Journal of Experimental Psychology: Applied
1998, Vol. 4, No. 2, 119–138

1076-898X/98/$3.00

Reading Bar Graphs: Effects of Extraneous Depth Cues and Graphical Context

Jeff Zacks, Ellen Levy, and Barbara Tversky
Stanford University

Diane J. Schiano
Interval Research Corporation

Manipulating the way a graph is drawn influences viewers' ability to extract information from it. In a series of experiments with simple bar graphs, the authors varied the rendering characteristics and relative heights of the bars and asked participants to estimate the quantities portrayed. The addition of 3-dimensional (3D) perspective depth cues lowered accuracy. This accuracy disadvantage diminished when a short delay was introduced before judgments were reported. The height of the judged bar relative to nearby graphical elements also affected accuracy; this effect was about 1 order of magnitude larger and remained intact when the delay was introduced. Nearby elements also affected viewers' bias (under- or overestimation). These effects do not seem to be due to misestimation of object depth. The results suggest that warnings about accuracy decrements due to 3D shading may be overstated, whereas distortions due to neighboring elements should be of more concern.

Recent advances in computing and printing technologies allow one to produce a dizzying array of different kinds of graphs—and have had a great impact on the kinds of graphs we see in newspapers, magazines, and technical journals. In each particular situation, how can one make a reasoned choice among all the possibilities? There is every indication that these choices matter: The visual characteristics of graphs affect the speed, accuracy, and difficulty of information extraction. They also affect memory for the appearance of graphs and for the information they convey (e.g., Cleveland, 1985; Gattis & Holyoak, 1996; Shah & Carpenter, 1995; Tversky & Schiano, 1989). Therefore, the psychological study of the effects of different rendering techniques seems particularly timely.

Studying graphs also provides an elegant means to study quantitative aspects of perception and conceptual inference. By manipulating features of graph rendering, we can learn about how the visual system combines depth cues and how visual elements interact in forming magnitude judgments.

Three-Dimensional Renderings of Two-Dimensional Data

Modern graphing programs provide the ability to render graphs with the appearance of three dimensions, using perspective cues. In some cases, the third dimension is used to depict a third variable. Research on three-dimensional (3D) data sets has indicated that 3D rendering is important for understanding the full structure of such data sets (Shah & Carpenter, 1995; Wickens, Merwin, & Lin, 1994). (Throughout this article, we will use *3D* to refer to the addition of perspective cues to give the impression of depth

Jeff Zacks, Ellen Levy, and Barbara Tversky, Psychology Department, Stanford University; Diane J. Schiano, Interval Research Corporation, Palo Alto, California.

This work was supported in part by Interval Research Corporation, and by National Science Foundation graduate fellowships. The authors would like to thank Gwo-Ing Lee for her assistance.

Correspondence concerning this article should be addressed to Jeff Zacks, Psychology Department, Stanford University, Stanford, California 94305-2130. Electronic mail may be sent to zacks@psych.stanford.edu.

and *two-dimensional* [*2D*] to refer to the absence of such cues.) However, in what seems to be an increasing number of other cases, the third dimension is not used to convey an additional dimension of the data but rather to enhance the visual appeal of the graphic. Because the addition of the third dimension adds visual complexity without adding information, its use has been decried by many, at least for depicting precise values (Kosslyn, 1985; Tufte, 1983; Wainer, 1984).

Effects of the Addition of Depth Cues

Adding the appearance of a third dimension not only adds extraneous visual clutter, it also adds conflicting depth cues. Pictorial cues such as linear perspective, shading, and occlusion suggest that the figure has a contour in depth, whereas binocular disparity, convergence, and motion parallax all indicate that the figure is flat. Both clutter and the conflict of depth information could have a deleterious effect on graph perception and comprehension. Several competing methods of depth-cue competition have been proposed (Bülthoff & Mallot, 1988; Johnston, Cumming, & Landy, 1994; Landy, Maloney, Johnston, & Young, 1995; Nakayama & Shimojo, 1990; Young, Landy, & Maloney, 1993). It seems likely that, whatever the depth-cue combination algorithm, it will be less accurate in reconstructing the 3D structure of an object when that object is represented by conflicting depth cues. Misperception of the structure of an object in depth (i.e., along the dimension orthogonal to the image plane) can affect not just judgments of the distance or depth of an object but also of its height or width. For example, given two objects that subtend the same vertical visual angle, the one that is seen as farther away will be perceived as taller than one seen as closer. This means that both inaccuracy and bias in depth perception lead to distorted estimates of the height of an object.

Considerations of depth-cue combination gives a theoretical grounding to the preferences of designers (Tufte, 1983; Wainer, 1984) and psychologists (Kosslyn, 1993) for area graphs over volume graphs, at least for making relative height judgments at the time of viewing. However, there is a simpler explanation for the presumed deficiencies of 3D graphs: Lower accuracy could be the result of distraction due to the irrelevant added graphical elements. This leaves us with two open questions. First, does adding depth cues to a graph lower viewers' accuracy for reading that graph? Second, if there is lower accuracy, is this lower accuracy due to depth-cue combination or is it simply a result of adding extraneous markings to a figure?

