent variable can likewise be fit (Hosmer & Lemeshow, 2000). The closely related variant of Poisson regression exists for counts, and probit and Tobit variants also can be used.

The application of models in these instances is central to EDA. Different models will have different assumptions and often describe the data well in one way, but fail in others. For example, in the world of item response theory, there is often great consternation regarding the choice of models to use. One-parameter models may misfit the data in some way, but have the desirable properties of sufficiency and consistent ordering of individuals. Two- and three-parameter models generally fit the data more tightly but without the conceptual advantages of the one-parameter model. From an EDA point of view, each model is "correct" in some respect insofar as each brings some value and loses others. Depending on the exact scientific or practical need, the decision maker may choose to emphasize one set of values or another.

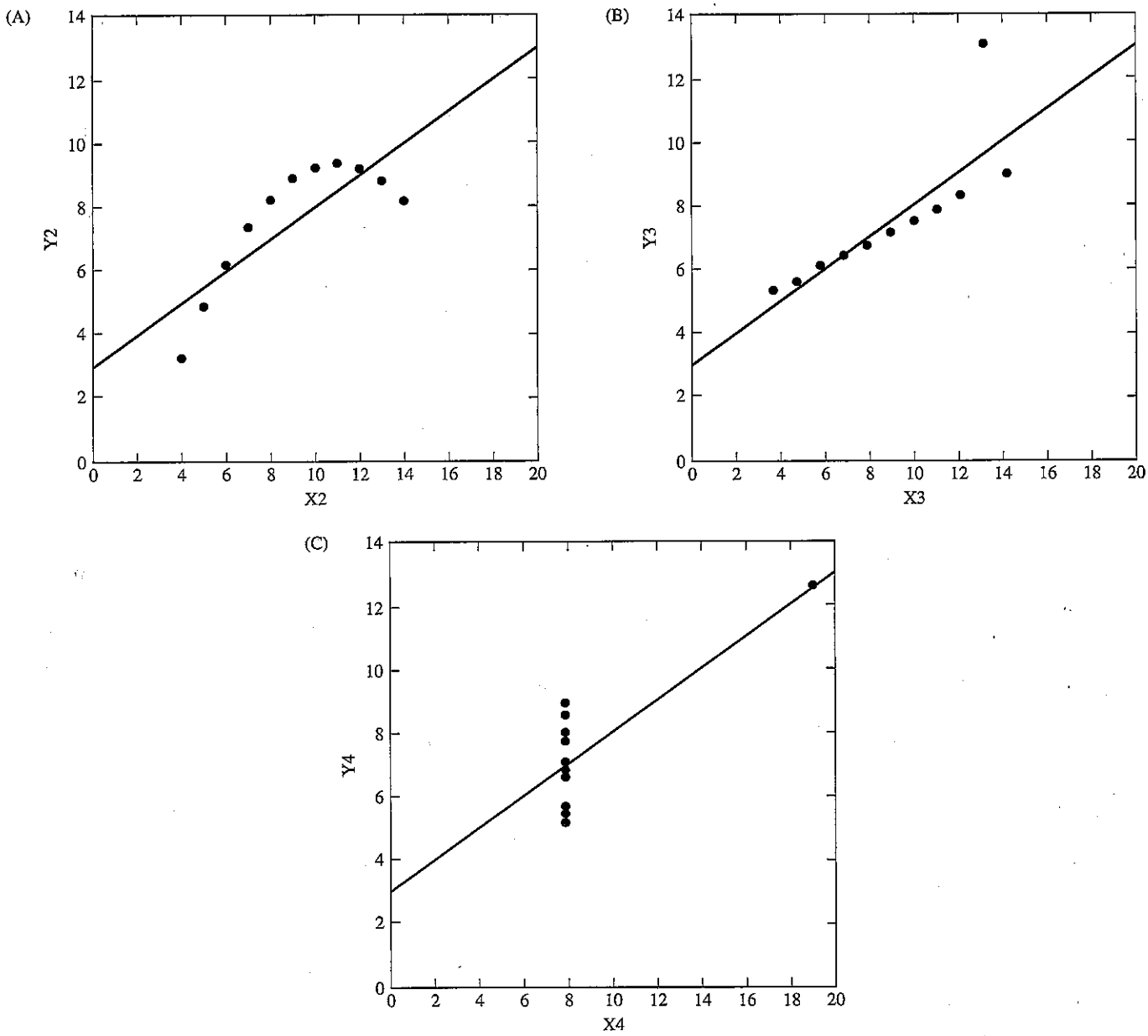

Figure 2.2 Additional data sets with same algebraic summaries as the data in Figure 2.1, with varying patterns and model fit.

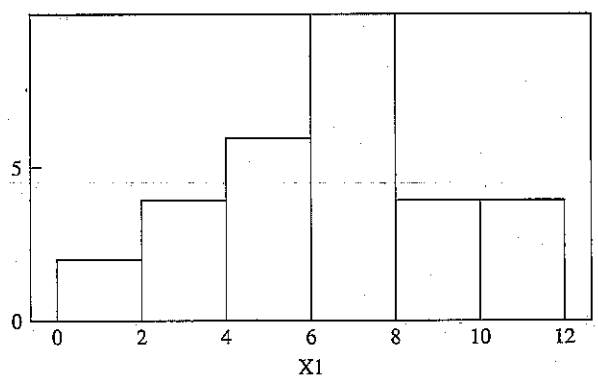

Figure 2.3 Histogram of small data set revealing slight skew.

Regardless of individual decisions about models, the most important issues are that one realizes (a) that there is always model being used (even if implicitly), (b) that for real-world data, there is no such thing as the perfect model, and (c) that the way in which the model is wrong tells us something about the phenomenon.

Abduction as the Logic of EDA

Because of the rich mathematical foundation of CDA, many researchers assume that the complete philosophical basis for inferences from CDA have been worked out. Interestingly, this is not the case. Fisher (1935, 1955) considered his

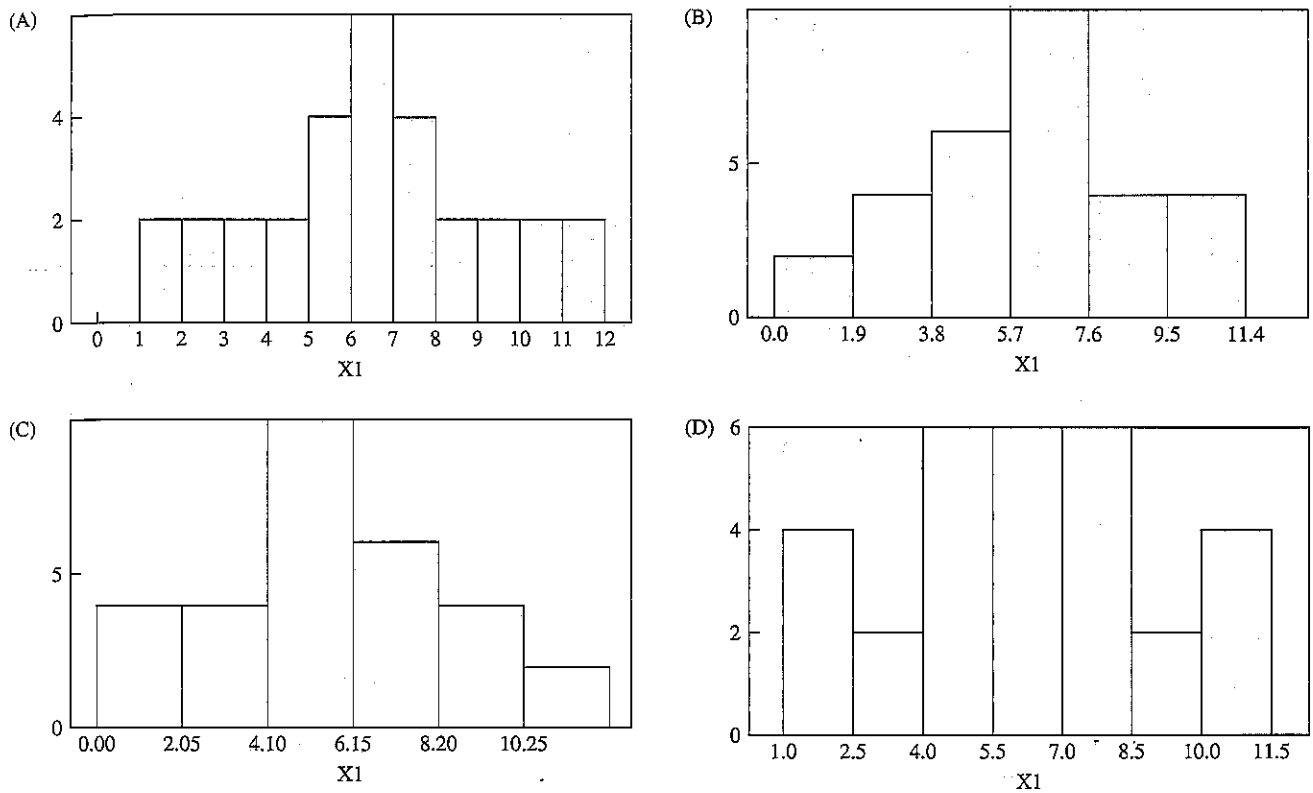

Figure 2.4 Additional histograms of the same data depicted in Figure 2.3 with varying appearances as a function of bin width and bin starting value.

approach to significance testing an as implementation of "inductive inference" and argued that all knowledge is gained in this way. Neyman and Pearson (1928, 1933a, 1933b), on the other hand, developed the concepts of power, type II error, and confidence intervals, to which Fisher objected (Gigerenzer, 1987, 1993; Lehmann, 1993). Neyman argued that only deductive inference was possible in statistics, as shown in the hypothesis testing tradition he developed. Others argue that classical statistics involves both logical modes, given that the hypothesis is generated deductively and data are compared against the hypothesis inductively (Lindsey, 1996).

Where, then, does this leave EDA? Because Tukey was primarily a mathematician and statistician, there has been little explicit work on the logical foundations of EDA from a formal philosophical viewpoint. A firm basis for understanding EDA, however, can be found in the concept of *abduction* proposed by the American philosopher Charles Sanders Peirce. Peirce, whose name is pronounced "pers," was a tour de force in American philosophy as the originator of modern semiotics, an accomplished logician in logic of probability, and the originator of pragmatism that was popularized by James and Dewey. Peirce (1934/1960) explained the three logical processes by arguing, "Deduction proves something must be. Induction shows that something actually is operative; abduction merely suggests that something may be" (vol. 5, p. 171). Put another way: *Abduction* plays the role of generating new ideas or hypotheses; *deduction* functions as evaluating the hypotheses; and *induction* justifies the hypotheses with empirical data (Staat, 1993).

Deduction involves drawing logical consequences from premises. The conclusion is true given that the premises are true also (Peirce, 1868). For instance,

First premise: All As are Bs (True).
Second premise: C is A (True).
Conclusion: Therefore, C is B (True).

Deductive logic confines the conclusion to a dichotomous answer (true-false). A typical example is the rejection or failure of rejection of the null hypothesis. To be specific, the formulated hypothesis is regarded as the first premise. When the data (the second premise) conform to the hypothesis, the conclusion must assert that the first premise is true.

Some have argued that deduction is incomplete because we cannot logically prove all the premises are true. Russell and Whitehead (1910) attempted to develop a self-sufficient logical-mathematical system. In their view, not only can mathematics be reduced to logic, but also logic is the foundation of mathematics. However, Gödel (1947/1986) found that

it is impossible to have such a self-contained system. Any lower order theorem or premise needs a higher order theorem or premise for substantiation, and it goes on and on; no system can be complete and consistent at the same time. Building on this argument, Kline (1980) held that mathematics had developed illogically with false proof and slips in reasoning. Thus, he argued that deductive proof from self-evident principles in mathematics is an "intellectual tragedy" (p. 3) and a "grand illusion" (p. 4).

For Peirce, inductive logic is based upon the notion that probability is the relative frequency in the long run and that a general law can be concluded based on numerous cases. For example,

A1, A2, A3 . . . A100 are B.
A1, A2, A3 . . . A100 are C.
Therefore, B is C.

Hume (1777/1912) argued that things are inconclusive by induction because in infinity there are always new cases and new evidence. Induction can be justified if and only if instances of which we have no experience resemble those of which we have experience. Thus, the problem of induction is also known as "the skeptical problem about the future" (Hacking, 1975). Take the previous argument as an example. If A101 is not B, the statement "B is C" will be refuted. We never know when a line predicting future events will turn flat, go down, or go up. Even inductive reasoning using numerous accurate data and high-power computing can go wrong, because predictions are made only under certain specified conditions (Samuelson, 1967).

Induction suggests the possible outcome in relation to events in the long run. This is not definable for an individual event. To make a judgment for a single event based on probability, such as saying that someone's chance of surviving surgery is 75%, is nonsensical. In actuality, the patient will either live or die. In a single event, not only is the probability indefinable, but also the explanatory power is absent. Induction yields a general statement that explains the event of observing, but not the facts observed. Josephson and Josephson (1994) gave this example:

> Suppose I choose a ball at random (arbitrarily) from a large hat containing colored balls. The ball I choose is red. Does the fact that all of the balls in the hat are red explain why this particular ball is red? No. . . "All A's are B's" cannot explain why "this A is a B" because it does not say anything about how its being an A is connected with its being a B. (p. 20)

The function of abduction is to look for a pattern in a surprising phenomenon and suggest a plausible hypothesis (Peirce, 1878). Despite the long history of abduction, it remains overlooked among many texts of logic and research methodology, while gaining ground in the areas of artificial intelligence and probabilistic computing (e.g., Josephson & Josephson, 1994; Schum, 1994). However, as logic is divided into formal types of reasoning (*symbolic logic*) and informal types (*critical thinking*), abduction is represented as informal logic. Therefore, unlike deduction and induction, abduction is a type of critical thinking rather than a formalism captured by symbolic logic. The following example illustrates the function of abduction, though illustrated with symbols for simplification:

The surprising phenomenon, X, is observed.
Among hypotheses A, B, and C, A is capable of explaining X.
Hence, there is a reason to pursue A.

At first glance, abduction may appear as no more than an educated guess among existing hypotheses. Thagard and Shelley (1997) addressed this concern. They argued that unifying conceptions are an important part of abduction, and it would be unfortunate if our understanding of abduction were limited to more mundane cases where hypotheses are simply assembled. Abduction does not occur in the context of a fixed language, since the formation of new hypotheses often goes hand in hand with the development of new theoretical terms such as *quark* and *gene*. Indeed, Peirce (1934/1960) emphasized that abduction is the only logical operation that introduces new ideas.

Although abduction is viewed as a kind of "creative intuition" for idea generation and fact explanation (Hoffmann, 1997), it is dangerous to look at abduction as impulsive thinking and hasty judgment. In *The Fixation of Belief,* Peirce explicitly disregarded the tenacity of intuition as the source of knowledge. Peirce strongly criticized his contemporaries' confusion of propositions and assertions. Propositions can be affirmed or denied while assertions are final judgments (Hilpinen, 1992). The objective of abduction is to determine which hypothesis or proposition to test, not which one to adopt or assert (Sullivan, 1991).

