Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods

Author(s): William S. Cleveland and Robert McGill

Source: *Journal of the American Statistical Association*, Vol. 79, No. 387 (Sep., 1984), pp. 531-554

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2288400

Accessed: 15-09-2019 14:13 UTC

Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods

WILLIAM S. CLEVELAND and ROBERT McGILL*

The subject of graphical methods for data analysis and for data presentation needs a scientific foundation. In this article we take a few steps in the direction of establishing such a foundation. Our approach is based on *graphical perception*—the visual decoding of information encoded on graphs—and it includes both theory and experimentation to test the theory. The theory deals with a small but important piece of the whole process of graphical perception. The first part is an identification of a set of *elementary perceptual tasks* that are carried out when people extract quantitative information from graphs. The second part is an ordering of the tasks on the basis of how accurately people perform them. Elements of the theory are tested by experimentation in which subjects record their judgments of the quantitative information on graphs. The experiments validate these elements but also suggest that the set of elementary tasks should be expanded. The theory provides a guideline for graph construction: Graphs should employ elementary tasks as high in the ordering as possible. This principle is applied to a variety of graphs, including bar charts, divided bar charts, pie charts, and statistical maps with shading. The conclusion is that radical surgery on these popular graphs is needed, and as replacements we offer alternative graphical forms—*dot charts, dot charts with grouping*, and *framed-rectangle charts*.

KEY WORDS: Computer graphics; Psychophysics.

1. INTRODUCTION

Nearly 200 years ago William Playfair (1786) began the serious use of graphs for looking at data. More than 50 years ago a battle raged on the pages of the *Journal of the American Statistical Association* about the relative merits of bar charts and pie charts (Eells 1926; Croxton 1927; Croxton and Stryker 1927; von Huhn 1927). Today graphs are a vital part of statistical data analysis and a vital part of communication in science and technology, business, education, and the mass media.

Still, graph design for data analysis and presentation is largely unscientific. This is why Cox (1978) argued, "There is a major need for a theory of graphical methods" (p. 5), and why Kruskal (1975) stated "in choosing, constructing, and comparing graphical methods we have little to go on but intuition, rule of thumb, and a kind of master-to-apprentice passing along of information. . . . there is neither theory nor systematic body of experiment as a guide" (p. 28–29).

There is, of course, much good common sense about how to make a graph. There are many treatises on graph construction (e.g., Schmid and Schmid 1979), bad practice has been uncovered (e.g., Tufte 1983), graphic designers certainly have shown us how to make a graph appealing to the eye (e.g., Marcus et al. 1980), statisticians have thought intensely about graphical methods for data analysis (e.g., Tukey 1977; Chambers et al. 1983), and cartographers have devoted great energy to the construction of statistical maps (Bertin 1973; Robinson, Sale, and Morrison 1978). The ANSI manual on time series charts (American National Standards Institute 1979) provides guidelines for making graphs, but the manual admits, "This standard . . . sets forth the best current usage, and offers standards 'by general agreement' rather than 'by scientific test'" (p. iii).

In this article we approach the science of graphs through human graphical perception. Our approach includes both theory and experimentation to test it.

The first part of the theory is a list of elementary perceptual tasks that people perform in extracting quantitative information from graphs. In the second part we hypothesize an ordering of the elementary tasks based on how accurately people perform them. We do not argue that this accuracy of quantitative extraction is the only aspect of a graph for which one might want to develop a theory, but it is an important one.

The theory is testable; we use it to predict the relative performance of competing graphs, and then we run experiments to check the actual performance. The experiments are of two types: In one, once the graphs are drawn, the evidence appears so strong that it is taken prima facie to have established the case. When a strong effect is perceived by the authors' eyes and brains, it is likely that it will appear to most other people as well. In

* William S. Cleveland and Robert McGill are statisticians at AT&T Bell Laboratories, Murray Hill, NJ 07974. The authors are indebted to John Chambers, Ram Gnanadesikan, David Krantz, William Kruskal, Colin Mallows, Frederick Mosteller, Henry Pollak, Paul Tukey, and the JASA reviewers for important comments on an earlier version of this article.

the other type, the case is not so clear; we must show the graphs to subjects, ask them to record their judgments of quantitative information, and analyze the results to test the theory. Both types of experiments are reported in this article.

The ordering of the elementary perceptual tasks can be used to redesign old graphical forms and to design new ones. The goal is to construct a graph that uses elementary tasks as high in the hierarchy as possible. This approach to graph design is applied to a variety of graphs, including bar charts, divided bar charts, pie charts, and statistical maps with shading. The disconcerting conclusion is that radical surgery on these popular types of graphs is needed, and as replacements we offer some alternative graphical forms: dot charts, dot charts with grouping, and framed-rectangle charts.

This is not the first use of visual perception to study graphs. A number of experiments have been run in this area (see Feinberg and Franklin 1975; Kruskal 1975,1982; and Cleveland, Harris, and McGill 1983 for reviews); but most have focused on which of two or more graph forms is better or how a particular aspect of a graph performs, rather than attempting to develop basic principles of graphical perception. Chambers et al. (1983, Ch. 8) presented some discussion of visual perception, along with a host of other general considerations for making graphs for data analysis.

Pinker (1982), in an interesting piece of work, developed a model that governs graph comprehension in a broad way. The model deals with the whole range of perceptual and cognitive tasks used when people look at a graph, borrowing heavily from existing perceptual and cognitive theory (e.g., the work of Marr and Nishihara 1978). No experimentation accompanies Pinker's modeling. The material in this article is much more narrowly focused than Pinker's; our theory deals with certain specific perceptual tasks that we believe to be critical factors in determining the performance of a graph.

2. THEORY: ELEMENTARY PERCEPTUAL TASKS

In this and the next section we describe the two parts of our theory, which is a set of hypotheses that deal with the extraction of quantitative information from graphs. The theory is an attempt to identify perceptual building blocks and then describe one aspect of their behavior.

The value of identifying basic elements and their interactions is that we thus develop a framework to organize knowledge and predict behavior. For example, Julesz's (1981) theory of textons identified the elementary particles of what is called preattentive vision, the instantaneous and effortless part of visual perception that the brain performs without focusing attention on local detail. He wrote that "every mature science has been able to identify its basic elements ('atoms,' 'quarks,' 'genes,' etc.) and to explain its phenomena as the known interaction between these elements" (Julesz in press).

Figure 1 illustrates 10 elementary perceptual tasks that people use to extract quantitative information from graphs. (Color saturation is not illustrated, to avoid the nuisance and expense of color reproduction.) The pictorial symbol used for each task in Figure 1 is meant to be suggestive and might not necessarily invoke only that task if shown to a viewer. For example, a circle has an area associated with it, but it also has a length, and a person shown circles might well judge diameters or circumferences rather than areas, particularly if told to do so.

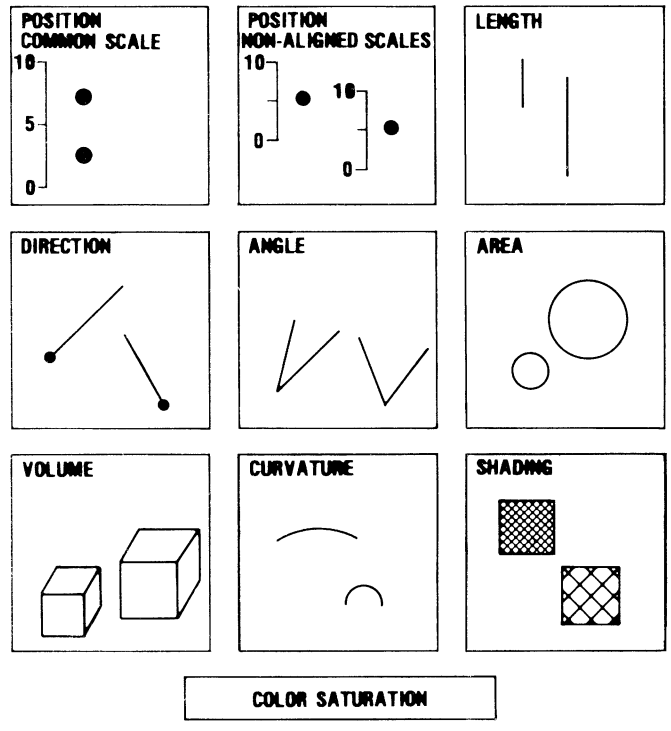

Figure 1. Elementary perceptual tasks.

We have chosen the term *elementary perceptual task* because a viewer performs one or more of these mental-visual tasks to extract the values of real variables represented on most graphs. We do not pretend that the items on our list are completely distinct tasks; for example, judging angle and direction are clearly related. We do not pretend that our list is exhaustive; for example, color hue and texture (Bertin 1973) are two elementary tasks excluded from the list because they do not have an unambiguous single method of ordering from small to large and thus might be regarded as better for encoding categories rather than real variables. Nevertheless the list in Figure 1 is a reasonable first try and will lead to some useful results on graph construction.

We will now show how elementary perceptual tasks are used to extract the quantitative information on a variety of common graph forms.

Sample Distribution Function Plot

Figure 2 is a sample distribution function plot of murders per 10^5 people per year in the continental United States. The elementary task that one carries out to perceive the relative magnitude of the values of the data is

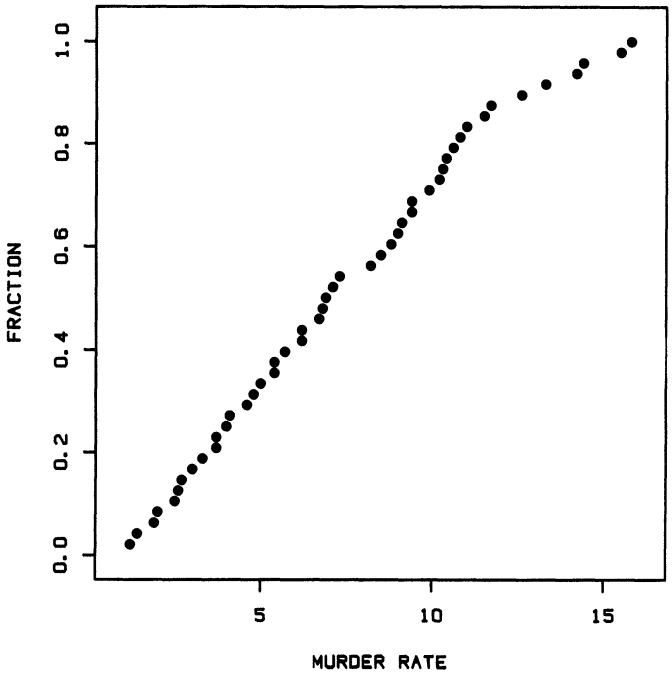

Figure 2. Sample distribution function of 1978 murder rate.

judging position along a common scale, which in this case is the horizontal scale.

Bar Charts

Figures 3 and 4 contain bar charts that were shown to subjects in perceptual experiments. The few noticeable peculiarities are there for purposes of the experiments, described in a later section.

Judging position is a task used to extract the values of the data in the bar chart in the right panel of Figure 3. But now the graphical elements used to portray the data—the bars—also change in length and area. We conjecture that the primary elementary task is judging position along a common scale, but judgments of area and length probably also play a role.

Pie Charts

The left panel of Figure 3 is a pie chart, one of the most commonly used graphs for showing the relative sizes of the parts of a whole. For this graph we conjecture that the primary elementary visual task for extracting the numerical information is perception of angle, but the areas and arc lengths of the pie slices are variable and probably are also involved in judging the data.

Divided Bar Charts

Figure 4 has three divided bar charts (Types 2, 4, and 5). For each of the three, the totals of A and B can be compared by perceiving position along the scale. Position judgments can also be used to compare the two bottom divisions in each case; for Type 2 the bottom divisions are marked with dots. All other values must be compared by the elementary task of perceiving different bar lengths; examples are the two divisions marked with dots in Type 4 and the two marked in Type 5.

Statistical Maps With Shading

A chart frequently used to portray information as a function of geographical location is a statistical map with shading, such as Figure 5 (from Gale and Halperin 1982), which shows the murder data of Figure 2. Values of a real variable are encoded by filling in geographical regions using any one of many techniques that produce gray-scale shadings. In Figure 5 the technique illustrated uses grids drawn with different spacings; the data are not proportional to the grid spacing but, rather, to a complicated function of spacing. We conjecture that the primary elementary task used to extract the data in this case is the perception of shading, but judging the sizes of the squares formed by the grids probably also plays a role, particularly for the large squares.

Curve-Difference Charts

Another class of commonly used graphs is curve-difference charts: Two or more curves are drawn on the graph, and vertical differences between some of the curves encode real variables that are to be extracted. One type of curve-difference chart is a divided, or aggregate, line chart (Monkhouse and Wilkinson 1963), which is typically used to show how parts of a whole change through time.

Figure 6 is a curve-difference chart. The original was drawn by William Playfair; because our photograph of the original was of poor quality, we had the figure redrafted, trying to keep as close to the original as possible. The two curves portray exports from England to the East Indies and imports to England from the East Indies. The vertical distances between the two curves, which encode the export–import imbalance, are highlighted. The quantitative information about imports and exports is extracted by perceiving position along a common scale, and the information about the imbalances is extracted by perceiving length, that is, vertical distance between the two curves.

Cartesian Graphs and Why They Work

Figure 7 is a Cartesian graph of paired values of two variables, x and y. The values of x can be visually ex-

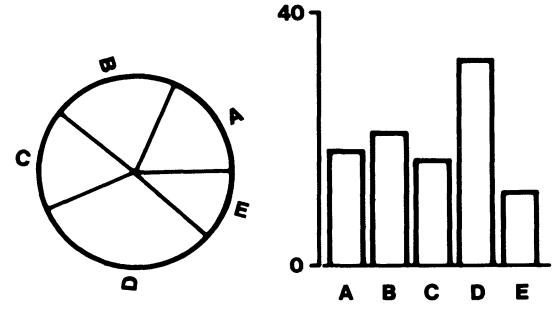

Figure 3. Graphs from position–angle experiment.

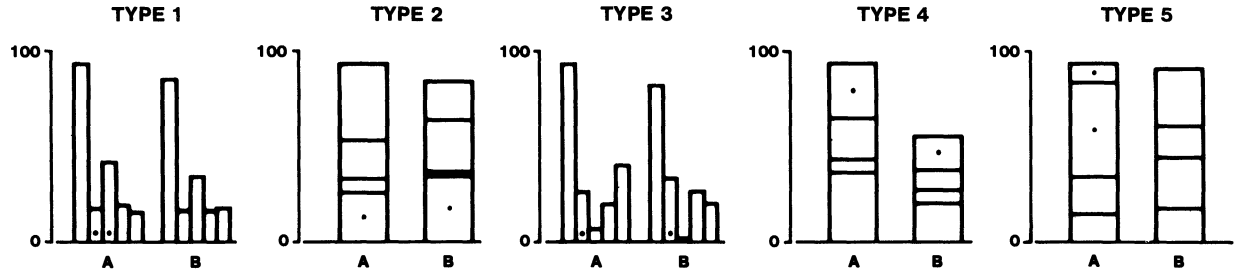

Figure 4. Graphs from position–length experiment.

tracted by perceiving position along a scale, in this case the horizontal axis. The y values can be perceived in a similar manner.

The real power of a Cartesian graph, however, does not derive only from one's ability to perceive the x and y values separately but, rather, from one's ability to understand the relationship of x and y. For example, in Figure 7 we see that the relationship is nonlinear and see the nature of that nonlinearity. The elementary task that enables us to do this is perception of direction. Each pair of points on the plot, (x_i, y_i) and (x_j, y_j), with $x_i \neq x_j$, has an associated slope

$$(y_j - y_i)/(x_j - x_i).$$

The eye–brain system is capable of extracting such a slope by perceiving the direction of the line segment joining (x_i, y_i) and (x_j, y_j). We conjecture that the perception of these slopes allows the eye–brain system to imagine a smooth curve through the points, which is then used to judge the pattern. For example, in Figure 7 one can perceive that the slopes for pairs of points on the left side of the plot are greater than those on the right side of the plot, which is what enables one to judge that the relationship is nonlinear.