A conclusive answer has not been forthcoming from the few studies that have examined effects of depth cues on accuracy with graph-like stimuli. In one study, Spence (1990) assessed accuracy judgments with seven different graph types and tables. He concluded that the apparent dimensionality of the graphs did not affect observers' accuracy. However, the graph types in this study were not selected systematically. One of the graphs in the 2D group was an unusual elliptical pie chart; errors with this graph type were much larger than for the other 2D graphs. Inspection of Spence's figures suggests that if this graph type had been excluded, a reliable disadvantage for the 3D graphs would have been observed. The choice in this study to omit the rectangular frame that typically surrounds a published graph raises another interesting question of interpretation. Two of us (Tversky & Schiano, 1989) have found that participants' interpretation of a figure as a graph, rather than something else (e.g., a map of a location), led to differences in the perceived orientation of a line in the figure. It could be that the inclusion or exclusion of a frame in these experiments will influence judgments by a similar mechanism.

In a similar experiment, Carswell and her colleagues did find extraneous depth cues to be associated with lower accuracy (Carswell, Frankenberger, & Bernhard, 1991). They used line, bar, and pie charts and created versions of each, both with and without depth shading. However, two aspects of the stimulus design cloud interpretation of their results. First, the pie chart with added depth shading was tilted so that the pie surface appeared as an ellipse, whereas the version without depth shading was drawn as a circle. Second, the pie and bar graph stimuli portrayed the data values as either an area (for the 2D versions) or as a volume (for the 3D versions), whereas the line graph stimuli portrayed the data with a simple line (for the 2D version) or a surface line (for the 3D version). (See the taxonomy provided below and Figure 2 for defini-

tions and examples of simple line and surface graphs.) The surface-line graph in particular is an unusual type of graph, and the line-graph comparison contrasts a surface with a simple line rather than comparing a volume to an area, as do the pie and bar graphs. The effect on accuracy was dominated by a large difference between the 2D and 3D line graphs. Comparing just the bar graphs with and without depth shading or the pie graphs with and without depth shading revealed no significant differences associated with added depth cues.

Taking into consideration the choice of graph types in these two experiments, there seems to be a small negative effect of adding extraneous depth cues on height-judgment accuracy. However, other factors in graph-rendering style (e.g., the choice of tilted pie graphs or surface-line graphs) had a more dramatic effect on observers' judgments.

Effects of the Relationships Between Graphical Elements

It is well known that the relationship between elements in a figure can affect the perception of those elements. The graphical context in which an element occurs can produce distortions in judgments of color, angle, size, and orientation (see, e.g., Goldstein, 1989, especially chap. 7; Howard, 1982). One particularly relevant example of such a distortion is given in Figure 1. This figure demonstrates the parallel lines illusion: When two parallel lines are viewed, the viewer tends to perceive assimilation (the lengths of the lines seem closer than they are) or contrast (the lengths of the lines seem more different than they are), depending on the ratio of the line lengths and the distance between them (Jordan & Schiano, 1986; Schiano, 1986).

The parallel lines figure is a very simple example, but the influence of perceptual assimilation and contrast is presumably at work in more complex figures as well. Bar graphs are examples of richer visual stimuli that encapsulate the key features of the parallel lines illusion. Accordingly, judgments of bar height should be affected by the relative heights of the bars in a figure and by the height of the judged bar or bars relative to the surrounding graphical frame.

This generates a third open question: What is the relative importance of the addition of depth cues compared with the relationships in the data elements? It is important to think of the effects of 3D rendering techniques in the context of the other factors influencing perceptual judgment. Beyond asking whether the addition of extraneous depth cues affects accuracy and bias in judgment, it makes sense to ask whether such effects are likely to be important "in the wild." One way to answer this question is to compare the size of effects due to the addition of conflicting depth cues with the size of effects due to the influence of nearby graphical elements. This allows an assessment of how well the depth-cue combination mechanism performs in the face of noise, relative to ubiquitous distortions due to graphical context.

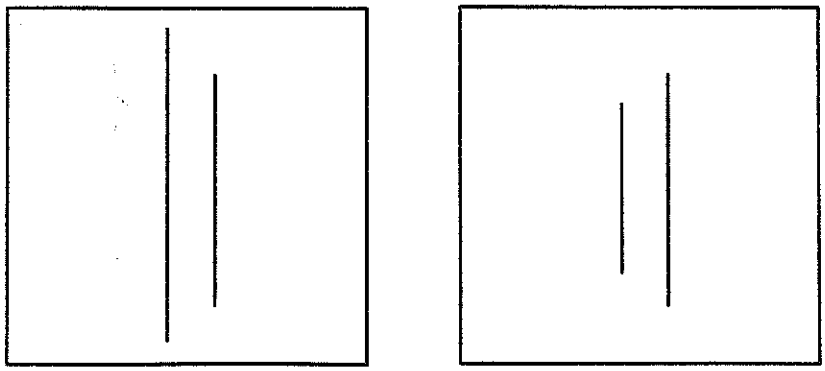


Figure 1. The parallel lines illusion. For most viewers, the right line in Panel A appears longer than the right line in Panel B.

This question is also of practical interest. The relative heights of bars in a graph come from the data being depicted, and the designer of the figure has little control over it. On the other hand, a designer has extensive control over the rendering style of the figure, including control over the inclusion of extraneous depth cues. If both factors have an impact on perception (and perceptual distortion), which is larger—the factor that is controllable or the one that is not?

The Role of Depth Cues in Encoding and Storage

Often, we use graphs to communicate information that is to be used at a later date. However, psychophysical studies of graphical perception have typically examined judgments made while viewing a figure (e.g., Cleveland, Harris, & McGill, 1983; Spence, 1990). This leaves us with a fourth open question: How do effects on perception combine with encoding and storage to influence later judgment?