In EDA, after observing some surprising facts, we exploit them and check the predicted values against the observed values and residuals (Behrens, 1997a). Although there may be more than one convincing pattern, we "abduct" only those that are more plausible for subsequent confirmatory experimentation. Since experimentation is hypothesis driven and EDA is data driven, the logic behind each of them is quite different. The abductive reasoning of EDA goes from data to hypotheses, whereas inductive reasoning of experimentation goes from hypothesis to expected data. In fact, closely

(and unknowingly) following Tukey (1969), Shank (1991), Josephson and Josephson (1994), and Ottens and Shank (1995) related abductive reasoning to detective work. Detectives collect related "facts" about people and circumstances. These facts are actually shrewd guesses or hypotheses based on their keen powers of observation.

In short, abduction can be interpreted as observing the world with appropriate categories, which arise from the internal structure of meanings. Abduction in EDA means that the analyst neither exhausts all possibilities nor makes hasty decisions. Researchers must be well equipped with proper categories in order to sort out the invariant features and patterns of phenomena. Quantitative research, in this sense, is not number crunching, but a thoughtful way of peeling back layers of meaning in data.

Exploration, Discovery, and Hypothesis Testing

Many researchers steeped in confirmatory procedures have appropriately learned that true hypothesis tests require true hypotheses and that unexpected results should not be treated with the same deference as hypothesized results. A corollary is that one should keep clear what has been hypothesized in a research study and not modify a hypothesis to match data. This is certainly true and is an essential aspect of confirmatory inference. In some researchers, however, a neurosis develops that extends the avoidance of hypothesis-modification based on knowledge of the data to an avoidance of intimacy with the data altogether. Sometimes this neurosis is exacerbated by the fear that every piece of knowledge has an amount of Type I error associated with it and, therefore, the more we know about the data the higher our Type I error. The key to appropriately balancing exploratory and confirmatory work is to keep clear what has been hypothesized in advance and what is being "discovered" for the first time. Discoveries are important, but do not count as confirmations.

After years of extensive fieldwork an entomologist develops a prediction that butterflies with a certain pattern of spotting should exist on top of a particular mountain, and sets off for the mountaintop. Clearly, if the entomologist finds such butterflies there will be evidence in support of her theory; otherwise, there is an absence of evidence. On her way to the mountain she traverses a jungle in which she encounters a previously unknown species of butterflies with quite unanticipated spottings. How does she handle this? Should she ignore the butterfly because she has not hypothesized it? Should she ignore it because it may simply be a misleading Type I error? Should she ignore it because she may change her original hypothesis to say she has really hypothesized this jungle butterfly?

For most individuals it is clear that a new discovery is valuable and should be well documented and collected. The entomologist's failure to have hypothesized it does not impugn its uniqueness, and indeed many great scientific conclusions have started with unanticipated findings (Beveridge, 1950). Should the entomologist worry about Type I error? Since Type I error concerns long-run error in decision making based on levels of specific cutoff values in specific distributions, that precise interpretation does not seem to matter much here. If she makes an inference about this finding then she should consider the probabilistic basis for such an inference, but nevertheless the butterfly should be collected. Finally, should she be concerned that this finding will contaminate her original hypothesis? Clearly she should continue her travel and look for the evidence concerning her initial hypothesis on the mountaintop. If the new butterfly contradicts the existing hypothesis, then the entomologist has more data to deal with and additional complexity that should not be ignored. If she is concerned about changing her hypothesis in midstream to match the new data, then she has confused hypothesis *generation* and hypothesis *testing*. With regard to any new theories, she must create additional predictions to be tested in a different location.

EDA and Exploratory Statistics

EDA and exploratory statistics (ES) have the same exploratory goals; thus, the question sometimes arises as to whether ES is simply a subset of EDA or EDA is a subset of ES. Because EDA is primarily an epistemological lens and ES is generally presented in terms of a collection of techniques, a more appropriate question is *Can ES be conducted from an EDA point of view?* To this question we can answer *yes*. Furthermore, EDA is a conceptual lens, and most research procedures can be undertaken with an EDA slant. For example, if one is conducting an "exploratory" factor analysis without graphing of data, examination of residuals, or attention to specific patterns of raw data underlying the correlations that are central to the analysis, then little seems to be consistent with the EDA approach. On the other hand, a clearly probabilistic analysis can be well augmented by plots of data on a number of dimensions (Härdle, Klinke, & Turlach, 1995; Scott, 1992), attention to residual patterns in a number of dimensions, and the use of detailed diagnostics that point to patterns of fits and misfits. Regardless of the software or specific statistical procedures used, such activity would clearly be considered EDA and ES. ES does not necessarily imply EDA, but ES can be conducted as EDA if the conceptual and procedural hallmarks of EDA are employed.

Summary

Exploratory data analysis is a rich data-analytic tradition developed to aid practical issues of data analysis. It recommends a data-driven approach to gaining intimacy with one's data and the phenomenon under study. This approach follows the analogy of the detective looking for clues to develop hunches and perhaps seek a grand jury. This counters the more formal and ambitious goals of confirmatory data analysis, which seeks to obtain and present formal evidence in order to gain a conviction. EDA is recommended as a complement to confirmatory methods, and in no way seeks to replace or eliminate them. Indeed, effective researchers should incorporate the best aspects of all approaches as needed.

HALLMARKS OF EDA

In this section, the techniques and attitudes that are standard aspects of EDA are discussed. The tools described here are only recommendations that have worked to allow researchers to reach the underlying goals of EDA. However, it is the underlying goals that should be sought, not the particular techniques. Following Hoaglin, Mosteller, and Tukey (1983), we discuss these tools under the *four Rs* of EDA: *revelation, residuals, reexpression,* and *resistance.*

Revelation

Graphics are the primary tool for the exploratory data analyst. The most widely cited reason for this is Tukey's (1977) statement that "The greatest value of a picture is when it forces us to notice what we never expected to see" (p. vi). In many ways, the graphics in Figures 2.1 and 2.2 illustrate all the rationale of graphics in EDA. First, even though the algebraic summaries are "sufficient statistics," they are sufficient for only the very limited purpose of summarizing particular aspects of the data. For specifying the exact form of the data without additional assumptions regarding distributional shapes, the summary statistics are not only "insufficient" but are downright dangerous. Second, the indeterminacy of the algebra calls us to fill in the details with possibly untenable assumptions. In the Anscombe data-thought experiment, participants almost universally imagine the data to be of the canonical form shown in Figure 2.1. In the absence of a skeptical mind and in the light of the history of statistics textbooks that are focused on mathematical idealizations at the expense of real-world patterns, many psychologists have developed schemas and mental models (Johnson-Laird, 1983) that lead to erroneous inferences.

Another psychological advantage of graphics is that it allows for a parsimonious representation of the data. The facts that are easily derivable from the image include all the individual values, the relative position of each data point to every other, shape-based characterizations of the bivariate distribution, and the relationship between the data and the proposed regression line. After some practice, the trained eye can easily discern and describe the marginal distributions as well as the distribution of residuals. The construction of a text-based representation of all of this information would require an extensive set of text-based descriptors. In short, visual images of the type shown here exploit the visual-spatial memory system to support efficient pattern recognition (Garner, 1974), problem solving (Larkin & Simon, 1987), and the construction of appropriate mental models (Bauer & Johnson-Laird, 1993).

Tukey's early work and concomitant advances in computing have led to an explosion in graphical methods over the last three decades. Numerous authors, including Tufte (1990, 1997, 1983/2001) and Wainer (1997; Wainer & Velleman, 2001) have worked to popularize data-based graphics. William Cleveland has had a large impact on the statistical community with his empirical studies of the use of graphics (Cleveland & McGill, 1984), the initiation of cognitive models of graph perception (Cleveland, 1985), and his application of these principles to statistical graphics (especially Cleveland, 1993). Wilkinson (1993, 1994, 1999) made substantial contributions to the study of proper use of statistical graphics, and has recently provided a comprehensive volume regarding graphics in software and statistical analysis (Wilkinson, 1999) that is required reading for anyone interested in the field. Kosslyn (1994) provided a discussion of numerous potential rules for graph construction from a psychological perspective, and Lewandowsky and Behrens (1999) provide a recent review of cognitive aspects of statistical graphs and maps.

Graphics Made for EDA

During the emergence of the EDA tradition, Tukey developed a large number of graphical tools, some of which have become commonplace, others of which have had little visibility outside specialized applications. It is important to remember that at the time of their original construction, much of what Tukey sought to do was to support quick summarization and analysis when data were available, and the analysis was to occur by hand.

Perhaps the best known of Tukey's graphical devices for EDA is the *box-and-whisker plot,* otherwise called the *box-plot.* The box-plot is a graph based on a five-number

summary of a distribution of data; these numbers are the median, the first and second hinges, and either the lowest and highest number or a similar measure of range number arrived at by separating very extreme values. The median is equal to the 50th percentile of the distribution. The hinges are either equal to or very close to the 25th and 75th percentiles—although they are found using a simpler rank-based formula for computation. To construct a box-plot, a scale is drawn, and a box is placed on the scale with one end of the box indicating the scale value of the lower hinge (25th percentile) and the other end of the box occurring at the scale position of the upper hinge (75th percentile). An additional line is drawn in the middle to indicate the scale value of the median. The scale-value difference between the two hinges is called either the *hinge spread* or the *interquartile range* (often abbreviated IQR), and in a normal distribution corresponds to approximately 0.67 standard deviations on each side of the mean.

Rules for the construction of the "whisker" portion of the display vary. In the most common version, lines are extended along the scale from the hinges to the farthest data value in each direction up to 1.5 hinge spreads. If there are data past that point, the whiskers extend to the farthest data point prior to the 1.5 hinge-spread cutoff. Data points beyond the whiskers are usually identified individually to bring attention to their extremeness and potential characterization as outliers.

An example of multiple box-plots is presented in Figure 2.5, panel A, with individual raw data values presented in panel B for comparison. These graphics depict the distributions of effect sizes from the meta-analysis of social memory conducted by Stangor and McMillan (1992). The categorical variable on the horizontal axis is the length of stimulus presentation in seconds. The continuous variable on the vertical axis is the size of the effect for each study included in the meta-analysis. As the reader may see, the box-plots provide a compact description of each distribution and allow relatively easy comparison of both the level and spread of each distribution. The distribution farthest to the left represents all the studies for which no presentation speed is reported. The range is approximately from −2 to +2, with a median slightly below zero. The second box-plot depicts the distribution of effect sizes from studies that used a 2-s presentation speed. It is the highest distribution of all, with some positive skew. The median of this distribution is higher than the 75th percentile of the remaining distributions, indicating a clear trend toward larger values. The median is also higher than the 75th percentile of the 6- and 10-s studies. The studies with presentation times of 6 s show very little variance with the exception of two outliers, which are indicated separately.

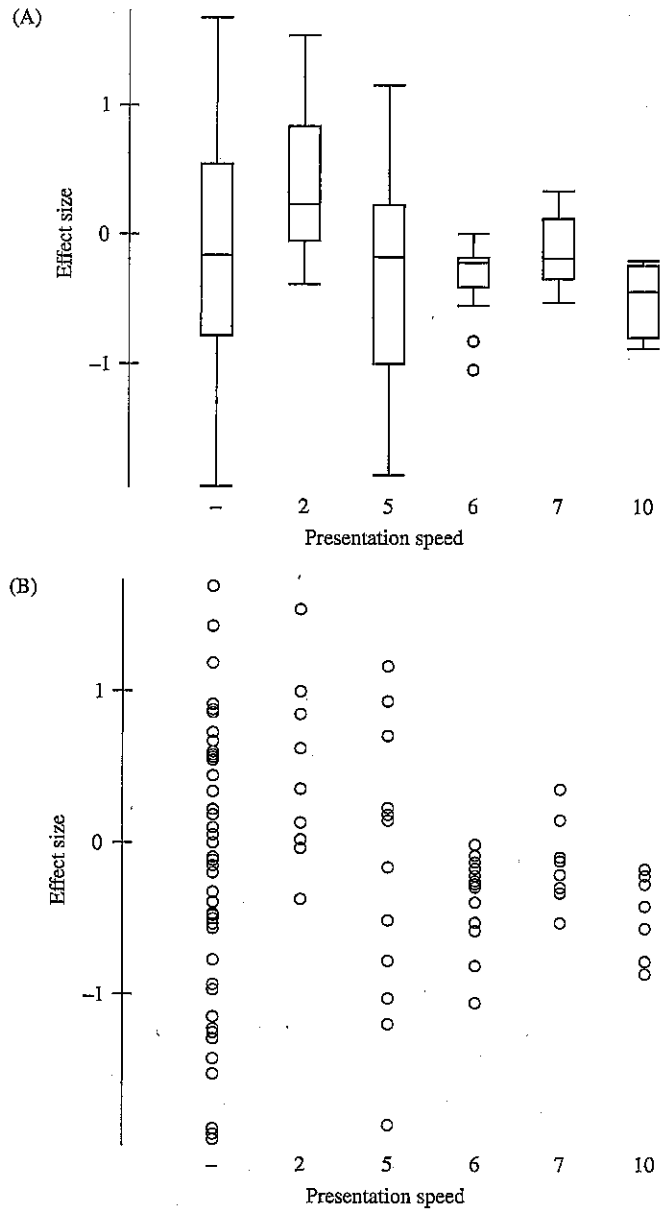

Figure 2.5 Panel A consists of multiple box-plots of effect sizes in social memory meta-analysis, organized by presentation speed (from Stangor & McMillan, 1992). Panel B depicts the same data by plotting individual values in a dot-plot.

When two box-plots are compared, the analyst is undertaking the graphical analog of the *t test.* Displays with additional boxes, as shown here, are analogous to the analysis of variance: Group-level measures of central tendency are displayed relative to the amount of within-group variability in the data.

Although the box-plots are very useful and informative in their current state, working in the exploratory mode raises additional issues. First, how might we be fooled by these displays? The answer to this is that there are times when the

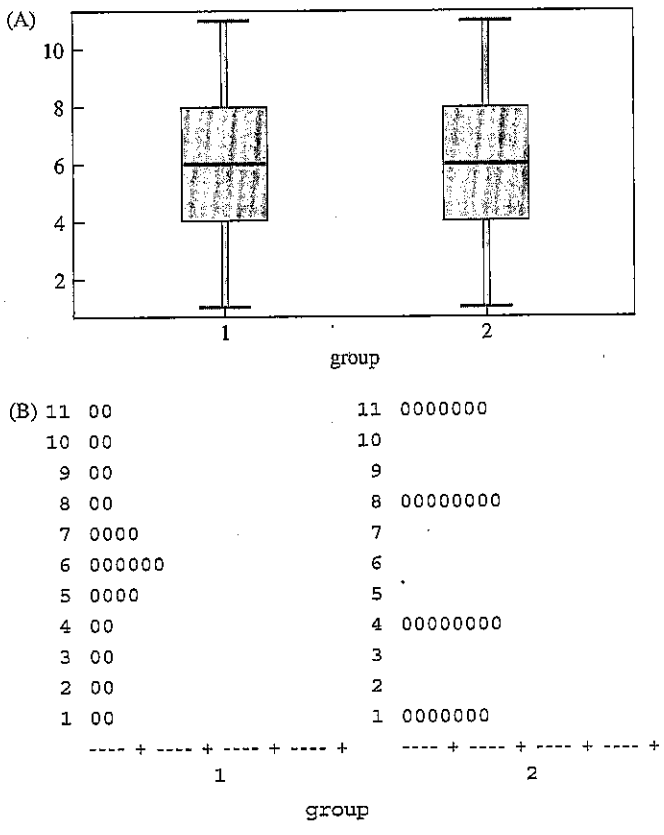

Figure 2.6 Panel A is two box-plots with identical summary statistics. Panel B depicts the underlying data used to make panel A, illustrating the possibility that different data may produce identical graphs.

five-number summaries are the same for different distributions. This leads to a case in which the box-plots look identical yet the data differ in structure. Consider the top panel of Figure 2.6, in which each of two box-plots is identical, indicating identical values of the five-number summary. The lower panel of Figure 2.6 depicts the underlying raw data values that vary in form. Distortions can also occur if there are very few levels of the variable being measured, because it will cause many data points to have the same value. In some cases the appearance of a plot may be distorted if the hinges are equal to the minimum or maximum, because no whiskers appear.

