Psychological Methods
1997, Vol. 2, No. 2, 131-160

1082-989X/97/$3.00

Principles and Procedures of Exploratory Data Analysis

John T. Behrens
Arizona State University

Exploratory data analysis (EDA) is a well-established statistical tradition that provides conceptual and computational tools for discovering patterns to foster hypothesis development and refinement. These tools and attitudes complement the use of significance and hypothesis tests used in confirmatory data analysis (CDA). Although EDA complements rather than replaces CDA, use of CDA without EDA is seldom warranted. Even when well-specified theories are held, EDA helps one interpret the results of CDA and may reveal unexpected or misleading patterns in the data. This article introduces the central heuristics and computational tools of EDA and contrasts it with CDA and exploratory statistics in general. EDA techniques are illustrated using previously published psychological data. Changes in statistical training and practice are recommended to incorporate these tools.

The widespread availability of software for graphical data analysis and calls for increased use of exploratory data analysis (EDA) on epistemic grounds (e.g. Cohen, 1994) have increased the visibility of EDA. Nevertheless, few psychologists receive explicit training in the beliefs or procedures of this tradition. Huberty (1991) remarked that statistical texts are likely to give cursory references to common EDA techniques such as stem-and-leaf plots, box plots, or residual analysis and yet seldom integrate these techniques throughout a book. A survey of graduate training programs in psychology corroborates such an impression (Aiken, West, Sechrest, & Reno, 1990). In this investigation, 37 (20%) of the 186 responding departments reported teaching some aspect of EDA in introductory graduate courses. However, the percentage of institutions indicating that most or all students could apply a learned technique was as follows: (a) detection and treatment of influential data, 8%; (b) modern graphical display, 15%; (c) data transformations, 31%; (d) alternatives to ordinary least squares (OLS) regression, 3%. These low levels of competencies and the generally bleak picture of methodological instruction presented by Aiken et al. (1990) indicate that little EDA makes its way into graduate training and even less makes its way out as usable skills.

This essay introduces researchers to the philosophical underpinnings and general heuristics of EDA in three sections. First, the background, rationale, and basic principles of EDA are presented. Next, a primer covers heuristics, prototypical beliefs, and procedures of EDA using examples from psychological research. The final section addresses implications of this analysis for psychological method and training.

Background and First Principles

What Is EDA?

Unaware of historical precedent, researchers may develop their own definition of EDA from denotations of its name. Sometimes the term is used to mean exploratory analysis in general. Mulaik (1984), for example, discussed a long history of generic "exploratory statistics" in response to an article concerning EDA (Good, 1983), and yet scarcely mentioned the specific tradition of EDA to be discussed in this essay. Sometimes the model-building approach of Box (e.g., 1980) is considered exploratory, although it relies more heavily on probabilistic measures than does EDA.

In this article, EDA refers to a specific tradition of data analysis that stems from the work of John Tukey and his associates, which dates back to the early 1960s. This tradition of EDA can be loosely characterized by (a) an emphasis on the substantive under-

I gratefully acknowledge comments and criticisms of earlier versions of this article, which were provided by Raymond Miller, Joe Rodgers, Larry Toothaker, Alex Yu, and Dan Huston.

Correspondence concerning this article should be addressed to John T. Behrens, Methodological Studies, Division of Psychology in Education, Arizona State University, Tempe, Arizona 85287-0611. Electronic mail may be sent via Internet to behrens@asu.edu.

standing of data that address the broad question of "what is going on here?" (b) an emphasis on graphic representations of data; (c) a focus on tentative model building and hypothesis generation in an iterative process of model specification, residual analysis, and model respecification; (d) use of robust measures, reexpression, and subset analysis; and (e) positions of skepticism, flexibility, and ecumenism regarding which methods to apply.

The goal of EDA is to discover patterns in data. Tukey often likened EDA to detective work. The role of the data analyst is to listen to the data in as many ways as possible until a plausible "story" of the data is apparent, even if such a description would not be borne out in subsequent samples. Finch (1979) asserted that "we claim for exploratory investigation no more than that it is an activity directed toward the formation of analogy. The end of it is simply a statement that the data look as if they could reasonably be thought of in such and such a way" (p. 189).

Classical works in this tradition are Tukey's *Exploratory Data Analysis* (1977); Mosteller and Tukey's *Data Analysis and Regression: A Second Course in Statistics* (1977); Hoaglin, Mosteller, and Tukey's studies (1983b, 1985, 1991); volumes three, four, and five of Tukey's collected works (Cleveland, 1988; Jones, 1986a, 1986b); and Velleman and Hoaglin's work (1981, 1992). Summaries of EDA have been presented by Hartwig and Dearing (1979), Leinhardt and Leinhardt (1980), Leinhardt and Wasserman (1979), and more recently by Behrens and Smith (1996) and Smith and Prentice (1993). Erickson and Nosanchuk's (1992) text is for a first course in data analysis that presents a balanced presentation of both EDA and confirmatory data analysis (CDA). Behrens (1996) provided on-line materials for teaching EDA. Although exploratory techniques have been developed by others, Tukey and his associates began the endeavor and continue to lead the articulation of the purpose and constraints necessary for reasonable EDA (cf. Hoaglin et al., 1991). Tukey (1969) recommended the EDA approach to psychologists at the 1968 meeting of the American Psychological Association in a paper entitled, "Analyzing Data: Sanctification or Detective Work?" Since that time, surprisingly few have responded.

The Need for EDA

Most psychologists are well trained in testing statistical hypotheses at the end of an investigation. Nevertheless, the scientific process of model building and testing often requires learning from the data at all stages of research. For example, while conducting a regression analysis, one may be interested in assessing the specific hypothesis that a particular $\beta_1 = 0$ in a model with X_1 and X_2. When assessing the status of prespecified statistical issues, the researcher is working in what Mayer (1980) called the confirmatory mode. More often, however, researchers are concerned with a broader range of questions about the data than the statistical significance of the partialed slopes: What if responses on X_2 occurred only at two levels rather than across all possible levels of the scale? Are there extreme values that unduly affect the estimation of the slopes? Is the shape of the data in the scatter plot like an ellipse, a horseshoe, or a banana? Is there something misleading me? When addressing such a broad set of questions, a researcher is working in an exploratory mode. Because the goals of the two modes of data analysis are different, the modes are complementary rather than antagonistic.

In contrast to EDA, most training in CDA fails to address the early and messy stages of data analysis. This practice constitutes what McGuire (1989) called the hypothesis testing myth. He argued that we do a disservice to training and practice by glossing over or ignoring preliminary data analyses during which we refine hypotheses, evaluate and clarify our auxiliary assumptions, and simply make sure our mental model of the data is well aligned with reality.

Exploratory and Confirmatory

CDA is often likened to Anglo-Saxon jury trials: Researchers play the role of prosecutor, data collection serves as the trial proceeding, and statistical analysis plays the role of jury decision (Kraemer & Thiemann, 1987; Tukey, 1977). The detective analogy for EDA fits well with this jurisprudence model because the role of the detective is to establish pretrial evidence and hunches, the veracity of which are tested at the trial. Kraemer and Thiemann pointed out that prosecutors examine preliminary evidence before deciding whether to prosecute or not. They equate this process with EDA and other pretrial evidence gathering such as power- or meta-analysis. By using both exploratory and confirmatory techniques, a data analyst collects complete pretrial evidence and brings the full weight of CDA to bear at the trial.

In a trial, rules of presenting and evaluating evi-

dence are as well established as the rules of statistical inference. To make the strong claim of innocence or guilt (significance or nonsignificance), one uses specific rules and procedures with strict interpretations of the data. In EDA, the goal is not to draw conclusions regarding guilt and innocence but rather to investigate the actors, generate hunches, and provide preliminary evidence. EDA is more like an interrogation in which clean and corrupted stories are told, whereas CDA is testimony regarding evidence that fits carefully laid-out trial procedures. The goal of EDA is indictment; the goal of CDA is conviction (Behrens & Smith, 1996).

There is, however, a point at which the trial analogy breaks down. In a jury trial a witness may be used to both formulate and test hunches. Alternatively, in scientific practice different data *must* be used for model formulation (EDA) and testing (CDA). Failure to recognize this important fact will lead to inflation of Type I error and overfitting. Along these lines Giere (1984, cited in Howson & Urbach, 1993) argued:

> If the known facts were used in constructing the model and were thus built into the resulting hypothesis . . . then the fit between these facts and the hypothesis provides no evidence that the hypothesis is true [since] these facts had no chance of refuting the hypothesis. (p. 408)

When sufficiently large samples are available, the exploratory data analyst is likely to conduct EDA on one data set to generate hypotheses and assess the model on another. The importance of distinguishing between model building and testing led Mosteller and Tukey (1977) to state that "we plan to cross-validate carefully wherever we can" (p. 40). Cross-validation means that when patterns are discovered, they are considered provisional (consistent with the EDA mode) until their presence is tested in different data.

Tukey (1972/1986b) discussed data analysis as a continuum from EDA to CDA. Between the two is an intermediate mode called rough confirmatory analysis. In EDA the researcher entertains numerous hypotheses, looks for patterns, and suggests hypotheses based on the data, with or without theoretical grounding. Working in this mode, the researcher begins to delineate a set of plausible models and seeks rich descriptions of the data. In rough CDA, the researcher undertakes initial assessment of the plausible models using probabilistic approaches such as confidence intervals or significance tests (cf. Behrens & Smith, 1996). In this step the researcher answers the question, "With what accuracy are the appearances already found to be believed?" (Tukey, 1972/1986b, p. 760). In the confirmatory mode, researchers work to test specific hypotheses using a strict probabilistic framework following a decision theoretic approach.

When trained in all three modes of data analysis, a researcher is likely to move fluidly between modes and work in multiple modes on the same problem. For example, a researcher may have strict hypotheses about main effects in a factorial analysis of variance (ANOVA) and yet have no hypotheses concerning possible interactions. Working in a strict confirmatory mode, the researcher would compute only the main effects test and ignore possible interactions. Working in multiple modes, a researcher would likewise state the hypothesis for main effects and test them using strict CDA. At the same time, however, the researcher working in multiple modes would explore possible interactions with statistical graphics, resistant summary statistics, and even loosely interpreted significance tests. Patterns of unexpected outcomes would be regarded as starting points for hypothesis generation and future testing rather than as statistical conclusions. In addition, the researcher familiar with EDA will also explore data patterns associated with the hypothesized main effect to make sure the CDA was not misled by unrecognized patterns that can lead to conclusions inconsistent with the data.

Tukey summarized the relation between these modes of data analysis, arguing "(a) *both* exploration and confirmation are important, (b) exploration comes *first,* (c) any given study can, and usually should, combine *both*" (Tukey, 1980/1986e, p. 822; cf. Tukey, 1980). Tukey (1982/1986d) presented a more detailed analysis of levels and types of data analysis following this framework. Jöreskog and Sörbom (1993) presented a similar discussion of situations of (a) model generating, (b) analysis of competing models, and (c) strictly confirmatory analysis of a single model, all in the context of structural equation modeling (p. 115).

Researchers who want to know more about their data than they have hypothesized sometimes use confirmatory methods while working in a pseudoconfirmatory mode. Examples of this kind of behavior include interpreting unexpected interactions as if hypothesized and computing t tests or chi-square homogeneity tests on myriad possibilities. This is not EDA. In these cases, researchers have exploratory goals but are approaching them by using confirmatory tools, assumptions, and conclusions improperly. EDA

helps avoid these improper approaches by being clear about hypothesis specificity and conclusion strength and by providing a language for different stages and purposes of data analysis.

EDA and Other Exploratory Methods

Exploratory statistics is an interesting point of convergence between classical CDA and EDA. A number of apparently confirmatory techniques are exploratory in their goals, including stepwise regression, some forms of factor analysis, cluster analysis, discriminant analysis, and many applications of structural equation modeling. These and other methods are exploratory when the researcher is trying to determine a "best" set of variables or the "best" model for a sample rather than testing a prespecified model for a specific population. For example, so-called confirmatory factor analysis via structural equation models becomes exploratory when a number of alternate models are assessed. The exploratory nature of these techniques underscores the idea that data exploration and the integration of empirical and theoretical knowledge are well-established aspects of scientific psychology.

Given that EDA is not simply a set of techniques but an attitude toward the data (Tukey, 1977), are researchers conducting EDA when they compute exploratory factor analysis or other exploratory statistics? The answer depends on how the analysis is conducted. A researcher may conduct an exploratory factor analysis without examining the data for possible rogue values, outliers, or anomalies; fail to plot the multivariate data to ensure the data avoid pathological patterns; and leave all decision making up to the default computer settings. Such activity would *not* be considered EDA because the researcher may be easily misled by many aspects of the data or the computer package. Any description that would come from the factor analysis itself would rest on too many unassessed assumptions to leave the exploratory data analyst comfortable. Henderson and Velleman (1981) demonstrated how an interactive (EDA based) approach to stepwise regression can lead to markedly different results than would be obtained by automated variable selection. This occurs because the researcher plots the data and residuals at each stage and thereby considers numerous patterns in the data while the computer program is blind to all aspects of the data except the R^2.

Although holding exploratory goals alone does not necessarily imply EDA, use of exploratory procedures such as the plotting of simple summaries or the tabulation of simple descriptive statistics does not necessarily imply EDA either. In many cases, simple descriptive statistics or plots may hide important patterns as much as they reveal others.

Summary

EDA emphasizes that at different stages of research there are different types of questions, different levels of hypothesis specificity used, and different levels of conclusion specificity that are warranted. EDA does not call for the abandonment of CDA but rather for the broadening of data analysis to incorporate a wide range of attitudes and techniques appropriate to the different stages and questions in scientific work. At the same time, EDA is seen as indispensable in any investigation: "Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone—as the first step" (Tukey, 1977, p. 3).

Beliefs, Heuristics, and Trademarks

Although Tukey often argues that EDA is an attitude rather than a set of tools, a number of heuristics have been devised for EDA. To find patterns, reveal structure, and make tentative model assessments, EDA emphasizes the use of graphics and the process of iterative model fit and residual analysis. To avoid being fooled by unwarranted assumptions about the data, EDA is a much more data-driven approach to data analysis than CDA. Because a complete cataloging of techniques is beyond the scope of this article, this section discusses major themes of EDA and presents examples.