The experiments reported next were designed to help answer the four open questions described earlier: (a) Does adding depth cues to a graph lower viewers' accuracy for reading that graph? (Experiments 1, 3, and 4); (b) if there is lower accuracy, is this lower accuracy due to depth-cue combination or is it simply a result of adding extraneous markings to a figure? (Experiment 5); (c) what is the relative importance of the addition of depth cues compared with the relationship between the judged data elements and the graphical context? (Experiments 1–5); and (d) how do effects on perception combine with encoding and storage to influence later judgment? (Experiment 2).

Types of Graphs

To allow for systematic comparisons among graph types, let us introduce a brief taxonomy. A good number of graphs can be classified on two dimensions: rendering style and graph type. Rendering style usually takes on one of four possible values. We call graphs that use lines (without shaded areas) to indicate the data values *simple*. Graphs that use the area of a region to depict the data values are called *area* graphs. Graphs that use a drawing of a volume (e.g., a rectangular box) to indicate data values are called *volume* graphs. Graphs that show the data by drawing floating surfaces are called *surface* graphs. Many common graphs take one of two possible values for graph type: *bar* graphs are figures that use an element oriented relative to the independent variable's axis to show each data point, whereas *line* graphs use a line that connects a set of data points. This two-way classification gives rise to eight ($4 \times 2 = 8$) possible kinds of graphs; examples of each are shown in Figure 2. The experiments described next used stimuli whose rendering style was simple, area, or volume; all were bar graphs. (For a more comprehensive taxonomy that combines visual features of the graph with the implicit task of the viewer, see Cleveland & McGill, 1984.)

Experiment 1: Perceptual Match

This experiment was designed to examine the effect of extraneous depth cues on height judgments, to measure the effect of neighboring elements on such judgments, and to compare these two effects. Observers made height judgments while looking at bar graphs. We chose a perceptual-match task for two reasons. First, it encapsulates one important use of data graphics: making quick, reasonably accurate quantitative estimates of data values, without necessarily reading an exact value from an ordinate scale. Second, this task has been productively used to study perceptual illusions (e.g., Jordan & Schiano, 1986).

Across trials, the rendering style of the graphs (area or volume) was varied to investigate the effects on perception of adding depth cues. The height of the test bar and the presence of a constant-height context bar were also varied, allowing for a parametric investigation of effects of graphical context, as in the parallel lines illusion.

Method

Participants. The 40 participants were undergraduate students at Stanford University. Each took part to fulfill a requirement in an introductory psychology course.

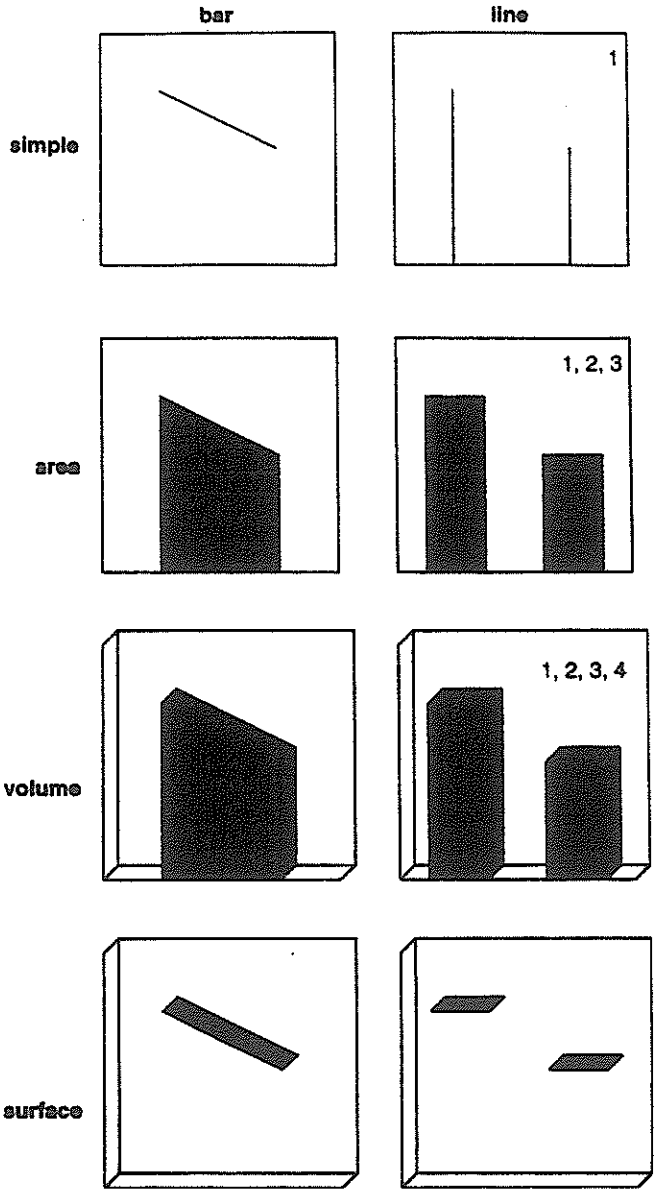


Figure 2. A brief taxonomy of some common graph types. Each row shows a different rendering style; each column shows a different graph type. The numbers in the upper right corners give the experiments (if any) in which a given kind of graph appears.

Stimuli. Two general types of graphs were prepared as stimuli: ones in which two elements were displayed and ones in which a single element was displayed within a graph frame. The two-element graphs consisted of one element of fixed height (context element) of 20 mm and an accompanying "test" element that varied in height from 20 mm to 100 mm by 20-mm increments, in order to create the ratio relationships between the elements of 1:1, 1:2, 1:3, 1:4, and 1:5. The test elements were placed to the right of the context element in the graph. In addition to these two-element graphs, five distractor graphs with noninteger ratio relationships between elements were created by introducing test elements of 16, 26, 46, 74, and 92 mm. These test elements were placed

to the left of a 20-mm context bar in the graph. In both cases, the left edges of the elements were separated by 23 mm. The single-element graphs included only the test elements described earlier without any context element. In these graphs, the test element was placed in the center of the graph frame. In all graphs, the test element was denoted by an asterisk, placed underneath the element and just below the graph frame. The frames were squares 110 mm to a side. See Figure 3 for examples of the stimuli.