That the elementary task of judging directions on a Cartesian graph is vital for understanding the relationship of x and y is demonstrated in Figure 8. The same x and y values are shown by paired bars. As with the Cartesian

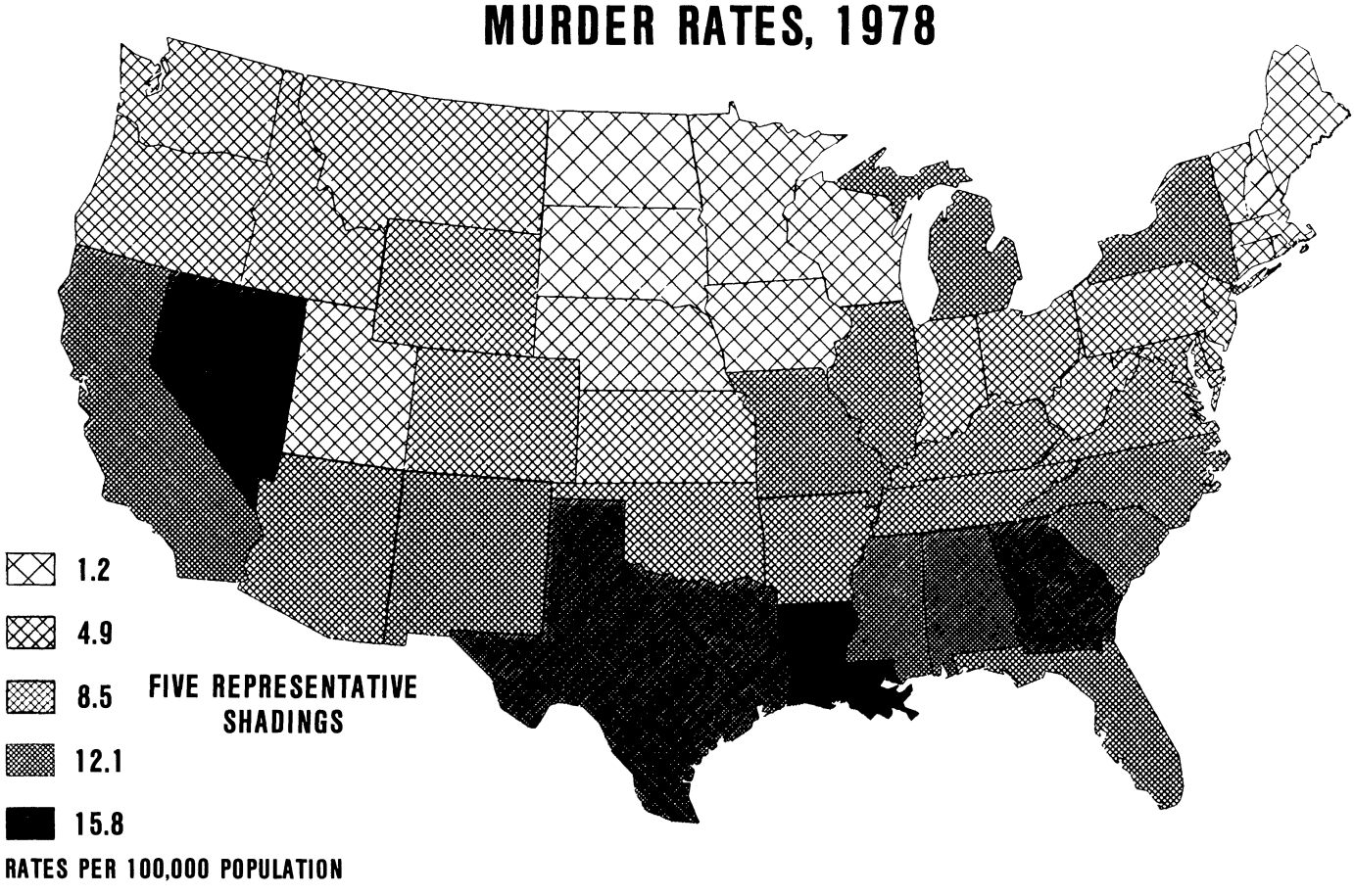

Figure 5. Statistical map with shading.

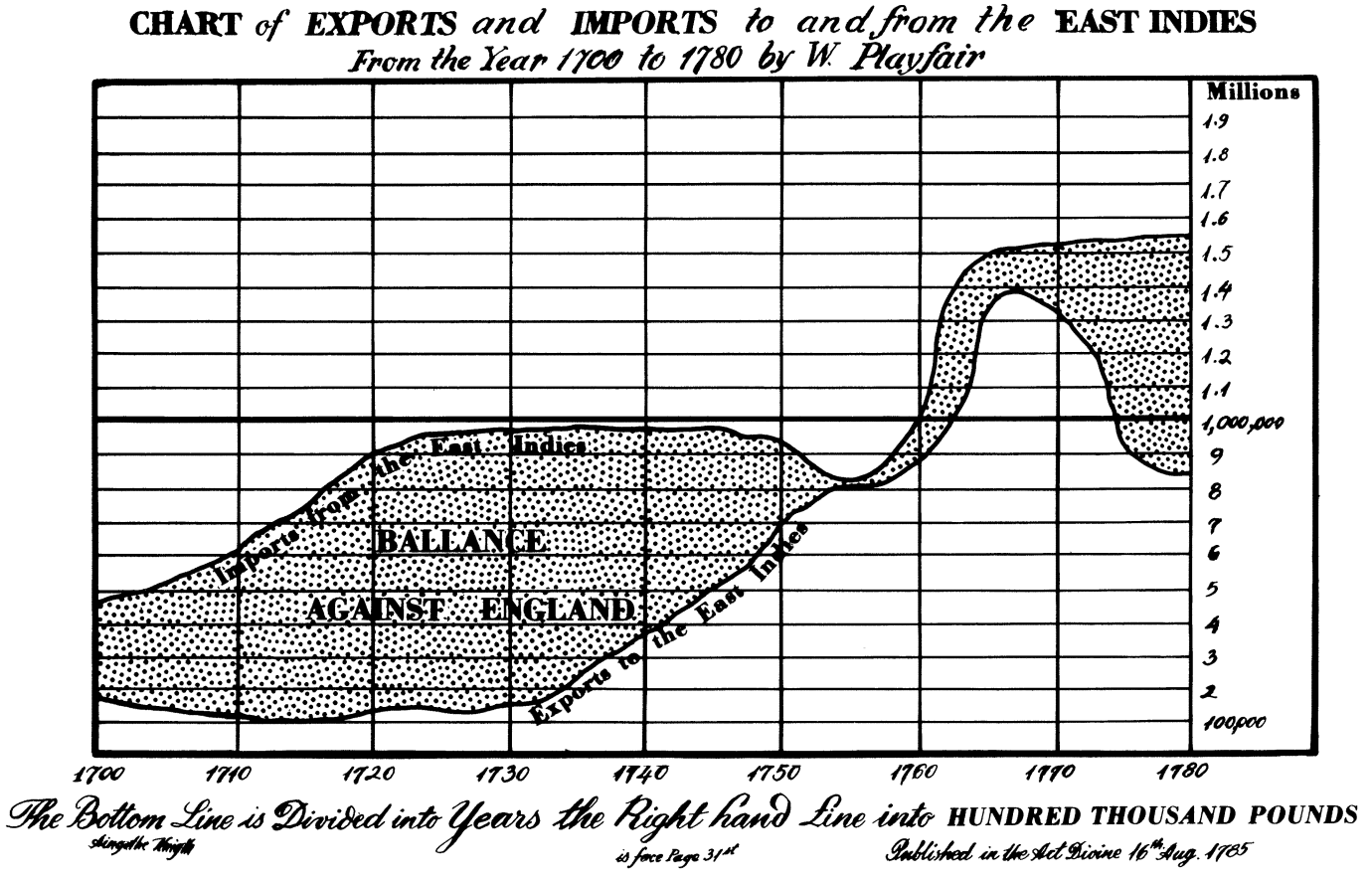

Figure 6. Curve-difference chart after Playfair.

graph, one can perceive the x and y values by perceiving their positions along a common scale. But with the ability to perceive slopes removed, the pattern of the nonlinear relationship is difficult to perceive.

Triple Scatterplots

Figure 9 is a triple scatterplot of three real variables, (x_i, y_i, z_i), for $i = 1$ to 25. (The name *triple scatterplot* was suggested by Anscombe 1973.) The x and y values are portrayed by the centers of the circles and so form an ordinary Cartesian graph. The third variable is encoded by the areas of the circles; thus the elementary perceptual task for extracting the z_i is area perception.

Volume Charts

The elementary task required in Figure 10 is volume perception. Such volume charts are used very infrequently in science and technology but are common in mass-media graphics (Tufte 1983).

Juxtaposed Cartesian Graphs

Frequently two or more panels of graphs are juxtaposed with the scales on the panels the same. Figure 11, which will be explained later, is an example of this; we juxtaposed the graphs because superimposing them would have resulted in an uninterpretable mess. In Figure 11, when we compare the log errors from two panels that are not in the same row, we must make judgments of positions along nonaligned scales.

3. THEORY: ORDERING THE ELEMENTARY PERCEPTUAL TASKS BY THE ACCURACY OF EXTRACTION

In this section we hypothesize an ordering of the 10 elementary perceptual tasks on the basis of the accuracy with which people can extract quantitative information by using them. One elementary perceptual task is taken to be more accurate than another if it leads to human judgments that come closer to the actual encoded quantities.

One must be careful not to fall into a conceptual trap by adopting accuracy as a criterion. We are not saying that the primary purpose of a graph is to convey numbers with as many decimal places as possible. We agree with Ehrenberg (1975) that if this were the only goal, tables would be better. The power of a graph is its ability to enable one to take in the quantitative information, organize it, and see patterns and structure not readily revealed by other means of studying the data.

Our premise, however, is this:

A graphical form that involves elementary perceptual tasks that lead to more accurate judgments than another graphical form (with the same quantitative in-

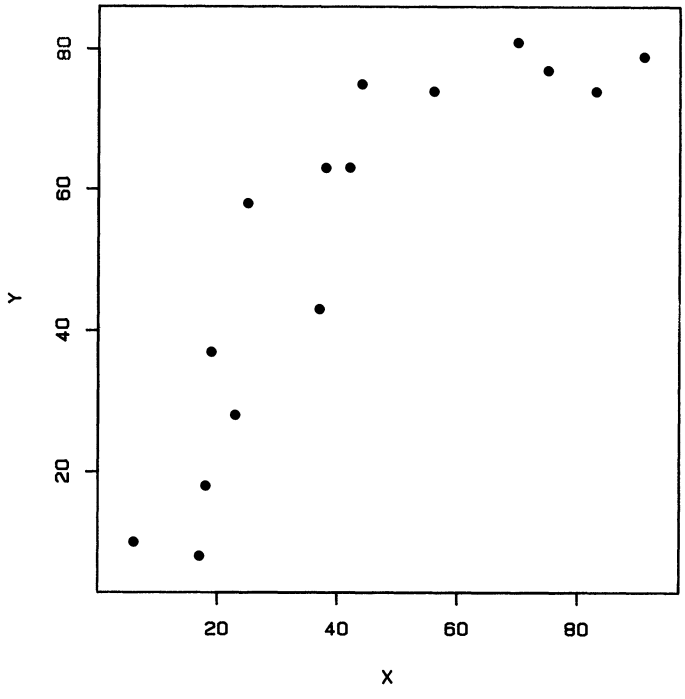

Figure 7. Cartesian graph.

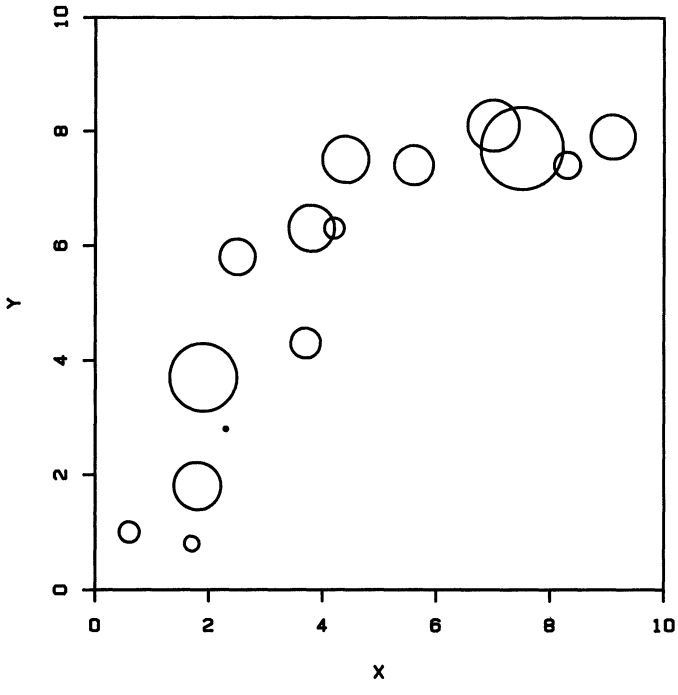

Figure 9. Triple scatterplot.

formation) will result in better organization and increase the chances of a correct perception of patterns and behavior.

In Section 5 we give examples of patterns emerging when elementary perceptual tasks are changed to increase the accuracy of judgments.

The following are the 10 elementary tasks in Figure 1, ordered from most to least accurate:

1. Position along a common scale
2. Positions along nonaligned scales
3. Length, direction, angle
4. Area
5. Volume, curvature
6. Shading, color saturation

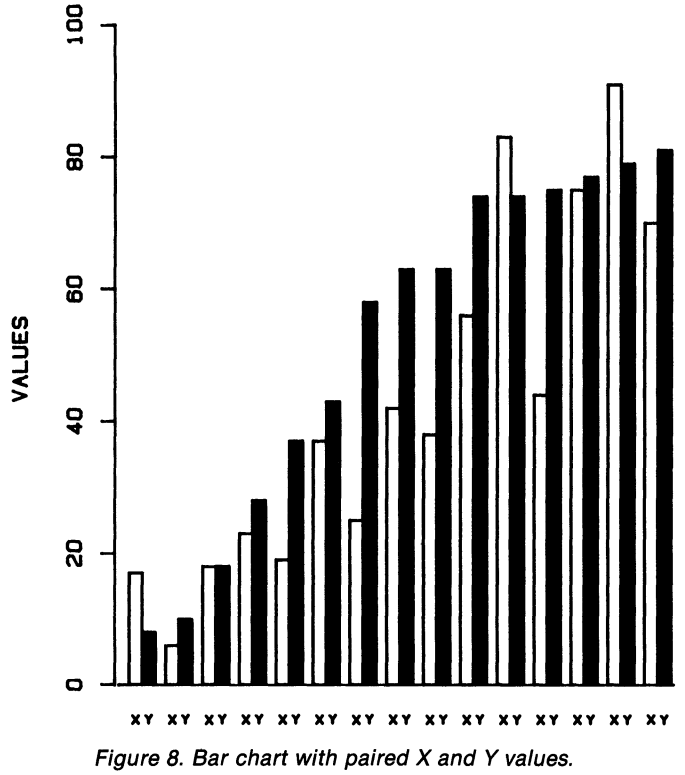

Figure 8. Bar chart with paired X and Y values.

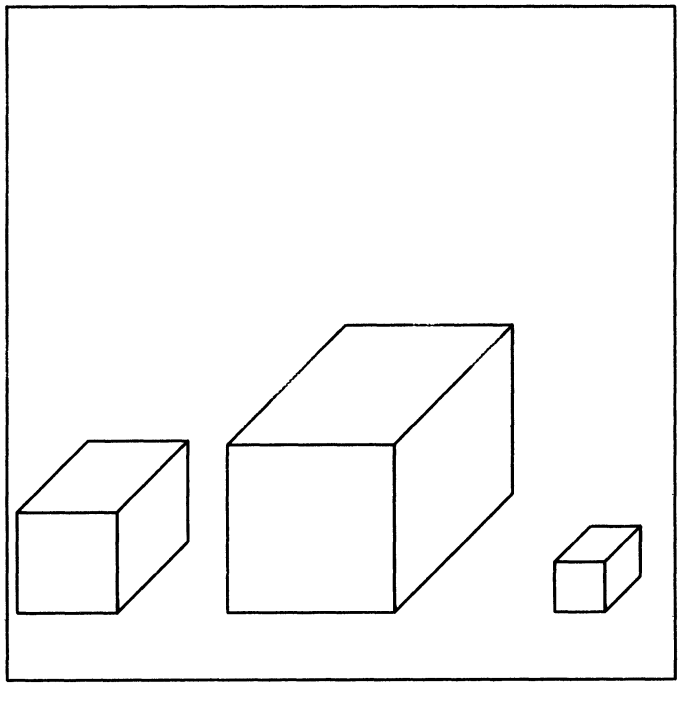

Figure 10. Volume chart.

