Int. J. Man-Machine Studies (1986) **25**, 491-500

An experiment in graphical perception

WILLIAM S. CLEVELAND AND ROBERT MCGILL
AT & T Bell Laboratories Murray Hill, New Jersey 07974, U.S.A.

(*Received 22 January 1986 and in revised form 7 August 1986*)

Graphical perception is the visual decoding of categorical and quantitative information from a graph. Increasing our basic understanding of graphical perception will allow us to make graphs that convey quantitative information to viewers with more accuracy and efficiency. This paper describes an experiment that was conducted to investigate the accuracy of six basic judgments of graphical perception. Two types of position judgments were found to be the most accurate, length judgments were second, angle and slope judgments were third, and area judgments were last. Distance between judged objects was found to be a factor in the accuracy of the basic judgments.

1. Introduction

When a graph is made, information is *encoded* on the graph. When a graph is studied, the information is *decoded* by the viewer's visual system. *Graphical perception* is the term we use to describe this visual decoding process (Cleveland & McGill, 1984; Cleveland, 1985). This paper reports the details of an experiment to study graphical perception.

Graphical data display has always played an important role in communicating data in science, engineering, business, and the mass media. The computer graphics revolution has provided low-cost, widely available hardware and software for making graphs; this has caused the amount of data graphing to increase at a tremendous rate. But despite the long tradition of graphing data and its recent increase, there has been relatively little study of how the quantitative information on a graph is visually decoded by the human visual system. The goal in running experiments in graphical perception is to learn how to enhance the decoding process and make it more accurate and efficient.

2. Basic judgments

Table 1 shows six basic judgments that people make to extract *quantitative* information from graphs.

TABLE 1
Basic judgments of graphical perception

(1) Position along a common scale.
(2) Position along identical but non-aligned scales.
(3) Length.
(4) Angle.
(5) Slope.
(6) Area.

0020-7373/86/050491+11$03.00/0

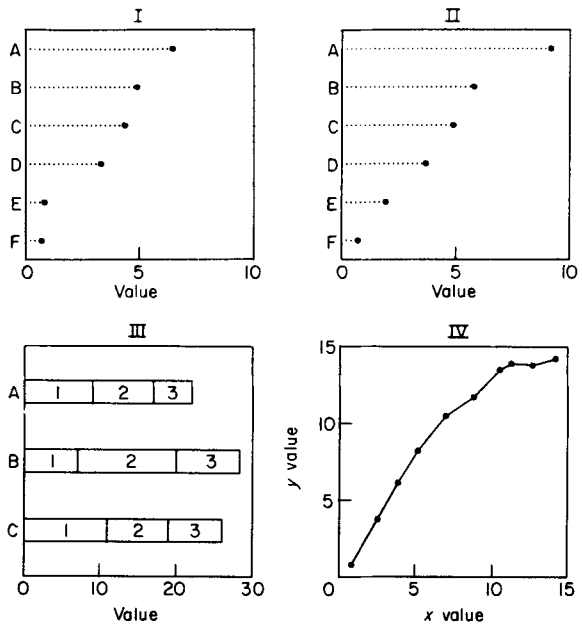

FIG. 1. Basic graphical judgments. We employ certain basic graphical judgments of attributes of geometric objects to visually decode quantitative information from graphs. Panels I and II require position judgments, Panel II requires length and position judgments, and Panel IV requires slope judgments.

Figure 1 illustrates the judgments. Panels I and II are dot charts. To compare visually the values in Panel I we make judgments of positions along a common scale. To compare a value in Panel I with a value in Panel II we make judgments of positions on identical but non-aligned scales. Panel III is a divided bar chart. To compare the three values of Item 1 or to compare the totals of groups A, B, and C we can judge positions along a common scale, but to compare any other set of values—for example, the values of Group A or the values of Item 2—we must make length judgments. Panel IV is an xy graph; the x values can be visually decoded by judgments of positions along a common scale; the same is true of the y values. But the power of such an xy graph comes, in part, from our ability to study the relationship of x and y, for example, how y changes as a function of x. The *local rate of change* of y as a function of x can be visually decoded by judging the slopes of the line segments connecting successive points; the overall visual impression is that the slope tends to decrease as x increases.