All told, there were five "test stimuli with context" graphs, five "test stimuli without context" graphs, and five "distractor" graphs. Two versions of each of these graphs were created, one using area bars (rectangles) and one using volume bars (boxes). The area bars were 7 mm wide. The 3D elements were created by using the corresponding area bar as the face of the box. The perception of three-dimensionality was created with orthogonal-perspective drawing in which 7-mm lines were drawn at 45° angles from the appropriate corners of the bar and then connected.

Two reference charts, an area version and a volume version, were also created to provide the "match" options available to the participants. Each chart depicted a series of elements increasing in height from 12–112 mm by 2-mm increments. Underneath each element was an identifying label (ranging from A to YY). The charts were pasted on two sides of a piece of tagboard.

Booklets. Each graph (2 rendering styles × 3 contexts × 5 ratio relationships) was tested twice, for a total of 60 trials. In addition, six filler pages were included, inserted every nine pages, on which participants were presented with a small data set and asked to draw the graph they felt best represented the information given. Booklets were prepared using one of two random orders of trials, modified only to eliminate the possibility of two trials of the exact same graph type (rendering style, context, and height of test element) occurring consecutively.

In all cases, the graphs were created in the drawing program MacDraw Pro 1.5v2 (Claris Corporation, 1992), and each was printed on a separate 5.5″ × 8.5″ piece of paper.

Procedure. Participants were given a booklet and told that they would be making perceptual judgments about elements in the graphs they were about to see. They were instructed to rely

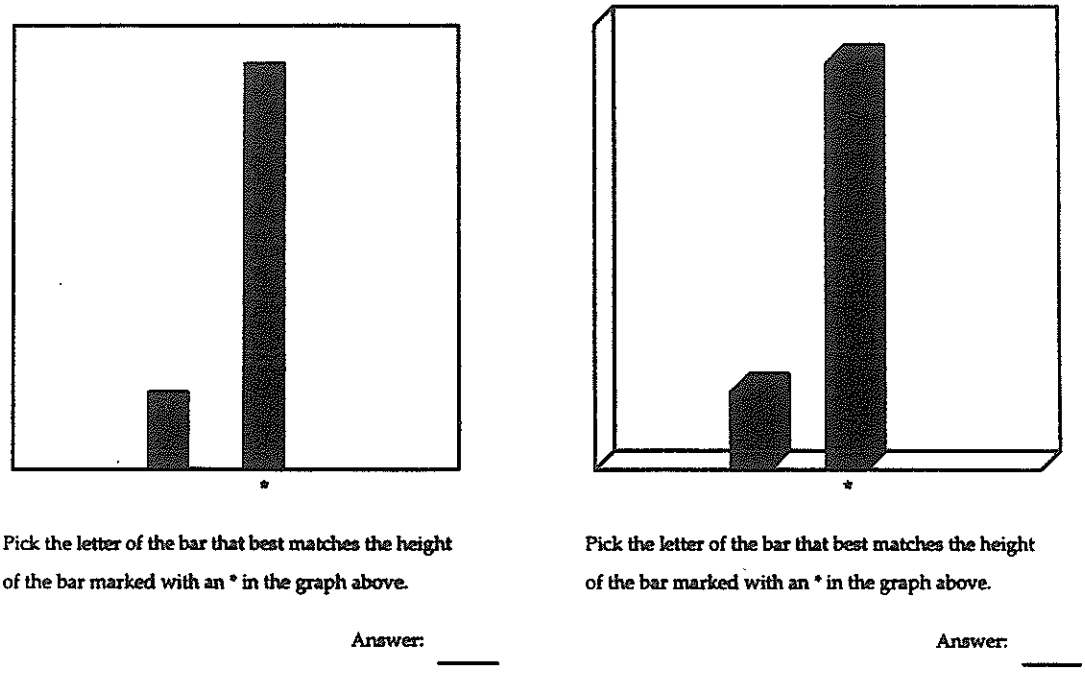


Figure 3. Examples of the perceptual-match stimuli used in Experiments 1 and 2. The left panel shows the area rendering style; the right panel shows the volume rendering style.

solely on their visual perception and not to use other means of assessing the goodness of match (e.g., using their fingers to map the height of the element onto the reference chart). The booklet was placed in a cardboard box with one end cut out, and participants were asked to take out one page at a time, complete the judgment, and then place the page face down on the side of the table before proceeding to the next page.

For each graph, participants were instructed to refer to the reference chart (either the area or volume version, whichever matched the element type shown in the graph) and to write down the letter of the bar or box that best "matched" the height of the element in the graph with the asterisk underneath.

The experiment was self-paced, with most participants taking roughly 25 min to complete the entire booklet.

Results

Participants estimated the height of bars by picking a match from a sample array. Two error measures were constructed from their judgments: *raw error,* which was the height of the chosen bar in millimeters subtracted from the correct bar height, and *error magnitude,* which was simply the absolute value of the raw error for a given trial. The aggregated raw errors show any systematic bias in participants' perceptions of the bar heights, whereas the error magnitudes describe how accurate the judgments were.