A second point of interest concerns how we can learn more from the data by enhancing the box-plot. Toward this end, recommendations abound. Tufte (1983/2001) recommended the omission of the box (an idea not well supported by empirical data; Stock & Behrens, 1991). Other suggested improvements include indicating the sample size of the subset below the box (e.g., Becker, Chambers, & Wilks, 1988), adding confidence intervals around the median in the box, or distorting the box shape to account both for sample size and the confidence intervals (McGill, Tukey, & Larsen, 1978). Berk (1994) recommended the overplotting of a dot-plot (as seen in the lower panel of Figure 2.6) on top of the box-plot so that two levels of detail can be seen simultaneously. Regardless of the exact implementation used, users must be wary that software packages vary on the algorithms used to calculate their five-number summaries, and they may not be looking at the summaries one expects (Frigge, Hoaglin, & Iglewicz, 1989).

Interactive Graphics

Although much can be gained by modifying the static appearance of plots such as the box-plot, substantial gains in data analysis can be made in computerized environments by using interactive graphics with brushing and linking (Cleveland & McGill, 1988). *Interactive graphics* are graphic displays that respond to the *brushing* (selection) action of a pointing device (such as a mouse) by modifying the graphics in real time. *Linking* is the connection of values on the same observation through brushing or selecting across different graphics. Highlighting an observation in one display (say, a scatter plot) causes the value of the same observation to appear highlighted in another display (say, a histogram) as well. In this way, an analyst working to analyze one graphic can quickly see how information in that graphic relates to information in another graphic. For example, in Figure 2.5 the conditional level of each distribution varies greatly. An analyst may wonder if this is primarily from the categorical variables listed on the horizontal axis, or if there are other variables that may also covary with these medians. One possibility is that different research laboratories tend to use different speeds and therefore, that laboratory covaries with speed.

Prior to the advent of interactive graphics, one would stop the analysis in order to look in a table to determine which data came from which laboratory. Such a process could easily become tedious and distracting from the main task. Using a program that has high graphical interactivity, in this case Data Desk (Data Description, 1997), highlighting the data of interest in one graphical display highlights the same data in other graphical displays or in the variable listings. To accomplish this in Data Desk we simply turn off the box-plots using the pull-down menu at the top of the graph window (thereby changing panel A of Figure 2.5 to panel B), indicate "Show identifying text" from a menu at the top of the screen, click on the identifying variable of interest (study name), and highlight observations to be linked. The final outcome of these few quick hand movements is presented in Figure 2.7. Here the graphics reveal some unexpected results. Eight of the nine effect sizes in this group come from only two studies

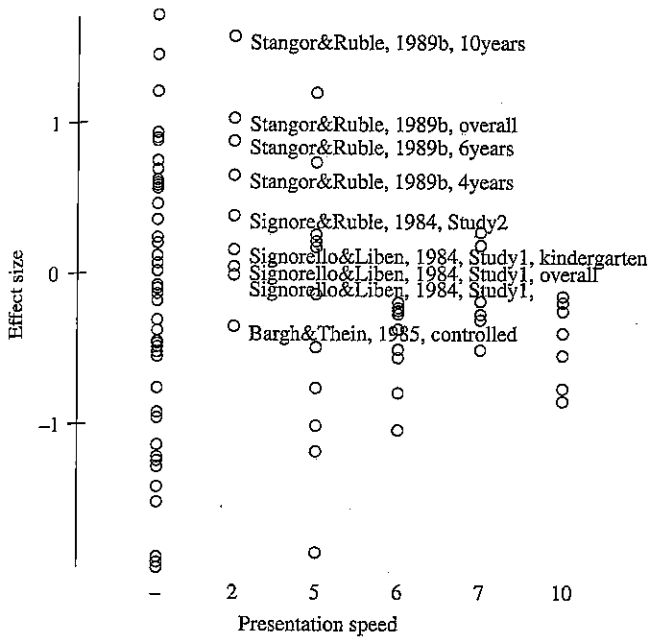

Figure 2.7 Dot-plot with identifying text obtained by selecting and linking.

and the studies are highly stratified by effect size. When the eight data points of the 7-s effects are circled, the study names indicate that all these effects come from a single study.

Moving Up the Path of Dimensionality

Whereas box-plots are often considered univariate graphics because they are often used to display a single variable, our simple example has demonstrated that the box-plot can easily function in three variables. In this case the variables are presentation speed, effect size, and study origin. In highly interactive environments, however, additional variables are easily added. For example, in Data Desk, palettes available on the desktop allow one to choose the shape or color of symbols for individual data points. This is accomplished through selecting the data points of interest by circling the points on the graphic and clicking on the desired shape or color. Symbol coloring can also be accomplished automatically by using a pull-down menu that indicates a desire to color symbols by the value of a specific variable. In our meta-analysis data, coloring the data points by sample size and varying the shape to indicate the value of another categorical variable of interest may aid in finding unanticipated patterns. In this way, we would have created a rather usable yet complex five-dimensional graphic representation of the data.

For combinations of categorical and measured data, the box-plot and corresponding dot-plot provide an excellent starting point. For analyses that focus more heavily on measured (continuous or nearly continuous) data, the scatter plot is the common fundamental graphic. Here variations abound, as well. Whereas the scatter plot is often used to understand two dimensions of data, when faced with high-dimensional data, one often uses a matrix of scatter plots to see the multidimensionality from multiple bivariate views. An example of a scatter plot matrix is presented in Figure 2.8. The data portrayed here are a subset of the data that Stangor and McMillan (1992) used in a weighted least-squares regression analysis of the effect sizes. The plot can be thought of as a graphical correlation matrix. Where we would have placed the value of the correlation, we instead put the graphical bivariate display. For each individual scatter plot, one can identify the variable on the vertical axis by looking at the variable named at the far left of the row. The horizontal axis is identified by looking down the column of the matrix to the variable identified at the bottom of the scatter plot. For example, the plot in the upper right corner has "*N*" (sample size) on the vertical axis and "con / incon" on the horizontal axis, indicating the ratio of congruent to incongruent stimuli. The top of the "con / incon" label is hidden due to plot size in the plot in the lower right corner. The plots in the diagonal cells are normal-probability plots whose interpretation is discussed below.

In this situation, as is often the case, the scatter plot matrix does an excellent job of revealing unexpected structure. For many of the bivariate relationships there is great departure from bivariate normality. Of particular concern is the combination of high skew in the congruent-incongruent ratio and the floor effect in the targets and traits variables. These issues lead to L-shaped distributions that will present a clear challenge to any continuous linear model. Outliers and combinations of missing data should also be considered carefully. Of particular note in these data is that the higher level of the dummy-coded delay variable exists in only two observations, but one of those observations has no matching data on many variables and thus functions as a single point. In a multivariate situation such as regression analysis, this is quite problematic because the estimation of the relationship of this variable with all others rests precariously on the value of the single point. Error at this point will thereby be propagated through the system of partial correlations used to estimate regression effects.

Plots with multiple straight lines indicate the 0 and 1 levels of dummy coding. A number of additional dummy-coded variables subjected to simultaneous regression by Stangor and McMillan (1992) were omitted because the number of plots became too large to present here clearly. Earlier versions of this matrix revealed additional unusual values that were traced back to the present authors' transcription process

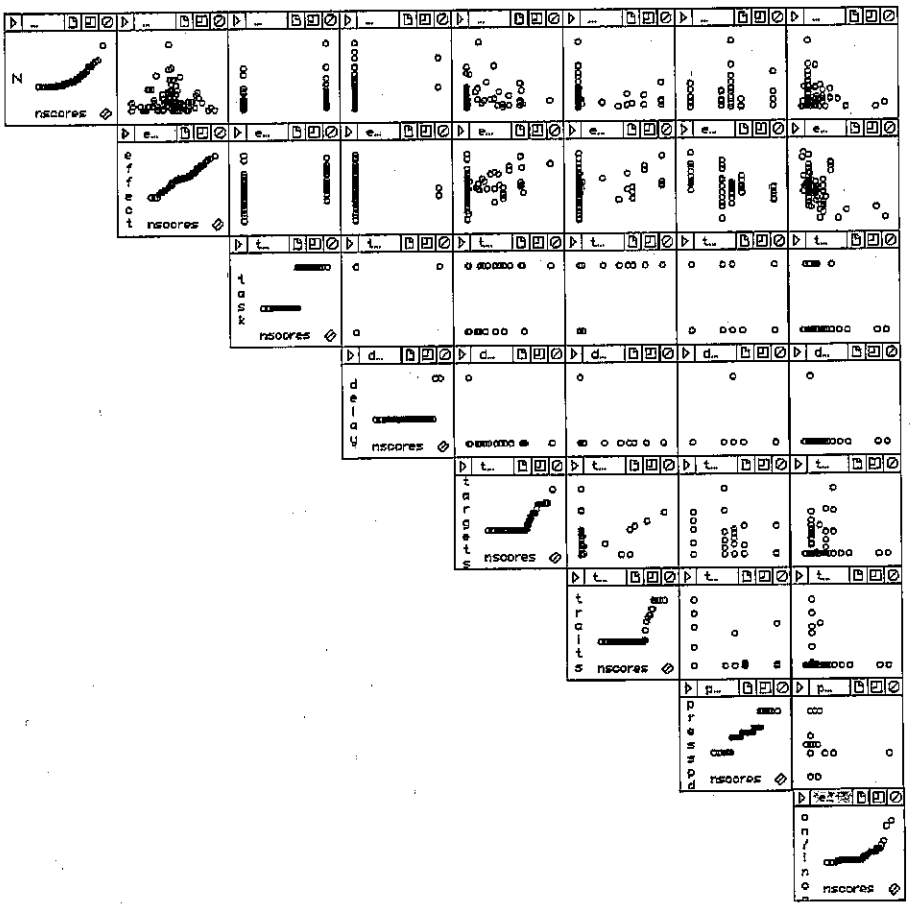

Figure 2.8 Scatter plot matrix of meta-analysis data from Stangor and McMillan (1992).

and have been since corrected. In this case, the graphics revealed structure and avoided error.

Going Deeper

Although the scatter plot matrix is valuable and informative, it is important that the reader recognize that a series of two-dimensional views is not as informative as a three-dimensional view. For example, when Stangor and McMillan computed a simultaneous regression model, the variables indicating the number of targets and traits used in each study reversed the direction of their slope, compared with their simple correlations. Although a classical "suppressor" interpretation was given, the exploratory analyst may wonder whether the simple constant and linear functions used to model these data were appropriate. One possibility is that the targets variable mediates other relationships. For example, it may be the case that some variables are highly related to effect size for certain levels of target, but have different relationships with effect size at other levels of targets.

To provide a quick and provisional evaluation of this possibility, we created a histogram of the target variables, selected those bins in the graphic that represent low levels of targets, and chose a unique color and symbol for the observations that had just been selected. From here, one can simply click on the pull-down menu on any scatter plot and choose "Add color regression lines." Because the observations have been colored by low and high levels of the target variable, the plots will be supplemented with regression lines between independent variables and the effect size–dependent variable separately for low and high levels of targets, as displayed in Figure 2.9.

Moving across the second row of Figure 2.9 (which corresponds to the response variable), first we see two regression lines with low identical slopes indicating little relationship between task and effect, which is constant across levels of target. The delay variable in the next column shows a similar pattern, whereas the next three variables show small indications of interaction. The interaction effect is very clear in the relationship between effect size and the congruent-incongruent ratio in the rightmost column. This relationship is positive for

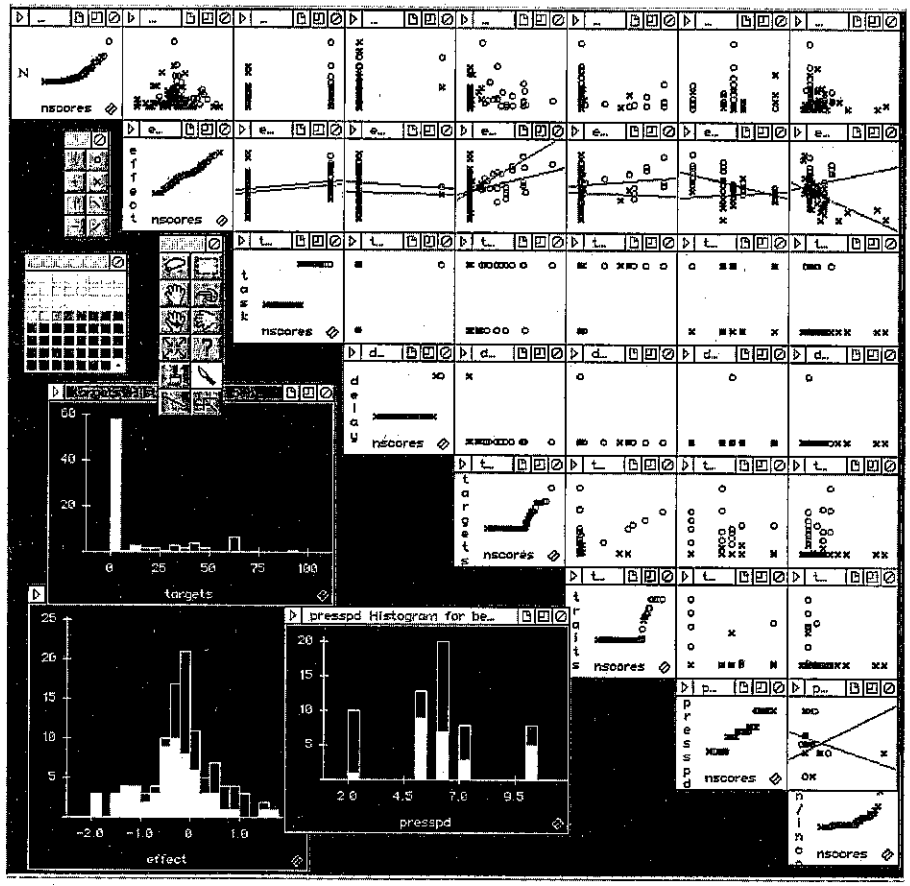

Figure 2.9 View of computer screen using Data Desk software for selecting, brushing, and linking across multiple plots. Several plots have been enhanced with multiple regression lines that vary by subsets of selected data.

observations with high numbers of targets, but negative for low numbers of targets. Unfortunately, in failing to recognize this pattern, one may use a model with no interactions. In such a case the positive slope observations are averaged with the negative slope observations to create an estimate of 0 slope. This would typically lead the data analyst to conclude that no relationship exists at all, when in fact a clear story exists just below the surface (one variable down!).