It is important to add several qualifications to our concept of basic judgments of quantitative information. First, by quantitative information we mean measurements on a more or less continuous scale such as the spatial frequency of a sinusoidal grating or the percentage of subjects that detect a stimulus near threshold; excluded from this are variables that take categorical values such as male and female or type of retinal cell.

A second qualification is that we do not argue that the six judgments in Table 1 are independent. For example, judgment of position along a common scale really consists of a collection of length judgments. In isolating the basic judgments we are not attempting to identify elementary particles of graphical perception in the same way that Julesz has identified textons as the basic units of preattentive vision (Julesz, 1981).

Rather, we are attempting to define the geometric aspects of objects that we must directly judge to extract quantitative information from a graph.

A third qualification is that the judgments in Table 1 are not meant to be an exhaustive list. Sometimes graphs require other judgments; for example, graphs made in the mass media sometimes require volume judgments based on perspective drawings of three-dimensional objects (Tufts, 1983). However, we believe the six judgments in Table 1 account for the vast majority of judgments that are made to extract quantitative information from graphs.

3. Design of an experiment in graphical perception

In the experiment, subjects judged seven types of stimuli, copies of which are shown in Fig. 2. On each stimulus the object in the upper left was a standard and subjects judged the sizes of the other three relative to the standard; subjects recorded what

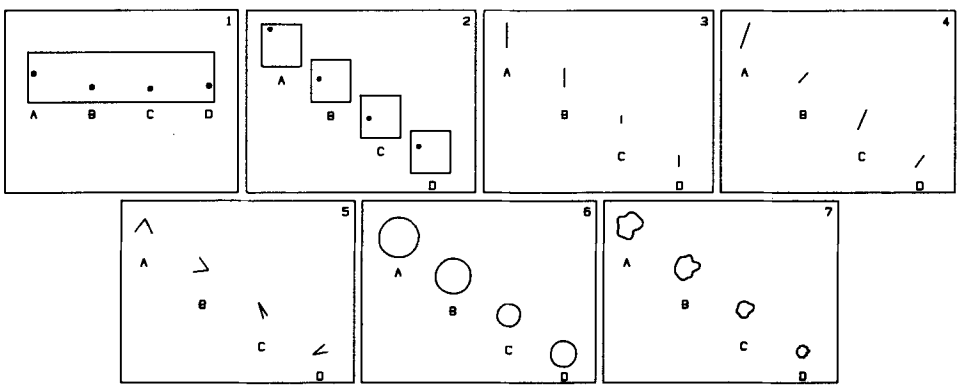

FIG. 2. Stimuli from experiment. An experiment was run to investigate the relative accuracy of basic graphical judgments. The seven types of displays in this figure were judged by subjects. The displays required the following judgments (proceeding from left to right and top to bottom). (1) position along a common scale; (2) position along identical, non-aligned scales; (3) length, (4) slope; (5) angle; (6) area; (7) area.

percentage each of the three sizes was of the size of the standard. The second column of Table 2 shows the seven attributes that were judged; the numbers in the first column correspond to the numbers in the legend of Fig. 2. The seven geometric objects are familiar ones except for blobs, which are irregular regions whose boundaries are given by trigonometric polynomials. The basic graphical judgment employed for each of the attribute judgments is shown in the third column of Table 2.

Subjects performed each of the seven basic graphical judgments 30 times; for half, the standard was "small" and for the other half it was "large." The fourth and fifth columns of Table 2 give information about the standards. The 15 values of the percentage that subjects judged for each half were equally spaced from 17.5% to 87.5%. Thus, subjects made $2 \times 7 \times 15 = 210$ judgments (two standards, seven types of judgment, and 15 true values). There were 70 visual displays like the ones in Fig. 2, and three judgments were made per display. The 70 displays were presented on 70 sheets of paper, each $8{\cdot}5 \times 11$ in; each display nearly filled the sheet.