It cannot be overemphasized that an appropriate technique for EDA is determined not by computation but rather by a procedure's purpose and use. Whether residuals are obtained from a computer program intended for CDA or EDA is not important. What is important is to obtain a rich description of the data and to understand the relationship between the model and patterns of residuals. The techniques described next have been helpful in EDA, but techniques are secondary to the goal of building rich mental models of the data. The reader may note that the procedures described are highly related, not simply a laundry list.

Each aspect of EDA is used in concert with other aspects so that a single isolated procedure is seldom used. Recommendations presented here are not necessarily unique to EDA. What is unique is the configuration of beliefs and procedures.

Understand the Context

To some, the analogy of the data analyst as detective connotes someone entering an unknown arena and cleverly finding patterns that may or may not reflect "true" effects. This connotation has led some to characterize EDA as naive empiricism gone amok (MacDonald, 1983). This description is inappropriate for both a detective and someone conducting data analysis in an exploratory mode.

EDA shares a view of the interaction of prior knowledge and data analysis similar to the postpositivist position put forward by Donald Campbell. Campbell (1988) argued that quantitative knowing is dependent on qualitative knowing. This view holds that, in quantitative data analysis, numbers map onto aspects of reality. Numbers themselves are meaningless unless the data analyst understands the mapping process and the nexus of theory and categorization in which objects under study are conceptualized. This approach closely matches Tukey's concern about sterilized, decontextualized approaches to data analysis. In his satirical list of "badmandments," Tukey (1986a) chided, "Never tell your statistical consultant about the two most important recent papers in the field of your own RESEARCH" (p. 205, emphasis in the original). Similarly, Tukey (1979) argued that substantive concerns must take precedence over statistical convenience.

Data analysts working in either a confirmatory or exploratory mode are sometimes considered technicians who provide recipes for "number crunching." Countering this characterization, Tukey and others (Bode, Mosteller, Tukey, & Winsor, 1986) suggested the training of scientific generalists. They pointed out that "statistics, as the doctrine of planning experiments and observations and of interpreting data, has a common relation to all sciences" (p. 3). They concluded that the complexity of science requires training of generalists with broad depth and experience in areas such as economics, psychology, and natural sciences. Boring (1919) expressed similar sentiments when he wrote that "statistical ability, divorced from a scientific intimacy with the fundamental observations, leads nowhere" (p. 332).

Use Graphic Representations of Data

Graphical analysis is central to EDA. Tukey (1977) summed up the role of graphics in EDA by saying that "the greatest value of a picture is when it forces us to notice what we never expected to see" (p. vi). Graphical summaries are almost universally sought to augment algebraic summaries because graphics can portray numerous data values simultaneously, while algebraic summaries often sum over important attributes of the data or fail to suggest important patterns. For example, Cleveland (1985, reprinted in Behrens & Smith, 1996) provided the data in Table 1, which relates the average intelligence quotient for fathers and sons at each level of a number of social classes as reported by Burt (1961). Although the general positive trend is evident in the table, the plot of the data in Figure 1 shows that the function underlying these data is so straight that it calls the veracity of the data into question. Even this simple plot is striking because it serves Tukey's function of showing what we did not expect. For an exploratory data analyst, graphical representation is the primary language.

Classical references in the area of statistical graphics include Bertin (1983); Chambers, Cleveland, Kleiner, and Tukey (1983); Cleveland (1985, 1993); Cleveland and McGill (1988); Wainer and Thissen (1981, 1993); and Tufte (1983, 1990). Cook and Weisberg (1994) presented a comprehensive treatment of graphics for regression analysis.

Tukey (1977) developed a number of graphical devices used by exploratory analysts that are gaining widespread use because of their incorporation in common statistical packages. For example, Figure 2 is a stem-and-leaf plot of effect sizes from studies examining sex differences reported by Feingold (1994). In this meta-analysis, negative effect sizes indicate studies in which males have average scores lower than

Table 1
Intelligence Quotient (IQ) Data Provided by Burt (1961)

Social class	Adult mean IQ	Child mean IQ
Higher professional	139.7	120.8
Lower professional	130.6	114.7
Clerical	115.9	107.8
Skilled	108.2	104.6
Semiskilled	97.8	98.9
Unskilled	84.9	92.6

Note. From *The Elements of Graphing Data* (p. 97), by W. S. Cleveland, 1985, Monterey, CA: Wadsworth. Copyright 1985 by W. S. Cleveland. Reprinted with permission.

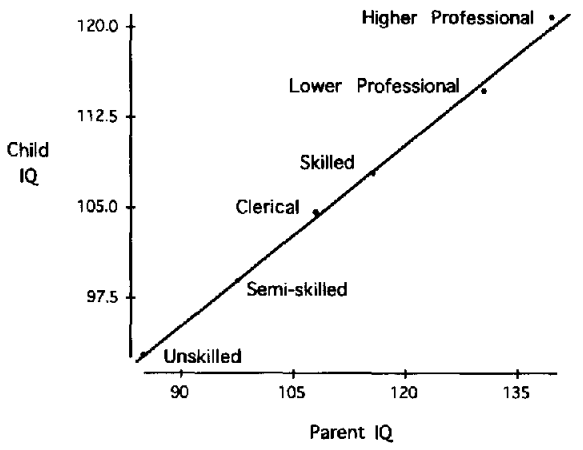

Figure 1. Plot of average intelligence quotient (IQ) for sons versus average IQ for fathers across different social classes. Original data are from Burt (1961), and the original plot is from *The Elements of Graphing Data* (p. 98), by W. S. Cleveland, 1985, Monterey, CA: Wadsworth. Copyright 1985 by W. S. Cleveland. Reprinted with permission.

females, and positive effect sizes indicate studies in which males have higher average scores than females. Meta-analytic studies are especially well suited for EDA because the small number of effects and variability in sample sizes can lead to wild imbalances in the influence of individual observations. The procedures for sensitivity analysis (Greenhouse & Iyengar, 1994) and visual analysis (Light, Singer, & Willett, 1994) of meta-analytic data described in *The Handbook of Research Synthesis* (Cooper & Hedges, 1994) rely extensively, and explicitly, on EDA.

```
-6   4
-5   4
-4   8 8
-3   1 3 5
-2
-1   3
-0
 0   0 2 2 3 5 5
 1   6 8
 2
 3   0
```

Figure 2. Stem-and-leaf plot of effect sizes for studies examining differences in anxiety across sexes as reported by Feingold (1994).

The stem-and-leaf plot shown in Figure 2 represents a type of frequency table organized graphically to resemble a histogram while retaining information about the exact value of each observation. The left side of the plot are the "stems" that mark intervals or bins; the right side of the plot contains "leaves," which represent the detail of numbers occurring in each bin. In this case the numbers on the left-hand side of the plot indicate the 10ths place value of effect sizes from the study, and those on the right indicate the 100ths place that occur in each bin. In this way the plot indicates that the smallest effect sizes are −0.64 and −0.54 whereas the largest is 0.30. The alignment of the numbers allows a quick sense of frequency and distribution shape, and close examination of individual values (the leaves) provides additional information about distributions within each bin. Stem-and-leaf plots will vary in appearance according to the number of bins used. In this case, as in all cases of graphic analysis, there is no single "plot of the data" but rather only one of many possible plots. When a large number of data points are examined, the stem-and-leaf plot may become cumbersome.

A dot plot can be an effective tool to examine a single distribution or compare a number of distributions. Figure 3 is a dot plot of effect sizes characterizing the differences between males and females across four personality traits from Feingold (1994). Each dot in the display represents the value of one observation. The general pattern is clear. Overall, males tend to be less anxious than females, are generally more assertive, and have higher locus of control and self-esteem scores. Great variability across these measures and an obvious high outlier for assertiveness are also easy to detect.

In the early stages of data analysis, it is often preferable to plot data directly because summaries may hide data or distort one's visual impression of data. This maxim, however, needs to be balanced against the need for multiplicity of graphic summaries and the need for parsimonious representation of numerous data points (Yu & Behrens, 1995). When seeking additional structure in univariate distributions or when a number of distributions need to be compared, a box plot is often used. This is a shortened name for the original box-and-whisker plot. Although box plots come in many varieties (see Frigge, Hoaglin, & Ig-

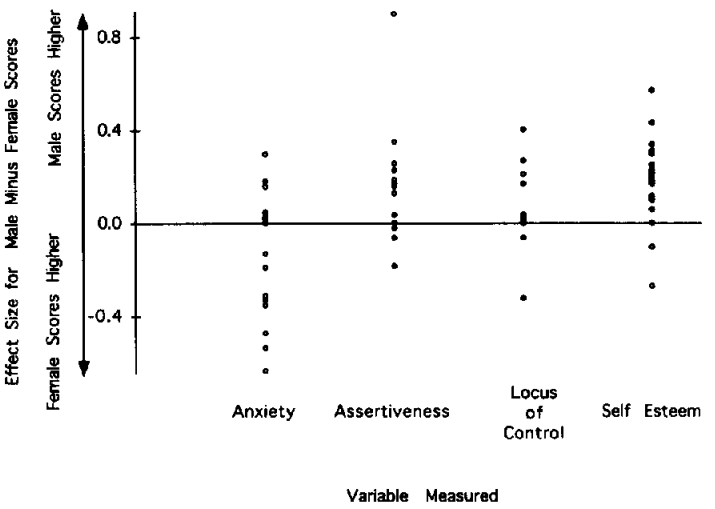

Figure 3. Dot plot of sex difference effect sizes for different affective and cognitive variables reported by Feingold (1994).

lewicz, 1989), a common form is shown in Figure 4, which portrays the data from Figure 3. The box plot offers a five-number summary in schematic form. The ends of a box mark the first and third quartiles, and the median is indicated with a line positioned within the box.[1] The ranges of most or all of the data in the tails of the distribution are marked using lines extending away from the box, creating "whiskers" or "tails." Rules governing the construction of the whiskers vary. One method suggested by Tukey (1977) was to extend the whisker to the most extreme value, not exceeding a distance of 1.5 times the interquartile spread (interquartile spread is the scale value of the 75th percentile minus the value at the 25th percentile). In this scheme the tails will cover the middle 99.3% of a Gaussian distribution. Data values occurring past this point are typically displayed individually, as shown in Figure 4.

Comparing the boxplot to the dotplot, one can see that the box plot offers information about the location of key elements in the distribution (including outliers) and omits more subtle details. The summarizing function of this plot is especially useful when a number of distributions are being compared. Other forms of the boxplot have been developed to indicate the confidence interval of the median by shading the center of the box or indenting the box along the length of the interval (cf. McGill, Tukey, & Larsen, 1978). Other modifications superimpose dots over the boxes (Berk, 1994) or alter the appearance of the box (Stock & Behrens, 1991). Emerson and Strenio (1983) presented a complete treatment of basic boxplot design.

Kernel density smoothers are graphic devices that provide estimates of a population shape, as seen in Figure 5. This smooth shape is arrived at by taking the relative frequency of data at each x value and averaging it with that of the surrounding data (Scott, 1992). Figure 5 is a kernel density smooth of the Feingold anxiety data depicted in Figures 3 and 4. What is clear from this graphic is the near bimodality of the data that is hidden in the boxplot and not obvious in the dotplot. By varying the type of averaging across the data at each point, the size of the window of averaging, or the weighting function used around each point, the appearance of the plot can be varied to make the plot appear more jagged or more smooth. Overlaying density functions from each distribution allows direct comparison of their shapes, as shown in Figure 6. From this we can see that the underlying distributions are quite similar, with the exception of the second mode of the anxiety distribution. Further analysis of these data is warranted to ascertain whether there are unique study characteristics associated with this group. This example underscores the importance of multiple depictions of data and the impor-

[1] More precisely, the key elements of the box plot are based on statistics that Tukey called "hinges." These are robust measures that generally match the quartiles, although slight differences may occur in some cases. See Frigge et al. (1989) for a discussion of these issues.

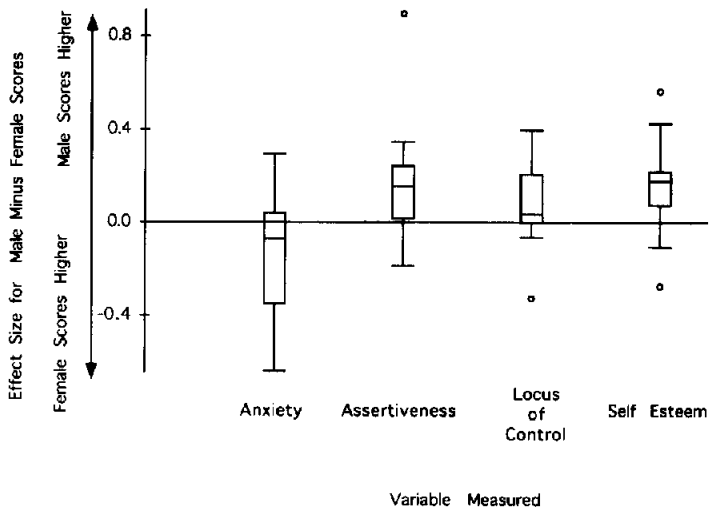

Figure 4. Box plot of effect sizes displayed in Figure 3. Data in the tails of the distribution are marked using lines extending away from the box. The ○ indicates outliers that deserve special attention.

tance of rejecting the notion of "the plot" of a set of data. Scott (1992) provided a full treatment of univariate and multivariate density-smoothing functions.

A major component of the detective work of EDA is the rough assessment of hunches, a quick look at the question "could it be that . . ." or "what if it is the case that. . . ." As Tukey and Wilk (1986) argued, citing Chamberlain (1965), "science is the holding of multiple working hypotheses." To hold and assess multiple working hypotheses, data analysts depend heavily on interactive computer graphics. Interactive graphics can be acted on directly by touching them with the cursor or other pointing device. Another important innovation, linked plots, are organized so that a change made to the color or shape of a point representing an observation in one plot automatically changes the appearance of the observation in all other plots.