Although the graphics employed so far have been helpful, we have essentially used numerous low-dimensional views of the data to try to develop a multidimensional conceptualization. This is analogous to the way many researchers develop regression models as a list of variables that are "related" or "not related" to the dependent variable, and then consider them altogether. Our brushing and coding of the scatter plot matrix has shown that this is a dangerous approach because "related" is usually operationalized as "linearly related"—an assumption that is often unwarranted. Moreover, in multidimensional space, variables may be related in one part of the space but not in the other.

Working in an exploratory mode, these experiences suggest we step back and ask a more general question about the meta-analytic data: In what way does the size and availability of effects vary across the variety of study characteristics? To begin to get such a view of the data, one may find three-dimensional plots to be useful. A graphic created using a nonlinear smoother for the effect size of each study as a function of the number of targets and presentation speed is presented in panel A of Figure 2.10. The general shape is similar to the "saddle" shape that characterizes a two-way interaction in continuous regression models (Aiken & West, 1991). The graphic also reveals that little empirical work has been undertaken with high presentation speed and a low number of targets, so it is difficult to assess the veracity of the smoothing function given the lack of data in that area. At a minimum, it suggests that future research should be conducted to assess those combinations of study characteristics. Panel B of Figure 2.10 shows an alternate representation of the data with a traditional linear surface function that is designed to provide a single additive prediction across all the data.

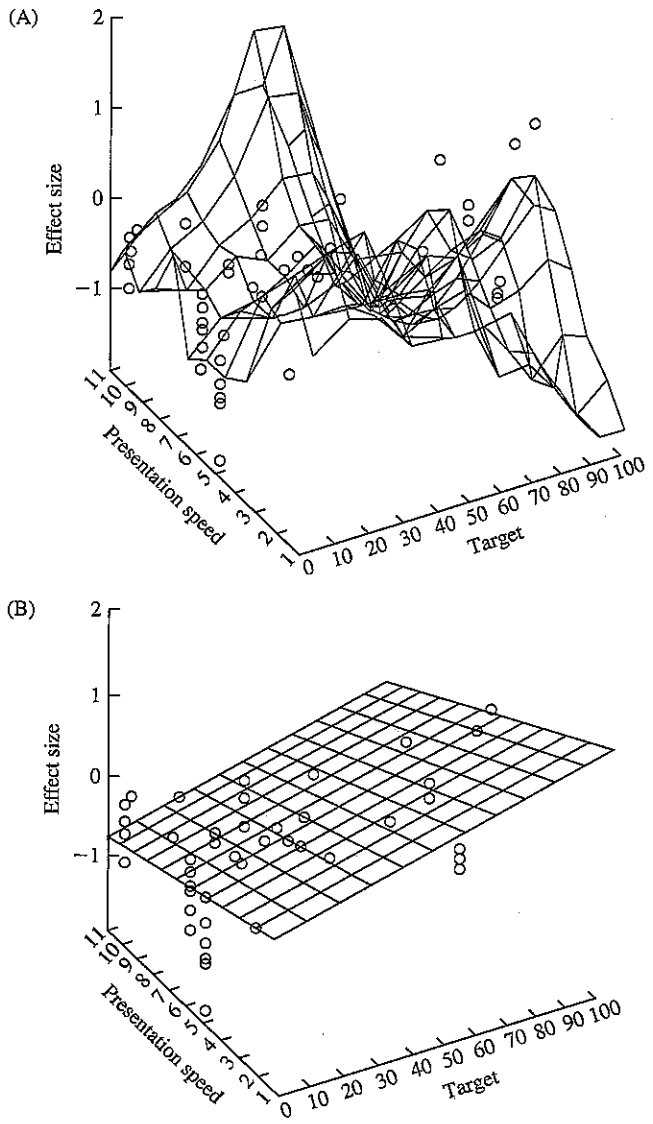

Figure 2.10 Panel A is a nonlinear surface estimate of the interaction of presentation speed, number of targets, and effect size in the Stangor and McMillan (1992) meta-analysis data. Panel B is a completely linear prediction surface as would be obtained using common least squares regression.

For another impression of the data, we can take a series of slices of the three-dimensional data cube and lay the slices out side by side. Such an arrangement is possible using a general framework called a *trellis plot* (Becker, Cleveland, & Shyu, 1996; Clark, Cleveland, Denby, & Liu, 1999) as implemented in Splus (Insightful, 2001) and shown in Figure 2.11. The three panels show progressive slices of the data with linear regression lines overlaid. As the reader can see, the plots correspond to three slices of the three-dimensional cube shown in Figure 2.10, panel A, with changes in the regression line matching different portions of the "hills" in the data.

The simple graphics we have used here provide an excellent start at peeling away the layers of meaning that reside in these data. If nothing else is clear, the data are more complex than can be easily described by a simple linear model in multiple dimensions. The theoretical concerns of Anscombe (1973) have proven to be realistic after all. Despite the interocular effect provided by these graphics, some readers will assure themselves that such difficulties appear primarily in data from meta-analyses and that the data they work with will not be so problematic. Unfortunately this is not often the case, and there is a cottage industry among EDA proponents of reanalyzing published data with simple graphics to show rich structure that was overlooked in original work.

Residuals and Models

In the EDA tradition, the second *R* stands for *residual,* yet this word signifies not simply a mathematical definition, but a foundational philosophy about the nature of data analysis. Throughout Tukey's writings, the theme of DATA = FIT + RESIDUALS is repeated over and over, often in graphical analog: DATA = SMOOTH + ROUGH. This simple formula reminds us that our primary focus is on the development of compact descriptions of the world and that these descriptions will never be perfect; thus there will always be some misfit between our model and the data, and this misfit occurs with every observation having a residual.

This view counters implicit assumptions that often arise in statistical training. First, many students acquire an unfortunate belief that "error" has an ontological status equivalent to "noise that can be ignored" and consequently believe the results of a model-fitting procedure (such as least squares regression) is the "true" model that should be followed. Such a view fails to emphasize the fact that the residuals are simply a byproduct of the model used, and that different models will lead to different patterns of residuals. As we saw in the previous section, different three-dimensional models provide different degrees of hugging the data, and hence, different amounts of residual. Second, in EDA the analyst focuses on the size and pattern of individual residuals and subsets of residuals. A curve that remains in a residual plot indicates the model has failed to describe the curve. Multiple modes remaining in the residuals likewise suggest that a pattern has been missed. On the other hand, if students are taught to focus on only the gross summary of the sums of squares, they will also miss much of the detail in the pattern that is afforded by a careful look at residuals. For example, as indicated by the common *r* among the Anscombe (1973) data sets, all four data sets have the same sums-of-squares residual, but dramatically different patterns of residuals.

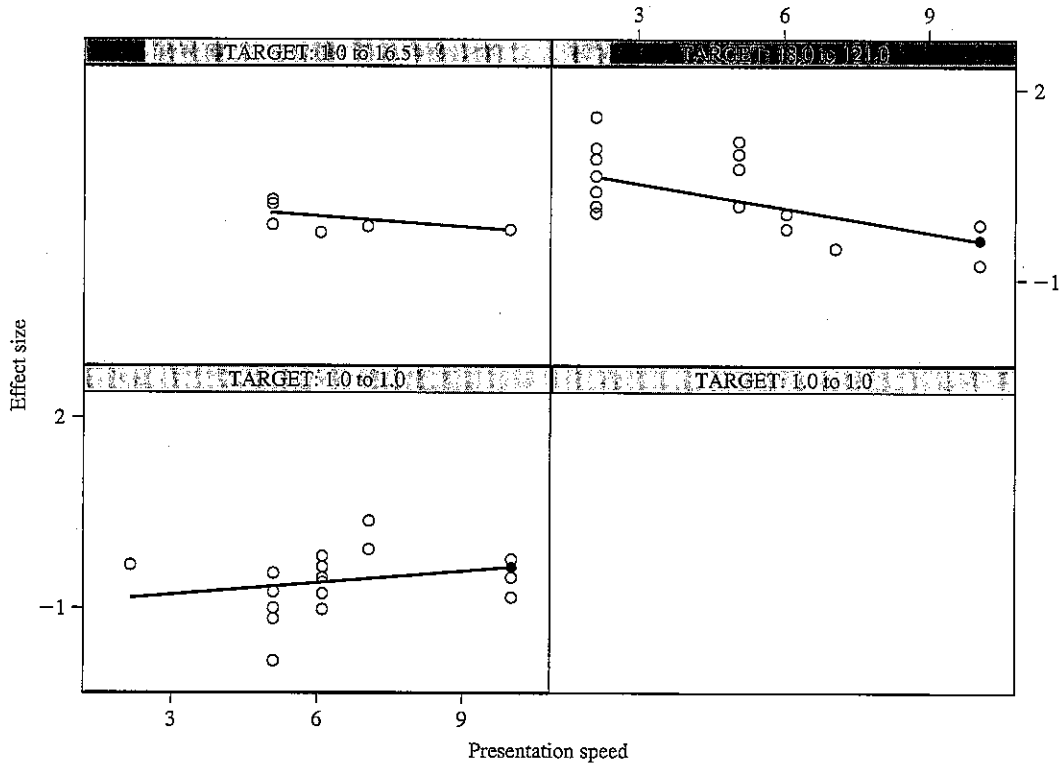

Figure 2.11 Plot of presentation speed by effect size at different ranges of number of targets using a trellis display. Data are identical with those presented in Figure 2.10.

This emphasis on residuals leads to an emphasis on an iterative process of model building: A tentative model is tried based on a best guess (or cursory summary statistics), residuals are examined, the model is modified, and residuals are reexamined over and over again. This process has some resemblance to forward variable selection in multiple regression; however, the trained analyst examines the data in great detail at each step and is thereby careful to avoid the errors that are easily made by automated procedures (cf. Henderson & Velleman, 1981). Tukey (1977) wrote, "Recognition of the iterative character of the relationship of exposing and summarizing makes it clear that there is usually much value in fitting, even if what is fitted is neither believed nor satisfactorily close" (p. 7).

The emphasis on examining the size and pattern of residuals is a fundamental aspect of scientific work. Before this notion was firmly established, the history of science was replete with stories of individuals that failed to consider misfit carefully. For example, Gregor Mendel (1822–1884), who is considered the founder of modern genetics, established the notion that physical properties of species are subject to heredity. In accumulating evidence for his views, Mendel conducted a fertilization experiment in which he followed several generations of axial and terminal flowers to observe how specific genes carried from one generation to another. On subsequent examination of the data, R. A. Fisher (1936) questioned the validity of Mendel's reported results, arguing that Mendel's data seemed "too good to be true." Using chi-square tests of association, Fisher found that Mendel's results were so close to the predicted model that residuals of the size reported would be expected by chance less than once in 10,000 times if the model were true.

Reviewing this and similar historical anomalies, Press and Tanur (2001) argue that the problem is caused by the unchecked subjectivity of scientists who had the confirmation of specific models in mind. This can be thought of as having a weak sense of residuals and an overemphasis on working for dichotomous answers. Even when residuals existed, some researchers tended to embrace the model for fear that by admitting any inconsistency, the entire model would be rejected. Stated bluntly, those scientists had too much focus on the notion of DATA = MODEL. Gould (1996) provides a detailed history of how such model-confirmation biases and overlooked residuals led to centuries of unfortunate categorization of humans.

The Two-Way Fit

To illustrate the generality of the model-residual view of EDA, we will consider the extremely useful and flexible

model of the two-way fit introduced by Tukey (1977) and Mosteller and Tukey (1977). The *two-way fit* is obtained by iteratively estimating row effects and column effects and using the sum of those estimates to create predicted (model or fit) cell values and their corresponding residuals. The cycles are repeated with effects adjusted on each cycle to improve the model and reduce residuals until additional adjustments provide no improvement. This procedure can be applied directly to data with two-way structures. More complicated structures can be modeled by multiple two-way structures. In this way, the general approach can subsume such approaches as the measures of central tendency in the ANOVA model, the ratios in the log-linear model, and person and item parameter estimates of the one-parameter item response theory model.

Consider the data presented in Table 2.1. It represents average effect sizes for each of a series of univariate analyses conducted by Stangor and McMillan (1992). Such a display is a common way to communicate summary statistics. From an exploratory point of view, however, we would like to see if some underlying structure or pattern can be discerned. Reviewing the table, it is easy to notice that some values are negative and some positive, and that the large number of -2.6 is a good bit larger than most of the other numbers which are between 0 and $+/-$ 1.0.

To suggest an initial structure with a two-way fit we calculate column effects by calculating the median of each column. The median of each column then becomes the model for that column, and we subtract that initial model estimate from the raw data value to obtain a residual that replaces the original data value in the data matrix. After this simple first pass, we have a new table in which each cell is a residual and the data from the original table are equal to the column effect plus the cell residual. The row effects are estimated next by calculating the median value of residuals in each row and subtracting the cell values (first-stage residuals) from these medians. The row effects are generally placed in the margin of the table and the new residuals replace the residuals from the previous stage. A similar calculation occurs on the row of medians that represents the column effects; the median of the column effects becomes the estimate of the overall or "grand" effect and the estimates of the column effects are likewise adjusted through subtraction.

This process is repeated iteratively until continued calculation of effects and residuals provides no improvement. The result of such a summary is provided in Table 2.2. Here we see an overall effect of $-.02$ as well as a characterization of each row and column. It is clear which columns are low, medium, and high, and likewise which rows stand out. Each cell in the original table can be reconstructed using the formula Data = grand effect + row effect + column effect + residual. For example, the memorization task for the bias condition can be recreated using the model of $1.01 = -0.02 + 0.69 + 0.14 - 0.02$.

To form a visual impression of these values, Tukey (e.g., Mosteller & Tukey, 1977) recommended a two-way fit plot such as that shown in Figure 2.12, panel A. In this figure,

TABLE 2.1 Average Effect Sizes by Dependent Variable and Study Characteristic. From Stangor and McMillan (1992).

Variable	Recall	Recognition	Bias
Strength of expectations			
a. Experimental session	−0.37	−0.47	0.32
b. Existing	0.32	−0.8	0.93
Content of the stimuli			
c. Behaviors	−0.21	−0.1	0.66
d. Traits	0.71	−2.16	1.98
Type of behavioral inconsistency			
e. Evaluative and descriptive	−0.27	0.1	0.29
f. Descriptive only	0.36	−0.54	0.85
Type of target			
g. Individual	−0.32	−1.14	1.04
h. Group	0.22	−0.38	0.33
Processing goal			
i. From impressions	−0.46	0.19	0.57
j. Memorize	0.12	−0.71	1.01
Interpolated task			
k. No	−0.44	−0.30	0.62
l. Yes	0.06	−1.26	0.75
Type of delay			
m. Within single session	−0.19	−0.65	0.82
n. Separate session	−0.02	−0.03	0.66

TABLE 2.2 Two-Way Decomposition of Average Effect Sizes by Dependent Variable and Study Characteristic. From Stangor and McMillan (1992).