TABLE 2
Seven types of stimuli used in the experiment

No.	Geometric aspect	Judgment	Large standard	Small standard
1	Distances of dots from baseline	Position along a common scale	4·4 cm	2·9 cm
2	Distances of dots from baselines	Position along identical, non-aligned scales	4·4 cm	2·9 cm
3	Lengths of lines	Length	3·8 cm	2·5 cm
4	Angles between lines	Angle	2·62 rad (150°)	1·22 rad (70°)
5	Slopes of lines	Slope	4·00	2·67
6	Areas of circles	Area	11·4 cm^2	5·1 cm^2
7	Areas of blobs	Area	11·4 cm^2	5·1 cm^2

We selected the order of the dispalys so that the type-of-judgment factor or the size-of-standard factor would not be confounded with time. The order of the 70 displays consisted of five randomized blocks of length 14; within each block each of the 14 combinations of type of judgment and type of standard appeared once. The true percentages being judged were randomly allocated to the displays

In the experiment there were 127 subjects from three groups: 24 high school students, 60 college students, and 43 technically trained professionals. Subjects were given written instructions which were then repeated verbally, with a different wording, by the experiment monitor. Subjects then practiced with seven displays, one for each of the seven types of judgments, and then judged the 70 displays of the experiment.

4. Analysis of experimental data

Our purpose in this section is to present an analysis of the data from the experiment and to compare the results with two earlier experiments in graphical perception (Cleveland & McGill, 1984).

SUBJECT ERRORS

For each subject we computed the average of the subject's absolute errors, the 210 values of |judged percentage − true percentage|. Figure 3 compares the absolute errors for the three groups of subjects by *box plots* (Tukey, 1977). (Appendix I describes the details of construction of box plots.) The figure shows that the overall performance of the three groups of subjects is nearly the same. The two subjects with the largest average errors had errors that were considerably larger than those of everyone else; an examination of their responses led us to believe that they did not understand the instructions, so we eliminated them from the ensuing analysis.

A MODEL FOR STIMULUS ERRORS

For each of the 210 experimental units—i.e. the judged objects—we have 125 values of the subjects' absolute errors. To summarize this distribution of 125 numbers we computed the average of all values between the 50th and 95th percentiles. Our summary statistic describes the upper tail of the distribution and will be called a trimmed upper

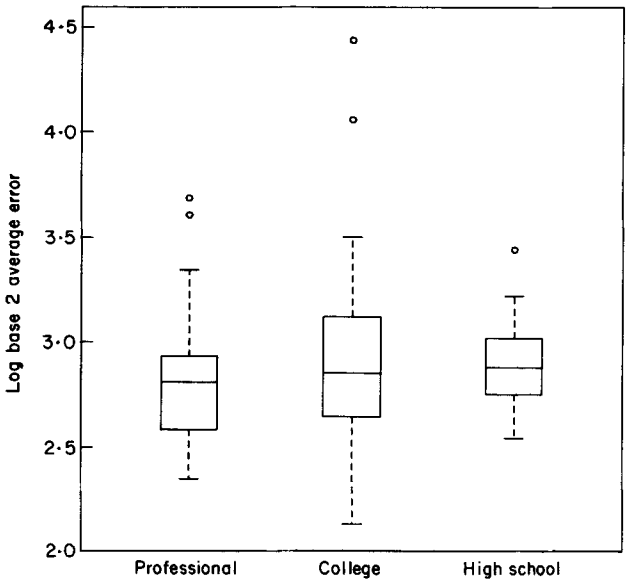

FIG. 3. Box plots. The data are log average errors for three groups of subjects in Experiment 3. Each box plot summarizes the distribution of one group; for example, the horizontal line segment inside the box is the 50th percentile and the top sides are the 25th and 75th percentiles. The box plots show that the distributions of the log average errors are very similar for the three groups, which implies that amount of technical training and experience is not a factor in the accuracy of the basic graphical judgments.

mean. [If we use ordinary means or 10% trimmed means (Tukey & Mosteller, 1977) our conclusions are unchanged].