Bar height judgments were less accurate for the volume graphs than for the area graphs. The mean error magnitude for the area graphs was 4.10 mm ($SEM = 0.150$), whereas for the volume graphs it was 4.62 mm ($SEM = 0.152$). This difference, although small (approximately half a millimeter), was statistically reliable, $F(1, 1553) = 7.57, p = .006$.

Participants were less accurate for taller bars than for shorter bars, $F(1, 1553) = 241, p < .001$. For the shortest bars, the mean error magnitude was 1.88 mm, whereas for the largest bars it was 5.97 mm (see Figure 4).

Under these viewing conditions, the presence or absence of a context bar had little or no effect on error magnitude, $F(1, 1553) = 0.106$, $p = .744$.

In this experiment, participants tended to slightly overestimate the height of the bar; the mean raw error was 0.879 mm, $t(1599) = 5.82$, $p < .001$. This overestimation was most pro-

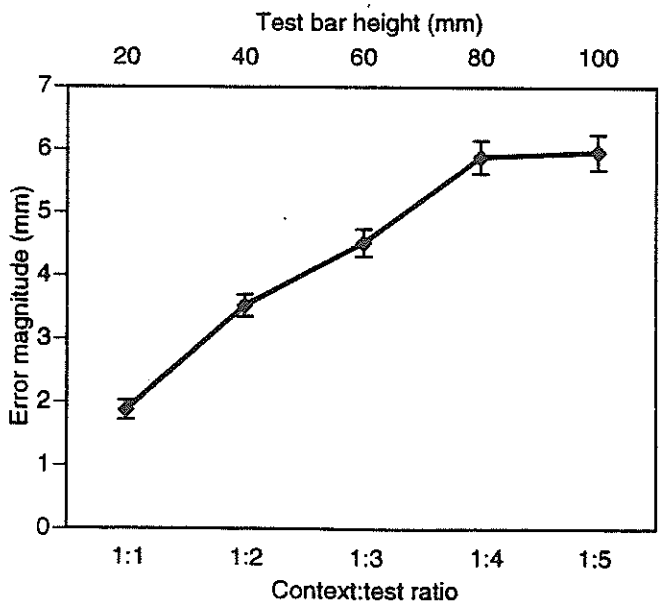


Figure 4. For the perceptual-match task, error magnitude depended on test bar height. Participants were more accurate for shorter bars. The figure shows data from Experiment 1. Plotted points represent mean error magnitude for each test bar height, and error bars show 1 standard error of the mean.

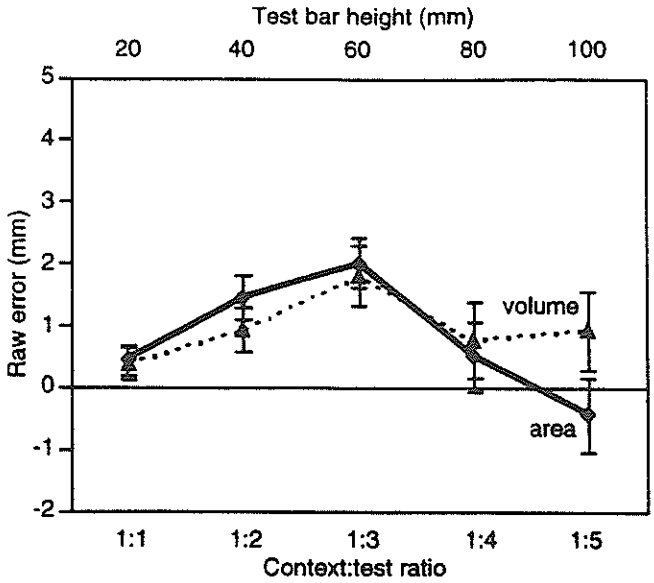


Figure 5. For the perceptual-match task, raw errors for the area and volume graphs did not differ significantly except for the tallest bars. For the tallest bars, area graphs were judged shorter than volume graphs. The figure shows data from Experiment 1. Plotted points represent mean raw error broken down by test bar height, and error bars show 1 standard error of the mean.

nounced for the intermediate-height graphs (see Figure 5). The raw errors were relatively insensitive to the experimental manipulations. There was an interaction between the rendering style (area vs. volume) and the height of the test bar that approached significance, $F(1, 1553) = 3.83$, $p = .051$: Despite the general tendency to overestimate the height of the bar, for the tallest bars this was true only for the volume graphs (see Figure 5).

Discussion

Judgments of bar height were approximately half a millimeter less accurate when 3D depth cues were added to the graphs. This suggests that, as predicted, adding extraneous depth cues does result in lowered accuracy for judgments about the depicted objects. However, neighboring graphical elements also affected judgments: Accuracy depended on the height of the judged bar, and this effect was approximately one order of magnitude larger than the effect of extraneous depth cues. (Another manipulation of neighboring graphical elements—the addition of a context bar—had no reliable effect on accuracy or bias.) There was also a general tendency to overestimate, which was again large relative to the effect of extraneous depth cues. The small relative size of the rendering-style effect suggests that either the visual system's depth-cue combination algorithms are robust in the face of conflicting information or that the effect of depth distortion on perceived height is relatively small. From a practical point of view, it also suggests that we should pause before making strict design recommendations based on the cognitive–visual problems with 3D graphs.

Rendering style, bar height, and the presence of a context bar all had little effect on systematic bias in participants' perceptions of the bar heights. What effects there were might be explained by the role graphical frames (provided by the graph-bounding box and the page) played in generating assimilation and contrast distortions of height judgments.