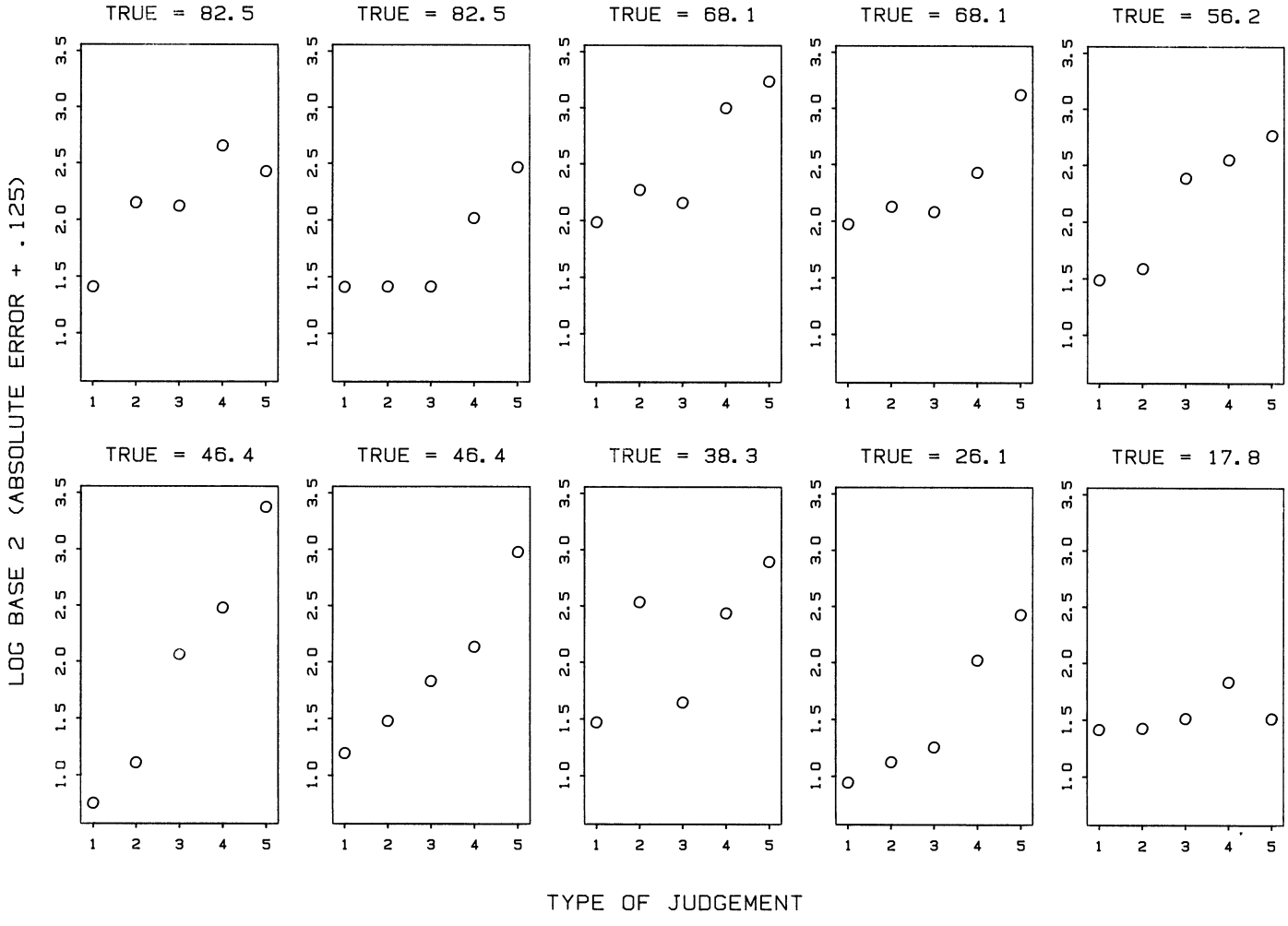

Figure 11. Position–length experiment: Midmeans of log absolute errors against judgment type for 10 pairs of judged values.

Three of the ranks—3, 5, and 6—have more than one task; at the moment there is not enough information to separate the ties.

The hypothesized ordering of the elementary tasks is based on information from a variety of sources: our own reasoning and experimentation with various graph forms, results of psychophysical experiments, and the theory of psychophysics. The following discussion attempts only a partial documentation. The sources of the theoretical ordering are not the most cogent factors in establishing it; rather, using the theory to predict the performance of graph forms and then running experiments to check the predictions is the cogent process for validating and revising the theory. It is only through such a procedure that we can claim to be establishing a science of graphical perception. A few comments about the sources of the ordering, however, will at least convey the process used to devise it.

In the ordering of perceptual tasks, length judgments are hypothesized to be more accurate than area judgments, which in turn are hypothesized to be more accurate than volume judgments. This ordering is based on a combination of psychophysical theory and experimental results.

Suppose an individual is asked to judge the magnitude of some aspect of a physical object such as length, area, volume, distance, loudness, weight, or pitch. The power law of theoretical psychophysics (Stevens 1975) says that if p is the perceived magnitude and a is the actual magnitude, then p is related to a by $p = ka^{\alpha}$. If a_1 and a_2 are two such magnitudes and p_1 and p_2 are corresponding perceived values, then $p_1/p_2 = (a_1/a_2)^{\alpha}$. Thus only if $\alpha = 1$ is the perceived scale the same as the actual physical scale. For visual perception this power law appears to be a good description of reality (Baird 1970).

Many psychophysical experiments have been conducted to estimate values of α. For judgments of length, area, or volume, average values of α from different experiments can vary according to how instructions are phrased and according to many experimental factors. And for a particular experiment, values of α can vary substantially for different subjects. Baird (1970) gave an excellent review of a large number of experiments; one pattern that emerges is that values of α tend to be reasonably close to 1 for length judgments, smaller than 1 for area judgments, and even smaller for volume judgments. This means that length judgments tend to be unbiased, whereas there tends to be distortion in area judg-

ments and even more in volume judgments. Partly for this reason we have set the order (as given previously) to be length, then area, and then volume.

Of course increased bias does not necessarily imply less overall accuracy. The reasoning, however, is that the mechanism leading to bias might well lead to other types of inaccuracy as well. We might try to combat bias and increase the accuracy of judgments by taking the areas or volumes to be proportional to power-transformed values of the data. Cleveland, Harris, and McGill (1983) gave reasons for not doing this, however, one of which is that the power coefficients vary from one person to the next.

The reason for putting position along nonaligned scales ahead of length is that there are additional visual cues on nonaligned scales to help in making judgments. We illustrate this with one particular graph form. The top of Figure 12 shows two bars, or rectangles, with equal widths and unequal heights. Suppose bar height encodes some real variable; the elementary perceptual task—judging length—is hard enough that we cannot easily perceive which bar is longer in Figure 12.

In the bottom of Figure 12, the same bars are drawn, but they are surrounded by frames of equal size and construction. Each symbol, called a *framed rectangle*, is actually a little graph with a scale and with one number portrayed. The elementary perceptual task is judging position along nonaligned scales, and now we can easily see that the right bar represents a larger quantity than the left. Actually, because the framed rectangle is such a simple graphical form, the task of judging position along nonaligned scales really amounts to two length judgments (as will be discussed shortly). In other circumstances, where the graph form is more complex (such as Figure 11, which was discussed in Section 2), a more complex set of visual tasks makes up the position-along-nonaligned-scales task because there are more visual cues.

Weber's Law (cf. Baird and Noma 1978), an important law of theoretical psychophysics, helps to explain how the frame of a framed rectangle increases accuracy. Suppose x is the length of some physical object, such as a line or bar. Suppose that $d_p(x)$, a positive number, is defined by the following: An object with length $x + d_p(x)$ is detected with probability p to be longer than the object with length x. Then Weber's Law states that for fixed p, $d_p(x) = k_p x$, where k_p does not depend on x. This law appears to hold up well for a variety of perceptual judgments, although Gregory (1966) argued that a modification for small values of x is needed.

The unfilled portion of a framed rectangle creates an unfilled bar with a length equal to the length of the frame minus the length of the filled bar. The lengths of the unfilled bars give additional visual cues to help in judging the encoded numerical quantities. Suppose two framed rectangles have filled bars that are long and close in length, such as in the bottom of Figure 12. Then the percentage difference of the lengths of the unfilled bars is much greater than that of the filled bars; by Weber's Law one can much more readily detect a difference in the short unfilled bars than in the long filled bars. Thus it is the unfilled bars in the bottom of Figure 12 that allow the perception of a difference that is not perceptible in the top.

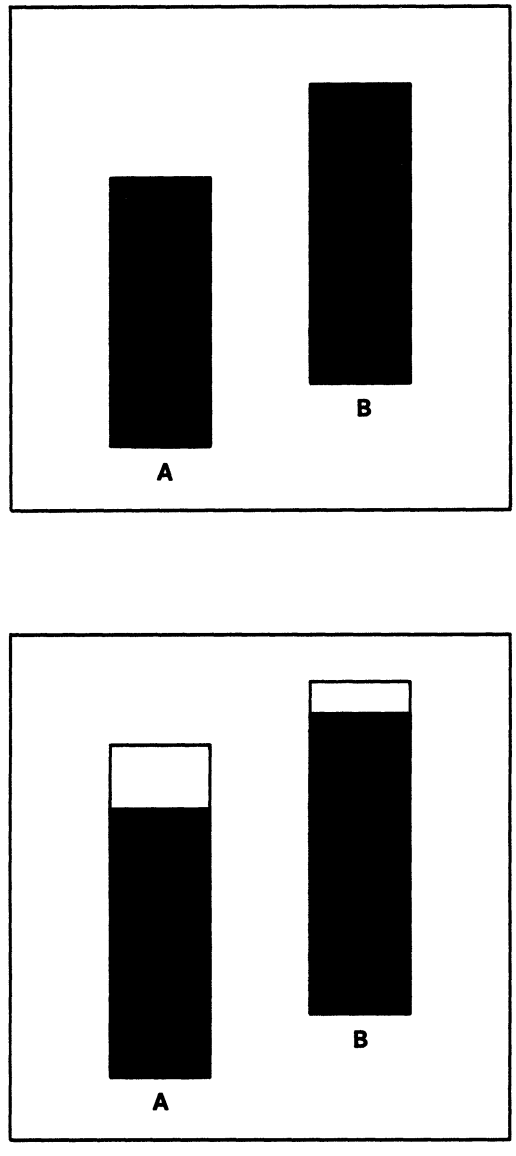

Figure 12. Bars and framed rectangles.

In Section 5.3 we put the framed rectangle to work to design a new type of statistical map.

4. EXPERIMENTATION

4.1 Introduction

We began checking the hypothesized ordering by running two experiments. The experiments demonstrated very clearly that some judgments of position along a common scale are more accurate than some judgments of length and of angle. Strictly speaking we cannot do more than assert that the results hold for the particular types of graphs in the experiment, but the important point is that the theory has correctly predicted the outcome. This section contains a detailed description of the experiment

and a detailed analysis of the data. (Those not interested in the details can read the summary in Section 4.5 and proceed to the application of the theory and experiments to graph design in Section 5.)

4.2 Design

In one experiment 55 subjects were shown the five types of graphs depicted in Figure 4. (The graphs used in the experiment were much larger than in Figure 4, each being on a separate $8\frac{1}{2} \times 11$ page and filling a large portion of the page.) Each graph was either a divided bar chart (as in the rightmost panel) or a grouped bar chart (as in the leftmost panel). A grouped bar chart can be used to show the same type of data as a divided bar chart by encoding the total by the left bar of each group and encoding the divisions by the remaining bars. On a grouped bar chart, unlike on a divided bar chart, all values can be extracted and compared by judging position along a common scale.

On each graph two bars or divisions were marked with dots, and subjects were asked to judge what percent the smaller is of the larger. For the grouped bar charts, the dots appeared either in the second and third bars of the left group or in the second bars of the two groups. For the divided bar chart, the dots appeared either in the bottom divisions of the two bars or in the top divisions of the two bars or in the top two divisions of the left bar. For Judgment Types 1–3, subjects had to judge position along a common scale, and for Judgment Types 4 and 5, subjects had to judge length. Hence we call this the position–length experiment.

In this position–length experiment, the values involved in the subjects' judgments were

$$s_i = 10 \times 10^{(i-1)/12}, \quad i = 1, \ldots, 10,$$

which are equally spaced on a log scale and range from 10 to 56.2. Subjects judged the ratios of 10 pairs of values; the ratios ranged from .18 to .83. Each pair of values was judged five times, once for each of the five judgment types.

Bar segments and heights not judged were chosen essentially at random, but subject to certain constraints. In particular, for Type 4 stimuli neither the top nor the bottom of the two topmost bar segments was permitted to have the same y value, since this would permit judgment along a common scale.

For each graph the subjects were asked to indicate which of the two bars or two segments was the smaller. Next they were to judge what percentage the smaller was of the larger. The instructions specifically stated that subjects were to make "a quick visual judgment and not try to make precise measurements, either mentally or with a physical object such as a pencil or your finger." Only four errors occurred in the choice of which bar or segment was smaller.

Graphs were presented in stapled packets. The instruction sheet was the first page. The next five were practice graphs, one of each type, followed by a page marked "STOP." The 50 graphs, in random order, completed the packet. All packets were identical. Answers were recorded on separate answer sheets, and subjects were instructed not to write on the graphs.

In the second experiment 54 subjects judged the two types of graphs shown in Figure 3; one type was a pie chart and the other was an ordinary bar chart. Ten sets of five numbers that added to 100 were generated, and each set was encoded by a bar chart and a pie chart, resulting in 20 graphs. For each graph, the answer sheet indicated which pie segment or bar was largest and subjects were asked to judge what percentage each of the other four values was of the maximum. Since subjects were judging position or angle, we call this the position–angle experiment.

The values were randomly generated by a uniform random-number generator, with results rescaled to sum to 100. Each set was constrained to meet three requirements: The minimum value had to be greater than 3; the maximum value had to be less than 39, and all differences between values in a set had to be greater than .1. Sets not meeting these requirements were rejected. For the values that actually arose in the constrained random selection, the ratios ranged from 10.0 to 99.7%.

The instruction sheet described the task to be performed on each stimulus—"to judge what percent each of the other segments or bars is of the largest." It also explained that on the answer sheet, the largest segment would be marked with an X. As in the previous experiment, subjects were instructed to make quick visual judgments, not measurements.

Graphs were put in stapled packets. The instruction sheet was the first page. The next two pages were practice graphs—one bar chart and one pie chart—followed by a page marked "STOP." The 20 graphs, in random order, completed the packets. All packets were identical. Answers were recorded on separate sheets.

4.3 Data Exploration

Subjects and Experimental Units

In the position–length experiment, the judgments of four people were deleted because it was clear from their answers that they had not followed instructions. In the position–angle experiment, the judgments of three subjects were deleted for the same reason. For both experiments, 51 subjects remained for analysis.

For each experiment the subjects fell into two categories: (1) a group of females, mostly housewives, without substantial technical experience; (2) a mixture of males and females with substantial technical training and working in technical jobs. Most of the subjects in the position–length experiment participated in the position–angle experiment; in all cases repeat subjects judged the position–angle graphs first.

We did not detect any differences in the accuracies of the judgments of the nontechnical and technical groups. This is not surprising, since the perceptual tasks that sub-

jects were asked to carry out were very basic ones carried out in everyday activities. Thus we treated the subjects as a homogeneous sample that could be used to make inferences about people in general.

It is important to think of each judgment made from a graph in the two experiments as an experimental unit. In the position–length experiment there were 50 judgments, which can be cross-classified into 10 ratios judged for each of five types of judgments: three length judgments and two position judgments. In the position–angle experiment there were 80 judgments, which can be cross-classified into 40 judged ratios for each of two types of jugments: one angle judgment and one position judgment.

Accuracy

To measure accuracy we used

$$\log_2(\,|\,\text{judged percent} - \text{true percent}\,|\, + 1/8).$$

A log scale seemed appropriate to measure relative error; we added 1/8 to prevent a distortion of the scale at the bottom end because the absolute errors in some cases got very close to zero. We used log base 2 because average relative errors tended to change by factors less than 10.

For a large number of the experimental units in each experiment, normal probability plots were made of the log errors; they showed substantial nonnormality in the empirical distribution of the log errors across subjects for each experimental unit. The deviations from normality were

1. Discrete data caused by subjects' tendencies to use multiples of five as answers
2. Mild skewness, sometimes to the left and sometimes to the right
3. Frequent outliers

Principally because of the outliers, we estimated the location of the distribution of the 51 log error values for each experimental unit by the midmean, a robust estimate of location (Mosteller and Tukey 1977).

Figure 13 shows plots of the 50 midmeans of the log absolute errors for the position–length experiment, and Figure 14 shows plots of the 80 midmeans for the position–angle experiment. In both figures the log absolute errors are plotted against the true percentages for each judgment type; superimposed on each plot are smooth curves computed by a scatterplot smoothing procedure called *lowess* (Cleveland 1979). For the position–length experiment, there appears to be a mild dependence of the log absolute errors on the true value for Judgment Types 1–4 and a larger dependence for Type 5. In the position–angle experiment, there is a dependence for the pie charts but very little for the bar charts.