For each of the seven basic judgments in the experiment the 30 trimmed upper means, y_1 to y_{30}, were modeled by:

$$y_i = \beta_0 + \beta_1 p_1(t_i) + \beta_2 p_2(t_i) + \alpha_i + \gamma_i + e_i$$

where:

t_i = true percentage being judged;

$$\alpha_i = \begin{cases} \alpha & \text{if standard is large} \\ 0 & \text{if standard is small} \end{cases}$$

$$\gamma_i = \begin{cases} 0 & \text{for the first position (closest to the standard) and the second position} \\ \gamma & \text{for the third position} \end{cases}$$

$p_1(x) = x - c_1$ where c_1 is the integral of x from 17.5 to 87.5

$p_2(x) = x^2 - c_2$ where c_2 is the integral of x^2 from 17.5 to 87.5

e_i = independent normal errors with mean 0 and variance σ^2.

The terms involving β_k prescribe a quadratic dependence of the upper trimmed mean on the true percent; this had been suggested from previous experiments (Cleveland & McGill, 1984) and exploratory analyses of the data from this experiments. The reason for parameterizing the polynomial in this particular way (i.e. subtracting c_i) is

to give β_0 a useful interpretation: the average error, averaged from 17·5% to 87·5%. The α_i allow for a dependence on the size of the standard. The γ_i allow for a dependence on the distance of the judged geometrical object from the standard. There were three distances—65-, 130-, and 195 mm—but since exploratory analyses suggested there was no difference between 65- and 130 mm, these two distances were grouped together. The unknown parameters—$\beta_0, \beta_1, \beta_2, \alpha, \gamma$ and σ^2—were estimated from the data using least squares.

We used a number of diagnostic procedures (Belsley, Kuh & Welsch, 1980; Chambers, Cleveland, Kleiner & Tukey, 1983; Daniel & Wood, 1980), many of them graphical, to check both the adequacy of the fitted model and the assumption of normality of the errors. The only problem we uncovered was that for the blob and circle judgments there were time trends; as the experiment progressed, subjects' absolute errors increased, presumably because these two judgments were the most difficult. To estimate the time trend for each type of judgment we computed the fitted errors:

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 p_1(t_i) - \hat{\beta}_2 p_2(t_i) - \hat{\alpha}_i - \hat{\gamma}_i \quad \text{for } i = \text{to } 30,$$

where the Greek letters with hats are the least squares estimates. Figure 4 is a graph of the 60 fitted errors for circles and blobs against the order number in the experiment.

The curve in Fig. 4 is the result of a method called *robust locally weighted regression* (Cleveland, 1985), often abbreviated to *lowess*. (The procedure is described in Appendix II; a listing of FORTRAN routines that carry out the procedure is available from the authors upon request). The fitted values of lowess, the y coordinates of the curve in

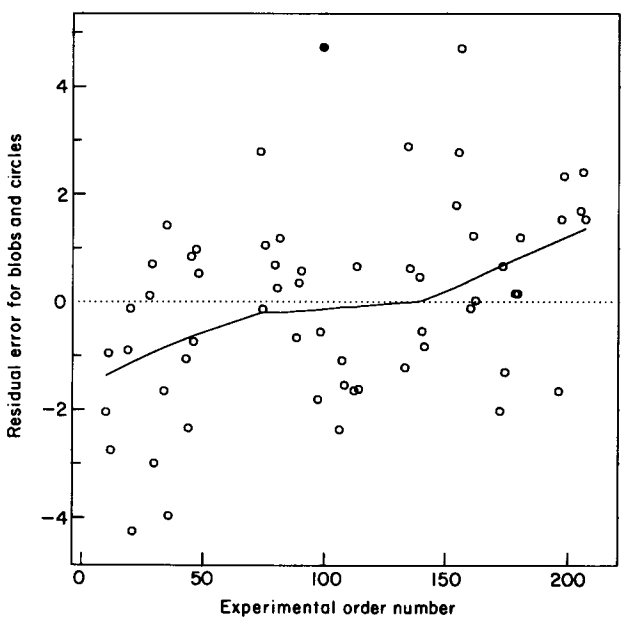

FIG. 4. Lowess. The data on the vertical scale are residuals from a model fitted to the circle and blob absolute error data from the experiment; the residuals are graphed against the experimental order numbers of the blob and circle judgments. The curve, which is the result of a data smoothing procedure called *lowess*, shows that there is an increasing trend throughout the course of the experiment in the magnitude of the errors.