Variable	Recall	Recognition	Bias	Row Effect
Strength of expectations				
a. Experimental session	0.00	0.43	0.00	−0.35
b. Existing	0.08	−0.51	0.00	0.26
Content of the stimuli				
c. Behaviors	−0.18	0.46	0.00	−0.00
d. Traits	0.00	−2.34	0.58	0.73
Type of behavioral inconsistency				
e. Evaluative and descriptive	0.00	0.90	−0.13	−0.25
f. Descriptive only	0.20	−0.17	0.00	0.19
Type of target				
g. Individual	0.00	−0.29	0.67	−0.30
h. Group	0.07	0.00	−0.51	0.17
Processing goal				
i. From impressions	−0.34	0.84	0.00	−0.10
j. Memorize	0.00	−0.30	0.20	0.14
Interpolated task				
k. No	−0.37	0.30	0.00	−0.05
l. Yes	0.00	−0.79	0.00	0.08
Type of delay				
m. Within single session	−0.07	0.00	0.25	−0.10
n. Separate session	0.00	0.52	−0.01	0.00
Column effects	0.00	−0.53	0.69	−0.02

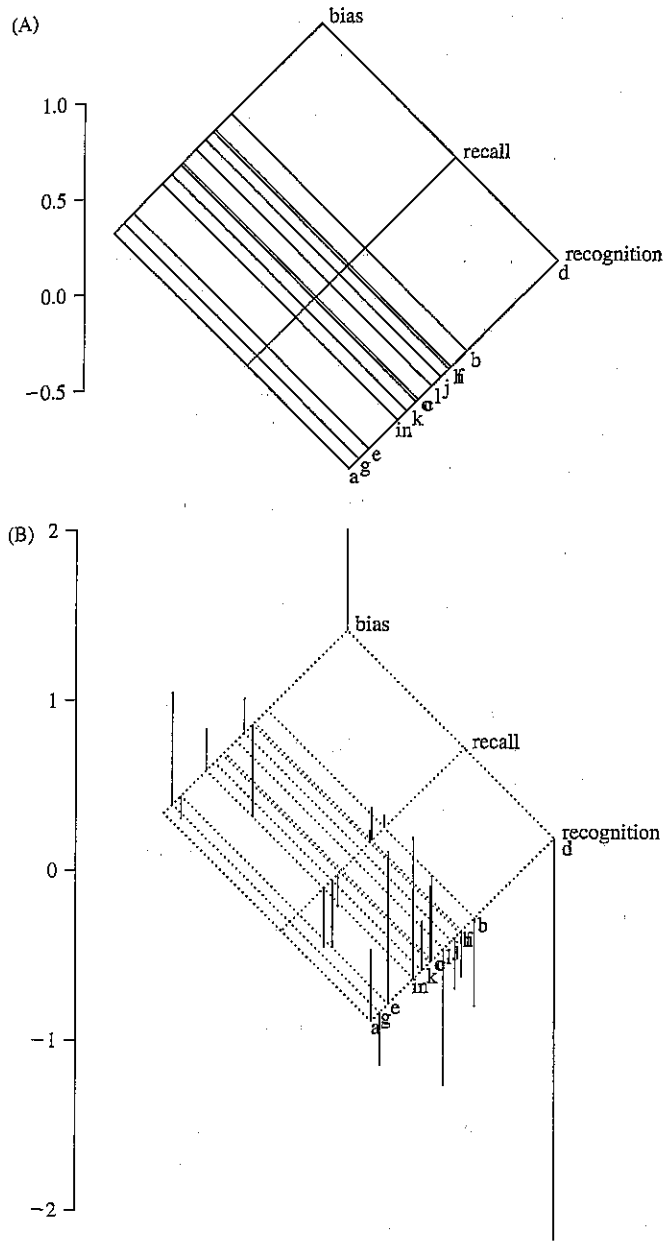

Figure 2.12 Panel A is a plot of main effects of a two-way fit. The height of the intersection of lines represents the sum of row, column, and grand effects and the corresponding predicted value for that cell. Panel B shows main effects of two-way fit with additional lines extending from predicted to actual values to highlight the location and size of residuals.

there is a line for each row and column effect in the model. For each row and column effect, the height of the intersection of the two lines is equal to the value of the predicted value in the model (overall effect + column effect + row effect). This plot clearly portrays the separation of the bias, recall, and recognition conditions, shows the clear separation of row d (trait levels), and displays a cluster of common small effects for rows a, g, and e. For us, this view was surprising because when we first characterized row d, it was with a focus on the large −2.16 value, which is the largest (negative) value in the table. This graphic, however, suggests more needs to be considered. Reexamining the data for that row we see that not only does that row have the largest negative value, but also two of the largest positive values. All together, we end up with a strong positive row effect. For individual cells, however, the effect may be low or high, depending on the column.

Because the grand, row, and column effects represent the model, an assessment of that model requires an examination of where the model fits and does not fit. The residuals for these models are presented in the center of Table 2.2. Examination of these values reveals that the extreme value observed in the raw data remains extreme in this model, and that this value is not simply the result of combining row and column effects. A graphical analog to the residuals can also be provided, as shown in Figure 2.12, panel B. In this diagram, lines are drawn from the value of the predicted values (the intersection of row and column lines), downward or upward to the actual data value. The length of each line thereby indicates the size of the residual. Clearly, the size of the trait residual for recognition tasks dwarfs the size of all other effects and residuals in the data. Other patterns of residuals may provide additional information about the data because they tell us what departs from a standard description.

In this example we used simple raw residuals. In other applications, the actual value of residuals may be modified in a number of ways. One common method is to report residuals reexpressed as *normal-deviates* in the distribution of residuals. This approach, often used in structural equation analysis, can help identify the locations of the extremes, but hides the scale values of the error. In the highly developed area of regression diagnostics, residuals may be adjusted for the size of the leverage associated with the value of the criterion variable (*studentized residuals*) or calculated using a model that obtained predicted values without the presence of the observation in the model (externally studentized). This prevents extreme values from distorting the model to the point that an aberrant value leads to a small residual, as displayed in panel C of Figure 2.2.

As illustrated previously, complementary to the notion of patterns of residuals and meaning in individual residuals is the emphasis on mathematical models of effects that provide rich and parsimonious description. This view is very much in line with the recently emerging view of the importance of effect sizes suggested by Glass (1976) and renewed by Cohen (1994) and the APA Task Force on Statistical Inference (Wilkinson, 1999). EDA reminds us that at the same time we focus on effects as a description of the data, we must also focus on the size and pattern of misfits between effects and the data.

Reexpression

The data examined previously remind us that data often come to the exploratory data analyst in messy and nonstandard ways. This should not be unexpected, given the common assumption that the data distributions are either always well behaved, or that statistical techniques are sufficiently robust that we can ignore any deviations that might arise, and therefore skip detailed examination. In fact, it is quite often the case that insufficient attention has been paid to scaling issues in advance, and it is not until the failure of confirmatory methods that a careful examination of scaling is undertaken (if at all). In the exploratory mode, however, appropriate scaling is considered one of the fundamental activities and is called *reexpression*. Although mathematically equivalent to what is called *transformation* in other traditions, reexpression is so named to reflect the idea that the numerical changes are aimed at appropriate scaling rather than radical change.

Because reexpression requires an understanding of the underlying meaning of the data that are being reexpressed, the EDA approach avoids using the common categorizations of data as nominal, ordinal, interval, and ratio that follow Stevens (e.g., 1951). Rather, Mosteller and Tukey (1977) discussed broad classes of data as (a) amounts and counts; (b) balances (numbers that can be positive or negative with no bound); (c) counted fractions (ratios of counts); (d) ranks; and (e) grades (nominal categories).

When dealing with common amounts and counts, Tukey suggested heuristics that hold that (a) data should often be reexpressed toward a Gaussian shape, and (b) an appropriate reexpression can often be found by moving up or down "the ladder of reexpression." A Gaussian shape is sought because this will generally move the data toward more equal-interval measurement through symmetry, will often stabilize variance, and can quite often help linearize trend (Behrens, 1997a). In EDA, the term *normal* is avoided in favor of *Gaussian* to avoid the connotation of prototypicality or social desirability.

The ladder of reexpression is a series of exponents one may apply to original data that show considerable skew. Recognizing that the raw data exists in the form of X^1, moving up the ladder would consist of raising the data to X^2 or X^3. Moving down the ladder suggests changing the data to the scale of $X^{1/2}$, $-X^{-1/2}$, $-X^{-1}$, $-X^{-2}$, and so on. Because X^0 is equal to 1, this position on the ladder is generally replaced with the reexpression of $\log_{10}(X)$. To choose an appropriate transformation, one moves up or down the ladder toward the bulk of the data. This means moving down the ladder for distributions with positive skew and up the ladder for distributions with negative skew. By far the most common reexpression for positively skewed data is the logarithmic transformation. For ratios of counts, the most common recommendation is to "fold" the counts around a midpoint (usually .5) so that equal fractions equal 0. This generally means using $P/1 - P$, where P is the proportion of the total that the count comprises. A second step is to take the log of this folded fraction to create a "flog" equal to $\log(P/1 - P)$. In more common parlance, this is a logit that serves as the basis for logistic regression, survival, or event-history analysis, and measurement via item response theory. Additional techniques recommend that balances should generally be left alone whereas grades and ranks should be treated much like counted fractions (see, e.g., Mosteller & Tukey, 1977).

Although reexpression is a long-standing practice in the statistical community, going back at least to Fisher's (1921) construction of the r to z transformation, only recently has its use become more widespread in psychological literature. In fact, it often continues to arise more out of historic tradition than as the result of careful and comprehensive analysis. Consider, for example, the subset of data from a word-recognition experiment recently reported by Paap, Johansen, Chun, and Vonnahme (2000) and depicted in Figure 2.13. The experiment reported in this paper concerns the percentage of times participants correctly identify word pairs (%C) from a memory task as a function of the word pair's correct-incorrect confusability (CIC), percentage correct-letter distinctiveness (CD), number of neighbors (N), percentage of friends in the lexical neighborhood (%F), number of higher frequency neighbors (H), log of frequency of the test word (LTF), and log of frequency formed by incorrect alternative (LAF).

As the reader may see, although the distributions associated with the logarithmic reexpression are quite Gaussian, the variables that are not reexpressed differ quite a bit in this respect and lead to quite non-Gaussian bivariate distributions. The CIC variable is the most skewed. This leads to quite distorted correlations that would suffer from variance compression. Distributional outliers in CIC are indicated with X symbols, and regression lines on the %C against CIC scatter plot are present both for all data (lower line) and for the data with the outliers removed (sloped line).

Because these authors have already reexpressed the two frequency (count) variables, it will be useful to reverse the reexpression to see the original data, which are presented in Figure 2.14. The top panels show the histograms of the original raw data along with the *quantile-quantile (QQ) plot,* which is often called the *normal-probability plot* in cases like these because the ranks of the data are plotted against the z scores of corresponding ranks in a unit-normal distribution. When the points of the QQ plot are straight, this reflects the

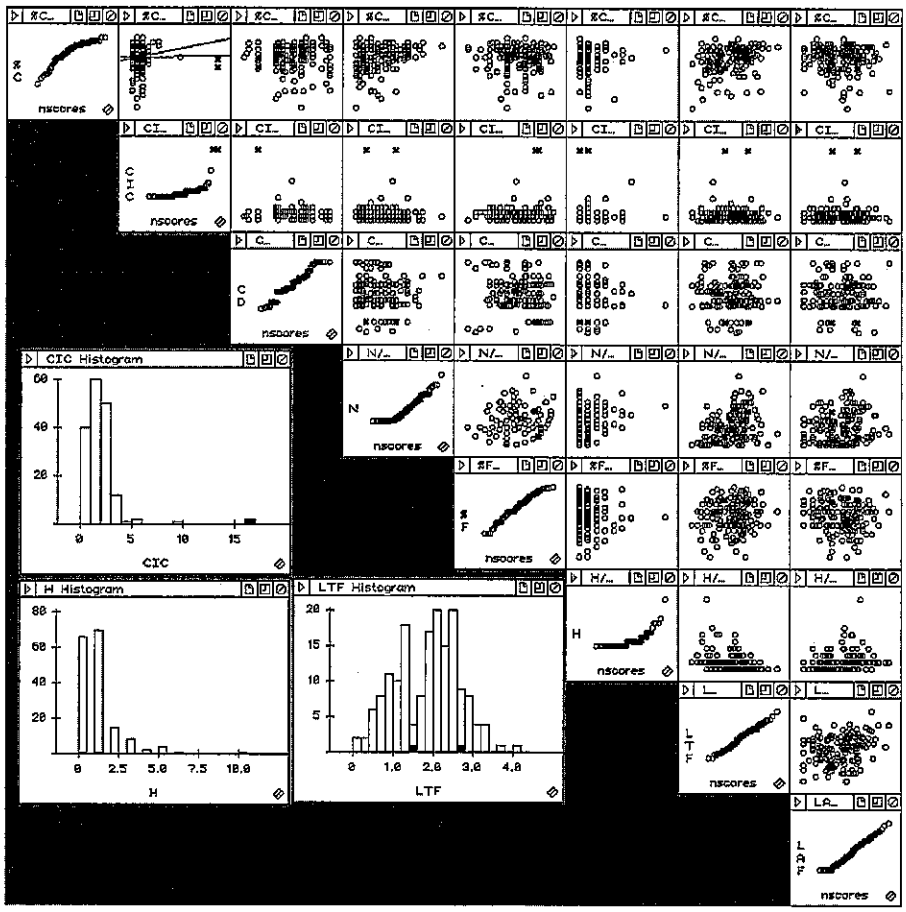

Figure 2.13 Scatter plot matrix of reaction time data from Paap, Johansen, Chun, and Vonnahme (2000). Patterns in the matrix reflect skew and outliers in some marginal distributions as well as nonlinearity in bivariate distributions. Outliers are highlighted in all graphics simultaneously through brushing and linking.

match between the empirical and theoretical distributions, which in this case is Gaussian. The data running along the bottom of these displays reflect the numerous data values at the bottom of the scales for the original frequency data. Panels E and F show the scatter plot of the raw data before reexpression and the corresponding simple regression residuals, which indicate that the spread of the error is approximately equal to the spread of the data. Although this logarithmic transformation is quite appropriate, it was chosen based upon historical precedent with data of this type rather than on empirical examination. Accordingly, in the view of EDA, the outcome is correct while the justification is lacking.

Turning to how remaining variables may be improved, we consider the four percentage variables, especially the highly distorted CIC variable. Working directly with percentages can be quite misleading because differences in values are not equally spaced across the underlying continuum. For example, it is generally easier to move from an approval rating of 50 to 55% than it is to move from 90 to 95%. All content aside, the variance for the underlying binomial distribution is largest around .5 and smallest near 0 and 1. As noted above, this mathematical situation leads to a general rule for substituting logits (a.k.a., flogs) for raw percentages. Accordingly, we move forward by converting each percentage into a proportion (using the sophisticated process of dividing by 100) and constructing the logit for each proportion. The effect of this reexpression on the %C and CIC variables is portrayed in Figure 2.15. As the reader can see, the distributions are greatly moved toward Gaussian, and the appearance of the scatter plot changes dramatically.