In the case of the Feingold data, interactive graphics allow the selection of the observations in the second mode of the anxiety effect size data by drawing a rectangle around the observations of interest. When this is done, the observations are highlighted in all the linked windows. To determine whether there is a common effect in these data based on the country in which the study occurred, a plot of the data organized by country is opened. Figure 7 is an illustration of how linking the two plots allows quick determination of covariation across variables. This figure is a picture of a computer screen obtained using Data Desk (Data Description, Inc., 1995), although linking is common in most EDA-oriented programs. The small window on top of the dot plot is a palette that allows rapid change of observation shape by selecting observations and pointing to the desired shape. The highlighted portions of the bar chart reflect the highlighted (second mode) portions of the dot plot. Of the nine highlighted observations, three are from the United States and two each are from Israel, Canada, and Sweden. All of the positive effect sizes are from studies conducted in the United States, suggesting the possibility of country-related effects. These data are limited in size but are consistent with the tentative hypothesis that sex differences in reported anxiety vary as a function of country of origin. Such an idea provides direction for conducting evaluations of other data sets and for conducting cross-cultural studies in the future.

Because it is impossible to anticipate all relevant aspects of data in either experimental or nonexperimental work, it is difficult to overstate the value of graphics. The multiplicity of data patterns that can match a single mean led early psychologists to consistently report means with histograms. Changes from this convention were heatedly discussed. By 1935, an editorial in *Comparative Psychology* (Dunlap, 1935) asked, ". . . Should we not exclude reports in which the group averages of performance are presented without interpretive distributions?" (p. 3). Even the staunchest proponents of CDA argued for balance in exploratory and confirmatory methods. Fisher's *Statistical Methods for Research Workers* (1925) included an entire chapter on "diagrams" that begins noting:

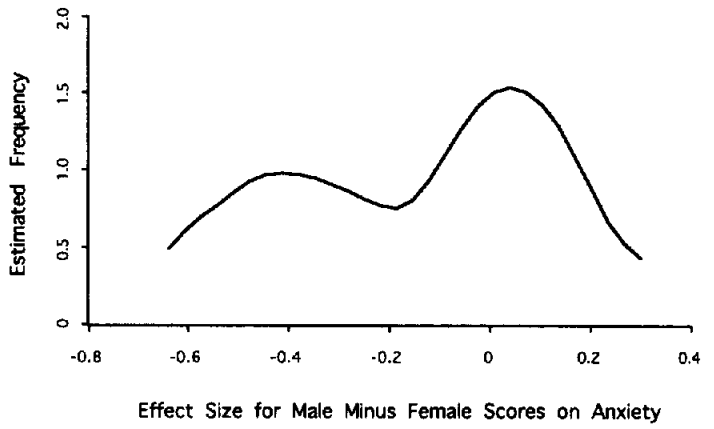

Figure 5. Density estimation plot of the anxiety effect size data indicating bimodality in the anxiety measures.

> The preliminary examination of most data is facilitated by the use of diagrams. Diagrams prove nothing, but bring outstanding features readily to the eye; they are therefore no substitute for such critical tests as may be applied to the data, but are valuable in suggesting such tests, and in explaining the conclusions founded upon them. (p. 24 of the 11th edition)

Develop Models in an Iterative Process of Tentative Model Specification and Residual Assessment

When working in the exploratory mode, the data analyst takes the goal of developing a plausible description of the data using the framework: data = fit + residual. Following a graphical analogy it is sometimes said that data = smooth + rough.

These formulas reflect the fact that the aim of data analysis is to fit or summarize the data and that all description fails to some degree as reflected in the residuals. Even the use of the mean or median is a fit in this view. The boxplot is valuable because it indicates both the fit and residual of a single set of data. The fit is the median or mean, single values that describe the data well. Residuals are deviations from that point. The singling out of outliers is part of the important process of identifying observa-

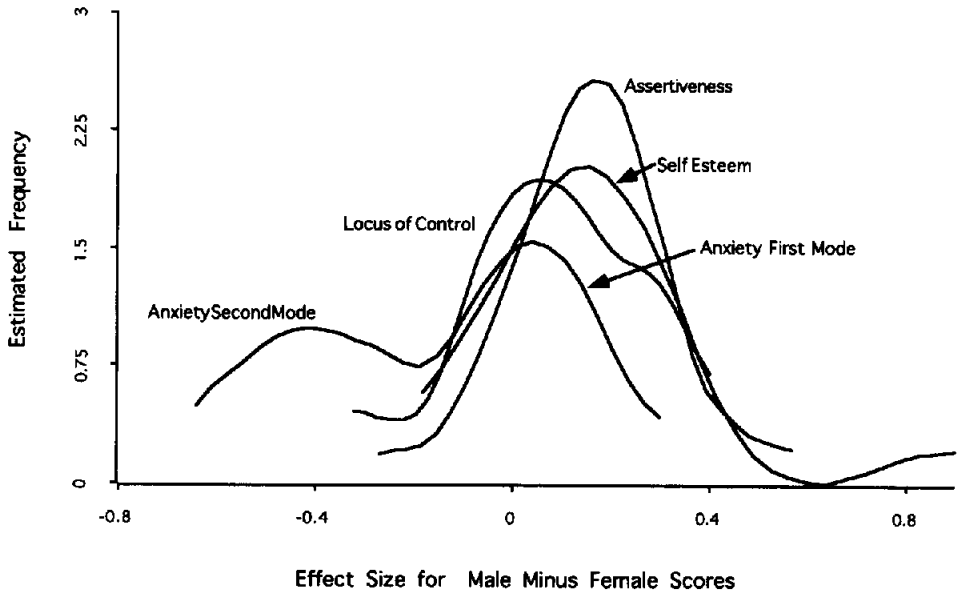

Figure 6. Density estimation plot for the distributions depicted in Figure 4.

tions that are very poorly described by summary statistics.

To create quantitative descriptions of data, the exploratory data analyst conducts an iterative process of suggesting a tentative model, examining residuals from the model to assess model adequacy, and modifying the model in view of the residual analysis. This occurs in a cyclical process that should lead the analyst, by successive approximation, toward a good description. This process was inherent in the examination of the meta-analysis data discussed previously. A single-mode model of the anxiety data was assumed as a starting point, but graphical analysis of residuals (data around the second mode) suggested that such a fit would hide important structure.

Although most psychologists are familiar with residual analysis from the regression literature, workers in EDA extend this framework to conceptualize all model development (Goodall, 1983a). Consider, for example, the data presented in Table 2 from Lauver and Jones (1991). These authors extended previous work in career–self-efficacy theory following Lent and Hackett (1987), who noted the need to collect occupational preference data from ethnically diverse groups. Lauver and Jones conducted a typical analysis consisting of a series of chi-square tests of homogeneity across ethnicities at each level of occupation. This approach ignores the effect of career differences, inflates Type I error, and does not address the possibility of interactions in addition to main effect. After finding occupations with "significant" differences, Lauver and Jones noted the ethnicity with the highest percentage of students considering that career an option. This overlooks the pattern of differences across the ethnicities and fails to build a descriptive model of the structure of the data. Because this type of data was being published for the first time, a rich description of the structure of the data from an exploratory view would have been

Table 2
Percentage of Respondents Perceiving Each Career as an Option by Ethnicity as Reported by Lauver and Jones (1991)

Occupation	Ethnicity		
	Native American %	Hispanic %	White %
X-ray technician	28	23	19
Medical technician	35	40	34
Physical therapist	37	40	41
Social worker	64	56	54
Bookkeeper	47	38	37
Secretary	52	56	48
Fashion shop manager	47	56	45
Receptionist	35	47	41
Librarian	36	23	18
Electrician	49	47	42
Electronics technician	54	51	43
Veterinarian	34	38	55
Probation officer	58	40	29
Armed forces	76	71	62
Accountant	47	51	49
Lawyer	63	68	62
Auto salesperson	34	33	26
Photographer	66	65	67

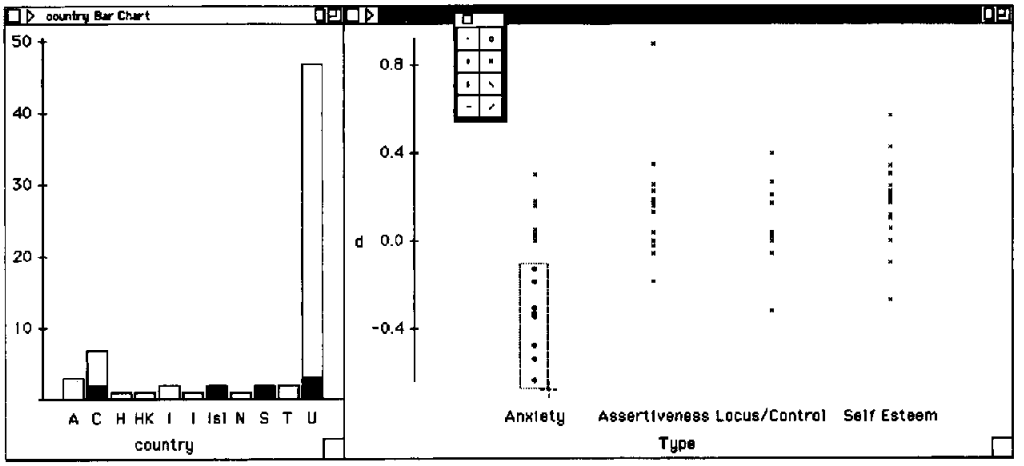

Figure 7. Image of computer screen during interactive graphic session with brushing and linking.

preferred. After these and other data have been published and specific hypotheses generated, more confirmatory approaches may be appropriate.

Building a Two-Way Fit

To build a tentative model of the two dimensions of the table, one may apply the fit-plus-residual framework iteratively in both dimensions to form a two-way fit. In this approach, each value in the table is modeled as the sum of ethnicity, occupation, and overall effects. First, a tentative fit or description of the level of options seen in each ethnicity is found by calculating the median percentage of options in each column. This provides fits of 47, 47, and 42.5 for the Native American, Hispanic, and White groups, respectively. After this first step, unexpected patterns begin to appear: On average, fewer White students rate occupations as options than their Native American or Hispanic counterparts. To complete the initial pass at decomposing data into fit and residual, residuals are computed by subtracting each data value from the median value for the corresponding ethnicity. The bottom of Table 3 displays fits for each ethnicity with residuals in the center. Each fit plus residual equals the original cell data.

Table 3
Residual Percentage of Individuals Viewing Each Career as an Option After a First Pass at Removing the Ethnicity Fit

Occupation	Ethnicity		
	Native American	Hispanic	White
X-ray technician	−19	−24	−23.5
Medical technician	−12	−7	−8.5
Physical therapist	−10	−7	−1.5
Social worker	17	9	11.5
Bookkeeper	0	−9	−5.5
Secretary	5	9	5.5
Fashion shop manager	0	9	2.5
Receptionist	−12	0	−1.5
Librarian	−11	−24	−24.5
Electrician	2	0	−0.5
Electronics technician	7	4	0.5
Veterinarian	−13	−9	12.5
Probation officer	11	−7	−13.5
Armed forces	29	24	19.5
Accountant	0	4	6.5
Lawyer	16	21	19.5
Auto salesperson	−13	−14	−16.5
Photographer	19	18	24.5
Ethnicity fits	47	47	42.5

Having subtracted out the ethnicity effects from the rows, patterns of occupation-related effects begin to emerge in the residuals. An occupation's residuals reflect how different it is from the median occupation within each ethnicity. Negative numbers indicate low-option occupations and positive residuals indicate high-option occupations. X-ray technician, librarian, and auto salesperson stand out as low-option professions. The armed forces leads as a high-option profession.

To model the values associated with these occupation effects, the two-way fit continues by calculating the median residual for each occupation. This provides a summary of the occupation effects, similar in form to the initial summary of the ethnicity effects. Next, cell values from Table 3 are subtracted from the occupation medians, allowing each cell of the original table to be recreated by adding the occupation and ethnicity fits to the residual. This process is then extended to find an overall fit by fitting the ethnicity and occupation fits. After these fits have been obtained, the process is repeated by iteratively refitting the residuals in each direction of the table until additional patterns cannot be extracted (Tukey, 1977).

The final results of such an analysis are presented in Table 4, with occupations reordered by the size of their fits. Each datum in the original table can be recreated by adding the overall, ethnicity, and occupation fits to the residual. When the overall, occupation, and ethnicity fits are perfectly additive, residuals equal zero. Residuals indicate interaction effects over and above the main effects modeled in the ethnicity and occupation fits. For example, 29% of the White respondents consider work as a probation officer an option. This is nine percentage points less than the predicted value of 38 obtained by adding the overall fit (45) plus the White fit (−2) plus the probation officer fit (−5). In contrast, the Native American group considers this occupation as an option 17% more often than one would expect given the overall fit (45), the Native American fit (+1), and the probation officer fit (−5).

The use of residuals in EDA differs from that in CDA in several ways. First, although the logic of reducing residuals by use of an improved model is inherent in ANOVA and regression, it is seldom discussed explicitly outside of model comparison approaches to these techniques (e.g., Maxwell &

Table 4
Additive Occupation and Ethnicity Fits With Residuals and Overall Fit From Median Smoothing of Table 2

Occupation	Residuals by ethnicity: Native American	Hispanic	White	Occupation fits
Armed forces	4	0	−7	26
Photographer	0	0	4	20
Lawyer	−2	4	0	19
Social worker	7	0	0	11
Secretary	0	5	−1	6
Accountant	−5	0	0	6
Electronics technician	2	0	−6	6
Electrician	1	0	−3	2
Fashion shop manager	−1	9	0	2
Receptionist	−9	4	0	−2
Probation officer	17	9	−9	−5
Physical therapist	−4	0	3	−5
Bookkeeper	7	−1	0	−6
Veterinarian	−5	0	19	−7
Medical technician	−2	4	0	−9
Auto salesperson	0	0	−5	−12
Librarian	12	0	−3	−22
X-ray technician	4	0	−2	−22
Ethnicity fits	1	0	−2	45 (Overall fit)

Note. Rows are reordered by size of fit.