Experiment 2: Perceptual Match From Memory

Experiment 1 showed that adding extraneous depth cues lowered accuracy for height judg-

ments during perception. Also, in Experiment 1 judgments depended on the relative height of the test element and neighboring graphical elements. Are these effects maintained in memory? To answer this question, we replicated Experiment 1 with one change to the procedure: Observers made their judgments after the test bar had been removed from sight.

Method

Participants. Forty Stanford undergraduates participated in partial fulfillment of a course requirement.

Stimuli and procedure. The design of this experiment was identical to that of Experiment 1, with one modification: Participants were required to turn the page on each graph before making their judgment, thus reporting the bar's height from memory. Booklets were assembled as before, with the addition of a page after each test graph consisting of that graph with the test bar deleted. The participants looked at each graph, then turned to this next page before choosing a bar from the appropriate reference chart.

Results

Under these delayed judgment conditions, participants were overall less accurate than in Experiment 1 (7.73-mm mean error magnitude, compared with 3.40 mm in Experiment 1).

When judging from memory, rendering style had no statistically reliable effect on the error magnitude of height judgments, $F(1, 1553) = 1.59$, $p = .208$ (area mean = 7.51 mm, $SEM = 0.28$ mm; volume mean = 7.94 mm, $SEM = 0.29$ mm). Although the trend was in the same direction, the magnitude of the difference was smaller and the overall variability was larger. (The errors themselves were also larger than in Experiment 1.)

On the other hand, the presence of a context bar, which had no effect on error magnitude in Experiment 1, under these conditions increased error magnitude from a mean of 6.85 mm ($SEM = 0.243$) to 8.61 mm ($SEM = 0.320$), $F(1, 1553) = 27.2$, $p < .001$. There was a significant interaction of rendering style and presence of a context bar, $F(1, 1553) = 4.49$, $p = .034$, such that the effect of context was larger for volume graphs than for area graphs: For area graphs the presence of a context bar increased the mean error by 1.05 mm, from 6.99 mm ($SEM = 0.363$ mm) to 8.04 mm ($SEM = 0.435$ mm), whereas for volume graphs the presence of a context bar increased the mean error by 2.48 mm, from 6.70 mm ($SEM = 0.324$ mm) to 9.18 mm ($SEM = 0.469$ mm).

As in Experiment 1, participants were less accurate with taller bars than with shorter bars, $F(1, 1553) = 453$, $p < .001$. Again, the difference in error magnitude between the shortest and tallest bars was about 10 mm. This effect was more pronounced for graphs in which there was a context bar than those in which there was no context bar, $F(1, 1553) = 11.1$, $p < 0.001$. Figure 6 shows this interaction of bar height and context.

Unlike in Experiment 1, participants in this study tended to underestimate the height of the bars, mean raw error = −3.95 mm, $t(1599) = -15.1$, $p < .001$. The underestimation was greater when there was a context bar present than when there was not, $F(1, 1553) = 42.9$, $p < .001$, and was more pronounced as the bars became taller, $F(1, 1553) = 551$, $p < .001$. Also, there was an interaction between the presence of the context bar and the bar height, $F(1, 1553)$ 24.8, $p < .001$. Examination of Figure 7 shows that the perceptual bias for short bars was small and did not differ greatly between the context bar and no-context bar conditions, whereas the perceptual bias for the tall bars was larger and more exaggerated for the condition in which there was a context bar.[1]

As in Experiment 1, rendering the graphs with depth cues had no effect on the raw errors (i.e., it had no systematic biasing effect on participants' judgments), $F(1, 1553) = .486$, $p = .486$.

Discussion

The procedure in the first two experiments differed only on one small point: In this experi-

[1] For readers familiar with the literature on visual illusions, we note that the "no-context" means are an estimate of the point of subjective equality for these stimuli. Thus, the extent of the illusion at each context: test ratio is given by the difference between the "context" mean and the "no-context" mean.

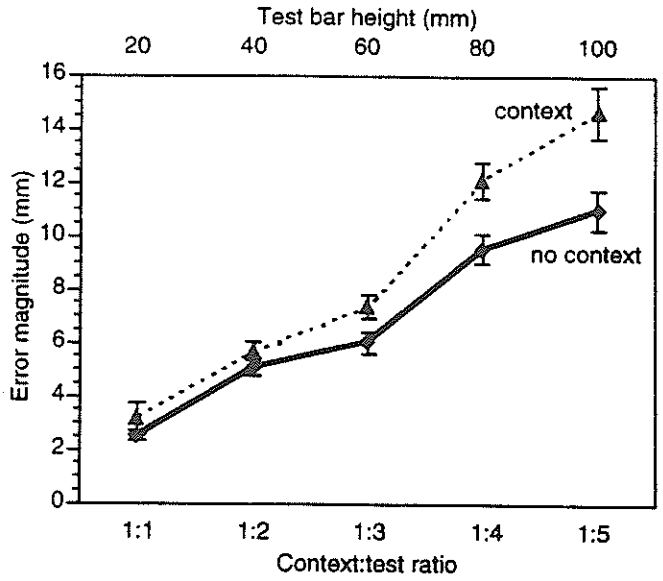


Figure 6. For delayed perceptual-match judgments, error magnitude depended on bar height. Participants were less accurate for taller bars than shorter bars. This effect was more pronounced when a context bar was added. The figure shows data from Experiment 2. Plotted points represent mean error magnitude for each test bar height, and error bars show 1 standard error of the mean.

ment, participants turned the page on a stimulus graph before making a height judgment about it. This small manipulation had two important results. First, it reduced the accuracy advantage for area (2D) graphs to undetectability. (This null result is particularly salient because Experiment 1 provides evidence that the design was powerful enough to detect a relatively small effect.) Second, it allowed a short context bar to exert an influence on participants' height judgments. With a brief delay, the presence of the context bar affected the accuracy (as well as the systematic bias) of participants' judgments. Thus, accuracy was affected by neighboring graphical elements in two ways: Accuracy depended on the height of the judged bar relative to the neighboring elements, and it was lowered by the introduction of a particular nearby element, the context bar.