Figure 11 is another plot of the 50 midmeans of the log absolute errors for the position–length experiment. Each panel shows the five midmeans for one of the 10 pairs of values whose ratio was judged; the five midmeans for the five types of judgments are plotted against the type number. Above each panel is the true percentage that the

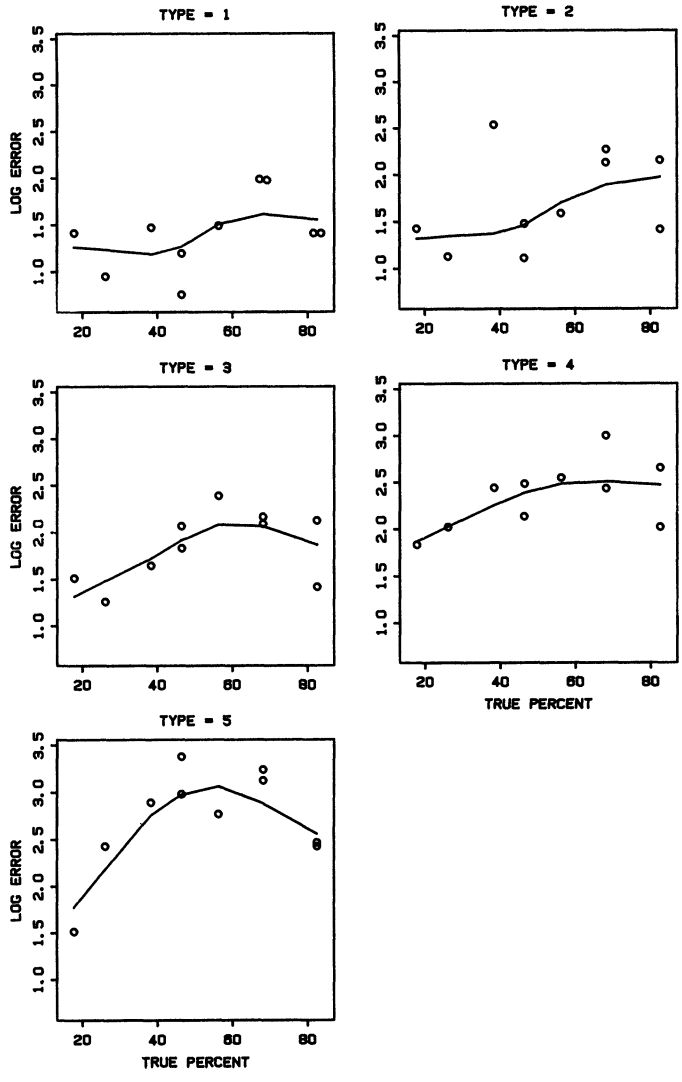

Figure 13. Position–length experiment: Midmeans of log absolute errors against true percentages for five judgment types.

subjects were judging. The striking pattern is that the log absolute errors almost always increase from Type 1 through Type 5. (The type numbers were chosen after the analysis to correspond to most accurate (1) to least accurate (5).) We will discuss this pattern in more detail later.

The midmeans from the left panel of Figure 14 minus the corresponding midmeans in the right panel are plotted in Figure 15 against the true percentage, with a lowess curve superimposed. In only 3 of the 40 cases was the pie chart more accurate on average than the bar chart.

Figure 16 shows average errors for each of the five judgment types in the position–length experiment (top) and each of the two judgment types in the position–angle experiment (bottom). The five values in the top panel are the means of the 10 midmeans for each judgment type (i.e., the means of the 10 midmeans in each panel of Figure 13). The two values in the bottom panel are the means of the 40 midmeans for each judgment type (i.e., the means of the 40 midmeans in each panel of Figure 14).

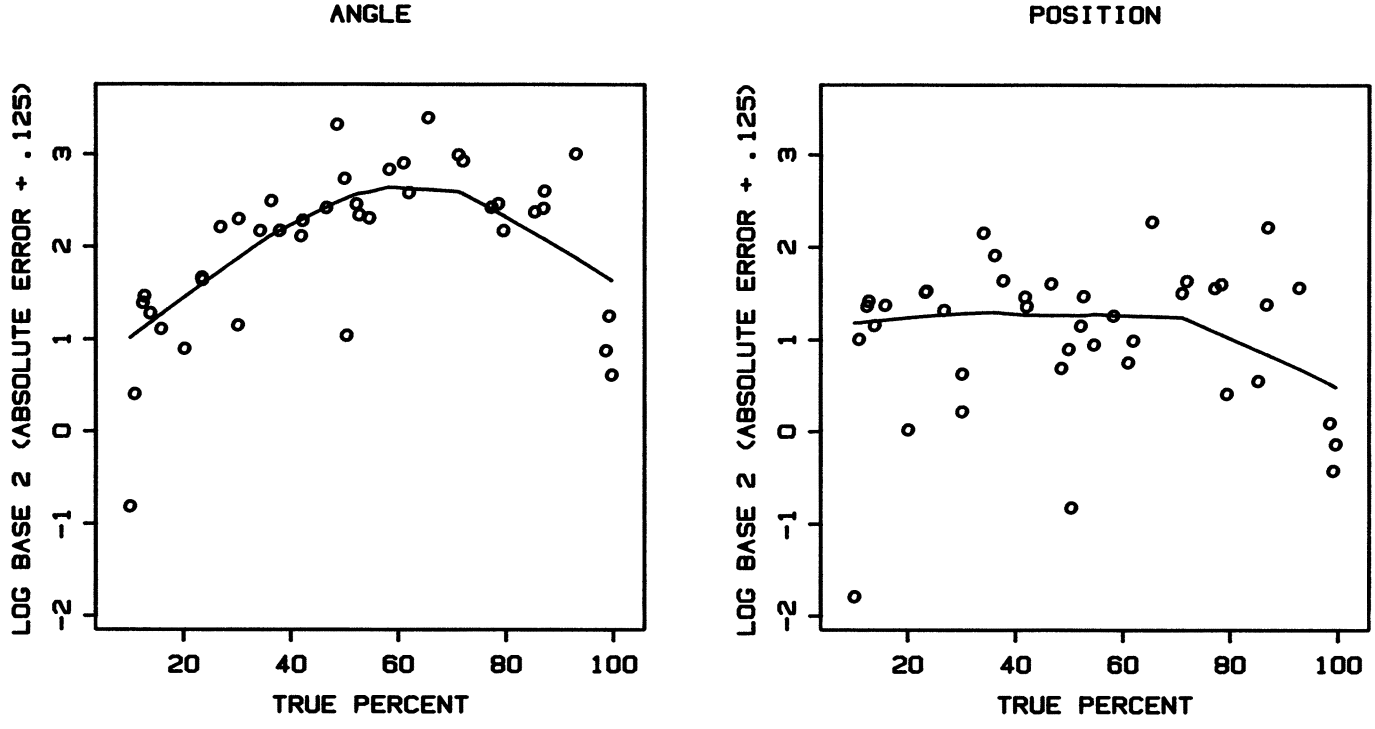

Figure 14. Position–angle experiment: Midmeans of log absolute errors against true percentages for two judgment types.

A 95% confidence interval is shown for each mean; these intervals are discussed in Section 4.4. The initial midmeans provide the requisite robustness to a small number of unusual observations. Since the midmeans are well behaved and have no distant outliers, we have taken a mean, rather than a robust statistic, to summarize them.

The means in Figure 16 provide convenient, but rough, summaries of the two experiments. The summaries are rough because it is clear from Figures 13 and 14 that there is some dependence of log error on the true percent. Within an experiment it is reasonable to compare the means of the judgments because the set of true percentages is the same for each judgment, but it would be inappropriate to compare the means of the first experiment with those of the second.

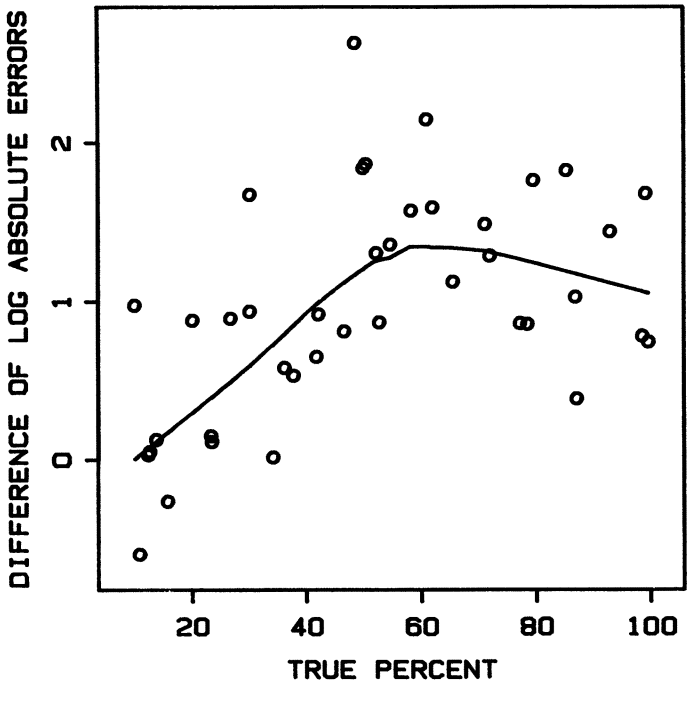

Figure 15. Positon–angle experiment: Angle midmeans minus position midmeans against true percentages.

The top panel of Figure 16 shows that average errors for length judgments are considerably larger than those for position judgments. A multiple comparison analysis (discussed in Section 4.4) showed that all pairs of the five averages are significantly different at the .05 level, except for Judgment Types 2 and 3. The larger of the two length values is 1.32 log units greater than the smallest of the three position values, which is a factor of $2^{1.32} = 2.5$. The smaller length value is .51 log units greater than the largest position value, which is a factor of 1.4. Thus the average errors for length judgments are 40%–250% larger than those for position judgments.

The bottom panel of Figure 16 shows that the average error for angle judgments is considerably larger than for position judgments. The difference is .97 on the log scale, which is a factor of $2^{.97} = 1.96$, and is statistically significant.

Large Absolute Errors

The top panel of Figure 17 shows a summary of the large errors for the position–length experiment. Of the 2,550 judgments made by the subjects, 136 had a log error greater than 4. The top panel of Figure 17 shows the percentage of these large errors that occurred for each of the

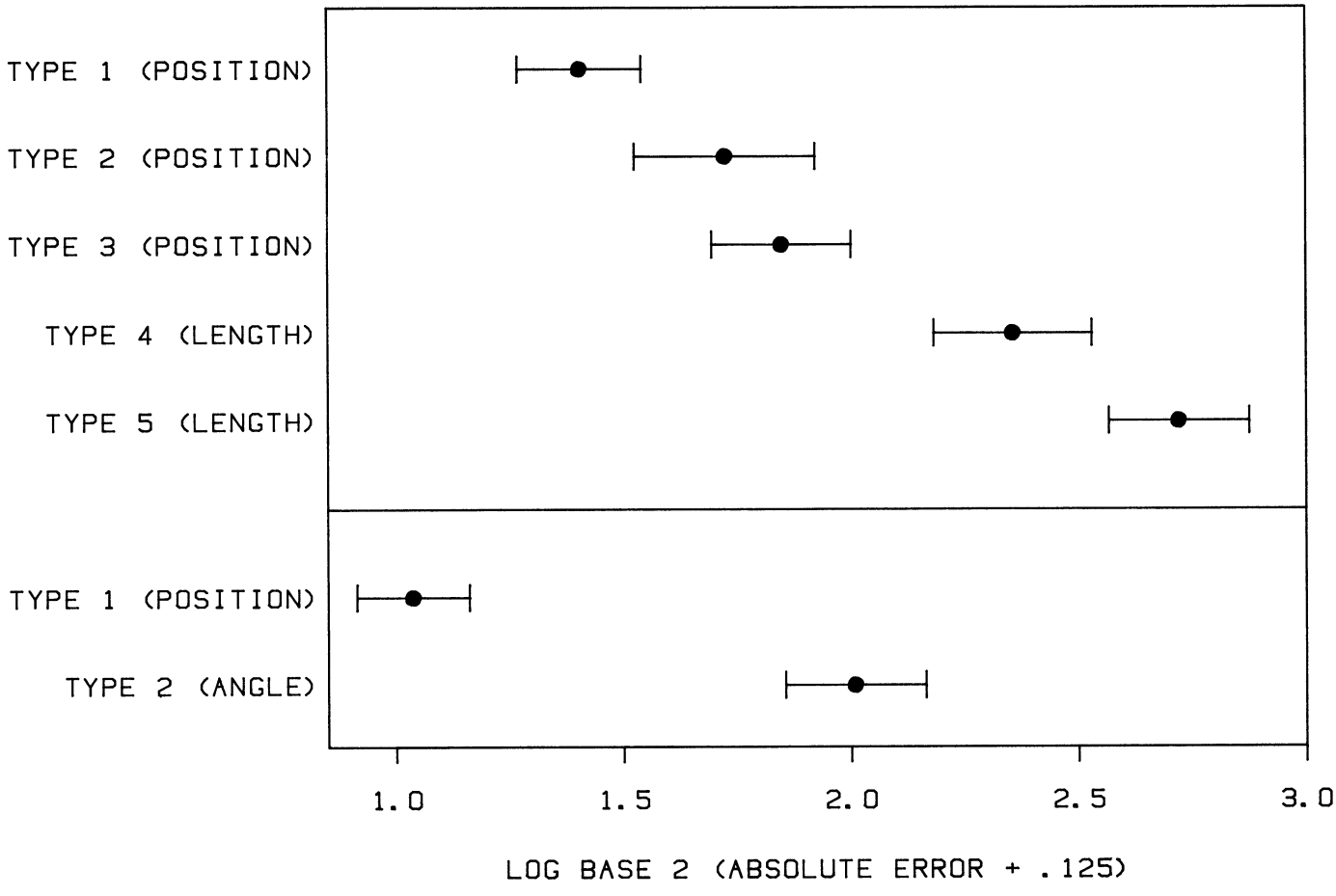

Figure 16. Log absolute error means and 95% confidence intervals for judgment types in position–length experiment (top) and position–angle experiment (bottom).

judgment types. Seventy-eight percent of the large errors occurred for the length judgments; since there were three position judgments for each two length judgments, the rate of occurrence of large errors for length judgments is 5.3 times that for position judgments.

The bottom panel of Figure 17 shows the percentage of large errors (those greater than 4) for the position–angle experiment; in this experiment 219 of the 4,080 judgments had large errors. Eighty-eight percent of the large errors occurred for the angle judgments; thus the rate of occurrence of large errors for the angle judgments is 7.3 times that for the position judgments.

Bias

Previously it was pointed out that subjective estimates of physical magnitudes can have systematic biases. To check for this in the two experiments, the errors,

$$\text{judged percentage} - \text{true percentage},$$

were analyzed. Just as for the log absolute errors, the midmeans of the errors across subjects were computed for each experimental unit in the two experiments. These midmeans are plotted against the true percentages for each judgment type in the position–length experiment (Figure 18) and the position–angle experiment (Figure 19), just as they were for the midmeans of the log absolute errors in Figures 13 and 14.

Figure 18 shows a convincing pattern for Judgment Type 5; there appears to be substantial negative bias for true percentages between 30 and 70. Figure 19 shows a pattern for the angle judgments on the pie charts; again, in the middle range of the true percentages, there are many experimental units with a negative bias.

Figure 20 shows the means of the midmeans for each judgment type in the two experiments; thus each value in the top panel is the mean of the midmeans in one panel of Figure 18, and each value in the bottom panel is the mean of the midmeans in one panel of Figure 19. As with the log absolute errors, these values are rough summaries because there appears to be a dependence of bias on the true percentage. Also shown are 95% confidence intervals for each mean, computed by a procedure described in the next section. The only source of significant bias appears to be the two length judgments and the angle judgment. The biases in these cases obviously contribute significantly to the log absolute errors. To see this, suppose that all subjects' judgments for an experimental unit had been identical; then we would have had

$$\log_2(\,|\,\text{bias}\,|\,+\,.125) = \text{log absolute error}.$$

The values of the log absolute bias for the Type 4 length judgment, the Type 5 length judgment, and the angle judgment are

.98 2.20 1.36,

respectively. The corresponding actual log absolute errors are

2.36 2.72 2.01.

Thus the log absolute biases are not small compared with the log absolute errors.

4.4 Confidence Intervals

The bootstrap (Efron 1982) proved to be a very convenient tool for estimating the sampling distributions of the means of the log absolute errors and the biases. Because each subject judged all of the experimental units in an experiment, the judgments of one unit are correlated with those of another, and modeling this correlation would have been a substantial chore. This correlation, the nonnormality of the log errors, and the use of the midmean make mathematical deviations of sampling distributions intractable.