Fig. 4, were subtracted from the y_i for blobs and circles, and the parameters were re-estimated using these trend-adjusted data; in fact, the new estimates differ only by small amounts from the initial ones because we had been careful in the design of our experiment not to allow time to be confounded with the other factors.

We will compare the results of this analysis with the results of a similar analysis of two earlier experiments (Cleveland & McGill, 1984), Experiments 1 and 2. (The experiment that was just discussed will be referred to as Experiment 3.) In Experiment 1, angle judgments and position judgments were compared. Ten sets of five numbers that added to 100 were generated and each set was encoded by a bar chart and a pie chart. To decode the values, subjects had to make position judgments for the bar charts and angle judgments for the pie charts. 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. Models similar to the one above were fit to the 40 upper trimmed means for the angle judgments and to the 40 upper trimmed means for the position judgments.

In Experiment 2, length and position judgments were compared by having subjects judge values encoded on divided bar charts. As in the other two experiments, subjects judged what percentage a smaller value was of a larger value. Ten percentages were judged for each of three types of position judgments; the three types varied in the distances between judged objects. The same 10 percentages were judged for each of two types of length judgments; the two types differed in the placement of the judged objects. Models similar to the one above were fit to the 30 upper trimmed means for the position judgments and to the 20 upper trimmed means for the length judgments.

Figure 5 shows the fitted polynomials for the three experiments; their domain is 17·5% to 87·5%, since this was the range of values covered by Experiment 3. (The ranges for the other two experiments were somewhat larger.) The fitted polynomials for Experiment 3 are for a small standard and the two closest positions; for Experiments 1 and 2 the polynomials are for a standard size and position that match "small" and "closest" in Experiment 3.

5. Discussion

Figure 3 shows that the three error distributions of the three groups of subjects are similar. Thus subject performance does not appear to depend on level of technical training. This was also found in Experiments 1 and 2 (Cleveland & McGill, 1984) and is not surprising; the visual tasks we are investigating are very basic judgments that the visual system of a person performs continually in everyday life.

The estimates of α are negative for six of the seven types of judgment but the absolute values are large only for position on a common scale ($-1{\cdot}92 \pm 0{\cdot}67$) and position on non-aligned scales ($-2{\cdot}27 \pm 0{\cdot}43$). The estimates of γ are positive for all types of judgment, so increasing distance decreases accuracy, but the values of $\hat{\gamma}$ are large only for circles ($2{\cdot}62 \pm 0{\cdot}66$) and blobs ($2{\cdot}61 \pm 0{\cdot}77$).

Figure 5 shows a number of interesting properties of the error measures. First, as a function of true percentage, the general pattern is for errors to be smallest at the extremes, which is not surprising; most subjects would judge true percentage of 0 or 100 with no error and would judge percentages very close to 0 or 100 with very high accuracy. Somewhat surprisingly, however, the positions of the curve maxima are not