Delaney, 1990). Second, CDA generally assesses the size of residuals in global summary statistics such as the mean squared error (*MSE*). Because the *MSE* is based on the sum of squared residuals, the size of individual residuals is aggregated and the pattern of residuals obscured. After data are well understood and CDA is asking the constrained question concerning the relative size of residuals compared with model effects, *F* statistics and related techniques may be appropriate. However, when the underlying form of the data is not well understood, an exploratory data analyst is more likely to ask "Where are the good and bad fits and why?" rather than the more specific questions addressed in CDA.

This analysis represents a valuable start for understanding how perceptions of occupations vary across ethnicity. A bivariate structure of the table is suggested that offers detail about the size of effects well beyond noting the ethnicity with the highest options in each of the six significant chi-square tests reported by Lauver and Jones (1991). In sharp contrast to most applications of CDA, detailed analysis of residuals was used both to assess the model and to understand the data by examining their departure from the model. In EDA these residuals represent important deviations from expectations that inform us about the structure of the data rather than simply "error" that should be minimized.

In this example, the table consisted of percentages, yet the two-way fit is general enough to apply to other types of values, including frequencies and means. Tukey (1986c) noted that such decompositions of tables based on multiplicative or, as in this case, additive models were long considered a standard tool in data analysis. He noted the additive model underlies ANOVA for crossed and nested factors, whereas the multiplicative model underlies the chi-square test of independence in contingency tables. This accounts for the fact that, when using mean smoothing on cell means, the two-way fit provides the same results as the procedures recommended by Rosenthal and Rosnow (1991) for interpreting interaction effects in ANOVA. From the perspective of EDA, Rosenthal and Rosnow are recommending the use of *F* ratios for hypothesis testing and two-way fits with residual analysis for parallel EDA to help

build a rich description. Most programs for computing log-linear models will give similar results of parameter estimates and cell residuals following a multiplicative model. Hoaglin et al. (1991) discussed the two-way fit in detail using mean smoothing for a number of ANOVA designs.

An elegant graphic representation of two-way fits and residuals is available, although its presentation is lengthy and beyond the scope of this article. Interested readers may consult Tukey (1977) for its original treatment or Becker, Chambers, and Wilks (1988) or Statistical Sciences, Inc. (1993) for some computer implementations. Behrens and Smith (1996) should be consulted for an example using data from instructional psychology.

Use Robust and Resistant Methods

In the analysis of the two-way table, fits were based on medians rather than means. In EDA, robust estimators such as the median are generally preferred. Hoaglin, Mosteller, and Tukey (1983a) defined robustness as a concern for the degree to which statistics are insensitive to underlying assumptions. Mallows (1979) discussed three aspects of robustness: resistance, smoothness, and breadth. Resistance concerns being insensitive to minor perturbations in the data and weaknesses in the model used. Smoothness concerns the degree to which techniques are affected by gradual introduction of bad data. Breadth is the degree to which a statistic is applicable in a wide range of situations. Robustness is important in EDA because the underlying form of the data cannot always be presumed, and statistics that can be easily fooled (like the mean) may mislead.

Several approaches are available to assess the resistance of a statistic (cf. Goodall, 1983b), including the breakdown point (Hampel, 1971). Hampel (1974) defined the breakdown point as "the smallest percentage of free contamination which can carry the value of the estimator over all bounds" (p. 388). Discussing the resistance of regression lines, Emerson and Hoaglin (1983) explained a breakdown point as follows:

> Operationally, we can think of dispatching data points "to infinity" in haphazard or even troublesome directions until the calculated slope and intercept can tolerate it no longer and break down by going off to infinity as well. We ask how large a fraction of the data—no matter how they are chosen—can be so drastically changed without greatly changing the fitted line. (p. 159).

Other statistics can be assessed in a similar manner. For example, the percentage of data points that can be arbitrarily changed in a set of data without changing the mean is 0. In contrast, half the data of a distribution can be altered to infinity before the median changes, thereby giving the median a breakdown point of 0.5.

Additional resistant measures include the trimean, which is a measure of central tendency based on the arithmetic average of the value of the first quartile, the third quartile, and the median counted twice. The median absolute distance from the median is a measure of dispersion that follows its name exactly. Winsorizing (pulling tail values of a distribution in to match a preset extreme score) or trimming (dropping values past a preset extreme score) may also be used. Some researchers object to the differential weighting afforded data in these cases. This differential weighting is, however, no different from procedures commonly used by instructors who drop a student's lowest score or Olympic judging that is based on a mean score following the elimination of the highest and lowest scores. As in psychological work, these strategies seem justified if the results downplay errant values while offering an otherwise expected summary. In one of his most influential papers, Fisher (1922) argued that "assuredly an observer need be exposed to no criticism, if after recording data which are not probably normal in distribution, he prefers to adopt some value other than the arithmetic mean" (p. 323). Lind and Zumbo (1993) presented an overview of robustness issues in psychological research as did Wainer (1977a).

Although problematic, data requiring resistant summaries are not uncommon in psychological work. For example, Paap and Johansen (1994) reported the results of reaction time (RT) experiments aimed at evaluating their memory model of word verification. Among the data reported is the frequency with which each word used in the experimental task occurs in a standard corpus. The distribution of this variable is depicted in Figure 8. In addition to the extreme skew, the distribution is marked by an extreme outlier representing the word "that" with word frequency of 10,595 in the reference corpus. The second most frequent word used from this corpus is "than" with a frequency of 1,789. The mean word frequency is 267, and the median frequency is 47. The failure of the mean and median to give a common indication underscores the value of resistant measures. The graphic display and these numbers suggest that the mean can

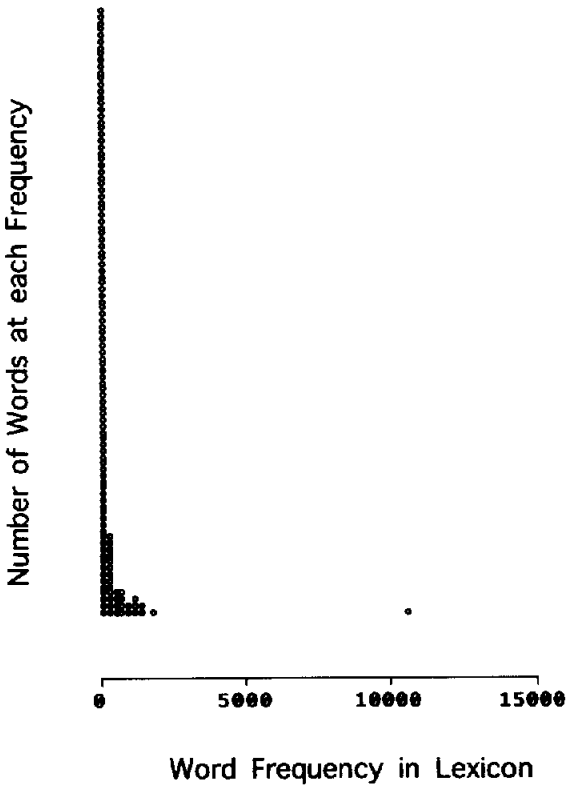

Figure 8. A dot plot of the distribution of word frequencies used by Paap and Johansen (1994).

easily mislead the researcher from the bulk of the data and that the median is a good fit for most of the data points.

Pay Attention to Outliers

Although resistant measures guard against misinformation from small perturbations, sometimes perturbations are so great that inclusion of the bulk of data along with well-documented oddities leads to meaningless summary statistics. In EDA, extreme or otherwise unusual data are noted as outliers so they may be treated differently or call increased attention to a phenomenon. The problem of outliers has a long history. Hampel, Ronchetti, Rousseeuw, and Stahel (1986) noted that discussion of the omission of outliers goes back as far as Bernoulli (1777/1961) and Bessel and Baeyer (1838). Hampel et al. provided additional references and notes, including Bernoulli's remark that rejection of outliers was commonplace among astronomers of his time. The discussion concerning the separation of extreme values has not ended (cf. Barnett & Lewis, 1994; Hawkins, 1980; Hoaglin & Iglewicz, 1987).

When working in an exploratory mode, comparison of patterns in data that include all or only a subset of data is considered acceptable if the actions taken and the rationale are documented. This intrusion of subjectivity is deemed important because failure to seek outliers supposes all data are of equal importance and similar to the underlying process being observed. Hawkins (1980) defined an outlier as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (p. 1), whereas Barnett and Lewis (1994) defined an outlier as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" (p. 7). As the detective analogy suggests, the outlying data are telling a different story from the rest of the data, and to try to summarize all of the data with a single model or statistic leads to a case of combining apples and oranges.

Temporarily setting aside an observation allows a diagnostic assessment of the role of the value in the summary statistics. For example, the effect of the word "that" in the Paap and Johansen experiments can be assessed by computing the mean both with and without the word included. When the word is removed, the mean of the data drops to 183 from the original 267. This change is considerable because the observation comprises only 1/128 or 0.8% of the data. This temporary diagnostic setting aside may lead the data analyst to set the observation aside for the remainder of the analysis or continue with it in the data set. A common extension of "setting one aside" is the generalized jackknife procedures (Efron, 1982; Mosteller & Tukey, 1977). When conducting a jackknife procedure, the data analyst repeatedly removes subsets of the data and recomputes a statistic of interest with the eye for deviations in the statistic across subsamples. Homogeneity of the statistics reflects homogeneity of information in the data, whereas variability in the statistics reflects variability in the data, as seen previously. Although once considered only as EDA techniques, such procedures have become mainstream methods in areas including regression diagnostics that use a "leave one out" approach in measures such as Cook's distance and diffitts (cf. Atkinson, 1985; Cook & Weisberg, 1994).

An idea closely related to outliers is that of fringeliers. Fringeliers are unusual points that are not as clearly deviant as outliers but may appear with unusual frequency in unexpected ways (Wainer, 1977a). A group of observations clumped three standard de-

viations from the mean would be one example. As with outliers, relating the structure of fringeliers to the phenomenon being studied is the best possible outcome. Hadi and Simonoff (1993) discussed a number of similar issues for outlier detection in multivariate models. Although the data analyst working in the exploratory mode is likely to set an observation aside if it allows sensible description of the remaining data, such a decision must consider other possible representations of the data, weigh gains and losses of information, and document the status of any data that were set aside.

Identification of outliers serve not only to improve the model of the remaining data but also to call attention to important aspects of the data that were not originally considered. Sometimes the outliers provide information about research-related processes such as typographical errors, malfunctions in recording machinery, or programming errors. On other occasions, outliers may point to important aspects of the phenomenon being studied that were unanticipated. Although outliers may be set aside from the bulk of the analysis, they are not simply dismissed. On the contrary, they should be considered in as much detail as possible to understand the process that generated them. Insofar as outliers tell us our original expectations were wrong, they provide feedback for correcting our mental and computational models. Beveridge (1950) discussed a number of examples of scientific discovery motivated by the quest to understand such anomalous results.

Reexpress the Original Scales

Reexpression is deemed an important part of the data analyst's toolbox because, as Mosteller and Tukey (1977) argued, "numbers are primarily recorded or reported in a form that reflects habit or convenience rather than suitability for analysis" (p. 89). The term "reexpression" is preferred in EDA to the more common usage of "transformation," because it avoids the connotation of radical change of the underlying information.

Reexpressions of data have long been used in experimental psychology. Arcsine transformations of percentages are typically recommended in the ANOVA context (e.g., Winer, 1971), and raw scores are often reexpressed as standard scores. Nevertheless, transformations are suspect in many subdisciplines of psychology and outright rejected in others. The most common concern is that reexpression leaves data analysis as a subjective process where a transformation can be chosen to "prove anything." These concerns can be mollified by noting (a) reexpression of numerical values is common in everyday life and psychological work and (b) the goal of reexpression is to find a scale that represents the phenomenon in a meaningful way.

Reexpressing data as standard scores or log functions is a familiar practice in psychological research. Daily life holds experience of reexpressed scales as well. Hoaglin (1988) noted a number of examples of reexpression from everyday experience, including the Richter scale for earthquakes (a logarithmic scale); gasoline consumption (the reciprocal of the gas used times the number of miles driven); and camera lenses (f-stops are spaced in a logarithmic scale). Many data analysts reexpress their data without concern. Sums are commonly multiplied by the reciprocal of the sample size to obtain a mean. When frequency data are not interpretable in a straightforward manner, the frequency in a particular category times the reciprocal of the total possible frequency (the relative frequency) is commonly used. Sometimes this scale is further altered by multiplying the relative frequencies by 100 to obtain percentages.

In a complete treatment of reexpression, Emerson and Stoto (1983; cf. Emerson, 1991) gave several reasons why transformations are desirable, including the fact that they may (a) "facilitate interpretation in a natural way," (b) "promote symmetry in a batch," (c) "promote stable spread in several batches," and (d) "promote a straight line relationship between two variables" (p. 104).

Most of the examples just given work toward the goal of facilitating interpretation in a natural way. Promoting symmetry is often desirable because it allows for comparison of scores across different parts of a single distribution. Stable spread across batches of data likewise allows comparison of scores across multiple distributions. Promoting a straight line relationship is also valuable because most regression analyses follow a linear form. In many cases, an appropriate reexpression solves several of these goals. Wainer (1977b) showed how different conclusions from very similar experimental data were likely caused by the presence of an outlier in data when organized as RT, although when it was reexpressed as speed data (1/RT) the extreme values were nonproblematic. By promoting symmetry, this reexpression also facilitated interpretation.

Reexpression is essential in EDA because it addresses the often neglected issue of scaling. Research-

ers may focus on precisely aligning their theoretical hypotheses with statistical tests and yet fail to collect preliminary evidence concerning the distributional form their measurements take. Although a relationship between two variables may be hypothesized on the basis of theoretical work, the statistical analysis occurs on an empirical realization that is a result of the underlying form of the constructs and the way in which the constructs are measured. Failure to delve into a detailed analysis of the form of the distributions and the reexpressions that make them most interpretable can lead to glaring misinterpretations of the data.