Bootstrap Distribution of Means for Log Absolute Errors

For each experiment we bootstrapped by drawing 1,000 random samples of size 51 with replacement from the 51 subjects. For each sample, the means of the midmeans of the log absolute errors were computed as in Figure 16. Thus in the position–length experiment, there were 1,000 values of the five judgment-type means for the log absolute errors; this multivariate empirical distribution in five dimensions appeared to be well approximated by a multivariate normal distribution. This was established by making probability plots of the five marginal distributions and a number of linear combinations. The standard deviations and the correlation coefficients computed from the five vectors of 1,000 numbers serve as estimates of the standard deviations and correlations of the five judgment-type means. Similarly in the position–angle experiment, there were 1,000 values of the two judgment-type means; for the log absolute errors, this bootstrap distribution was well approximated by a bivariate normal one. The 95% confidence intervals in Figure 16 are simply plus and minus 1.96 times the bootstrap standard deviation estimates.

Using the normal approximation to the bootstrap distribution of the means in the position–angle experiment, a 95% confidence interval for the difference (angle − position) in the log absolute error means is (.79, 1.15). For the position–length experiment, the bootstrap distribution can be used to generate simultaneous confidence intervals for all pairs of differences of the means without being tied to any specific multiple comparison

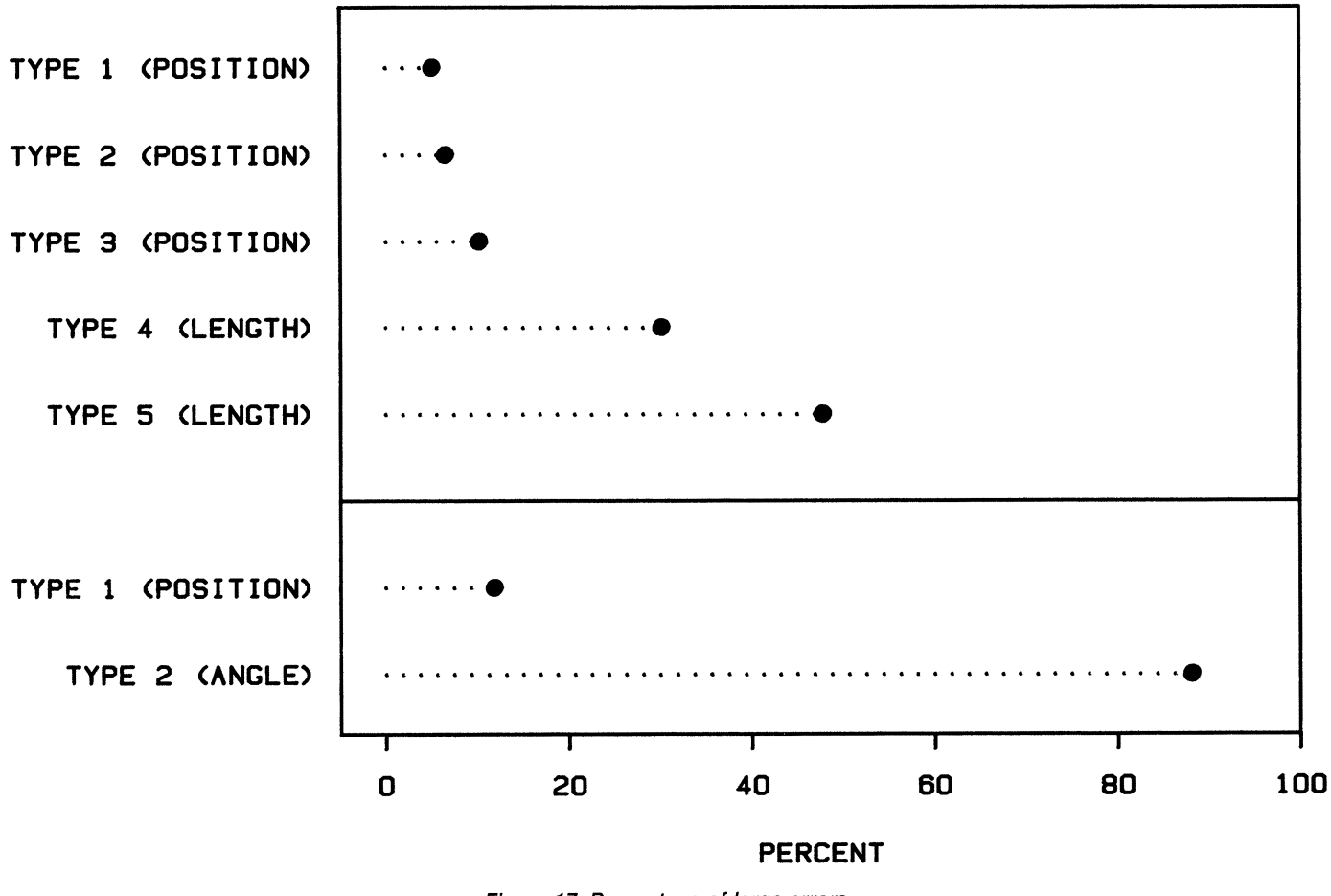

Figure 17. Percentage of large errors.

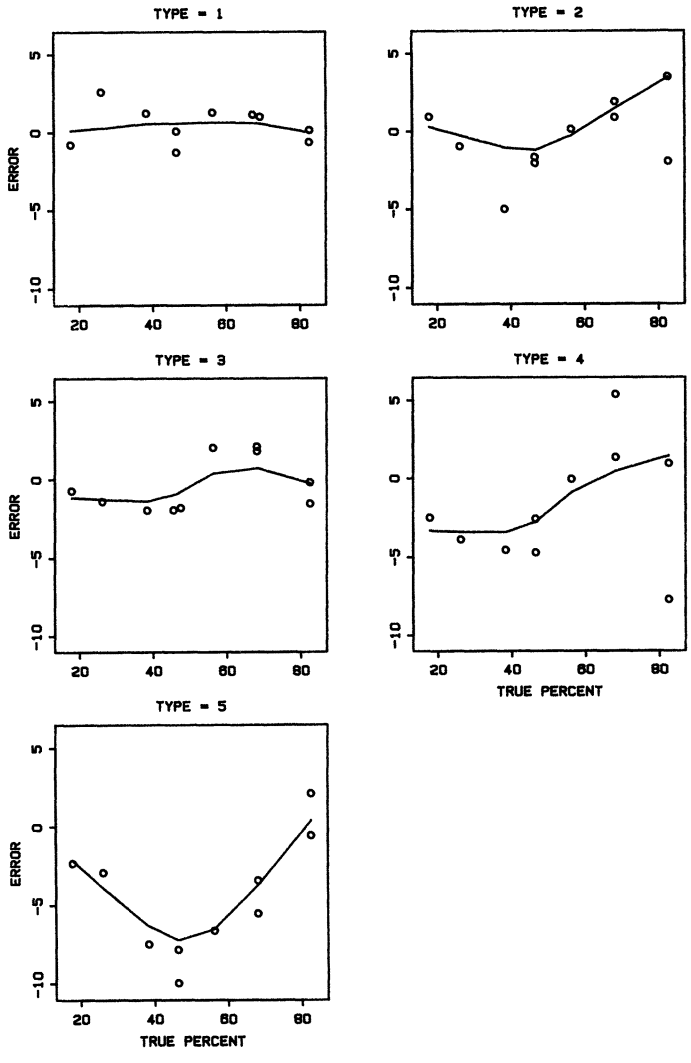

Figure 18. Position–length experiment: Midmeans of errors against true percentages for five judgment types.

method, such as the S or T method (Scheffé 1959). Let $\bar{\theta}_i$, for $i = 1, \ldots, 5$, be the judgment-type means for the log absolute errors. Let θ_{ik}^*, for $k = 1, \ldots, 5$ and $j = 1, \ldots, 1{,}000$, be the 5,000 bootstrap values, and let s_{ij} be the standard deviation of $\theta_{ik}^* - \theta_{jk}^*$, for $k = 1, \ldots, 1{,}000$. We found the c such that for 95% of the 1,000 bootstrap 5-tuples,

$$| (\bar{\theta}_i - \bar{\theta}_j) - (\theta_{ik}^* - \theta_{jk}^*) | \leq c s_{ij}.$$

This turned out to be 2.79. Thus

$$\bar{\theta}_i - \bar{\theta}_j \pm 2.79\, s_{ij} \qquad \text{for } i, j = 1, \ldots, 5$$

are a set of simultaneous 95% intervals for the differences of the means; these intervals are displayed in Figure 21. Note that only the means for Types 2 and 3 are not significantly different at the .05 level.

Bootstrap Distribution of Means for Errors

The bootstrap was used to assess the sampling distribution of the error means displayed in Figure 20; the bootstrap distribution was generated by 1,000 samples in a manner analogous to that described for the log absolute errors. Again, the multivariate normal was found to be a good approximation, and the confidence intervals in Figure 20 show plus and minus 1.96 times the bootstrap standard errors.

4.5 Summary of the Experiments

Two experiments were run in which subjects judged bar charts and pie charts. In the first experiment, five types of judgments were made—two length judgments and three judgments of position along a common scale. In the second experiment, there were two types of judgment—position and angle. For all types of judgments, subjects made visual assessments of what percentage one value was of a larger value; thus all recorded values were between 0 and 100. In both experiments there were 51 subjects with usable data.

Figure 16 summarizes the accuracy of the judgments. The top panel shows the first experiment and the bottom panel shows the second. The scale is the log base 2 of the absolute errors plus 1/8. In the first experiment, position judgments were more accurate than length judgments by factors varying from 1.4 to 2.5. In the second experiment, position judgments were 1.96 times as accurate as angle judgments. The 95% confidence intervals shown in Figure 16 were computed by using the bootstrap. An important part of the contribution to the errors for length and angle judgments is consistent bias. When the true percentages are in the range of 25–50, subjects tend to underestimate values for these types of judgments.

The first experiment suggested that the position task should be expanded to a whole range of tasks. As the distance between the two values being judged increased along an axis perpendicular to the common scale, the accuracy decreased. Type 1 judgments had the smallest distance, Type 2 the next smallest, and Type 3 the largest. Not surprisingly, after just two experiments a revision of the theory seems appropriate.

5. APPLYING THE THEORY TO ANALYZE AND REDESIGN SEVERAL MUCH-USED GRAPH FORMS

The mode of graph design that we advocate is the construction of a graphical form that uses elementary perceptual tasks as high in the hierarchy as possible. The hypothesis is that by selecting as high as possible, we will elicit judgments that are as accurate as possible, and therefore the graph will maximize a viewer's ability to detect patterns and organize the quantitative information.

In this section we use this mode of graph design to analyze several much-used graph forms and to construct replacements for some of them. The comparison of old graph forms and new ones provides another type of experiment that can be used to decide the validity of our approach.

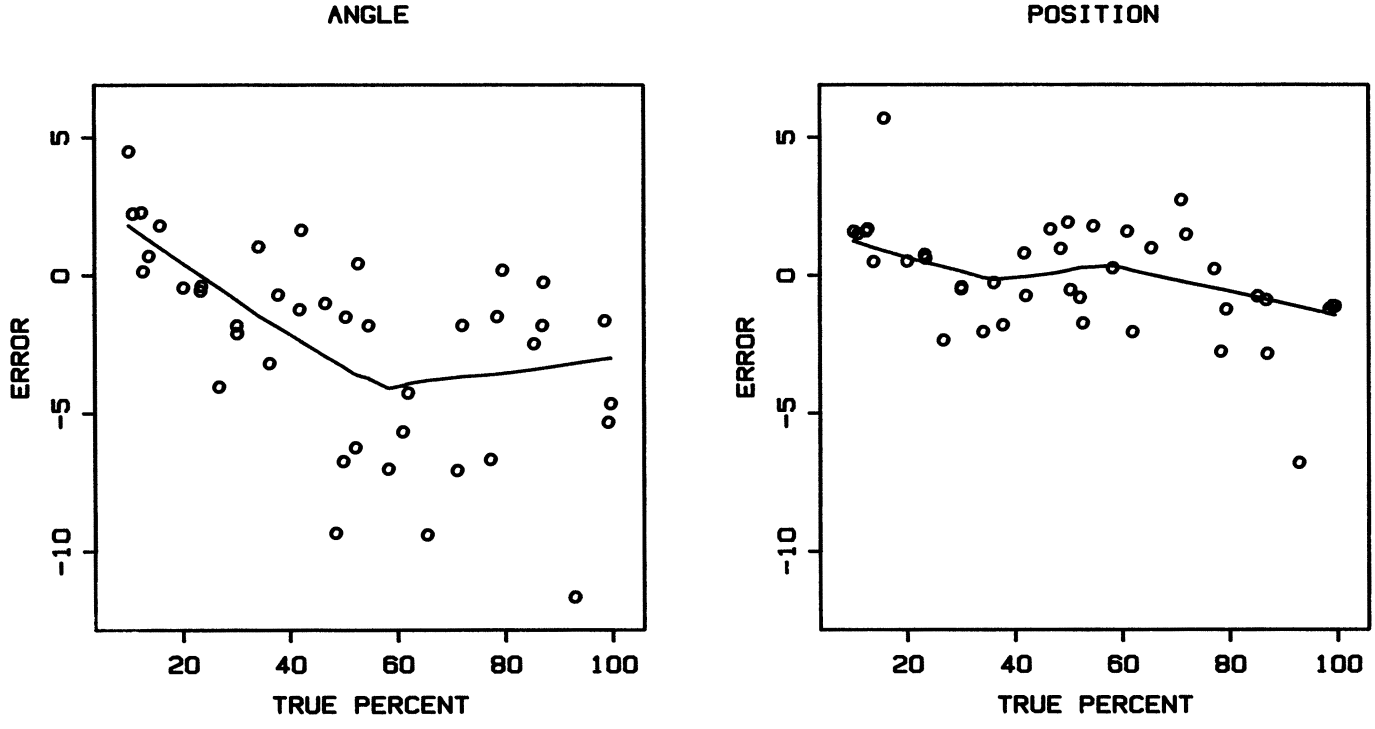

Figure 19. Position–angle experiment: Midmeans of errors against true percentages for two judgment types.

5.1 Dot Charts and Bar Charts as Replacements for Divided Bar Charts and Pie Charts; Grouped Dot Charts and Grouped Bar Charts as Replacements for Divided Bar Charts

For certain types of data structures, one cannot always use the most accurate elementary task, judging position along a common scale. But this is not true of the data represented in divided bar charts and pie charts; one can always represent such data along a common scale.

A pie chart can always be replaced by a bar chart, thus replacing angle judgments by position judgments. In so doing it might be sensible in many cases to make the scale go from 0 to 100% so that the viewer can more readily appreciate the fraction that each bar is of 100%; but 0 to 25 or 50% are also reasonable simple choices.

Actually we prefer dot charts, which are introduced and discussed in Cleveland (1983), to bar charts. Figures 16, 17, and 20 are dot charts. (The reasons for our preference depart somewhat from our theme, so we refer the reader to Cleveland 1983.)

Figure 22 is a pie chart. What is the ordering of the values of the five categories? The answer is not easy to find from the pie chart. From the dot chart in Figure 23, it is clear that the ordering from smallest to largest is A to E. This demonstrates the increase in ability to perceive patterns that results from the increased accuracy of perceptions based on position relative to that based on angle judgments.

A divided bar chart can always be replaced by a grouped bar chart; again, we prefer a grouped dot chart (discussed in Cleveland 1982) to a grouped bar chart. To illustrate the replacement of divided bar charts, consider the graph in Figure 24. What is the ordering of the five items in category A? As with the pie chart, making the judgments is not easy. Figure 25 is a grouped dot chart of the data in Figure 24. For each of the categories A, B, and C, the totals and the item values are shown. Thus the many length judgments in the divided bar chart have been replaced by position judgments. It is clear that the order of the items in category A from smallest to largest is 1 to 5. Again, there is an increased ability to perceive patterns as a result of the increased accuracy of perceptions.

Our analysis has provided, in a sense, a resolution of the "Bar–Circle Debate," as Kruskal (1982) refers to it. This was a controversy (Eells 1926; Croxton 1927; Croxton and Stryker 1927; von Huhn 1927) about whether the divided bar chart or the pie chart was superior for portraying the parts of a whole. The contest appears to have ended in a draw. We conclude that neither graphical form should be used because other methods are demonstrably better.