The impact of this reexpression is considerable. Using the correlation coefficient in the original highly skewed and variance-unstable scale of percentages resulted in a measure of association of $r = .014$, suggesting no relationship between these two values. However, when the scale value and corresponding variance are adjusted using the logistic reexpression, the measure of relationship is $r = .775$—a dramatic difference in impression and likely a dramatic effect on

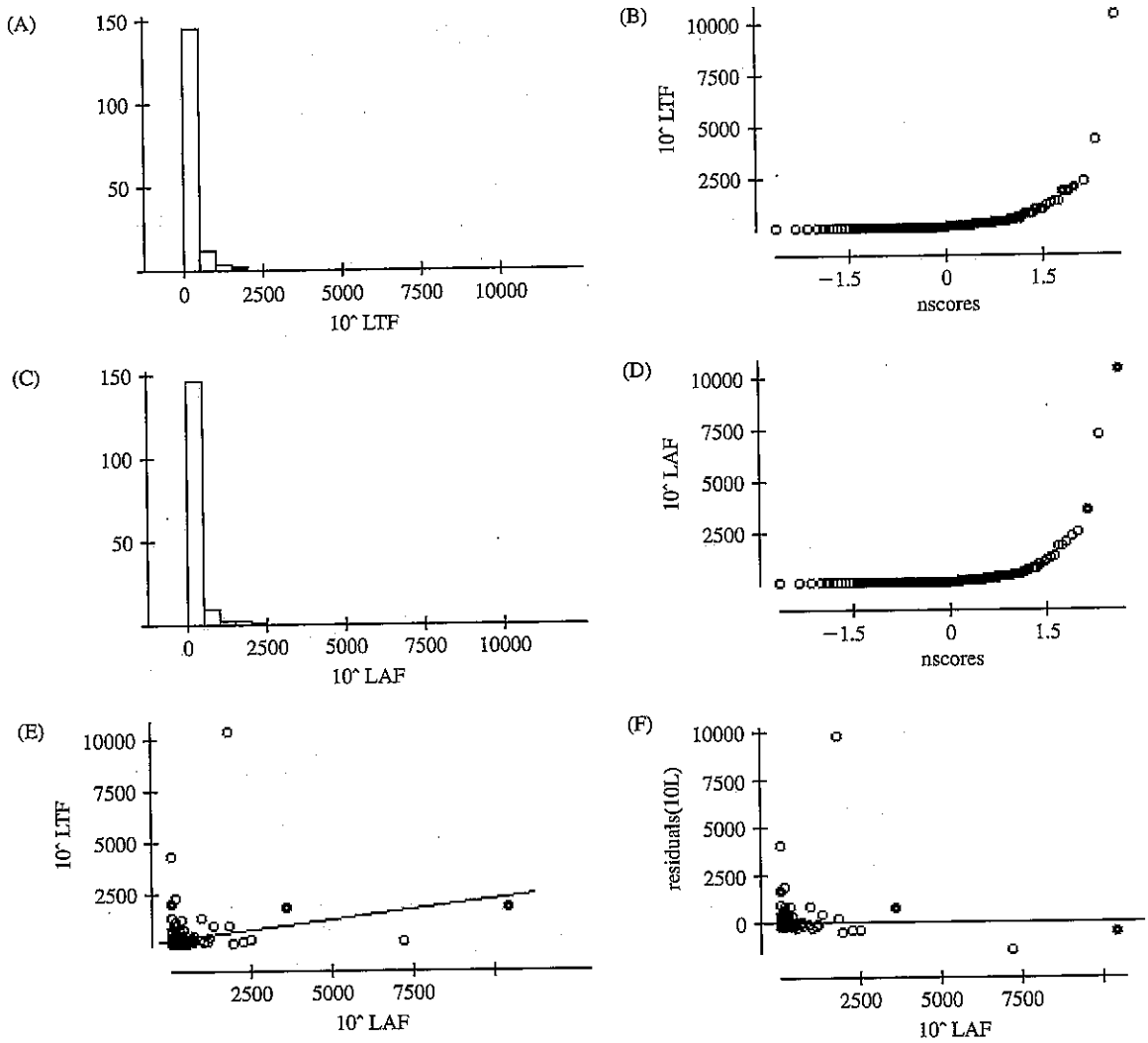

Figure 2.14 Histograms, normal-probability plots, and scatter plots for reaction time data as they would appear without the logarithmic reexpression used by the original authors.

theory development and testing. In a similar vein, Behrens (1997a) demonstrated how failure to appropriately reexpress similar data led Paap and Johansen (1994) to misinterpret the results of a multiple regression analysis. As in this case, a simple plotting of the data reveals gross violations of distributional assumptions that can lead to wildly compressed correlations or related measures.

The H variable is likewise of interest. Because it is a count, general heuristics suggest a square-root or logarithmic reexpression. Such reexpressions, however, fail to improve the situation substantially, so another course of action is required. Because the H variable is a count of high-frequency neighbors, and this number is bounded by the number of neighbors that exist, a logical alternative is to consider H as the proportion of neighbors that are high frequency rather than the simple count. When such a proportion is computed and converted to a logit, the logit-H variable becomes very well behaved and leads to much clearer patterns of data and residuals. The revised scatter plot matrix for these variables is presented in Figure 2.16. As the reader may see, a dramatic improvement in the distributional characteristics has been obtained.

Although some researchers may reject the notion of reexpression as "tinkering" with the data, our experience has been that this view is primarily a result of lack of experience with the new scales. In fact, in many instances individuals often use scale reexpressions with little thought. For example, the common practice of using a proportion is seldom questioned, nor is the more common reexpression to z scores. In daily language many people have begun to use the $\log_{10}$

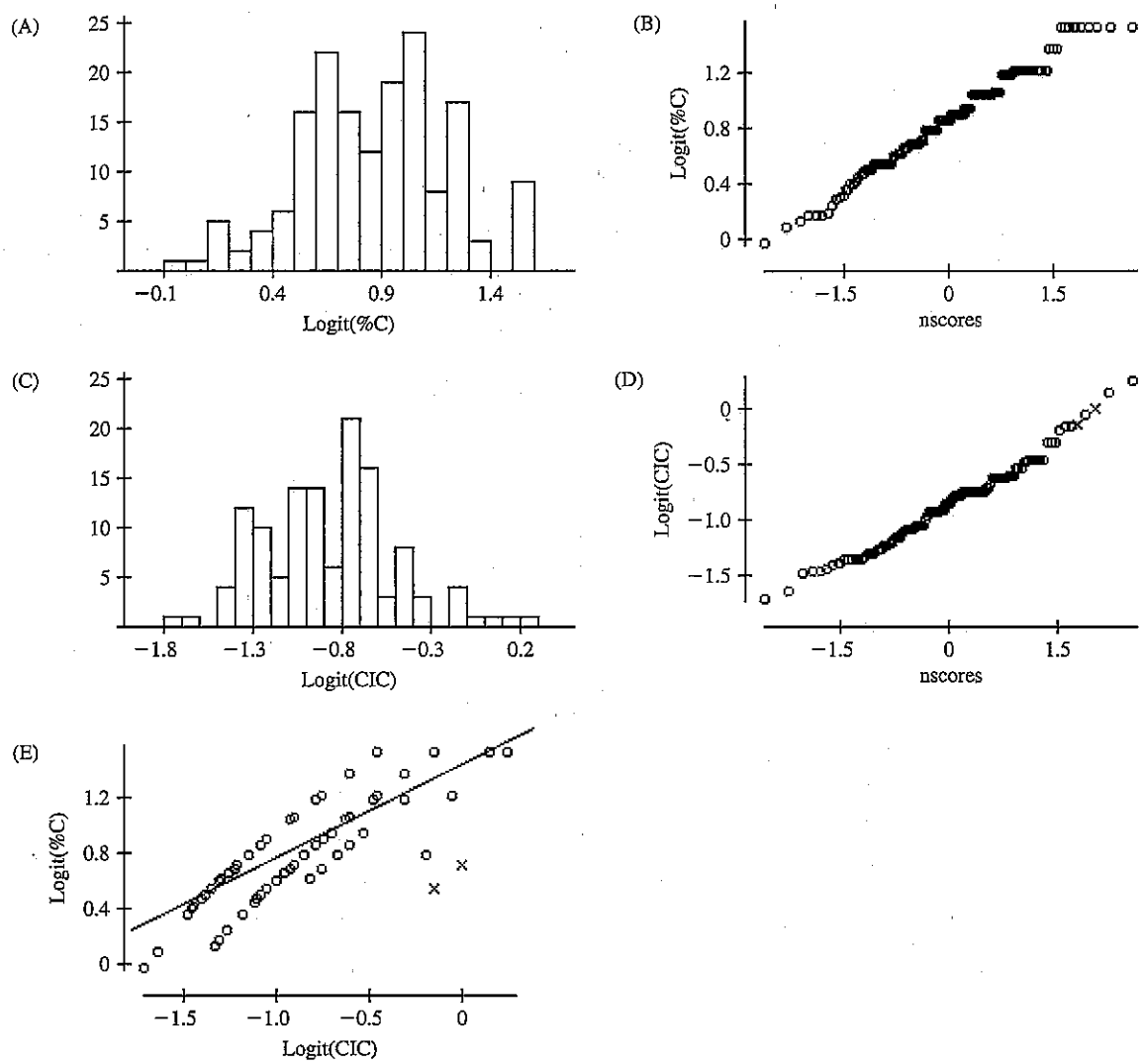

Figure 2.15 Histograms, normal-probability plots, and scatter plot for percent correct (%C) and percent congruent-incongruent ratio (CIC) following logistic reexpression.

reexpression of dollar amounts as "five-figure," "six-figure," or "seven-figure." Wainer (1977) demonstrated that the often recommended reexpression of $1/X$ for reaction-time tasks simply changes the scale from seconds per decision (time) to decisions per second (speed). Surely such tinkering can have great value when dramatic distributional improvements are made and sufficient meaning is retained.

Resistance

Because a primary goal of using EDA is to avoid being fooled, resistance is an important aspect of using EDA tools. *Resistant methods* are methods that are not easily affected by extreme or unusual data. This value is the basis for the general preference for the median rather than the mean. The mean has a smaller standard error than the median, and so is an appropriate estimator for many confirmatory tests. On the other hand, the median is less affected by extreme scores or other types of perturbations that may be unexpected or unknown in the exploratory stages of research.

In general, there are three primary strategies for improving resistance. The first is to use rank-based measures and absolute values, rather than measures based on sums (such as the mean) or sums of squares (such as the variance). Instead, practitioners of EDA may use the tri-mean, which is the average of Q1, Q3, and the median counted twice. For measures of spread, the interquartile range is the most common, although the median absolute deviation (MAD) from the median is available as well. The second general resistance-building strategy is to use a procedure that emphasizes more

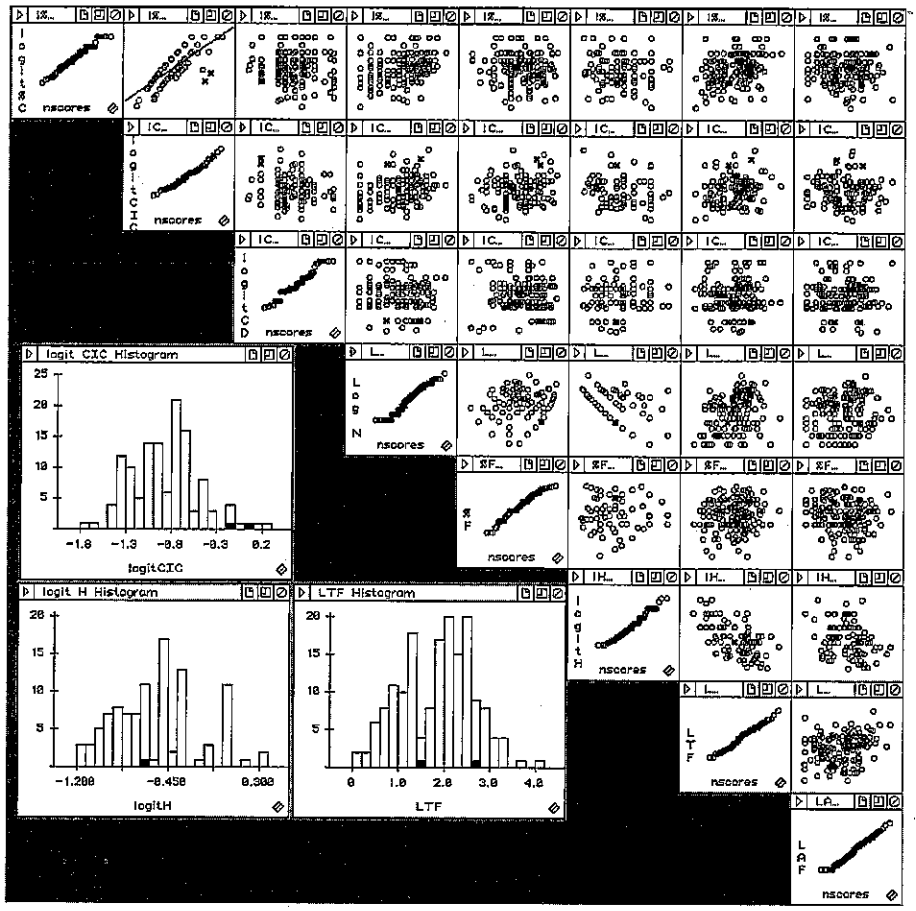

Figure 2.16 Scatter plot matrix of reaction time data after full set of reexpressions. When compared with the original data shown in Figure 2.13, this matrix reflects the improved ability to model the data using linear models.

centrally located scores and that uses less weight for more extreme values. This category includes trimmed statistics in which values past a certain point are weighted to 0 and thereby dropped from any estimation procedures. Less drastic approaches include the use of the biweight, in which values are weighted as an exponential function of their distance from the median. A third approach is to reduce the scope of the data one chooses to model on the basis of knowledge about extreme scores and the processes they represent.

Dealing with Outliers

The goal of EDA is to develop understandings and descriptions of data. This work is always set in some context and always presents itself with some assumptions about the scope of the work, even when these assumptions are unrecognized. Consider, for example, the task described in Behrens and Smith (1996) of developing a model of state-level economic aspects of education in the United States. In this analysis, simple use of a scatter plot matrix revealed three consistent outliers in distributions of variables measured in dollars. The first outlier was the observation associated with the District of Columbia. How should this extreme value be approached? If the original intention was state-level analysis, the outlier in the data simply calls attention to the fact that the data were not prescreened for non-states. Here the decision is easy: Reestablish the scope of the project to focus on state-level data.

The remaining two outliers were observations associated with the states of Hawaii and Alaska. These two states had values that were up to four times higher than the next highest values from the set of all states. In many cases, the mean of the data when all 50 states were included was markedly different from the mean computed using values from only the contiguous 48 states. What should the data analyst do about this problem? Here again, the appropriate role of the exploratory work has led to a scope-clarification process that many data analysts encounter in the basic question *Do I model all the data poorly, or do I model a specific subset that I can*

describe well? Although this question needs to be answered on a case-by-case basis, in the situation described here there is little doubt. Alaska and Hawaii should be set aside and the researcher should be content to construct a good model for the 48 contiguous states. Furthermore, the researcher should note that he or she has empirical evidence that Alaska and Hawaii follow different processes. This is a process of setting aside data and focusing scope. Clearly this process of examination and scope revision would need to be reported.