Putting It All Together: A Reexamination of the Paap and Johansen Data

The preceding sections have examined a number of foundational tools in the EDA toolbox. The theme common to all the procedures described previously is not the use of canonical technique but a willingness to use any technique that helps ensure a rich mental model of the data that fits closely with the true form of the data. This requires a high degree of interactivity with the data and a familiarity with a wide range of techniques. To illustrate how these principles and procedures interact, data published in the *Journal of Experimental Psychology: Human Perception and Performance,* which were introduced previously in the discussion of resistance, are now reexamined from an EDA perspective. In this article, Paap and Johansen (1994) reported numerous analyses of several data sets used to support their theory of word verification, including RT data for 128 words from a standard experimental lexicon. Variables associated with each word include the average RT to respond to each word in an experimental task, the number of high-frequency neighbors (HFN), neighborhood size (NS), word frequency in the lexicon (WF), and the summed bigram frequency (SBF). A neighbor is a word created by changing a single letter in an original word. HFN are words similar to the original word that occur often in standard use. The SBF is a measure of position-specific bigrams that occur in the word. The authors summarized their expectations toward these data arguing "the only variable that directly determines word RT is the number of HFNs . . . Thus, NS, SBF and WF are all indirectly effects that should not account for any of the variance in word RT once the effects of HFN were partialed out" (pp. 144–145).

Paap and Johansen tested this hypothesis by using OLS multiple linear regression, which is commonly referred to as "multiple regression." The fuller name, however, reminds us of the assumption of linearity inherent in that procedure and the sensitivity of the regression line to extreme values when the OLS approach is used. While keeping an eye on their original hypothesis, working in an exploratory mode allows broader questions such as: How are the independent variables related to each other and the dependent variable? What patterns underlie the results reported by Paap and Johansen? What can be done to improve the model? What can we find that we did not expect? How might we be fooled by the summaries?

A first look. When working with multivariate data such as these, a common strategy is to examine variables individually and then in bivariate and higher order configurations. Figure 9 depicts the shapes of distributions from this analysis using boxplots. An analyst working on these data should view histograms, density plots, and dot plots as well. Before suggesting first aid for these messy distributions using outlier handling or reexpression, it is often helpful to assess how the shapes of these distributions affect assessment of bivariate and higher order relationships in the data. This can be done graphically using a scatter plot matrix (also called a generalized draftsman's display) described in Chambers et al. (1983) and shown in Figure 10. The plot presents all pairwise combinations of the five variables of interest. The graphic may be thought of as a pictorial correlation matrix with scatter plots replacing correlation coefficients. In this version of a scatter plot matrix, normal probability plots are presented in the matrix diagonals, with variable labels indicating the associated row and column scales. For example, on the top row of plots, the RT measure is plotted on each vertical axis while the x-axes vary from HFN to NS to WF and SBF as one moves from left to right in that row. The HFN variable is plotted on the vertical axis of all plots in the second row and on the horizontal axis of plots in the second column. The top right-most plot indicates RT on the vertical axis and SBF on the horizontal axis.

Normal probability plots are a diagnostic aid used to assess the degree to which the empirical distribution matches the Gaussian distribution. This is accomplished by calculating the fraction of data below each data value (i.e., the quantile) and computing the z

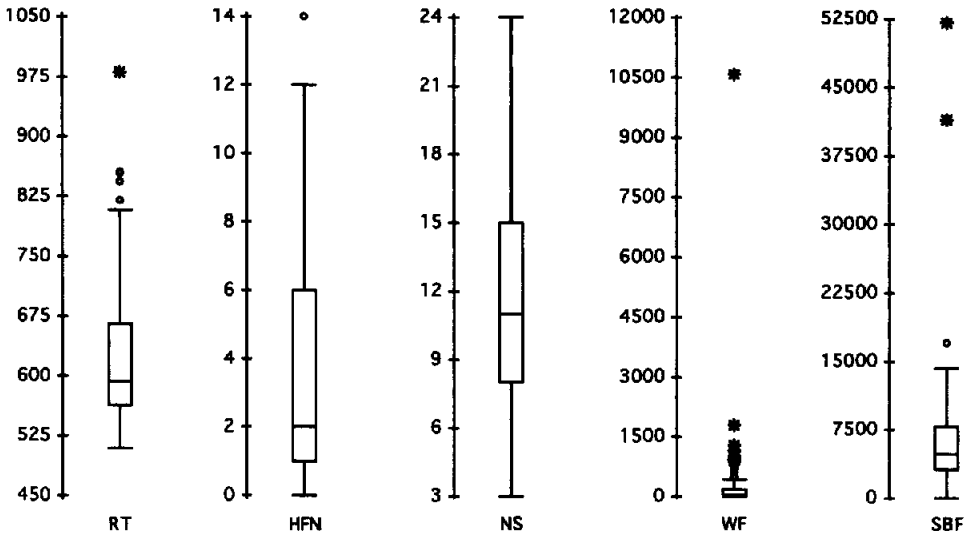

Figure 9. Box plots of word verification variables from Paap and Johansen (1994) indicating severe nonnormality and outliers. The locations of data in the upper and lower quartiles are marked using lines extending away from the box. The O indicates outliers that deserve special attention. The * indicates extreme outliers. (RT = reaction time; HFN = high-frequency neighbors; NS = neighborhood size; WF = word frequency; SBF = summed bigram frequency.)

score for points with corresponding quantiles in the Gaussian distribution. When the scale values are plotted against the expected z score, a straight line is obtained if the distribution is Gaussian. Curves in the normal probability plot indicate skew, whereas S shapes indicate shorter than expected tail regions. Cleveland (1993) presented a complete discussion of the normal probability plot and the more general quantile-quantile plot. The farthest left plots in Rows 1 and 2 of Figure 10 indicate moderate positive skew in RT and HFN while the NS plot indicates relative normality, and WF and SBF plots indicate marked deviation from Gaussian shape. Individual outliers are marked as "*xs*." Although these patterns were visible in the box plots, normal probability plots are an important adjunct because they are compared directly against the normal distribution and display each piece of datum rather than the five-number summary of the box plot. The extremity of the word "that" in WF can be seen in the bivariate plots.

One natural method for summarizing the bivariate relationships between variables is to use the formula for a line as a fit from which to derive residuals. If an OLS fit is used (as is the default in most computer packages), the line is easily affected by extreme values such as the outliers in WF and SBF, which represent the common words "that" and "than." In interactive data analysis environments common to EDA, quick assessment of such effects is straightforward. In this case, regression lines were added to a number of the scatter plots in Figure 10 to indicate the OLS predictions that would be computed with and without the two outlying values. This was accomplished by selecting options from pull-down menus accessible on the scatter plots themselves (Data Description, Inc., 1995). In each case, the regression line nearest the outlier indicates prediction lines with the outliers included.

It is clear from these plots that the extreme outliers are dramatically different from the bulk of the data and disproportionately influence the fit from the least squares line. Likewise, these extreme points are artificially inflating or deflating the correlation that holds in the mass of the data. For the case of WF, the outliers pull the line toward a slope of zero when compared against the negative slope that exists when the outliers are set aside. In addition to the difficulty with the regression lines being disproportionally affected by these points, their presence compresses the variability in the bulk of the data and may hide important patterns. Because these two outliers appear to be qualitatively different from the bulk of the data, unduly influence the OLS summary, and may distort the visual impression of the data, it is advisable to set them aside for some portion of the analysis.

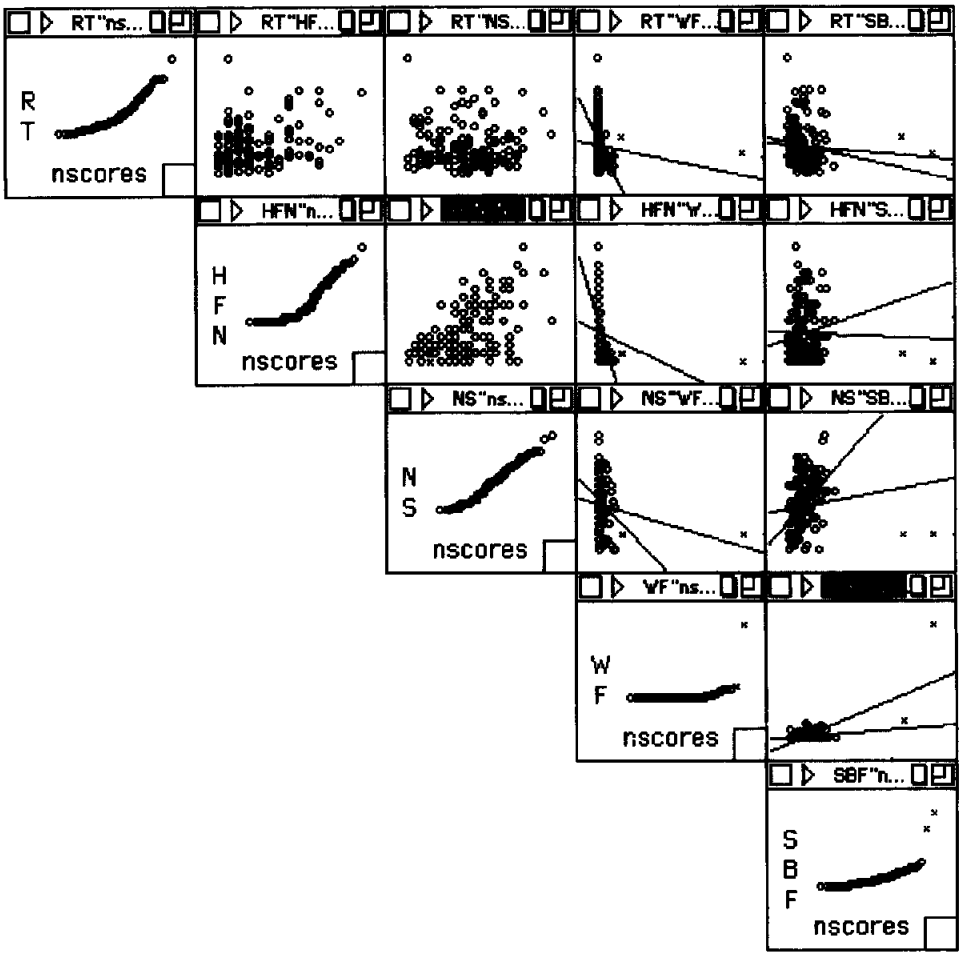

Figure 10. Scatter plot matrix of word verification data from Paap and Johansen (1994). Outliers are indicated with the "x" symbol. Regression lines have been added to indicate predicted values as they would occur with and without the outliers included. (RT = reaction time; NS = neighborhood size; HFN = high-frequency neighbors; SBF = summed bigram frequency, WF = word frequency.)

A Better Description. Temporarily setting aside the two outlying data points and reconstructing the scatter plot matrix leads to the display in Figure 11. In this plot the relationships with SBF are clearer (although not very strong), and curvilinear relationships between WF and both RT and HFN are visible. These curvilinear relationships are not completely unexpected. The curved form of the data is reflected in the bunching up of the data in the lower left corner of the two-dimensional plot. This is likely to occur given the bunching of data in the lower part of each of the univariate distributions.

A straightforward way to find an appropriate description for the curved function is to find a reexpression of the univariate distributions that leads them to a roughly Gaussian shape. By finding the degree to which the univariate distributions need to be reexpressed to be Gaussian, one also finds the degree to which the line of fit must be bent to meet the data. When reexpressing variables in EDA, one may use the notion of a ladder of reexpression. A number of versions of the ladder exist. In each case the rungs of the ladder represent an exponential value to which scores may be raised. In the simplest case, movement up the ladder refers to raising scores to a higher power. Moving down the ladder refers to raising scores to decreasing negative exponents (reciprocals). Exponents along this ladder and the corresponding reexpression are listed below for the range of exponents from −2 to +2.

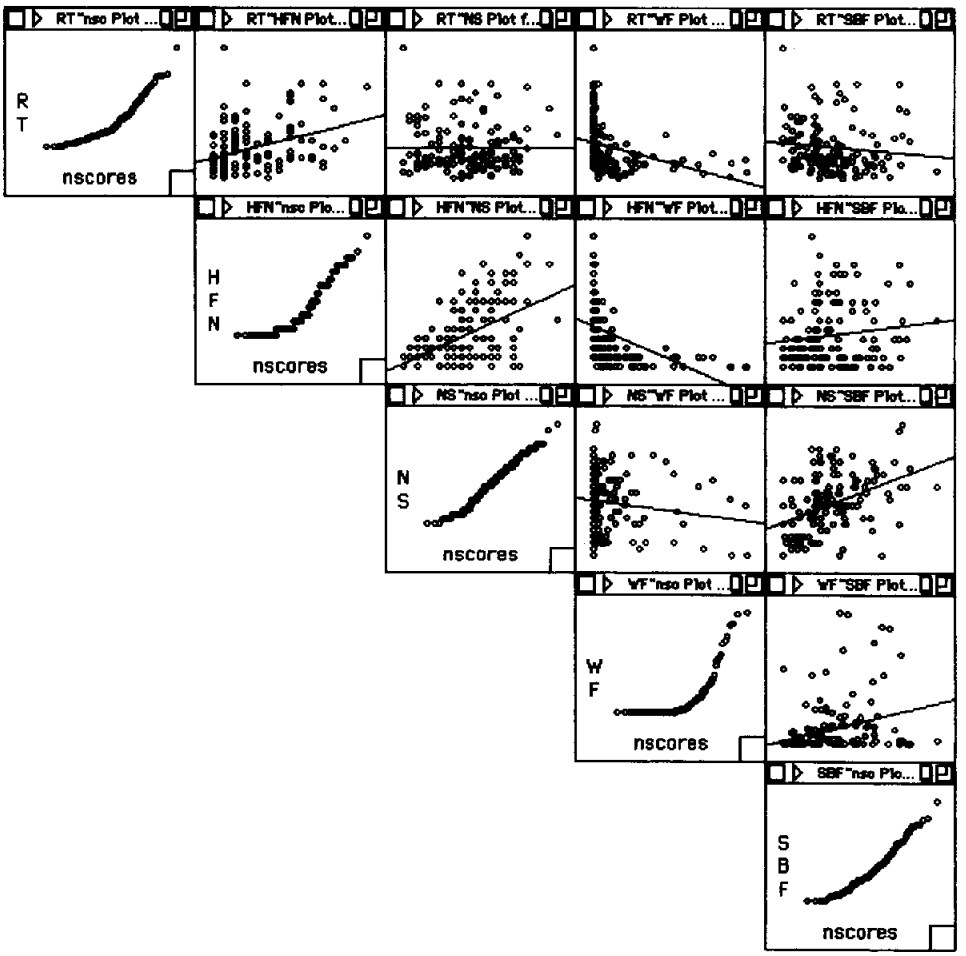

Figure 11. Scatter plot matrix of word verification variables from Paap and Johansen (1994) with two outliers removed and curvilinear relation between WF and RT as well as WF and HFN exposed. (RT = reaction time; NS = neighborhood size; HFN = high-frequency neighbors; SBF = summed bigram frequency, WF = word frequency.)