5.2 Showing Differences Directly for Curve-Difference Charts

In the Playfair chart of Figure 6, the vertical distances between the two curves encode pictorially England's balance of payments with the East Indies. Thus the elementary task in extracting the curve differences is perceiving length. It turns out that making such length judgments is inaccurate and even more difficult than on a divided bar chart. In fact the situation is so striking that

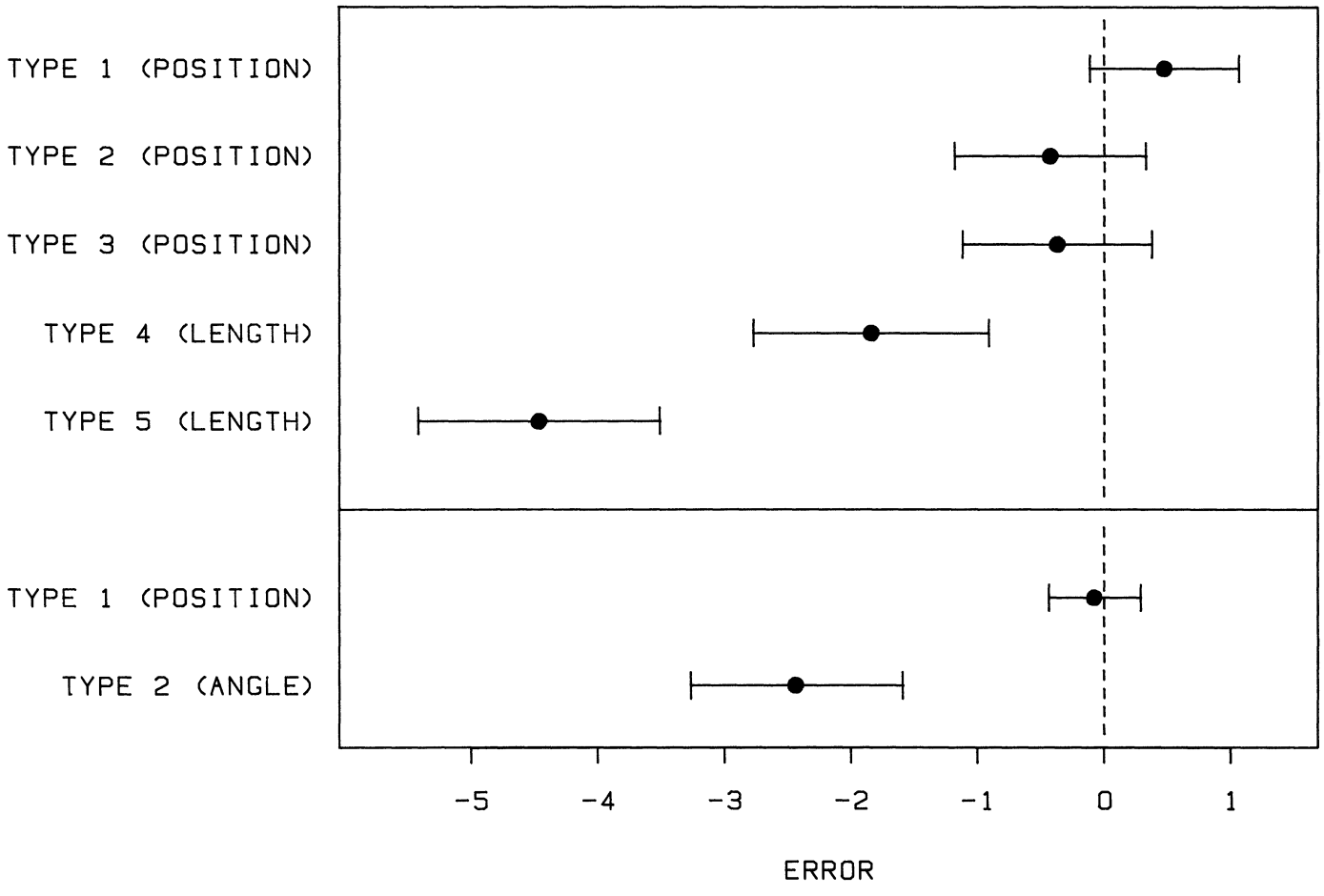

Figure 20. Error means and 95% confidence intervals for judgment types in position–length experiment (top) and position–angle experiment (bottom).

an experiment with subjects recording judgments is not necessary; it has taken only a few examples to convince us. One is shown in Figure 26. It is almost impossible to get even a rough idea of the behavior of the differences of the curves in the nine panels. The problem is that the brain wants to judge minimum distance between the curves in different regions, and not vertical distance. Thus in each panel of Figure 26, one tends to see the curves getting closer, going from left to right. The actual vertical differences are plotted in Figure 27; it is clear that Figure 26 has not conveyed even the grossest qualitative behavior of the differences.

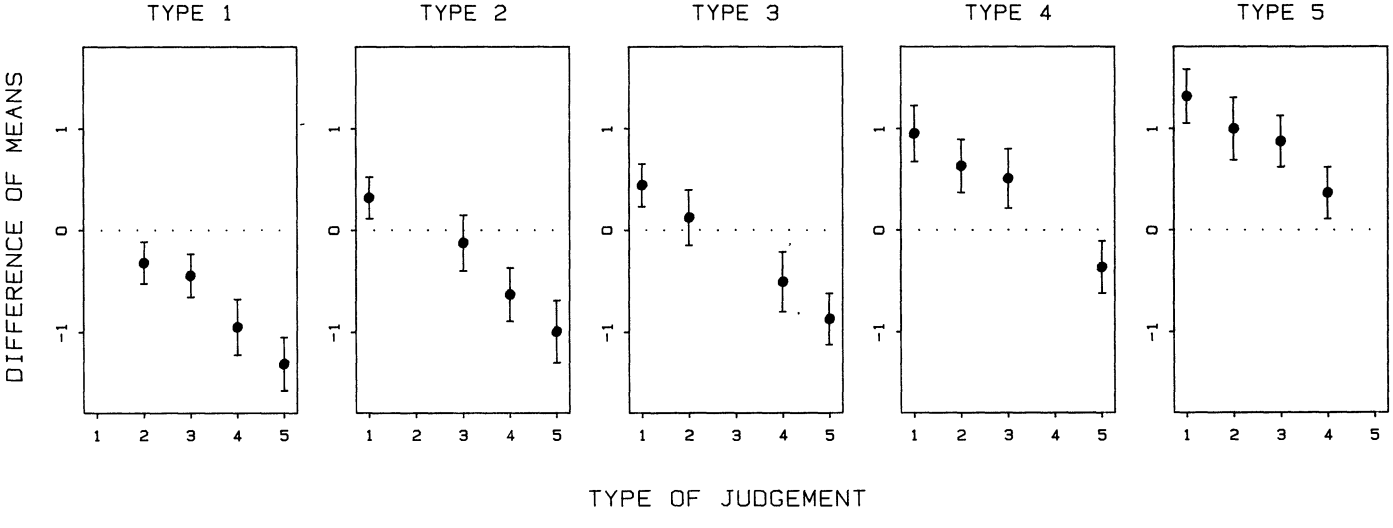

Figure 21. Simultaneous 95% confidence intervals for differences of judgment-type means in position–length experiment.

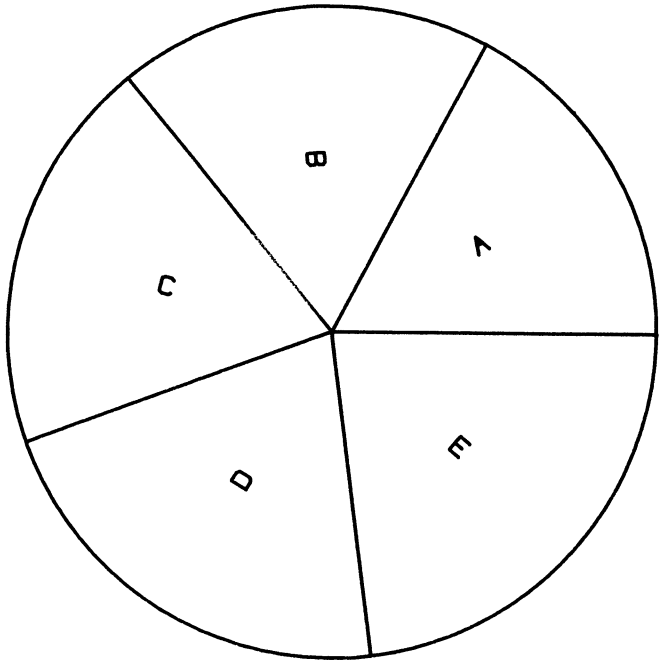

Figure 22. Pie chart.

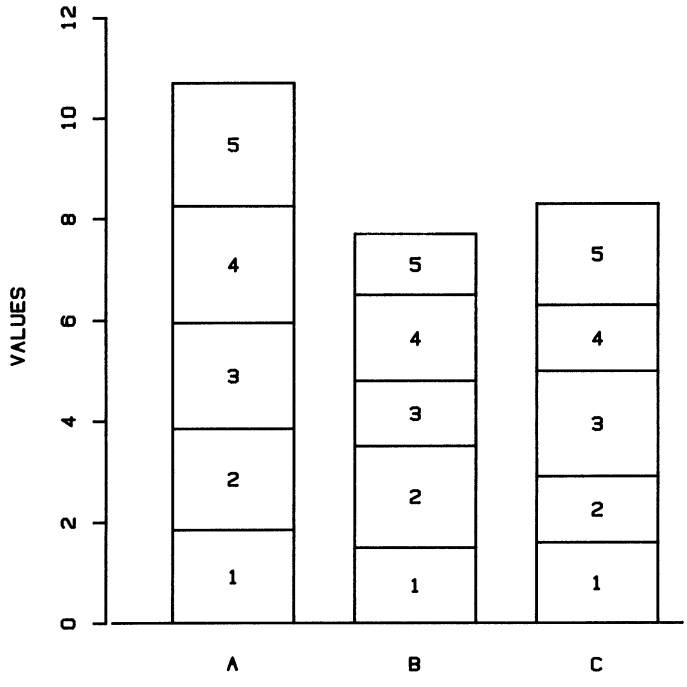

Figure 24. Divided bar chart.

The same problem exists in the Playfair chart of Figure 6, although a little less severely. Figure 28 contains a Cartesian graph of the differences, which does a far better job of portraying them because the elementary perceptual task is judging position along a common scale. For example, Figure 28 does a far better job of showing the occurrence of the rapid rise and descent of the balance against England around 1760; in Figure 6 this peak goes almost unnoticed unless considerable cognitive mental effort is expended. A sensible graphing of these data would show the import–export curves and the differences graphed directly, as in Figure 28.

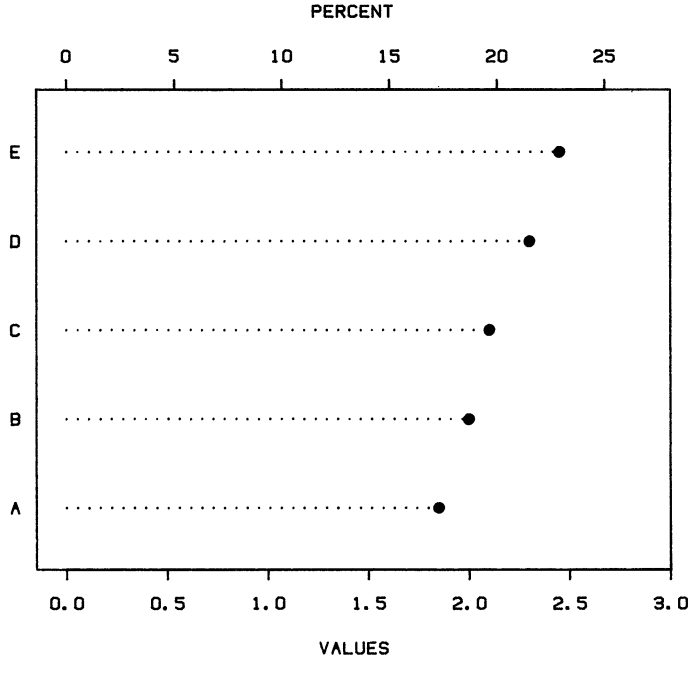

Figure 23. Dot chart.

The remedy in this case seems simple: If differences are to be conveyed, they should also be plotted on their own Cartesian graph. This applies equally to the much-used divided line chart, sometimes called an aggregate line chart (Monkhouse and Wilkinson 1963). In such a graph the amounts in various categories, say A to D, are portrayed through time by plotting A, A + B, A + B + C, and A + B + C + D against time as four curves. Thus only A and the total, T = A + B + C + D, can be judged by perception along a common scale, whereas B, C, and D must be judged by perceiving vertical lengths between two curves. Our perceptual theory and examples strongly indicate that abandoning divided line charts and plotting A, B, C, D, and T directly will lead to far more accurate judgments.

5.3 Framed-Rectangle Charts as Replacements for Statistical Maps With Shading

Statistical maps that use shading (or color saturation or color hue) to encode a real variable, which Tukey (1979) called patch maps, are commonly used for portraying measurements as a function of geographical location. Figure 5 is one example. Murder rate is encoded by the grid spacing, forming a kind of graph-paper collage.

To judge the values of a real variable encoded on a patch map with shading, one must perform the elementary perceptual task of judging shading, which is at the bottom of our perceptual hierarchy. One can move much farther up the hierarchy by using the framed rectangles discussed earlier to form a framed-rectangle chart. This

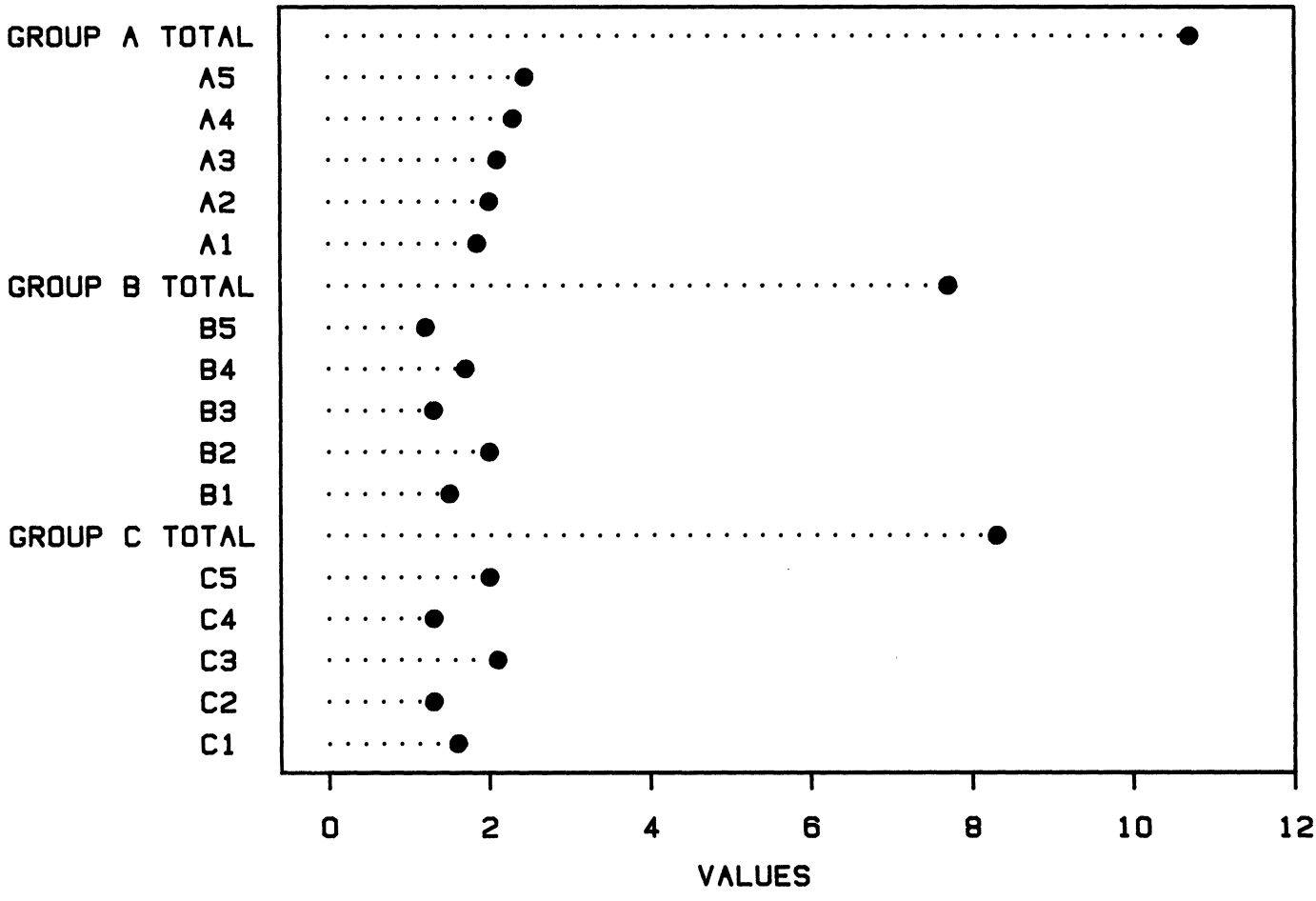

Figure 25. Dot chart with grouping.

is illustrated in Figure 29 with the murder data portrayed in Figure 5. Had we merely shown the bars without the frames, we would have had what Monkhouse and Wilkinson (1963) called a statistical map with located bars; the elementary task would then have been perceiving length. The framed rectangles, which are one step higher in the hierarchy, lead to more accurate judgments, for the reasons discussed in Section 3.