As in Experiment 1, the most dramatic influence on error magnitudes was simply the height of the test bar. Furthermore, with a brief delay there was a strong linear relationship between the bar height and the raw errors, which was augmented somewhat by the presence of a context bar.

Taken together, Experiments 1 and 2 indicate that adding 3D rendering effects to bar graphs does have a small but significant effect on accuracy in judging the height of the bars. However, this effect is small compared with the effect of other factors, such as the height of the test bar and the presence of another bar next to it. Furthermore, the effect diminishes to statistical undetectability if one simply removes the stimulus from visibility before making a height judgment. Adding depth cues does not appear to systematically bias perception of the height of these stimuli.

In this experiment, there was evidence for increasing underestimation of the bar height as that height increased. Furthermore, as the height of the test bar increased, the presence of a shorter context bar had a growing tendency to lower participants' height judgments. This indicates that as the test bar grew taller, it tended to assimilate more to the height of the shorter bar. It is somewhat surprising that we observed consistent assimilation for delayed judgments but not for the immediate judgments of Experiment 1, which was closer to a procedure that has produced reliable assimilation and contrast with the parallel-lines figure (Jordan & Schiano, 1986; Schiano, 1986).

With a delay, viewers tended to underestimate

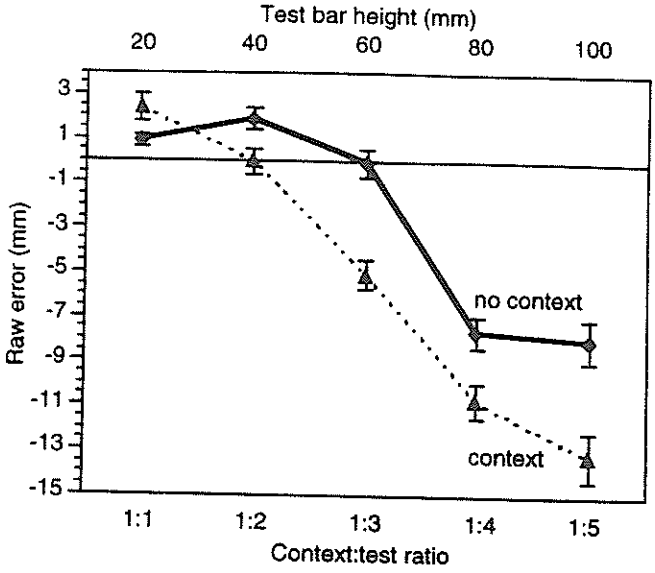


Figure 7. For delayed perceptual-match judgments, raw error depended on bar height. Viewers tended to slightly overestimate the height of short bars and slightly underestimate the height of tall bars. The presence of a context bar amplified this effect. The figure shows data from Experiment 2. Plotted points represent mean raw error broken down by test bar height, and error bars show 1 standard error of the mean.

the height of the bars. In perception, participants tended to overestimate (Experiment 1). Taken together, these results show a general shrinkage in memory similar to that which has been observed with other stimuli (Kerst & Howard, 1978) and has been explained in terms of reperception at the time of memory, transformations to the memory trace, or increased uncertainty leading to regressive estimates (Radvansky, Carlson-Radvansky, & Irwin, 1995).

Experiment 3: Magnitude Estimation

Experiments 1 and 2 demonstrated an effect of rendering style on height judgment accuracy. This effect was smaller than effects due to the height of the judged bar relative to the surrounding frame and held only for immediate perception. Experiment 3 was designed to replicate Experiment 1 using a judgment procedure that differed from that of Experiment 1 in two ways. First, participants reported the height of the bar as a percentage of the context bar, whereas in Experiment 1 they had picked a matching bar from an array of samples. Second, the context bar (which was always present) was always taller than or equal to the height of the test bar, whereas in Experiment 1 the context bar (when present) was shorter than or equal to the height of the test bar. In this experiment, observers gave judgments by expressing the height of a test bar (which varied from trial to trial) as a fraction of a constant-height context bar. (As in the previous experiments, the height of the test bar and the rendering style of the graphs were varied across trials.)

Like the perceptual match procedure, magnitude estimation is recommended by considerations both of ecological validity and comparability to prior work. It distills another common use of data graphics: the comparison of two or more data points to examine the relative size of an effect or quantity. Also, it closely resembles techniques used in other studies of data graphics (Carswell et al., 1991).

Method

Participants. Forty Stanford undergraduates participated in partial fulfillment of a course requirement.

Stimuli. The stimuli consisted of two bars: one context bar at a fixed height of 78 mm and the other a bar varying in height from 1–99% of the context bar by 7% increments. This yielded ratio relationships between elements of 1%, 8%, 15%, . . . , 92%, 99%. These ratios are the same as those used by Carswell et al. (1991). For each ratio relationship, both rendering style and location were varied within participant.