In this case, the rationale is clear and the data have semantic clarity. In other cases, however, quite extreme values may be found in data that are not simply victims of poor measurement models (e.g., the end of a long tail awaiting logarithmic reexpression). Under these circumstances the fundamental question to ask is *Do we know something about these observations that suggest they come from a different process than the process we are seeking to understand?* In experimentally oriented psychology, rogue values could be caused by numerous unintended processes: failure to understand instructions (especially during opening trials), failure to follow the instructions, failure to pay attention to the task (especially during closing trials), or equipment or data transcription failures. Under such circumstances, it is clear the data are not in the domain of the phenomenon to be studied, and the data should be set aside and the situation noted.

In other cases, extreme values present themselves with little auxiliary information to explain the reason for the extremeness. In such a situation we may first assess how much damage the values create in the model by constructing the model with all the data involved as well as with the questionable data set aside. For example, Behrens (1997b) conducted a meta-analysis of correlations between subscales of the White Racial Identity Attitude Scale (Helms, 1997). Initial review of the data suggested the distributions were not homogeneous and that some study results differed dramatically from the average. To assess the effect of these extreme values, Behrens calculated the average correlations, first, using all the data, and second, using a 20% trimmed mean. Results were consistent across approaches, suggesting the data could remain or be set aside with little impact on the inferences he was to make. What would have happened if, on the other hand, the trimmed results deviated from the full-data results? In such a case both sets of analysis should be conducted and reported and the difference between the two results considered as a measure of the effect of the rogue values. The most important aspect in either case is that a careful and detailed description of the full data, the reduced data, and the impact of the rogue data be reported. Unfortunately, the extremely terse and data-avoidant descriptions of much research reporting is inconsistent with this highly descriptive approach.

Summary

The tools described in this section are computational and conceptual tools intended to guide the skilled and skeptical data analyst. These tools center on the four *R*s of revelation, residuals, reexpression, and resistance. The discussion provided here provides only a glimpse into the range of techniques and conceptualizations in the EDA tradition. In practice, the essential elements of EDA center on the attitudes of flexibility, skepticism, and ingenuity, all employed with the goals of discovering patterns, avoiding errors, and developing descriptions.

CURRENT COMPUTING AND FUTURE DIRECTIONS

Computing for EDA

While EDA has benefited greatly from advances in statistical computing, users are left to find the necessary tools and appropriate interfaces spread across a variety of statistical packages. To date, the software most clearly designed for EDA is Data Desk (Data Description, 1997). Data Desk is highly interactive and completely graphical. Graphical windows in Data Desk do not act like "static" pieces of paper with graphic images, but rather consist of "live" objects that respond to brushing, selecting, and linking. The philosophy is so thoroughly graphical and interactive that additional variables can be added to regression models by dragging icons of the variables onto the regression sums-of-squares table. When this occurs, the model is updated along with all graphics associated with it (e.g., residual plots). Data Desk has a full range of EDA-oriented tools as well as many standards including multivariate analysis of variance (MANOVA) and cluster analysis—all with graphical aids. Because of its outstanding emphasis on EDA, Data Desk has been slower to develop more comprehensive data management tools, as has been the case with more popular tools like those offered by SAS, SPSS, and SYSTAT.

Both SAS (SAS Institute, 2001) and SPSS (SPSS, Inc., 2001) have improved their graphical interactivity in recent years. The SAS Insight module allows numerous interactive graphics, but linking and other forms of interaction are less comprehensive than those found in Data Desk. SYSTAT serves as a leader in graphical and exploratory tools, placing itself between the complete interactivity of Data Desk and the more standard approach of SAS and SPSS.

The *S* language and its recent incarnation as S-PLUS (Insightful, 2001) has long been a standard for research in statistical graphics, although it appears also to be emerging as

a more general standard for statistical research. Venables and Ripley's 1999 book titled *Modern Applied Statistics with S-PLUS* (often called MASS) provides an outstanding introduction to both the Splus language and many areas of modern statistical computing of which psychologists are often unaware, including projection pursuit, spline-based regression, classification and regression trees (CART), *k*-means clustering, the application of neural networks for pattern classification, and spatial statistics.

High-quality free software is also available on the Internet in a number of forms. One consortium of statisticians has created a shareware language, called *R*, that follows the same syntax rules as S-PLUS and can therefore be used interchangeably. A number of computing endeavors have been based on Luke Tierney's XLISP-STAT system (Tierney, 1990), which is highly extensible and has a large object-oriented feature set. Most notable among the extensions is Forrest Young's (1996) ViSta (visual statistics) program, which is also free on the Internet.

Despite the features of many of these tools, each comes with weaknesses as well as strengths. Therefore, the end user must have continuing facility in several computing environments and languages. The day of highly exchangeable data and exchangeable interfaces is still far off.

Future Directions

The future of EDA is tightly bound to the technologies of computing, statistics, and psychology that have supported its growth to date. Chief among these influences is the rise of network computing. Network computing will bring a number of changes to data analysis in the years ahead because a network allows data to be collected from throughout the world, allows the data to have extensive central or distributed storage, allows computing power a distributed or centralized location, and allows distribution of results quickly around the world (Behrens, Bauer, & Mislevy, 2001). With regard to the increase in data acquisition, storage, and processing power, these changes will lead to increasing availability and need for techniques to deal with large-scale data. With increasingly large data sets, data analysts will have difficulty gaining detailed familiarity with the data and uncovering unusual patterns. Hopefully, the capacity for ingenuity and processing power will keep up with the increase in data availability.

Data Projections

Currently, most existing visualization tools are based upon *variable space,* in which data points are depicted within the Cartesian coordinates. With the advent of high-powered computing, more and more statistical software packages incorporate graphical tools that utilize other spatial systems. For example, several statistical packages implement the *biplot* (Gabriel, 1981; Gower & Hand, 1996), which combines variable space and subject space (also known as *vector space*). In *subject space* each subject becomes a dimension, and vectors are displayed according to the location of variables mapping into this subject space. In addition, SYSTAT implements additional projections, including triangular displays, in which both the Cartesian space and the barycentric space are used to display four-dimensional data.

Interactive methods for traversing multivariate data have been developed as well. Swayne, Cook, and Buja (1998) developed the X-GOBI system, which combines the high-dimensional search techniques of projection pursuit (Friedman & Tukey, 1974) and grand tour (Asimov, 1985). The *grand tour* strategy randomly manipulates projections for high-dimensional data using a biplot or similar plotting system, so that the user has an experience of touring "around" three- (or higher) dimensional rotating displays. *Projection pursuit* is a computing-intensive method that calculates an "interestingness function" (usually based on nonnormality) and develops search strategies over the multidimensional gradient of this function. Grand tour can provide interesting views but may randomly generate noninteresting views for quite some time. Projection pursuit actively seeks interesting views but may get caught in local minima. By combining these two high-dimensional search strategies and building them into a highly interactive and visual system, these authors leveraged the best aspects of several advanced exploratory technologies.

Data Immersion

To deal with the increasing ability to collect large data sets, applications of EDA are likely to follow the leads developed in high-dimensional data visualization used for physical systems. For example, orbiting satellites send large quantities of data that are impossible to comprehend in an integrated way without special rendering. To address these issues, researchers at the National Aeronautics and Space Administration (NASA) have developed tools that generate images of planetary surface features that are rendered in three-dimensional virtual reality engines. This software creates an imaginary topology from the data and allows users to "fly" through the scenes.

Although most psychologists are unlikely to see such huge (literally astronomical!) quantities of data, desktop computers can provide multimedia assistance for the creation of interactive, three-dimensional scatter plots and allow the animation of multidimensional data (e.g., Yu & Behrens, 1995).

Distributed Collaboration

Because data analysis is a social process and groups of researchers often work together, EDA will also be aided by the development of computer–desktop sharing technologies. Internetworking technologies currently exist that allow individuals to share their views of their computer screens so that real-time collaboration can occur. As statistical packages become more oriented toward serving the entire data-analytic process, developers will consider the social aspects of data analysis and build in remote data-, analysis-, image-, and report-sharing facilities. Such tools will help highly trained data analysts interact with subject-matter experts in schools, clinics, and businesses.

Hypermedia Networks for Scientific Reporting

While the natures of scientific inquiry, scientific philosophy, and scientific data analysis have changed dramatically in the last 300 years, it is notable that the reporting of scientific results differs little from the largely text-based and tabular presentations used in the eighteenth century. Modern print journals, under tight restrictions for graphics and space, have largely omitted the reporting of exploratory results or detailed graphics. Although a textual emphasis on reporting was necessary for economic reasons in previous centuries, the rise of network-based computing, interactive electronic information display, and hypertext documents supports the expansion of the values of EDA in scientific reporting. In a paper-based medium, narrative development generally needs to follow a linear development. On the other hand, in a hypertext environment the textual narrative can appear as traditionally implemented along with auxiliary graphics, detailed computer output, the raw data, and interactive computing—all at a second level of detail easily accessed (or ignored) through hypertext links. In this way, the rich media associated with EDA can complement the terse reporting format of the American Psychological Association and other authoring styles (Behrens, Dugan, & Franz, 1997).

To illustrate the possibility of such a document structuring, Dugan and Behrens (1998) applied exploratory techniques to reanalyze published data reported in hypertext on the World Wide Web. Figure 2.17 is an image of the interface

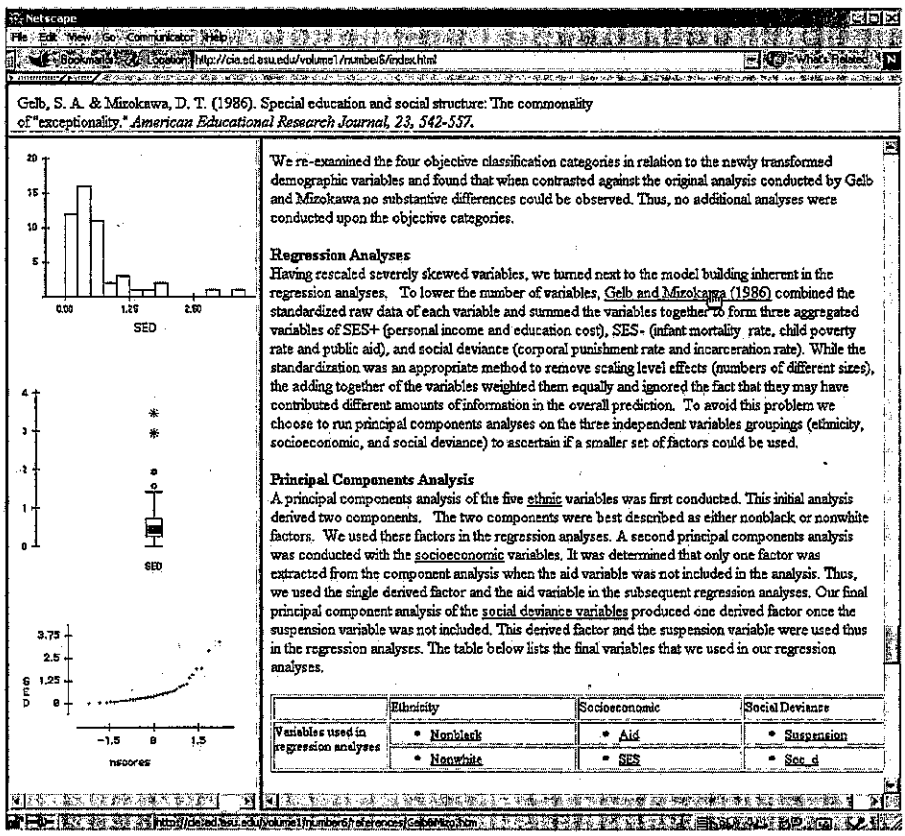

Figure 2.17 Screen appearance of electronic document created from EDA perspective by Dugan and Behrens (1998).

used by these authors. The pages were formatted so that the names of variables were hyperlinked to graphics that appeared on a side frame, and the names of references were hyperlinked to the references that appeared on the top frame of the page. References to F tests or regression-results linked to large and well-formatted result listings, and the data were hyperlinked to the paper as well.

While we wait for arrival of widespread hypertext in scientific journals, personal Web sites for improving the reporting of results can be used. For example, Helms (1997) criticized the analysis of Behrens (1997b), which questioned the psychometric properties of a commonly used scale in the counseling psychology racial-identity literature. Part of the concern raised by Helms was that she expected large amounts of skew in the data, and hence, likely violations of the statistical assumptions of the meta-analyses and confirmatory factor analyses that Behrens (1997b) reported. In reply, Behrens and Rowe (1997) noted that the underlying distributions had been closely examined (following the EDA tradition) and that the relevant histograms, normal-probability plots, scatter plot matrices (with hyperlinks to close-up views), and the original data were all on the World Wide Web (Behrens & Dugan, 1996). This supplemental graphical archive included a three-dimensional view of the data that could be navigated by users with Web browsers equipped with commonly available virtual-reality viewers. Such archiving quickly moves the discussion from impressions about possibilities regarding the data (which can be quite contentious) to a simple display and archiving of the data.

Summary

Emerging tools for EDA will continue to build on developments in integration of statistical graphics and multivariate statistics, as well as developments in computer interface design and emerging architectures for collecting, storing, and moving large quantities of data. As computing power continues to increase and computing costs decrease, researchers will be exposed to increasingly user-friendly interfaces and will be offered tools for increasingly interactive analysis and reporting. In the same way that creating histograms and scatter plots is common practice with researchers now, the construction of animated visualizations, high-dimensional plots, and hypertext reports is expected to be commonplace in the years ahead. To offset the common tendency to use new tools for their own sake, the emergence of new technologies creates an increased demand for researchers to be trained in the conceptual foundations of EDA. At the same time, the emergence of new tools will open doors for answering new scientific questions, thereby helping EDA evolve as well.

CONCLUSION

Despite the need for a wide range of analytic tools, training in psychological research has focused primarily on statistical methods that focus on confirmatory data analysis. Exploratory data analysis (EDA) is a largely untaught and overlooked tradition that has great potential to guard psychologists against error and consequent embarrassment. In the early stages of research, EDA is valuable to help find the unexpected, refine hypotheses, and appropriately plan future work. In the later confirmatory stages, EDA is valuable to ensure that the researcher is not fooled by misleading aspects of the confirmatory models or unexpected and anomalous data patterns.

There are a number of missteps the reader can make when faced with introductory materials about EDA. First, some readers may focus on certain aspects of tradition and see their own activity in that area as compelling evidence that they are already conducting EDA. Chief among these aspects is the use of graphics. By showing a broad range of graphics, we sought to demonstrate to the reader that statistical graphics has become a specialization unto itself in the statistics literature, and that there is much to learn beyond what is commonly taught in many introductory courses. Whereas the exploratory data analyst may use graphics, the use of graphics alone does not make an exploratory data analyst.

A second pitfall the reader should be careful to avoid is rejecting the relevance of the examples used in this chapter. Some might argue that the pathological patterns seen herein exist only in data from meta-analyses or reaction time experiments, or in educational data. Our own work with many types of data sets, and in conversations with psychologists in numerous specializations, suggests that these are not isolated or bizarre data sets but are quite common patterns. The reader is encouraged to reanalyze his or her own data using some of the techniques provided here before making judgments about the prevalence of messy data.