Exponent	Reexpression
$-x^{-2}$	$\frac{-1}{x^2}$
$-x^{-1}$	$\frac{-1}{x}$
$-x^{-1/2}$	$\frac{-1}{\sqrt{x}}$
x^0	$\log_{10}(x)$
$x^{1/2}$	$\sqrt{x}$
x^1	(no change)
x^2	x^2

The center of the ladder is an exponent of 1, which leaves a score unchanged. An exponent of 0 always returns a value of 1 so this is typically replaced by a $\log_{10}$ transformation, which fits in well between $X^{-1/2}$ and $X^{1/2}$ transformations (Mosteller & Tukey, 1977). Because negative exponentiation alone reverses the ordering of observations, most analysts prefer using $-X^{-n}$ rather than simply X^{-n}, as reflected above. To facilitate reexpression, the most advanced EDA programs provide menu options for transformations along the ladder of reexpression as well as slider bars that can be moved up and down similar ladders (such as the Box-Cox family of reexpressions). Histograms, box plots, normal probability plots, and any other graphics or summary tables are updated automatically

as the slider values change. Such systems allow quick assessment of a large number of reexpressions.

A choice of transformation is recommended by moving up or down the ladder in the direction of the bulk of the data on the scale. Positively skewed distributions with the bulk of the data lower on the scale can be normalized by moving down the ladder of reexpression; distributions with the bulk of the data high on the scale can be normalized by moving up the ladder of reexpression. In the present case, the WF variable has the bulk of the data in the lower portion of the distribution, so moving down the ladder is appropriate. Starting with WF^1 (the unchanged data), we move down to $WF^{1/2}$, which is the square root of WF, followed by WF^0, which is assigned the value of $\log_{10}(WF)$, and $-WF^{-1/2}$, which is equal to minus one over the square root of WF. Box plots of each of these transformations for all data, including the outliers, are presented in Figure 12. As the reader may see, reexpression to a log transformation provides an approximately Gaussian distribution, whereas more extreme reexpression leads to distortion in the opposite direction and less extreme reexpression fails to correct the shape. In practice, normal probability plots rather than box plots would be used to assess normality. Box plots, however, effectively and compactly communicate the effect of the reexpressions.

Panel a of Figure 13 is a scatter plot of RT regressed on $\log_{10}(WF)$ with the two outliers indicated by "*X*s" and regression lines for models with all the data as well as from outlier-deleted data only. Panel b is a plot of the regression residuals versus $\log_{10}(WF)$. These plots indicate that the log reexpression improves the fit dramatically, whereas the similarity of regression lines with and without the outliers (Panel a) indicates that setting aside the outliers does little to change the regression line. In this dimension the outliers may be considered natural extensions of the tail of this log-normal distribution. A very misleading fit would occur in any model assuming RT = $a + b$(WF) rather than RT = $a + b$(log(WF)). The residual plots also indicate three words whose residuals are exceptionally large as indicated by their positions above the bulk of the data in the left, center, and right side of the residual plots. These points indicate the values of the words "oaf," "mere," and "came" respectively, with RTs much longer than otherwise expected given their word frequency. Analyses of the SBF variable lead to the conclusion that the SBF variable is roughly Gaussian with the exception of the two outliers.

To properly specify a linear model using the WF variable, it should be reexpressed to log(WF). To appropriately include SBF, the two extreme points should be set aside and noted for their impact on the analysis. Including the two outliers in subsequent analyses would serve no purpose but to demonstrate that the majority of the SBF pattern cannot be well modeled because of two rogue points. Setting them aside will allow appropriate modeling of the bulk of the data. This is a practical application of the principle that it is better to be somewhat right than precisely wrong. All of this information suggests the corpus of words used in this study requires additional attention.

How did we do? To assess the total effect of the

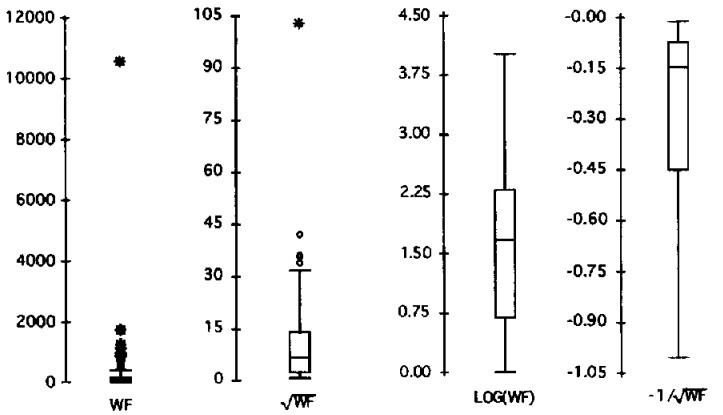

Figure 12. Box plots of reexpressions of word frequency (WF) moving down the ladder of reexpression. The location of data in the upper and lower quartiles are marked using lines extending away from the box. The ○ indicates outliers that deserve special attention. The * indicates extreme outliers. (RT = reaction time; HFN = high-frequency neighbors; NS = neighborhood size; WF = word frequency; SBF = summed bigram frequency.)

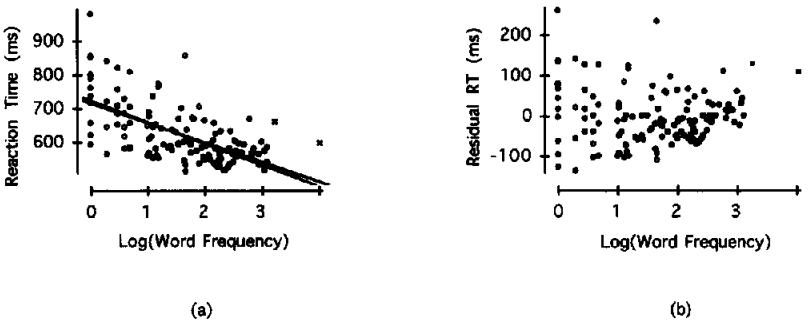

Figure 13. Panel a: Scatter plot of reaction time (RT) versus log (word frequency[WF]) with regression lines for all data and outlier deleted data. Panel b: Residuals from regression of data in Panel a.

work we have done up to this point, we replot the scatter plot matrix with the SBF outliers removed and the reexpressed WF variable as shown in Figure 14. Comparing the quality of regression lines for predicting data in Figure 14 with that of Figure 10 underscores the value of EDA in steering researchers toward an appropriate model. The reexpression corrected the curvilinearity in the WF–RT relationship as well as in the WF–HFN relationship. Because it is sometimes difficult to understand the reexpression being used, readers may benefit from seeing predicted values from reexpressed variables plotted in the scale of the original variables as shown in Figure 15. Panel a portrays the predicted values of RT from log(WF) plotted against log(WF), and Panel b portrays the same values plotted against their corresponding WF values. When viewed in conjunction with the RT versus WF plot in Figure 11, Figure 15 reveals the relation between log(WF) and WF is a natural reexpression that catches the curve in the data that is otherwise missed by failing to reexpress the data.

Interestingly, Paap and Johansen (1994) noted that log transformations have been computed in other research labs and have led to substantive conclusions different from their own. They therefore reanalyzed the data described here using the log(WF) transformation on the grounds of historical precedence in RT experiments. Focusing primarily on the size of the correlations and the role of the logged variable in a multiple regression analysis predicting RT, the improvement in fit observed here was interpreted quite differently.[2]

> In summary, when plain WF was entered as a predictor, the number of HFNs and NS were significant predictors. However, when log WF was used, only log WF was a significant predictor. The skittishness of the variables in these analyses may have occurred because the collinearity problem between the predictors actually became worse when the log transform was applied. The correlation of −.23 between plain WF and the number of HFNs ballooned to −.65 for the log WF. The greater the collinearity between two predictors, the less confident one can be that the statistical model has identified the real winner. . . . Because of the collinearity problem, some will see the hole (effects of log WF) where others see the doughnut (effects of NS and the number of HFNs) in our data. (pp. 1145–1146)

Without the log transformation, these authors found what they predicted: a significant relationship between RT and HFN and a nonsignificant relation between RT and WF. Alternatively, the logarithmic reexpression led to a nonsignificant correlation between RT and HFN and a significant correlation between RT and log(WF), results inconsistent with their theory. Without understanding the shapes of the distributions involved and the effect of curvilinearity and outliers, these authors were left to hypothesize "skittishness" and "ballooning" variables, collinearity, and a positive-thinking bakery theory for choosing among statistical models. The simple graphics used here, however, explain the situation quite well. WF has a curvilinear (logarithmic) relationship with RT and HFN. This curvilinearity is a violation of an assumption of the linear regression model used. Therefore, no significant slopes can be found, as indicated in Figure

[2] The analysis discussed in the passage quoted here discusses a model with a term for the summed log bigram frequency rather than the summed bigram frequency discussed in this article. This difference did not affect the relationship among RT, HFN, and WF discussed here and is omitted for the sake of simplicity.

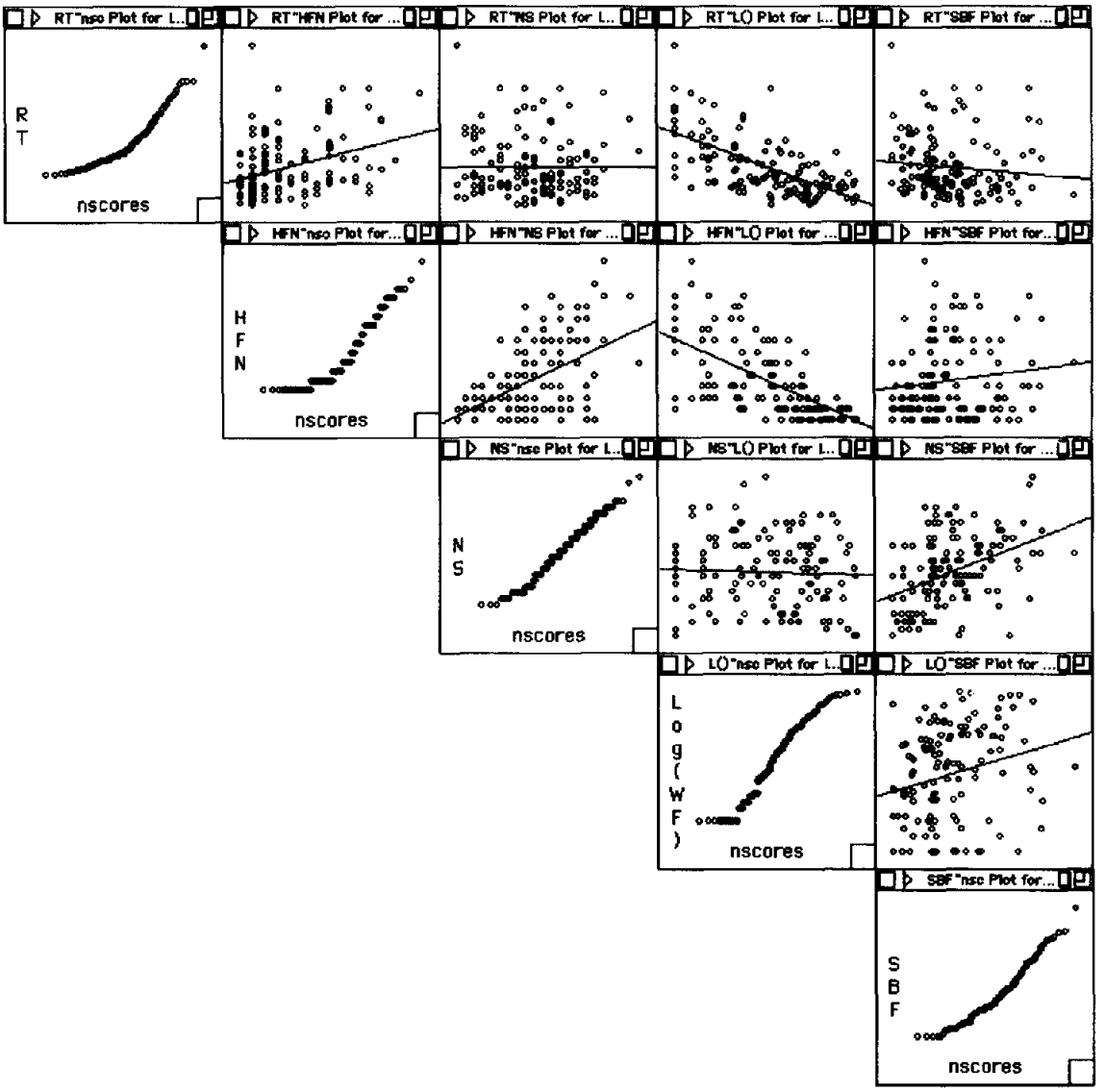

Figure 14. Scatter plot matrix of the Paap and Johansen (1994) data with log reexpression of WF and two outliers removed. (RT = reaction time; NS = neighborhood size; HFN = high-frequency neighbors; SBF = summed bigram frequency, WF = word frequency.)

10. The log transformation specifies the degree of bend in the data so it can be accommodated by the regression model RT = a + b(HFN) + b(NS) + b(SBF) + b(log(WF)). The correct log(WF) model specification reveals a strong curvilinear relationship between RT and WF as well as HFN and WF.