The framed-rectangle chart also solves another serious problem of statistical maps with shading. On such patch maps the states are treated in a very uneven way because of their different areas. For example, in Figure 5 the total amount of black for each state is actually encoding

$$\frac{\text{number of murders}}{\text{number of people}} \times \text{area}.$$

The result is that Texas is imposing and Rhode Island is hard to see.

There is another, more subtle perceptual problem that arises on a patch map with shading. In Figure 5, for example, one tends to see contiguous clusters of states: The two most prominent clusters are the north central states (North Dakota, South Dakota, Nebraska, Minnesota, Iowa, and Wisconsin) and New England (Maine, New Hampshire, Vermont, Massachusetts, Connecticut, and Rhode Island).

Part of the reason why the clustering occurs so strongly on the patch map is the reduction in the accuracy of the perceived quantitative information; values group together because we cannot visually differentiate them. Thus the encoding of the data on the patch map provides a kind of visual data reduction scheme in which noise is reduced and a signal comes through. Unfortunately the signal is of poor quality, since the clustering is subject to the vagaries of the shading scheme. For example, the deep South states (Texas, Louisiana, Mississippi, Alabama, and Georgia) deserve to cluster together as forcefully as the New England states but do not because our sensitivity to differences at the high end of the scale appears to be greater than at the low end of the scale. The deep South states contain five of the six largest rates, and their range is 3.2. The range for New England is 2.7. Furthermore the largest deep South value (Louisiana) is 1.4 units larger than the next largest value in the cluster, and the smallest New England value (New Hampshire) is 1.3 units less than the next smallest value; but Louisiana appears to stand out in its cluster much more forcefully than does New Hampshire.

If we want to perform data reduction, eliminating noise to allow a signal to come through, then we can use a

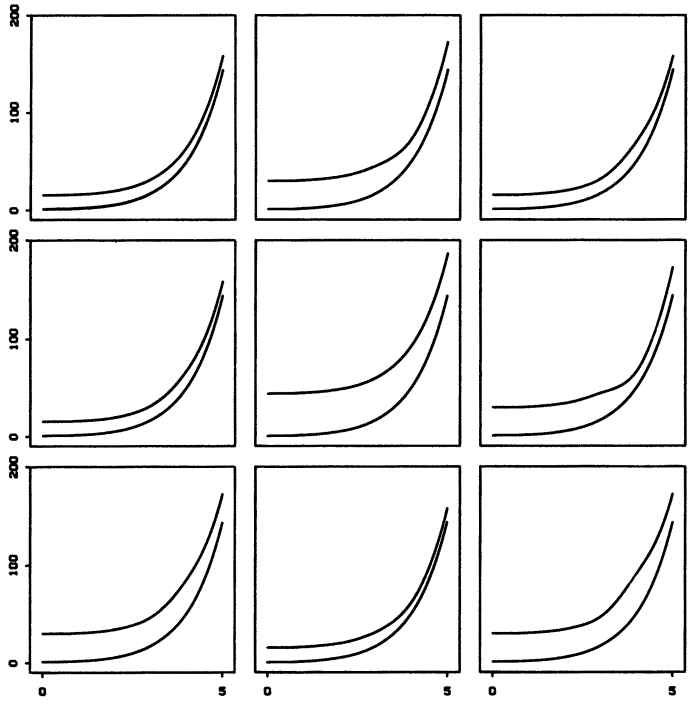

Figure 26. Curve-difference chart.

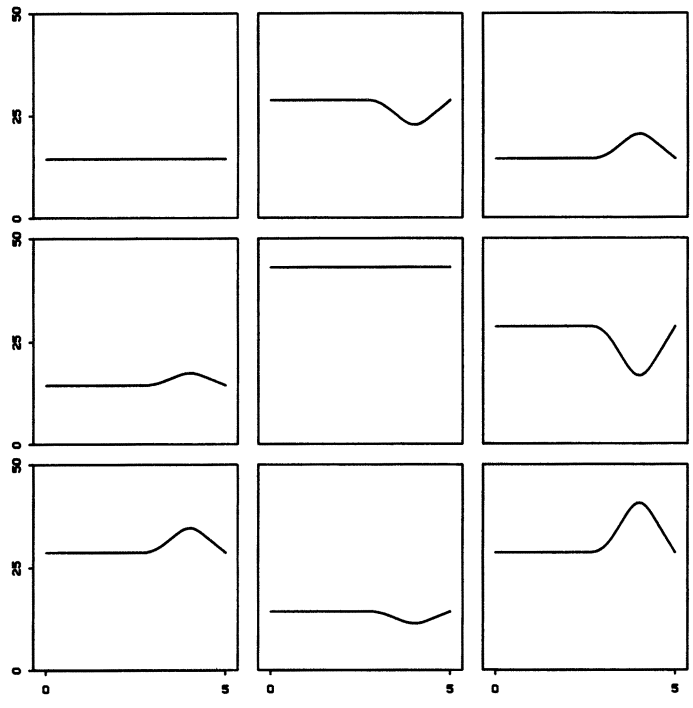

Figure 27. Curve differences.

sensible numerical scheme together with a higher accuracy chart such as the framed-rectangle chart. One procedure, suggested by Tukey (1979), would be to smooth the variable as a function of geographical location and portray smoothed values. Another somewhat crude, but reasonable data reduction procedure is to group the data

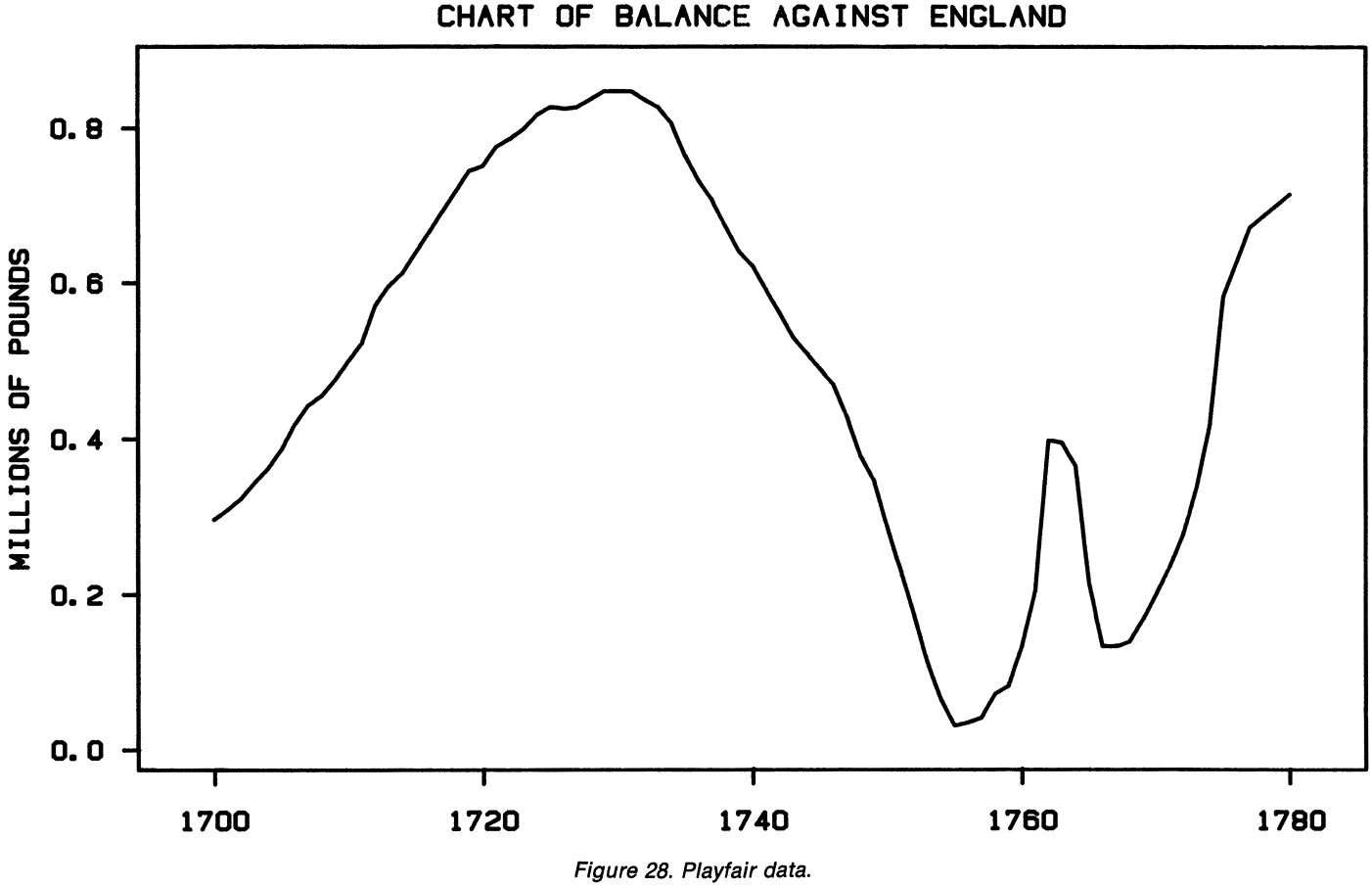

Figure 28. Playfair data.

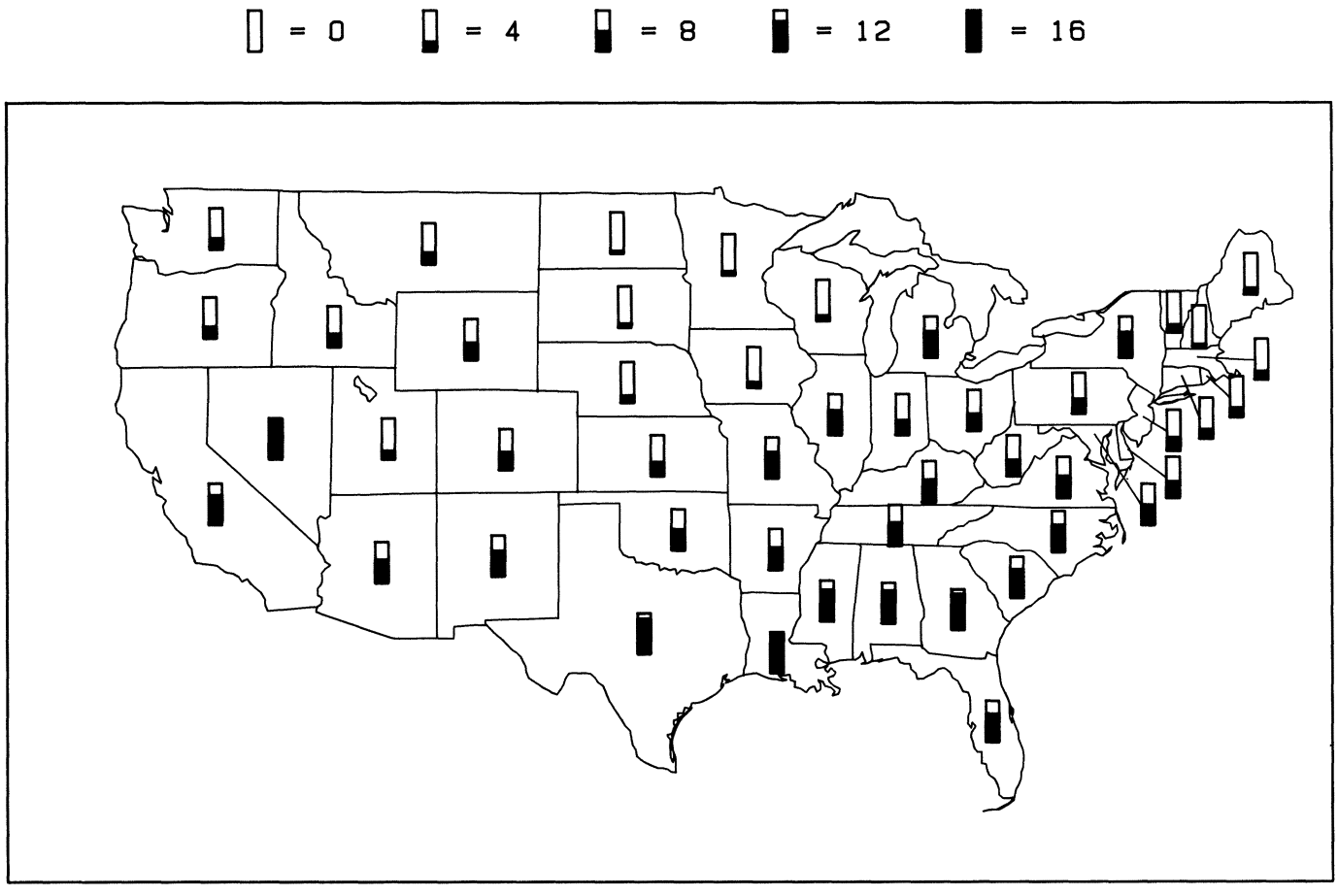

Figure 29. Framed-rectangle chart.

into equal-length intervals and portray the midpoints. This has been done in Figure 30, and now the north central states, northern New England, and the deep South form more clear-cut visual clusters than in Figure 29.

Another data reduction technique, a visual one, that results in effective but somewhat fuzzier clusters is simply to reduce the vertical resolution of the framed rectangles by reducing their heights. This has been done in Figure 31; clusters of states now appear to form more readily than in Figure 29. It should be noted that this technique works because the reduction prevents one from *optically* detecting certain differences. In general one would not expect graph size to be a major factor in graphical perception until things were so small that differences would be optically blurred. Because the graph elements in our experiments were sufficiently large, as graph elements usually are, size was not a factor that we needed to take into account. It is fortunate that this was so; otherwise the distance the viewer held the graph from his or her eyes would have been a factor.

Our conclusion about patch maps agrees with Tukey's (1979), who left little doubt about his opinions by stating, "I am coming to be less and less satisfied with the set of maps that some dignify by the name *statistical map* and that I would gladly revile with the name *patch map*" (p. 792).

5.4 Graphs for Data Analysis

The graphical forms discussed so far in this section are used more in data presentation than in data analysis. But our perceptual theory can serve equally well as a guide for designing graphical methods for statistical analyses.

Triple Scatterplots

The triple scatterplot is a useful tool in data analysis for understanding the structure of three-dimensional data. Figure 9 shows one implementation; perceiving the values encoded by the circles requires the elementary task of judging area. Anscombe (1973) has suggested another scheme for typewriter terminals and printers in which overplotted characters, increasing in size and amount of black, encode the third variable.

In a sense the framed-rectangle chart is a triple scatterplot; thus one might think in terms of a general triple scatterplot procedure in which the third variable is coded by framed rectangles. But for general data analytic purposes, this is unlikely to work well because of a practical

difficulty—overlapping symbols. For the statistical map it was easy to avoid overlap, but for general scatterplots, where points can get very crowded, the problem would often be insurmountable.

Circles can overlap a lot and still permit perception of circle. Part of the reason for this is that overlapping circles tend to form regions that do not look like circles, so the individuals stand out. Since squares do not have this property, overlap becomes a problem much more quickly.

Our perceptual theory suggests that the third variable be encoded by line length so that a more accurate elementary perceptual task can be performed. We have not experimented with this procedure enough to know whether line overlap is a lesser or greater problem than circle overlap.

Hanging Rootograms and Slopes of Normal Probability Plots

It can be said that John Tukey has already implemented a part of our perceptual theory by recommending the redesign of two common statistical graphical methods. The hanging rootogram (Tukey 1972) modifies the usual method of superimposing a normal density on a histogram, converting the perceptual task from judging length to judging position along a common scale. Tukey (1962) also suggested modifying normal probability plots by plotting the slopes of lines connecting the median point (i.e., data median vs. the median of the normal, which is usually taken to be zero) with other points on the plot; the slope from the median point to the point associated with the ith largest order statistic is plotted against i. The viewer of an ordinary normal probability plot must judge whether the points form a straight line pattern, so Tukey's modification converts judgment of direction (slope) to judgment of position along a common scale.