A third pitfall to avoid is overlooking the fact that embracing EDA may imply some confrontation with traditional values and behaviors. If EDA is added to the methodology curriculum then other aspects may need to be deemphasized. If new software is desired, changes in budgets may need to occur, with their associated social conflicts. Additionally, conflict may arise within the researcher as he or she works to balance the value of EDA for scientific advancement while finding little explicit value for EDA in manuscript preparation.

Psychological researchers address complex and difficult problems that require the best set of methodological tools available. We recommend EDA as a set of conceptual and

computational tools to supplement confirmatory statistics, and expect psychological research will increase in efficiency and precision by its wider applications.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A review of Ph.D. programs in North America. *American Psychologist, 45,* 721–734.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27,* 17–21.

Asimov, D. (1985). The grand tour: A tool for viewing multidimensional data. *SIAM Journal of Statistical Computing, 6,* 138–143.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science, 4*(6), pp. 372–378.

Becker, R. A., Chambers, J. M., & Wilks, A. (1988). *The new S language: A programming environment for data analysis and graphics.* New York: CRC Press.

Becker, R. A., Cleveland, W. S., & Shyu, M. J. (1996). The visual design and control of Trellis Display. *Journal of Computational and Statistical Graphics, 5,* 123–155.

Behrens, J. T. (1997a). Principles and procedures of exploratory data analysis. *Psychological Methods, 2,* 131–160.

Behrens, J. T. (1997b). Does the white racial identity attitude scale measure racial identity? *Journal of Counseling Psychology, 44,* 3–12.

Behrens, J. T. (2000). Exploratory data analysis. In A. E. Kazdin (Ed.), *Encyclopedia of psychology* (Vol. 3, pp. 303–305). New York: Oxford University Press.

Behrens, J. T., Bauer, M., & Mislevy, R. (2001, April). *Future prospects for assessment in the on-line world.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Behrens, J. T., & Dugan, J. G. (1996). *A graphical tour of the White Racial Identity Attitude Scale data in hypertext and VRML.* Retrieved from http://research.ed.asu.edu/reports/wrias.graphical.tour/graphtour.html.

Behrens, J. T., Dugan, J. T., & Franz, S. (1997, August). *Improving the reporting of research results using the World Wide Web.* Paper presented at the 106th Annual Convention of the American Psychological Association, Chicago, IL.

Behrens, J. T., & Rowe, W. (1997). Measuring White racial identity: A reply to Helms (1997). *Journal of Counseling Psychology, 44,* 17–19.

Behrens, J. T., & Smith, M. L. (1996). Data and data analysis. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 949–989). New York: Macmillan.

Berk, K. N. (1994). *Data analysis with Student SYSTAT.* Cambridge, MA: Course Technology.

Beveridge, W. I. B. (1950). *The art of scientific investigation.* New York: Vintage Books.

Box, G. E. P. (1979). Robustness in scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201–236). New York: Academic Press.

Braun, H. I. (Ed.). (1994). *The collected works of John W. Tukey: Vol 8. Multiple comparisons, 1948–1983.* New York: Chapman & Hall.

Breckler, S. J. (1990). Application of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107,* 260–273.

Brillinger, D. R. (Ed.). (1984). *The collected works of John W. Tukey: Vol 1. Time series: 1949–1964.* Monterey, CA: Wadsworth.

Brillinger, D. R. (Ed.). (1985). *The collected works of John W. Tukey: Vol 2. Time series: 1965–1984.* Monterey, CA: Wadsworth.

Brillinger, D. R., Fernholz, L. T., & Morgenthaler, S. (1997). *The practice of data analysis: Essays in honor of John W. Tukey.* Princeton, NJ: Princeton University Press.

Clark, L. A., Cleveland, W. S., Denby, L., & Liu, C. (1999). Competitive profiling displays: Multivariate graphs for customer satisfaction survey data. *Marketing Research, 11,* 25–33.

Cleveland, W. S. (1985). *The elements of graphing data.* Monterey, CA: Wadsworth.

Cleveland, W. S. (Ed.). (1988). *The collected works of John W. Tukey: Vol. 5. Graphics.* Belmont, CA: Wadsworth.

Cleveland, W. S. (1993). *Visualizing data.* Summit, NJ: Hobart Press.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 79,* 531–554.

Cleveland, W. S., & McGill, R. (1988). *Dynamic graphics for statistics.* Monterey, CA: Wadsworth.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Cook, R. D., & Weisberg, S. (1994). *An introduction to regression graphics.* New York: Wiley.

Cox, D. R. (Ed.). (1992). *The collected works of John W. Tukey: Vol. 7. Factorial and ANOVA, 1949–1962.* Monterey, CA: Wadsworth.

Data Description, Inc. (1997). Data Desk (version 6.1) [computer software]. Ithaca, NY: Data Description.

Dugan, J. G., & Behrens, J. T. (1998, November 18). A Hypermedia exploration of the classification problem in special education. *Current Issues in Education, 1,* (6). Retrieved from http://cie.ed.asu.edu/volume1/number6/.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1,* 3–32.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society, 98,* 39–82.

Fisher, R. A. (1936). Has Mendel's work been rediscovered? *Annals of Science, 1,* 115–117.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society B, 17,* 69–78.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers, 23,* 881–889.

Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the boxplot. *The American Statistician, 43,* 50–54.

Gabriel, K. R. (1971). The biplot graphical display of matrices with application to principal component analysis. *Biometrics, 58,* 453–467.

Garner, W. R. (1974). *The processing of information and structure.* Potomac, MD: Erlbaum.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Kroger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 11–33). Cambridge, MA: MIT Press.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Gödel, C. (1986). *Collected works.* New York: Oxford University Press. (Original work published 1947)

Gould, S. J. (1996). *The mismeasure of man* (2nd ed.). New York: Norton.

Gower, J. C., & Hand, D. J. (1996). *Biplots.* London: Chapman & Hall.

Hacking, I. (1975). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference.* New York: Cambridge University Press.

Härdle, W. (1991). *Smoothing techniques with implementation in S.* New York: Springer-Verlag.

Härdle, W., Klinke, S., & Turlach, B. A. (1995). *XploRe: An interactive statistical computing environment.* New York: Springer-Verlag.

Helms, J. E. (1997). Implications of Behrens for the validity of the White Racial Identity Attitude Scale. *Journal of Counseling Psychology, 44,* 13–16.

Henderson, H. V., & Velleman, P. F. (1981). Building multiple regression models interactively. *Biometrics, 37,* 391–411.

Hilpinen, R. (1992). On Peirce's philosophical logic: Propositions and their objects. *Transaction of the Charles S. Peirce Society, 28,* 467–488.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). *Understanding robust and exploratory data analysis.* Reading, MA: Addison-Wesley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1985). *Exploring data tables, trends, and shapes.* Reading, MA: Addison-Wesley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1991). *Fundamentals of exploratory analysis of variance.* Reading, MA: Addison-Wesley.

Hoffmann, M. (1997). *Is there a logic of abduction?* Paper presented at the 6th congress of the International Association for Semiotic Studies, Guadalajara, Mexico.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression.* New York: Wiley.

Hume, D. (1912). *An enquiry concerning human understanding, and selections from a treatise of human nature.* Chicago: Open Court. (Original work published 1777)

Insightful. (2001). Splus 6 [Computer software]. Available at http://www.insightful.com/.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard University Press.

Jones, L. V. (Ed.). (1986a). *The collected works of John W. Tukey: Vol. 3. Philosophy and principles of data analysis: 1949–1964.* Belmont, CA: Wadsworth.

Jones, L. V. (Ed.). (1986b). *The collected works of John W. Tukey: Vol. 4. Philosophy and principles of data analysis (1965–1986).* Belmont, CA: Wadsworth.

Josephson, J. R., & Josephson, S. G. (1994). (Ed.). *Abductive inference: Computation, philosophy, technology.* Cambridge, England: Cambridge University Press.

Keselman, H. J., Huberty, C., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., & Keselman, J. C. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350–386.

Kline, M. (1980). *Mathematics: The loss of certainty.* New York: Oxford University Press.

Kosslyn, S. (1994). *Elements of graph design.* New York: W. H. Freeman.

Larkin, J., & Simon, H. (1987). Why a diagram is (sometimes) worth a thousand words. *Cognitive Science, 11,* 65–99.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88,* 1242–1249.

Lewandowsky, S., & Behrens, J. T. (1999). Statistical graphs and maps. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi (Eds.), *The handbook of applied cognition* (pp. 514–549). New York: Wiley.

Lindsey, J. K. (1996). *Parametric statistical inference.* Oxford, England: Clarendon Press.

Mallows, C. L. (Ed.). (1990). *The collected works of John W. Tukey: Vol. 6. More mathematical: 1938–1984.* Monterey, CA: Wadsworth.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *American Statistican, 32,* 12–16.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics.* Reading, MA: Addison-Wesley.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I and II. *Biometrika, 20,* 174–240, 263–294.

Neyman, J., & Pearson, E. S. (1933a). The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of Cambridge Philosophical Society, 20,* 492–510.

Neyman, J., & Pearson, E. S. (1933b). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of Royal Society; Series A, 231,* 289–337.

Norusis, M. J. (2000). *SPSS 10.0 Guide to data analysis.* New York: Prentice Hall.

Ottens, J., & Shank, G. (1995). The role of abductive logic in understanding and using advanced empathy. *Counselor Education & Supervision, 34,* 199–213.

Paap, K. R., & Johansen, L. S. (1994). The case of the vanishing frequency effect: A retest of the verification model. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 1129–1157.

Paap, K. R., Johansen, L. S., Chun, E., & Vonnahme, P. (2000). Neighborhood frequency does affect performance in the Reicher task: Encoding or decision? *Journal of Experimental Psychology: Human, Perception, and Performance, 26,* 1691–1720.

Peirce, C. S. (1868). Some consequences of four incapacities. *Journal of Speculative Philosophy, 2,* 140–157.

Peirce, C. S. (1878). How to make our ideas clear. *Popular Science Monthly, 12,* 286–302.

Peirce, C. S. (1960). *Collected papers of Charles Sanders Peirce.* Cambridge, MA: Harvard University Press. (Original work published 1934)

Press, S. J., & Tanur, J. M. (2001). *The subjectivity of scientists and the Bayesian approach.* New York: Wiley.

Russell, B., & Whitehead, A. N. (1910). *Principia mathematica.* Cambridge, MA: Cambridge University Press.

Salsburg, D. S. (1985). The religion of statistics as practiced in medical journals. *American Statistician, 39,* 220–223.

Samuelson, P. (1967). Economic forecast and science. In P. A. Samuelson, J. R. Coleman, & F. Skidmore (Eds.), *Reading in economics* (pp. 124–129). New York: McGraw-Hill.

SAS Institute (2001). SAS/Insight [Computer software]. Retrieved from http://www.sas.com.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning.* New York: Wiley.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization.* New York: Wiley.

Shank, G. (1991, October). *Abduction: Teaching to the ground state of cognition.* Paper presented at the Bergamo Conference on Curriculum Theory and Classroom Practice, Dayton, OH.

SPSS, Inc. (2001). SYSTAT [Computer software]. Retrieved from http://www.spss.com.

Staat, W. (1993). On abduction, deduction, induction and the categories. *Transactions of the Charles S. Peirce Society, 29,* 225–237.

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social development literatures. *Psychological Bulletin, 111,* 42–61.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

Stock, W. A., & Behrens, J. T. (1991). Box, line, and mid-gap plots: Effects of display characteristics on the accuracy and bias of estimates of whisker length. *Journal of Educational Statistics, 16,* 1–20.

Sullivan, P. F. (1991). On falsification interpretation of Peirce. *Transactions of the Charles S. Peirce Society, 27,* 197–219.

Swayne, D. F., Cook, D., & Buja, A. (1998). XGobi: Interactive dynamic data visualization in the X Window System. *Journal of Computational and Graphical Statistics, 7*(1), 113–130.

Thagard, P., & Shelley, C. (1997). *Abductive reasoning: Logic, visual thinking, and coherence.* Retrieved from http://cogsci.uwaterloo.ca/Articles/Pages/%7FAbductive.html.

Tierney, L. (1990). *Lisp-Stat: An object-oriented environment for statistical computing and dynamic graphics.* New York: Wiley.

Tufte, E. R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.

Tufte, E. R. (1997). *Visual explanations: Images and quantities, evidence and narrative.* Cheshire, CT: Graphics Press.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press. (Original work published 1983)

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24,* 83–91.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Tukey, J. W. (1986a). Data analysis, computation and mathematics. In L. V. Jones (Ed.), *The collected works of John W. Tukey: Vol. 4. Philosophy and principles of data analysis: 1965–1986* (pp. 753–775). Pacific Grove, CA: Wadsworth. (Original work published 1972)

Tukey, J. W. (1986b). Exploratory data analysis as part of a larger whole. In L. V. Jones (Ed.), *The collected works of John W. Tukey: Vol. 4. Philosophy and principles of data analysis: 1965–1986* (pp. 793–803). Pacific Grove, CA: Wadsworth. (Original work published 1973)

Tukey, J. W. (1986c). The future of data analysis. In L. V. Jones (Ed.), *The collected works of John W. Tukey: Vol. 3. Philosophy and principles of data analysis: 1949–1964* (pp. 391–484). Pacific Grove, CA: Wadsworth. (Original work published 1962)

Tukey, J. W., & Wilk, M. B. (1986). Data analysis and statistics: An expository overview. In L. V. Jones (Ed.), *The collected works of John W. Tukey: Vol. 4. Philosophy and principles of data analysis: 1965–1986* (pp. 549–578). Pacific Grove, CA: Wadsworth. (Original work published 1966)

Venables, W. N., & Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS.* (3rd ed.). New York: Springer-Verlag.

Wainer, H. (1977). Speed vs. reaction time as a measure of cognitive performance. *Memory and Cognition, 5,* 278–280.

Wainer, H. (1997). *Visual Revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot.* Mahwah, NJ: Erlbaum.

Wainer, J., & Velleman, P. (2001). Statistical graphs: Mapping the pathways of science. *Annual Review of Psychology, 52,* 305–335.

Wilcox, R. (1997). *Introduction to robust estimation and hypothesis testing.* San Diego, CA: Academic Press.

Wilcox, R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy.* New York: Springer.

Wilkinson, L. (1993). Comments on W. S. Cleveland, a model for studying display methods of statistical graphs, *Journal of Computational and Graphical Statistics, 2,* 355–360.

Wilkinson, L. (1994). Less is more: Two- and three-dimensional graphs for data display. *Behavior Research Methods, Instruments, & Computers, 26,* 172–176.

Wilkinson, L. (1999). *The grammar of graphics.* New York: Springer.

Wilkinson, L. (2001). Graphics. In *Systat User's Guide.* Chicago: SPSS Inc.

Wilkinson, L., & Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Young, F. W. (1996). *ViSta: The Visual Statistics System.* Chapel Hill, NC: UNC L.L. Thurstone Psychometric Laboratory Research Memorandum.

Yu, C. H., & Behrens, J. T. (1995). Applications of scientific multivariate visualization to behavioral sciences. *Behavior Research Methods, Instruments, and Computers, 27,* 264–271.

HANDBOOK of PSYCHOLOGY

VOLUME 2
RESEARCH METHODS IN PSYCHOLOGY

John A. Schinka
Wayne F. Velicer

Volume Editors

Irving B. Weiner

Editor-in-Chief

John Wiley & Sons, Inc.

Contents