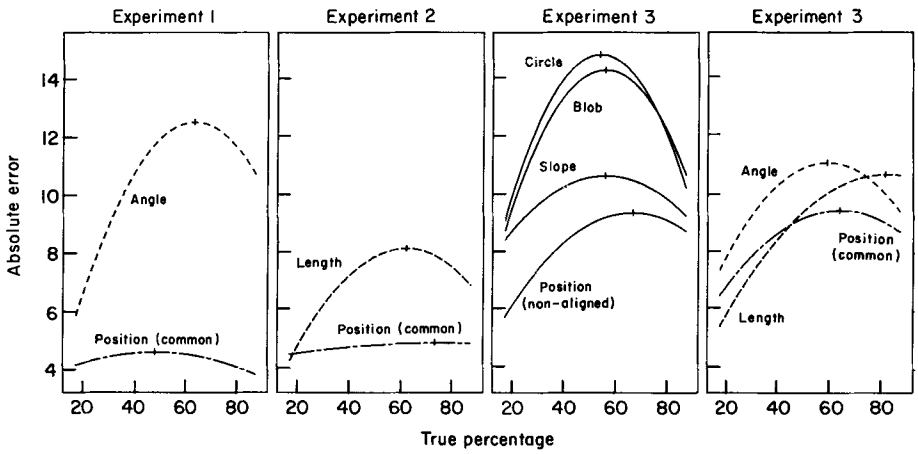

FIG. 5. Absolute judgment error as a function of true percentage. The dependence of the absolute error of judgments was modeled as a quadratic function of the true percentage for each of the basic graphical judgments in three experiments. The coefficients of the polynomials were estimated by least squares. The figure shows the fitted quadratics with the maxima marked by the short vertical lines. Position judgments are the most accurate, length judgments are second, angle and slope judgments are third, and area judgments are last.

clustered at 50% but rather at a higher value. In Fig. 5 the short vertical lines through the curves show the positions of the curve maxima. In Fig. 6, the positions of the curve maxima are shown by a dot chart. All but one of the percentages at which the maxima occur is greater than 50%; the median of these 11 percentages is 61·2%. One might expect that the error at p% might be the same as the error at $(100-p)$%, but there is a consistent deviation from this symmetric behavior.

Figure 5 also reveals some consistency in the behavior of the errors for the different types of judgments. The two position judgments are generally the most accurate within an experiment. Length is next. Angle and slope are comparable and appear less accurate

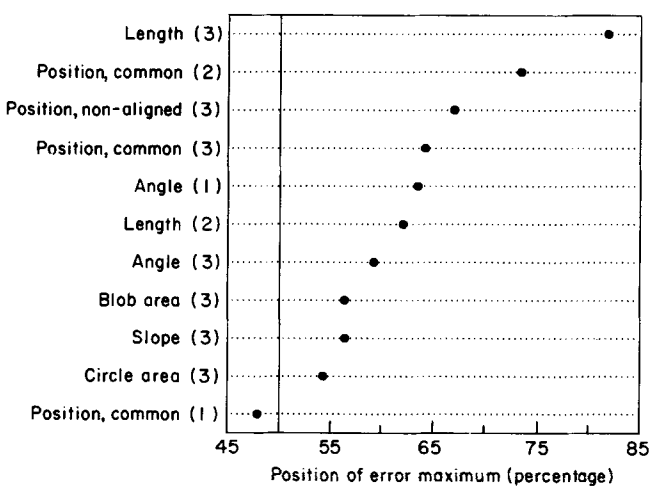

FIG. 6. Positions of maxima. The positions of the maxima of the curves in Fig. 5 are shown by a dot chart. The distribution of the percentage is not centered at 50%, as one might expect, but rather at a higher value.

than the length and position judgments. Finally, the area judgments are the least accurate.

Croxton & Stein (1932) ran an experiment in which comparison judgments were made of lengths, areas, and volumes; percentages ranging from 2 to 90 were judged. Accuracy as a function of judged percentage followed a quadratic behavior similar to that of the curves in Fig. 5. Furthermore, length judgments were most accurate, area judgments were next, and volume judgments were last. Thus, overall, the three experiments presented here and the Croxton-Stein experiment are in general agreement with respect to the dependence of accuracy on the basic judgment employed and the true percent being judged. [Other experiments in graphical perception also have been run; a number of them are reviewed by Cleveland, Harris & McGill (1983).] However, these experiments have tended to compare whole graph forms and do not focus on basic judgments in a way that adds understanding to the issues investigated here].

It is important to add one qualification about the slope judgments in Experiment 3. The error of a slope judgment will depend on its value. In Experiment 3, slopes ranged from 4·00 to 0·47. This means that the angles of the line segments whose slopes were judged ranged from 1·33 rad (76°) to 0·44 rad (25°). Thus the line segments were not close to 0° or 90°, and we must make the qualification that the slope accuracy measures from Experiment 3 are valid for slopes not close to 0° or 90°. The reason for this qualification is that it is quite clear that the accuracy of *relative* slope judgments must completely degrade as the angle of the line segment approaches 0° to 90°. Suppose we have two variable line segments with positive slopes and with acute angles θ and $\theta + \phi$ with the horizontal, where $\phi \geq 0$. Suppose the ratio of the slopes is fixed:

$$r = \frac{\arctan(\theta + \phi)}{\arctan(\theta)}.$$

A simple proof shows that as θ tends to 0° or 90°, ϕ must tend to 0. Thus as the segments get close to 0° or 90°, we lose all ability to judge r. A similar argument has been used by Marr (1982) and Stevens (1981) to argue that judgments of slant and tilt of three-dimensional objects are based on angle and not slope.