Symbols for Multidimensional Data

One area of statistical graphics that has received a lot of attention is designing symbols for representing multidimensional data. Examples are polygons, Anderson glyphs, faces, profiles, and Kleiner-Hartigan trees (Chambers et al. 1983). Let us consider faces. Judging the values of the individual encoded variables requires five elementary perceptual tasks: position along nonaligned scales, length, direction, area, and amount of curvature. Thus extracting the quantitative information requires substantial perceptual processing; and there is no

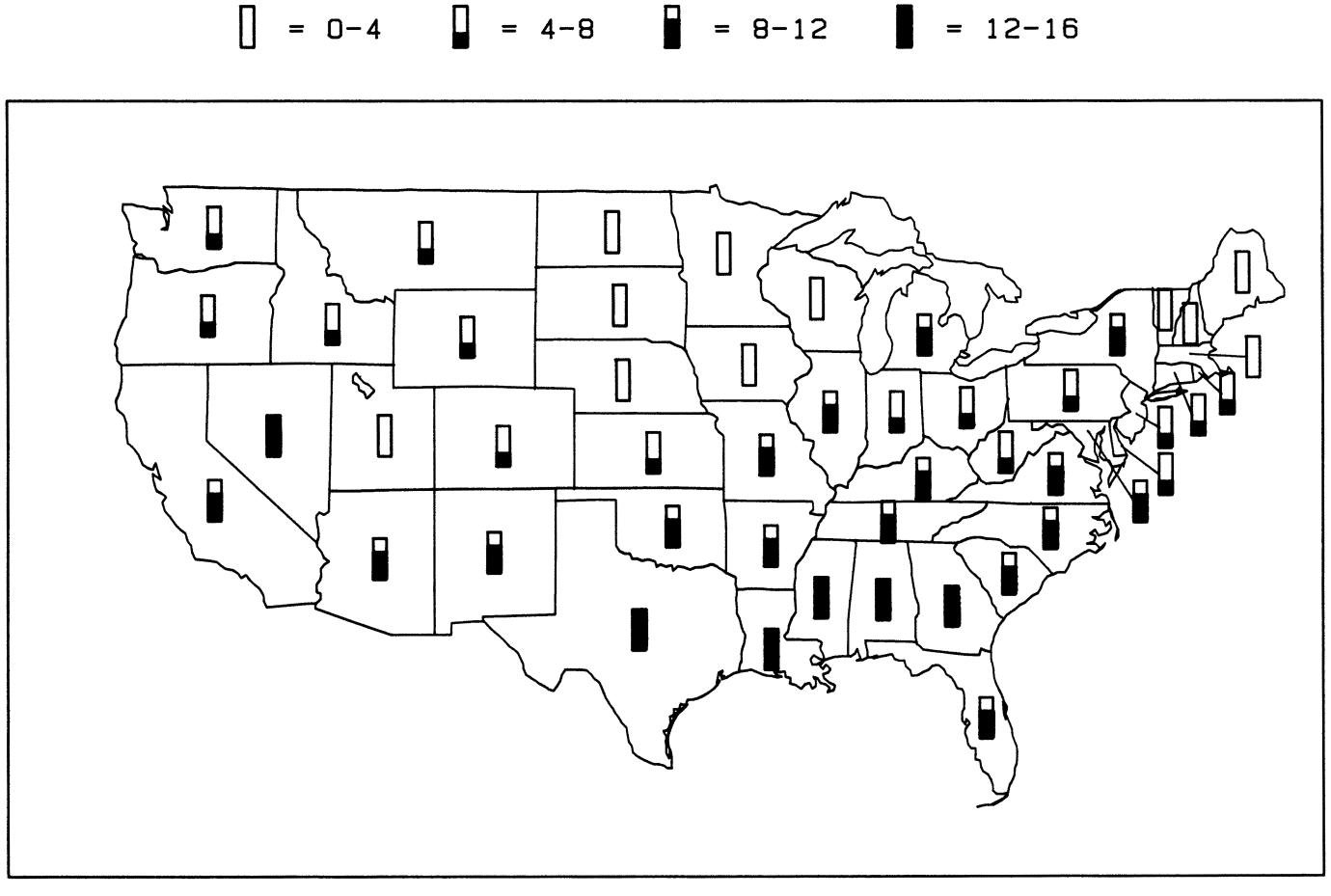

Figure 30. Framed-rectangle chart.

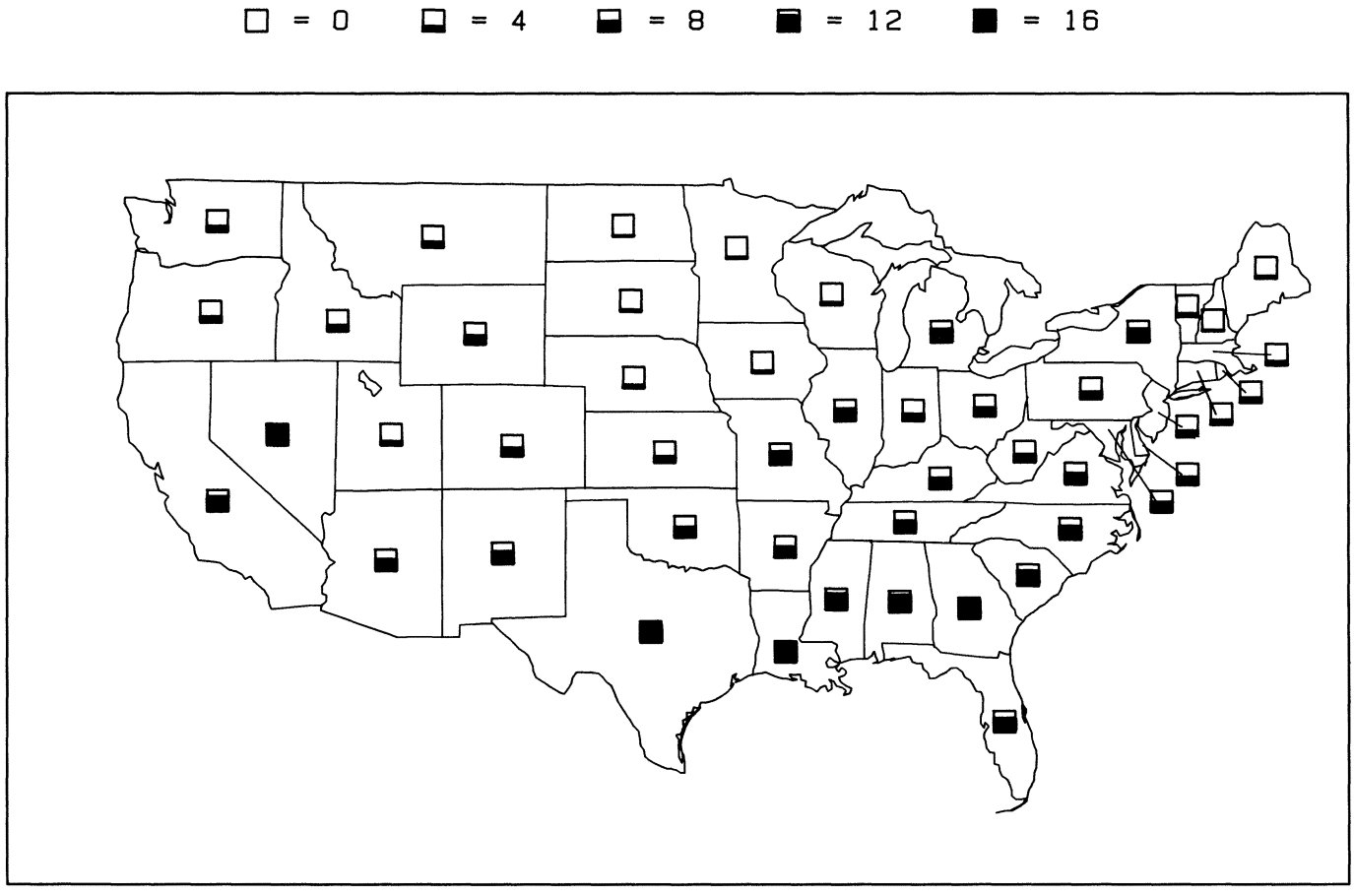

Figure 31. Framed-rectangle chart.

easy and direct elementary task that one can perform to allow the study of the relationship between two variables, as for a Cartesian plot. For this reason faces and the other symbolic displays tend not to tell us much about the geometry of the multidimensional point cloud. Fixed or data-driven projections of the points onto planes (Tukey and Tukey 1981) appear to be more useful; each projection is portrayed by a scatterplot, so the elementary tasks performed are judgments of position along a common scale and direction (slope). Of course the integration of the projections requires complex perceptual and cognitive mental tasks.

6. PERSPECTIVES, REALISM, AND CRITICISM

For some the word *theory* implies a detailed, systematic, and comprehensive description of a subject. Such a meaning would, of course, be ludicrous for the tentative first step in this article. For us the identification and ordering of the perceptual tasks is a theory in a less restrictive sense: It is a set of plausible statements that describe a phenomenon—the relative accuracy with which various graphical forms convey quantitative information.

We expect that our theory, like all theories, will undergo much revision as new experimental information is accumulated. The outcomes of the two experiments reported here were correctly predicted by the theory; position judgments were more accurate than length judgments and angle judgments. The position–length experiment suggests, however, that a revision in the theory might be appropriate. Although Judgment Types 1–3 involved judgments of position along a common scale, namely the vertical scale of the bar charts, the horizontal distance between the graphical elements being judged varied from 0 cm for Type 1 to 2.8 cm for Type 2 to 5.6 cm for Type 3; Figures 16 and 17 show that errors increased in going from Type 1 to Type 2 to Type 3. This suggests that the elementary task of judging position be expanded into a continuum of tasks for which accuracy is conjectured to decrease with increasing distance between the graphical elements encoding the data, where distance is measured perpendicular to the axis along which the data are plotted. Not surprisingly, after just two experiments a revision in the theory appears necessary.

The ordering of the perceptual tasks does not provide a complete prescription for how to make a graph. Rather, it provides a set of guidelines that must be used with judgment in designing a graph. Many other factors, such as what functions of the data to plot, must be taken into

account in the design of a graph. A discussion of this is given in Chambers et al. (1983, Ch. 8).

We have used *elementary perceptual task* to describe the basic elements involved in our theory. It may have been more appropriate to call them *elementary graphical encodings*, emphasizing that they are basic ways of encoding data on graphs. We cannot realistically claim to have isolated 10 basic, independent perceptual tasks. Each task is really a complex set of tasks, and there is much overlap. For example, it might be argued that judging positions along nonaligned scales really involves making two length judgments, one from each end of the axis. Despite these shortcomings, we have used *elementary perceptual task* to emphasize that we are studying the decoding process of the human–graph interface.

One substantial danger in performing graphical perceptual experiments is that asking people to record judgments will make them perform judgments differently from the way they perform them when they look at graphs in real life. Subjects will try to get the right answer and might perform much more highly cognitive tasks than the basic perceptual tasks they perform in real life. We tried to guard against this in various ways in our experiments: One way was to encourage subjects to work quickly, much as they might in looking at a graph in real life. Another was to omit tick marks and labels on axes except at the extremes. For example, consider the bar chart in the right panel of Figure 3. Had we put many tick marks and labels on the vertical axis, subjects could have judged ratios by reading values off the axis and performing a mental division. Although some people may perform such an operation in real life, it is not the basic perceptual processing from geometrical information that we wanted to study and that we conjecture is the main way viewers judge ratios in real life. We have no proof that our laboratory results are realistic and work in the field, but it appears plausible that this is so.

Whatever the limitations of the current theory, it appears to have led to some useful results. Its application to some of the most-used charts in graphical communication (bar charts, divided bar charts, pie charts, and statistical maps with shading) has led to replacements (dot charts, dot charts with grouping, and framed-rectangle charts). We do not lightly recommend the dismissal of some of the most popular graph forms, but it appears to be the inescapable conclusion of this analysis of graph design. If progress is to be made in graphics, we must be prepared to set aside old procedures when better ones are developed, just as is done in other areas of science.

[*Received May 1983. Revised October 1983.*]

REFERENCES

AMERICAN NATIONAL STANDARDS INSTITUTE (1979), *Time-Series Charts*, New York: The American Society of Mechanical Engineers.

ANSCOMBE, F.J. (1973), "Graphs in Statistical Analysis," *The American Statistician*, 27, 17–21.

BERTIN, J. (1973), *Semiologie Graphique* (2nd ed.), Paris: Gauthier-Villars.

BAIRD, J.C. (1970), *Psychophysical Analysis of Visual Space*, New York: Pergamon Press.

BAIRD, J.C., and NOMA, E. (1978), *Fundamentals of Scaling and Psychophysics*, New York: John Wiley.

CHAMBERS, J.M., CLEVELAND, W.S., KLEINER, B., and TUKEY, P.A. (1983), *Graphical Methods for Data Analysis*, Belmont, Calif.: Wadsworth.

CLEVELAND, W.S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

——— (1983), "Graphical Methods for Data Presentation: Dot Charts, Full Scale Breaks, and Multi-based Logging," Bell Laboratories memorandum.

CLEVELAND, W.S., HARRIS, C.S., and McGILL, R. (1983), "Experiments on Quantitative Judgments of Graphs and Maps," *Bell System Technical Journal*, 62, 1659–1674.

COX, D.R. (1978), "Some Remarks on the Role in Statistics of Graphical Methods," *Applied Statistics*, 27, 4–9.

CROXTON, F.E. (1927), "Further Studies in the Graphic Use of Circles and Bars II: Some Additional Data," *Journal of the American Statistical Association*, 22, 36–39.

CROXTON, F.E., and STRYKER, R.E. (1927), "Bar Charts Versus Circle Diagrams," *Journal of the American Statistical Association*, 22, 473–482.

EELLS, W.C. (1926), "The Relative Merits of Circles and Bars for Representing Component Parts," *Journal of the American Statistical Association*, 21, 119–132.

EFRON, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: Society for Industrial and Applied Mathematics.

EHRENBERG, A.S.C. (1975), *Data Reduction: Analyzing and Interpreting Statistical Data*, New York: John Wiley.

FEINBERG, B.M., and FRANKLIN, C.A. (1975), *Social Graphics Bibliography*, Washington, D.C.: Bureau of Social Science Research.

GALE, N., and HALPERIN, W.C. (1982), "A Case for Better Graphics: The Unclassed Choropleth Map," *The American Statistician*, 36, 330–336.

GREGORY, R.L. (1966), *Eye and Brain, the Psychology of Seeing*, New York: McGraw-Hill.

JULESZ, B. (1981), "A Theory of Preattentive Texture Discrimination Based on First-Order Statistics of Textons," *Biological Cybernetics*, 41, 131–138.

——— (in press), "Toward an Axiomatic Theory of Preattentive Vision," *Dyanmic Aspects of Neocortical Function*, eds. Edelman, Cowan, and Gall, New York: John Wiley.

KRUSKAL, W.H. (1975), "Visions of Maps and Graphs," in *Auto-Carto II, Proceedings of the International Symposium on Computer Assisted Cartography*, ed. J. Kavaliunas, Washington, D.C.: U.S. Bureau of the Census and American Congress on Survey and Mapping, 27–36.

——— (1982), "Criteria for Judging Statistical Graphics," *Utilitas Mathematica*, Ser. B, 21B, 283–310.

MARCUS, A., MARCUS, S., REINECK, J., and REINECK, G. (1980), *Graphic Design and Information Graphics*, Siggraph Tutorial Notes, New York: Association for Computing Machinery.

MARR, D., and NISHIHARA, H.K. (1978), "Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes," *Proceeding of the Royal Society of London*, Ser. B, 200, 269–294.

MONKHOUSE, F.J., and WILKINSON, H.R. (1963), *Maps and Diagrams*, London: Methuen.

MOSTELLER, F., and TUKEY, J.W. (1977), *Data Analysis and Regression*, Reading, Mass.: Addison-Wesley.

PINKER, S. (1982), "A Theory of Graph Comprehension," Occasional Paper No. 15, Cambridge, Mass.: MIT Center for Cognitive Sciences.

PLAYFAIR, W. (1786), *The Commercial and Political Atlas*, London.

ROBINSON, A.H., SALE, R.D., and MORRISON, J. (1978), *Elements of Cartography*, New York: John Wiley.

SCHEFFE, H. (1959), *The Analysis of Variance*, New York: John Wiley.

SCHMID, C.F., and SCHMID, S.E. (1979), *Handbook of Graphic Presentation*, New York: John Wiley.

STEVENS, S.S. (1975), *Psychophysics*, New York: John Wiley.

TUFTE, E. (1983), *The Visual Display of Quantitative Information*, Cheshire, Conn.: Graphics Press.

TUKEY, J.W. (1962), "The Future of Data Analysis," *Annals of Mathematical Statistics*, 33, 1–67.

——— (1972), "Some Graphic and Semi-Graphic Displays," in *Statis-*

tical Papers in Honor of George W. Snedecor, ed. T.A. Bancroft, Ames, Iowa: Iowa State University Press, 292–316.

——— (1977), *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley.

——— (1979), "Methodology and the Statistician's Responsibility for BOTH Accuracy AND Relevance," *Journal of the American Statistical Association*, 74, 786–793.

TUKEY, P.A., and TUKEY, J. W. (1981), "Graphical Display of Data Sets in 3 or More Dimensions," in *Interpreting Multivariate Data*, ed. V. Barnett, Chichester, U.K.: John Wiley, 189–275.

VON HUHN, R. (1927), "Further Studies in the Graphic Use of Circles and Bars I: A Discussion of Eells' Experiment," *Journal of the American Statistical Association*, 22, 31–36.