The disappearance of the HFN effect needs to be understood in the context of the multiple regression models used. In such models, relationships between each predictor variable and the criterion are adjusted for the presence of all the other predictor variables. When WF is included in the equation in its original form, the suppressed measure of relationship leads to little correction in the HFN–RT relationship. When, however, WF is appropriately reexpressed to account for the curvilinearity, its relation with HFN is properly expressed as high linear relationship and the relation between HFN and RT is adjusted downward to take into account the now large correlation between HFN and log(WF).

These important aspects of the analysis can be inferred from Figure 14, correlation matrices, and the

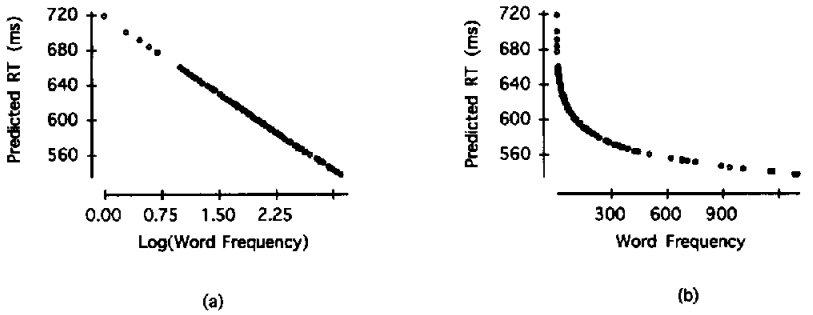

Figure 15. Plot of predicted values from linear regression of reaction time (RT) on high-frequency neighbor, neighborhood size, log(word frequency), and summed bigram frequency on (Panel a) scale of log(WF) and (Panel b) scale of WF. Note how the predicted values properly model the curve of the data in the original scale.

slope estimates of the multiple regressions. Nevertheless, the exploratory analyst may want additional information because partial correlations and conditional slopes may also be distorted by aberrant data patterns. To obtain a more detailed description of the data as represented in the machinery of the multiple regression and to assess the validity of the computational model, partial regression plots can be used. A partial regression plot takes advantage of the fact that adjusting one predictor, such as HFN, for its relationship with another predictor, such as log(WF), is equivalent to regressing HFN onto log(WF) and using the residuals for subsequent analyses. Because the residuals are the data after the effect of the model have been subtracted, the residual from HFN $= a + b(\log(\text{WF}))$ are WF-corrected HFN data. Any analysis with these residuals would be equivalent to a partial regression of HFN. According to Velleman (1992), "A partial regression plot *graphs y with the linear effects of the other x-variables removed* against *x with the linear effects of the other variables removed*" (pp. 23–24).

Panel a of Figure 16 is the partial regression plot between RT and HFN when each is adjusted for NS, WF, and SBF. Note that the slope of the line indicates that a relationship exists between these two variables after their adjustment for relations with other variables. Panel b is a partial regression plot between RT and HFN when both are adjusted for NS, log(WF), and SBF. The absence of relationship between the two sets of residuals in Panel b reflects the small partial correlation between these variables that has occurred because the properly specified model leads to appropriate measures of relationship with WF and RT and HFN. The WF variable is not skittish but strongly curvilinear in a world in which these data analysts assumed all relationships are linear.

Life without EDA. The interactive analysis described here contrasts Paap and Johansen's (1994) use of CDA alone. Working in an exploratory mode, an initial model was sought, transformations were attempted, residuals were used for model evaluation, and the cycle of model searching continued. This pro-

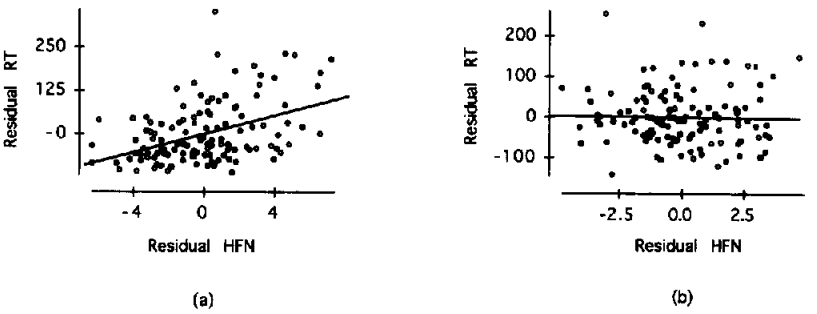

Figure 16. Partial regression plots of reaction time (RT) and high-frequency neighbor (HFN). Each variable is adjusted for linear relations with all other predictors. Additional explanatory variables are neighborhood size (NS), word frequency (WF), and summed bigram frequency (SBF) for Panel a and NS, log(WF), and SBF for Panel b.

cess quickly revealed numerous unexpected aspects of the data with important consequences for model development. Without the detailed description and open attitude available in EDA, Paap and Johansen were left with seemingly conflicting statistics from the black box of the hypothesis tests. They began with specific hypotheses concerning what variables would be related under what conditions, but, because of a lack of detailed familiarity with the data, they failed to specify a model even close to the empirical outcome. This underscores the idea that theoretical hypotheses need to be balanced with rich knowledge of the data being examined. Even when firm hypotheses are held a priori, working in the exploratory mode is always useful to find out what we did not expect. The only caveat required is that conclusions obtained as the result of exploratory analyses are considered exploratory and that confirmation of such conclusions will occur only when CDA is undertaken on different data.

The analysis reported here is a small part of an exploratory analysis of this data. Although the logarithmic transformation appears to lead to quite good model specification, the RT variable has some departure from symmetry and may benefit from a 1/RT reexpression that would put it in the scale of speed. Other types of regression diagnostics and plots could have been used such as three-dimensional rotating plots and the assessment of other outliers.

Conclusion: Psychological Method and EDA

EDA is a well-established tradition in the statistical literature. The goal of EDA is to find patterns in the data that allow researchers to build rich mental models of the phenomenon being examined. Examining the Feingold (1994) and Lauver and Jones (1991) data, we found EDA useful when there is little explicit theoretical background to guide prediction and the first stages of model building is desired. Examining the Paap and Johansen data, we also saw that, even when a priori hypotheses exist, EDA can perform a valuable service by providing rich descriptions of the data that can inform the research whether their mental models are even close enough to the underlying data patterns to consider CDA. In either case, tools for EDA provide a much wider range of information than the answers to a specific probabilistic question.

Theory development and testing are hallmarks of scientific psychology. Good theory development and testing integrate a wide range of information about the data being evaluated, so researchers can be certain they have not missed important aspects of the phenomenon and have not been fooled by pathological data patterns or model misspecification. EDA promotes good theory development and testing by helping researchers ensure their models are aligned with reality and they are not being misled by more removed summaries. As long as researchers are clear about what activity is exploratory and what is confirmatory, and the strength of conclusions from each mode are appropriate, EDA will facilitate, rather than retard, theory development and testing. An increase in our knowledge about the data is always beneficial as long as its limits are clear. From this analysis a number of recommendations can be made.

First, EDA should be recognized as an important aspect of data analysis whose conduct and publication are valued. By admitting EDA as an acceptable set of procedures, researchers can avoid the improper use of CDA techniques for the purposes of data exploration. As long as EDA remains a covert activity, researchers will continue to improperly use CDA for data exploration through model underspecification and overtesting. An increase in EDA will focus more resources at the preliminary stages of investigations and less at the advanced stages. In so doing, the number of irreproducible results may be reduced by the substitution of adequate model building for the cataloging of significant effects. Further, the detail in modeling afforded by EDA may improve our understanding of phenomenon otherwise hidden behind simple summary statistics and tests, as seen in the Paap and Johansen (1994) data. In this regard, editors and reviewers should follow the lead of Loftus (1993), whose first editorial statement for *Memory and Cognition* included headings of "Figures are Good" and "Data Analysis: A Picture is Worth a Thousand Words."

This is not to say that all exploratory work should be published, but rather that all published and initial work should be explored. The field would greatly benefit if all published reports included the statement "we examined the data in detail and found the patterns underlying the summary statistics were not obviously pathological." More detailed reporting would also be welcome. When auxiliary exploratory analysis cannot fit into a standard journal format, additional graphics and reports may be distributed over the Internet or by other electronic means. Behrens and Dugan (1996) provides an example of such supplemental graphic reporting.

Second, quantitative analysis should be thought of "more as applied epistemology and less as applied

mathematics" (Behrens & Smith, 1996). When considering statistics as applied mathematics rather than applied epistemology, many messy real-world issues are often swept under the rug. Instruction addressing the assumptions that must be met for a statistic to be meaningful almost always focuses on assumptions about theoretical distributions rather than assumptions about the world. Sharing this value with EDA, Box (1976) labeled the overemphasis on theoretical issues "mathematistry" for which he prescribed practical experience and trust of the scientist's intuitions. By focusing on understanding the data in whatever way is reasonable (not only probabilistically), EDA opens the data analyst to consider the wide range of ways of knowing about data. This ecumenical view leaves researchers considering mathematics as an epistemic tool rather than a complete answer in itself. Mathematics should be used based on how helpful it is in understanding data, not simply on its syntactical correctness. Such a position will minimize what has been referred to as Type III error: "precisely solving the wrong problem, when you should have been working on the right problem" (Mitroff, Kilmann, & Barabba, 1979, p. 140, cited in Barabba, 1991).

Third, graduate programs should integrate instruction in confirmatory statistics with alternative data analytic methods. Instruction in EDA offers students a view of data analysis from outside traditional statistics. Such an alternate view may allow new appreciation and understanding of CDA. Other complementary methods include meta-analysis (Glass, 1976; Glass, McGaw, & Smith, 1981), Bayesian analysis (Howson & Urbach, 1993; Winkler, 1993), interval estimation approaches, and hybrid combinations (Box, 1980). Just as history and systems of psychology are taught in psychology, might students not benefit from a history and systems of data analysis? The appropriate size of such curricular additions will vary across programs. At the very least, the idea of multiconceptual approaches could be incorporated in already existing classes.

Fourth, data analysts should recognize that subjectivity and potential bias are inherent in all data analysis, exploratory or otherwise. One great danger in overmathematizing data analysis is believing that the reliability and precision of mathematics itself imbue reliability and precision to the data and the data analysis. The artifactual nature of psychological investigation has been well established by Rosenthal (1966), Rosnow (1981), Danziger (1990), and others. Understanding of the role of cognitive, historical, and social artifact in data analysis is also emerging. Rosenthal (Cooper & Rosenthal, 1980; Rosenthal & Gaito, 1963, 1964) demonstrated a consistent overweighing of "significance" in light of varying sample sizes, and Bar-Hillel (1989; Bar-Hillel & Falk, 1982) illustrated the subjectivity inherent in translating mathematical concepts into natural language. Flow charts and expert systems suggest data analysis is a purely rational process, yet choice of data analytic behavior is ultimately dependent on the same psychological factors that affect cognition and behavior in other spheres of life. Bias in data analysis will not be mollified by assent to stricter design and control of Type I error, but by the detailed analysis of data that excludes alternate statistical explanations as demonstrated previously.

Fifth, psychologists should consider the possibility that their craft can improve the conduct of EDA and data analysis in general. For example, Simon (1973; Simon, Langley, & Bradshaw, 1981) has long held that logics of discovery are possible and psychologically tractable, a position supported by the construction of the BACON program (cf. Langley, Simon, Bradshaw, & Zytkow, 1987). Gigerenzer (1991) noted that the heuristics encoded in BACON are quite similar to those of EDA and mentioned Tukey (1977) specifically. Investigations are still needed to examine the processes involved in comprehending common statistical graphics (cf. Simkin & Hastie, 1987; Kosslyn, 1989; Lewandowsky & Spence, 1989, 1990) as well as those specific to EDA (cf. Behrens, Stock, & Sedgwick, 1990; Stock & Behrens, 1991). The statistical community recognizes the potential of transdisciplinary work and has provided open invitations to the psychological community (Kruskal, 1982; Mosteller, 1988; Tukey & Wilk, 1986).

Given dramatic improvements in computational ability and increased sensitivity to the psychological and social aspects of data analysis, the time is ripe for a broad conceptualization of data analysis that includes the principles and procedures of EDA. Lest these recommendations seem dogmatic, the final word is left for Neyman and Pearson (1928) from "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference." This article represented the first great break from the Fisherian view (introducing alternative distributions and Type II error) and the beginning of current practice. Their attitude toward mechanized inference can easily be deduced. It is, in fact, good counsel for consideration of any method:

> The process of reasoning, however, is necessarily an individual matter, and we do not claim that the method which has been most helpful to ourselves will be of greatest assistance to others. It would seem to be a case where each individual must reason out for himself his own philosophy. (p. 230).

References

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.

Atkinson, A. C. (1985). *Plots, transformation, and regression: An introduction to graphical methods of diagnostic regression analysis.* Oxford, England: Oxford University Press.

Bar-Hillel, M. (1989). Discussion: How to solve probability teasers. *Philosophy of Science, 56,* 348–358.

Bar-Hillel, M., & Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition, 11,* 109–122.

Barabba, V. P. (1991). Through a glass *less* darkly. *Journal of the American Statistical Association, 86,* 1–8.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (2nd ed.). New York: Wiley.

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The New S Language: A programming environment for data analysis and graphics.* Monterey, CA: Wadsworth & Brooks/Cole.

Behrens, J. T. (1996). *Course materials for EDP 691: Graphical and exploratory data analysis.* Available http://research.ed.asu.edu/classes/eda.

Behrens, J. T., & Dugan, J. G. (1996). *A graphical tour of the White Racial Identity Attitude Scale data in hyper-text and VRML.* Available http://research.ed.asu.edu/reports/wrias.

Behrens, J. T., & Smith, M. L. (1996). Data and data analysis. In D. Berliner & B. Calfee (Eds.), *The handbook of educational psychology* (pp. 945–989). New York: Macmillan.

Behrens, J. T., Stock, W. A. & Sedgwick, C. E. (1990). Judgment errors in elementary box-plot displays. *Communications in Statistics B: Simulation and Computation, 19,* 245–262.