In our experiments, subjects compared two magnitudes, m_1 and m_2. In the previous paragraph we have argued that for slopes, judgments of $r = m_1/m_2$ depend not only on r but also on m_1 and m_2. Angle judgments are also likely to depend to some extent on m_1 and m_2 in so far as angles of 90° and 180° can be made with very high accuracy; in our experiment we avoided these special angles. Current evidence, however, is that position, length, and area judgments are unlikely to depend on m_1 and m_2 for a fixed r; that is, overall size is unlikely to be a factor until values are so small that they cannot be readily seen. Weber's Law (Baird & Noma, 1978), which says that detection of a difference between m_1 and m_2 is independent of overall size, suggests, although does not prove this.

As the discussion has so far indicated, the factors for which we controlled in the experiment were type of judgment, distance, judged percentage, and the values of m_1 and m_2 for angle and slope judgments. Another factor that might affect judgments is the surrounding visual stimuli. We do not know if this is a major factor in the judgment of actual graphs, but to minimize any possible effect we did not have subjects judge parts of actual graphs, but rather, as shown in Fig. 2, used stimuli that contained little else except the geometric objects being judged.

6. Conclusions

The main goal of our experimentation has been to understand the relative accuracy of the basic judgments shown in Table 1. Figure 5 shows how accuracy varies according to the basic judgment employed and the true percentage being judged. The two position judgments are the most accurate, length judgments are second, angle and slope judgments are third, and area judgments are last. The experiments also showed that accuracy decreases as the distance between judged objects increases, but that level of technical training of subjects does not affect accuracy.

These experiments in graphical perception were carried out to help improve data display. By our choice of the graphical methods we use to show data, we can control, to some extent, the basic judgments that are required to decode quantitative information. Choosing methods that involve accurate basic judgments leads to more effective data display (Cleveland, 1985; Cleveland & McGill, 1985). We do not mean to imply that the direct goal of a graph is to show the data to as many decimal places as possible; the direct goal is to see patterns in the data and understand the overall behavior, but typically the more accurately the data are visually decoded, the better our chance to detect and understand properly the patterns and behavior of the data.

We are grateful to two referees for helpful comments on the manuscript.

References

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

BELSLEY, D. A., KUH, E. & WELSCH, R. E. (1980). *Regression Diagnostics.* New York: Wiley.

CHAMBERS, J. M., CLEVELAND, W. S., KLEINER, B. & TUKEY, P. A. (1983). *Graphical Methods for Data Analysis.* Monterey, California: Wadsworth Advanced Books and Software.

CLEVELAND, W. S. (1985). *The Elements of Graphing Data.* Monterey, California: Wadsworth Advanced Books and Software.

CLEVELAND, W. S., HARRIS, C. & MCGILL, R. (1983). Experiments on quantitative judgments of graphs and maps. *Bell System Technical Journal* **62**, 1659–1674.

CLEVELAND, WILLIAM S. & MCGILL, R. (1984). Graphical perception: theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, **79**, 531–554.

CLEVELAND, W. S. & MCGILL, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, **229**, 828–833.

DANIEL, C. & WOOD, F. S. (1980). *Fitting Equations to Data.* New York: Wiley.

JULESZ, B. (1981). Textons, the elements of texture perception and their interactions. *Nature (London)*, **290**, 91–97.

MARR, D. (1982). *Vision.* San Francisco: W. H. Freeman.

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

STEVENS, K. A. (1981). The visual interpretation of surface contours. *Artificial Intelligence* **17**, 47–73.

STONE, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5**, 595–620.

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

TUKEY, J. W. (1977). *Exploratory Data Analysis.* Reading, Massachusetts: Addison-Wesley.