Berk, K. N. (1994). *Data analysis with student Systat.* Cambridge, MA: Course Technology, Inc.

Bernoulli, D. (1961). The most probable choice between several discrepant observations and the formation therefrom of the most likely induction. *Biometrika, 48,* 3–13. (Original work published 1777)

Bertin, J. (1983). *Semiology of graphics* (W. J. Berg, Trans.). Madison: University of Wisconsin Press.

Bessel, F. W., & Baeyer, J. J. (1838). *Gradmessung in Ostpreussen und ihre Verbindung mit Preussischen und Russischen Dreiecksketten.* Berlin: Druckerei der Königlichen Akademie der Wissenshaften.

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

Bode, H., Mosteller, F., Tukey, J. W., & Winsor, C. (1986). The education of scientific generalist. In L. V. Jones (Ed.), *The collected works of John W. Tukey, Volume III: Philosophy and principles of data analysis: 1949–1964.* Pacific Grove, CA: Wadsworth. (Original work published 1949)

Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin, 16,* 335–339.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association, 71,* 791–799.

Box, G. E. (1980). Sampling and Bayes' inference in scientific modeling and robustness. *Journal of the Royal Statistical Society (A), 143,* 383–430.

Burt, C. (1961). Intelligence and social mobility. *British Journal of Statistical Psychology, 14,* 3–23.

Campbell, D. T. (1988). Descriptive epistemology: Psychological, sociological, and evolutionary. In E. S. Overman (Ed.), *Methodology and epistemology for social science: Selected papers of Donald T. Campbell* (pp. 435–486). Chicago: University of Chicago Press.

Chamberlain, T. C. (1965). The method of multiple working hypotheses. *Science, 148,* 754–759.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis.* Belmont, CA: Wadsworth.

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. V. Graphics.* Belmont, CA: Wadsworth.

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

Cleveland, W. S., & McGill, M. E. (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.

Cooper, H. M., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87,* 422–449.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research.* New York: Cambridge University Press.

Data Description, Inc. (1995). *Data desk, 5.0* (computer software). Ithaca, NY: Data Description.

Dunlap, K. (1935). The average animal. *Journal of Comparative Psychology, 19,* 1–3.

Efron, B. E. (1982). *The jackknife, the bootstrap, and other resampling methods.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Emerson, J. D. (1991). Introduction to transformation. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Fundamentals of exploratory analysis of variance* (pp. 365–400). New York: Wiley.

Emerson, J. D., & Hoaglin, D. C. (1983). Resistant lines for y versus x. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 129–165). New York: Wiley.

Emerson, J. D., & Stoto, M. A. (1983). Transforming data. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 97–128). New York: Wiley.

Emerson, J. D., & Strenio, J. (1983). Boxplots and batch comparisons. In D. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 58–96). New York: Wiley.

Erickson, B. H., & Nosanchuk, T. A. (1992). *Understanding data* (2nd ed.). Toronto, Ontario, Canada: University of Toronto Press.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116,* 429–456.

Finch, P. D. (1979). Description and analogy in the practice of statistics. *Biometrika, 66,* 195–208.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, 222,* 309–368.

Fisher, R. A. (1925). *Statistical methods for research workers.* London: Oliver and Boyd.

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

Giere, R. N. (1984). *Understanding Scientific Reasoning.* (2nd ed.). New York: Holt, Rinehart & Winston.

Gigerenzer, G. (1991). From tools-to-theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98,* 254–267.

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

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Good, I. J. (1983). The philosophy of exploratory data analysis. *Philosophy of Science, 50,* 238–295.

Goodall, C. (1983a). Examining residuals. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 211–246). New York: Wiley.

Goodall, C. (1983b). M-Estimators of location: An outline of the theory. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 339–403). New York: Wiley.

Greenhouse, J. B., & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 383–398). New York: Russell Sage Foundation.

Hadi, A. S., & Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association, 88,* 1264–1272.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics, 42,* 1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association, 69,* 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions.* New York: Wiley.

Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis.* Beverly Hills, CA: Sage.

Hawkins, D. M. (1980). *Identification of outliers.* New York: Chapman & Hall.

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

Hoaglin, D. C. (1988). Transformations in everyday experience. *Chance, 1,* 40–45.

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association, 82,* 1147–1149.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983a). Introduction to more refined estimators. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 283–296). New York: Wiley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983b). *Understanding robust and exploratory data analysis.* New York: Wiley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1985). *Exploring data tables, trends, and shapes.* New York: Wiley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1991). *Fundamentals of exploratory analysis of variance.* New York: Wiley.

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach* (2nd ed.). Peru, IL: Open Court.

Huberty, C. J. (1991). Introduction to the practice of statistics [Review]. *Journal of Educational Statistics, 16,* 77–81.

Jones, L. V. (Ed.). (1986a). *The collected works of John W.*

Tukey. Volume III: Philosophy and principles of data analysis: 1949–1964. Belmont, CA: Wadsworth.

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

Jöreskog, K., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Chicago: Scientific Software International.

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology, 3,* 185–225.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects?: Statistical power analysis in research.* Beverly Hills, CA: Sage.

Kruskal, W. H. (1982). Criteria for judging statistical graphics. *Utilitas Mathematica, 21B,* 283–310.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery.* Cambridge, MA: MIT Press.

Lauver, P. J., & Jones, R. M. (1991). Factors associated with perceived career options in American Indian, White, and Hispanic rural high school students. *Journal of Counseling Psychology, 38,* 159–166.

Leinhardt, G., & Leinhardt, S. (1980). Exploratory data analysis: New tools for the analysis of empirical data. In D. Berliner (Ed.), *Review of Research in Education (Vol. 8,* pp. 85–157).

Leinhardt, S., & Wasserman, S. S. (1979). Exploratory data analysis: An introduction to selected methods. In K. F. Schuessler (Ed.), *Sociological methodology* (pp. 311–365). San Francisco: Jossey-Bass.

Lent, R. W., & Hackett, G. (1987). Career self-efficacy: Empirical status and future directions. *Journal of Vocational Behavior, 30,* 347–382.

Lewandowsky, S., & Spence, I. (1989). Discriminating strata in scatterplots. *Journal of the American Statistical Association, 84,* 682–688.

Lewandowsky, S., & Spence, I. (1990). The perception of statistical graphs. *Sociological Methods and Research, 18,* 200–242.

Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analyses. In H. Cooper & L. V. Hedges, (Eds.), *The handbook of research synthesis* (pp. 439–454). New York: Russell Sage Foundation.

Lind, J. C., & Zumbo, B. D. (1993). The continuity principle in psychological research: An introduction to robust statistics. *Canadian Psychology, 34,* 407–414.

Loftus, G. R. (1993). Editorial comment. *Memory and Cognition, 21,* 1–3.

MacDonald, K. I. (1983). Exploratory data analysis: A process and a problem. In D. McKay, N. Schofield, & P. Whiteley (Eds.), *Data analysis and the social sciences* (pp. 256–284). London: Pinter.

Mallows, C. L. (1979). Robust methods—Some examples of their use. *The American Statistician, 33,* 179.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data.* Belmont, CA: Wadsworth.

Mayer, L. S. (1980). The use of exploratory methods in economic analysis: Analyzing residential energy demand. In J. Kmenta & J. B. Ramsey (Eds.), *Evaluation of econometric models* (pp. 15–45). New York: Academic Press.

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

McGuire, W. J. (1989). A perspectivist approach to the strategic planning of programmatic scientific research. In B. Gholson, W. R. Shadish, Jr., R. A. Neimeyer, & A. C. Houts (Eds.), *Psychology of science: Contributions to metascience.* Cambridge, England: Cambridge University Press.

Mitroff, I. I., Kilmann, R. K., & Barabba, V. P. (1979). Management information versus misinformation systems. In G. Zaltman (Ed.), *Management principles for nonprofit agencies and organizations* (p. 104). New York: AMACOM.

Mosteller, F. (1988). Broadening the scope of statistics and statistics education. *The American Statistician, 42,* 93–99.

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

Mulaik, S. A. (1984). Empiricism and exploratory statistics. *Philosophy of Science, 52,* 410–430.

Neyman, J., & Pearson, W. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20a,* 175–240.

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.

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century-Crofts.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55,* 33–38.

Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports, 15,* 570.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and analysis.* New York: McGraw-Hill.

Rosnow, R. L. (1981). *Paradigms in transition: The methodology of social inquiry.* New York: Oxford University Press.

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

Simkin, D., & Hastie, R. (1987). An information processing analysis of graph perception. *Journal of the American Statistical Association, 82,* 454–465.

Simon, H. A. (1973). Does scientific discovery have a logic? *Philosophy of Science, 40,* 471–480.

Simon, H. A., Langley, P. W., & Bradshaw, G. L. (1981). Scientific discovery as problem solving. *Synthese, 47,* 1–27.

Smith, A. F., & Prentice, D. A. (1993). Exploratory data analysis. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral science: Statistical issues* (pp. 349–390). Hillsdale, NJ: Erlbaum.

Statistical Sciences, Inc. (1993). *S-PLUS for Windows User's Manual, Version 3.1.* Seattle, WA: Statistical Sciences, Inc.

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

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

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

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

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. (1979). Methodology, and the statistician's responsibility for both accuracy and relevance. *Journal of the American Statistical Association, 74,* 786–793.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician, 34,* 23–25.

Tukey, J. W. (1986a). Data analysis and behavioral science or learning to bear the quantitative man's burden by shunning badmandments. In L. V. Jones (Ed.), *The collected works of John W. Tukey. Volume III: Philosophy and principles of data analysis: 1949–1964* (pp. 187–389). Belmont, CA: Wadsworth.

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

Tukey, J. W. (1986c). The future of processes of data analysis. In L. V. Jones (Ed.), *The collected works of John W. Tukey: Vol. 4. Philosophy and principles of data analysis: 1965–1986* (pp. 517–547). Monterey, CA: Wadsworth & Brooks/Cole.

Tukey, J. W. (1986d). Introduction to styles of data analysis techniques. In L. V. Jones (Ed.), *The collected works of John W. Tukey. Volume IV: Philosophy and principles of data analysis: 1965–1986* (pp. 969–983). Belmont, CA: Wadsworth. (Original work published 1982)

Tukey, J. W. (1986e). Methodological comments focused on opportunities. In L. V. Jones (Ed.), *The collected works of John W. Tukey. Volume IV: Philosophy and principles of data analysis: 1965–1986* (pp. 819–867). Belmont, CA: Wadsworth. (Original work published 1980)

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. Volume IV: Philosophy and principles of data analysis: 1965–1986* (pp. 549–578). Belmont, CA: Wadsworth.

Velleman, P. F. (1992). *Data desk: The new power of statistical vision.* Ithaca, NY: Data Description, Inc.

Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis.* Boston: Duxbury Press.

Velleman, P. F., & Hoaglin, D. C. (1992). Data analysis. In D. C. Hoaglin & D. S. Moore (Eds.), *MAA notes number 21: Perspectives on contemporary statistics* (pp. 19–39). Washington, DC: Mathematical Association of America.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval and ratio typologies are misleading. *The American Statistician, 47,* 65–68.

Wainer, H. (1977a). Robust statistics: A survey and some prescriptions. *Journal of Educational Statistics, 1,* 285–312.

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

Wainer, H., & Thissen, D. (1981). Graphical data analysis. *Annual Review of Psychology, 32,* 191–241.

Wainer, H., & Thissen, D. (1993). Graphical data analysis. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical issues* (pp. 391–457). Hillsdale, NJ: Erlbaum.

Winer, B. J. (1971). *Principles of experimental design.* New York: McGraw-Hill.

Winkler, R. L. (1993). Bayesian statistics: An overview. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 201–232). Hillsdale, NJ: Erlbaum.

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

(Appendix follows)

Appendix

Software for EDA

As computer graphic capabilities widen in commonly used machines, software for graphic and exploratory analyses gain in popularity. Nevertheless, a few programs have a decisively strong EDA emphasis. In this article, all the graphics with the exception of kernel density estimates were produced in Data Desk on an Apple Power Macintosh computer. As illustrated previously, Data Desk is an exceptional tool for EDA, having been designed from its inception to be an EDA technology. The documentation that accompanies the software may be the single best source of technical and practical information concerning EDA. Data Desk offers a completely graphical interface for EDA and requires no programming. Another heavily EDA oriented program is S-plus, the only software package I know of that supports graphics for the two-way fit. S-plus is a completely extensible object-oriented programming language available for UNIX and MS-Windows environments; however, it has less graphical interactivity than Data Desk. S-plus is commonly used in the statistical graphics research community, and there is a large archive of user-created S-plus functions available at the Carnegie Mellon Statlib at http://lib.stat.cmu.edu.

The kernel density estimate plots shown previously were produced in XLISP-STAT (Tierney, 1990), a LISP-based system of statistical functions and graphics that is completely extensible and highly interactive. XLISP-STAT is also gaining a wide following in the statistical graphics community. XLISP-STAT does not require LISP programming skills for most tasks and is available for UNIX, Macintosh, and Windows environments for free. Copies of the program can be obtained at http://stat.umn.edu.

Other programs incorporate EDA procedures as well. Systat has a wide variety of graphics and some interactivity as does SAS-JMP, although SAS-JMP reflects some more CDA philosophies than may be convenient for strong EDA work such as the strong association of levels of measurement with types of analyses (see Velleman & Wilkinson, 1993, for a discussion of this concept). Almost all software packages are now emphasizing the strength and beauty of graphical analysis with access to box plots, stem-and-leaf plots, and so on. Consumers should beware that EDA functions best in highly interactive environments that support quick question assessment. This involves complex interface issues that cannot be solved by the simple inclusion of a new plot in the list of options.

Readers interested in learning more about current issues in data visualization and advances in statistical computing should consult the *Journal of Computational and Graphical Statistics,* which was inaugurated in 1992. When considering specific products, readers may want to consult *The American Statistician,* which periodically contains software reviews by statistical computing experts.

Received June 7, 1996
Revision received August 26, 1996
Accepted October 12, 1996 ■