Varying the apparent rendering style of the graphs resulted in three versions of each ratio relationship: a simple bar graph, an area bar graph, and a volume bar graph (see Figure 8 for examples). The simple graph was created by drawing a pair of vertical lines within a rectangular frame to represent the appropriate quantities. The bars in the area graphs were 19 mm wide. The elements in the volume bar graphs were created by using the corresponding area bar as the face of the volume. The perception of three-dimensionality was created with orthogonal perspective: 7-mm lines were drawn at 45° angles from the appropriate corners of the bar and then connected. In all graphs, the test element was denoted by an asterisk, placed underneath the element and just below the graph frame. The left edges of the two elements were 46 mm apart. The elements were drawn in a 92 mm square frame (with added lines for the volume graphs, as shown in Figure 8).

Finally, for each graph of a particular ratio relationship and dimensionality, the location of the test element (right or left) was counterbalanced.

Booklets. Booklets contained 90 pages, each showing a different graph (15 ratios × 3 rendering styles × 2 locations). In all cases, the graphs were created with the graphing program DeltaGraph 2.0 (DeltaPoint, Inc., 1992), edited with MacDraw Pro 1.5v2 (Claris Corporation, 1992), and printed on a separate 5.5 × 8.5″ piece of paper. Graphs were organized in the booklet using one of four random orders of trials, modified only to eliminate the possibility that the same rendering style would occur three consecutive times and to ensure that graphs with the same ratio relationship were separated by at least one trial.

Procedure. Participants were given a booklet and told that they would be making perceptual judgments about the relationship between the elements in the graphs they were about to see. They were instructed to rely solely on their visual perception and not to use other means of assessing the goodness of match (e.g., drawing tick marks on the bars). A booklet was placed face down in a box in front of the participant, who was instructed to look at one page at a time, complete the judgment, and then place the page face down on a separate pile before proceeding to the next page. Participants were instructed to look at the bar marked with an asterisk and to provide an estimate of how tall that bar was relative to the height of a second bar in the graph. Participants were asked to provide estimates in the form of integer percentages rather than in fractions (e.g., 66% rather than ⅔). In addition, the example contained estimates that were not "rounded numbers" (ending in a 5 or a 0), a fact that was pointed out by the experimenter in order to encourage participants to use the full scale and be as accurate as possible.

The experiment was self-paced, with most

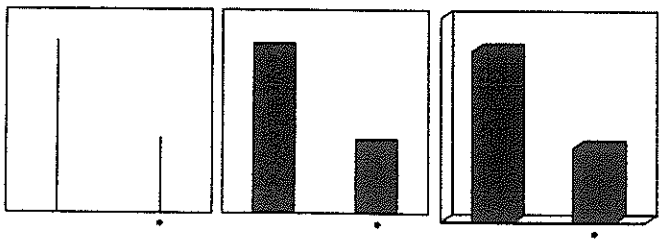


Figure 8. Examples of simple (on the left), area (in the middle), and volume (on the right) bar graphs used in Experiment 3. In each stimulus, an asterisk (*) marked the test bar; participants reported the height of the test bar as a percentage of the height of the context bar (which was a constant height).

participants taking roughly 25 min to complete the entire booklet.

Results

This experiment replicated the manipulations of Experiment 1, using a magnitude estimation procedure rather than a perceptual match procedure. As expected, the results generally reproduced those seen in Experiment 1. Participants' error magnitudes were higher for volume graphs (*M*, 3.60% = 2.81 mm; *SEM*, 0.134% = 0.105 mm) than for area graphs (*M*, 3.17% = 2.47 mm; *SEM*, 0.109% = 0.085 mm), and mean error magnitude for simple graphs was intermediate (*M*, 3.42% = 2.66 mm; *SEM*, 0.116% = 0.090 mm). An analysis of covariance (ANCOVA) showed that the three means differed significantly, $F(2, 3549) = 3.33$, $p = .036$, and a planned comparison showed that the area and volume rendering styles in particular differed by a significant margin, $t(2398) = 2.46$, $p = .014$. Error magnitude also depended on height of the test bar, $F(1, 3549) = 4.01$, $p = .045$. The significant ANCOVA coefficient was due to a small decreasing linear trend, but inspection of the data (see Figure 9A) shows that accuracy was better for the shortest and tallest bars and worse for the intermediate-height bars. The unusually low error magnitude for the 50%-ratio (39 mm) graphs was probably due to participants' tendency to use round numbers (multiples of 5% or 10%) to report the relative bar height. (Apparently, the instructions intended to discourage this rounding were ineffective with some of the participants.)

The difference in error magnitude between the area and volume graphs was .426% (.332 mm) of the height of the context bar, whereas the difference between the bar heights for which participants were most and least accurate was 3.37% (2.63 mm). As in the perceptual-match procedure, the effect of the bar height manipulation on accuracy was about one order of magnitude larger than the effect of adding depth cues.

Under these viewing and judgment conditions, participants tended to slightly overestimate the height of the bars, mean raw error = .596% = .465 mm, $t(3599) = 6.70$, $p < .001$. As Figure 10 shows, this overestimation generally decreased with the height of the bar and was slightly negative for the two tallest bar heights, $F(1, 3549) = 35.0$, $p < .001$. As in the perceptual-match experiments, adding depth cues to the graphs had no effect on the raw errors. The simple, area, and volume graphs were all very similar, $F(2, 3549) = .821$, $p = .440$.

A brief note about effects of which side of the figure the test bar was on: Viewers judged the height of the test bar to be higher when it was on the right than when it was on the left, $F(1, 3549) = 7.71$, $p = .006$. This effect was replicated in Experiments 4 and 5, described later, $F(1, 3459) = 237$, $p < .001$, and $F(1, 4261) = 8.70$, $p = .001$, respectively. This could be due to the order in which viewers tend to scan these figures. Also, in Experiments 3–5 there were other effects and interactions that involved the side of the test bar, which for brevity we will not report here.

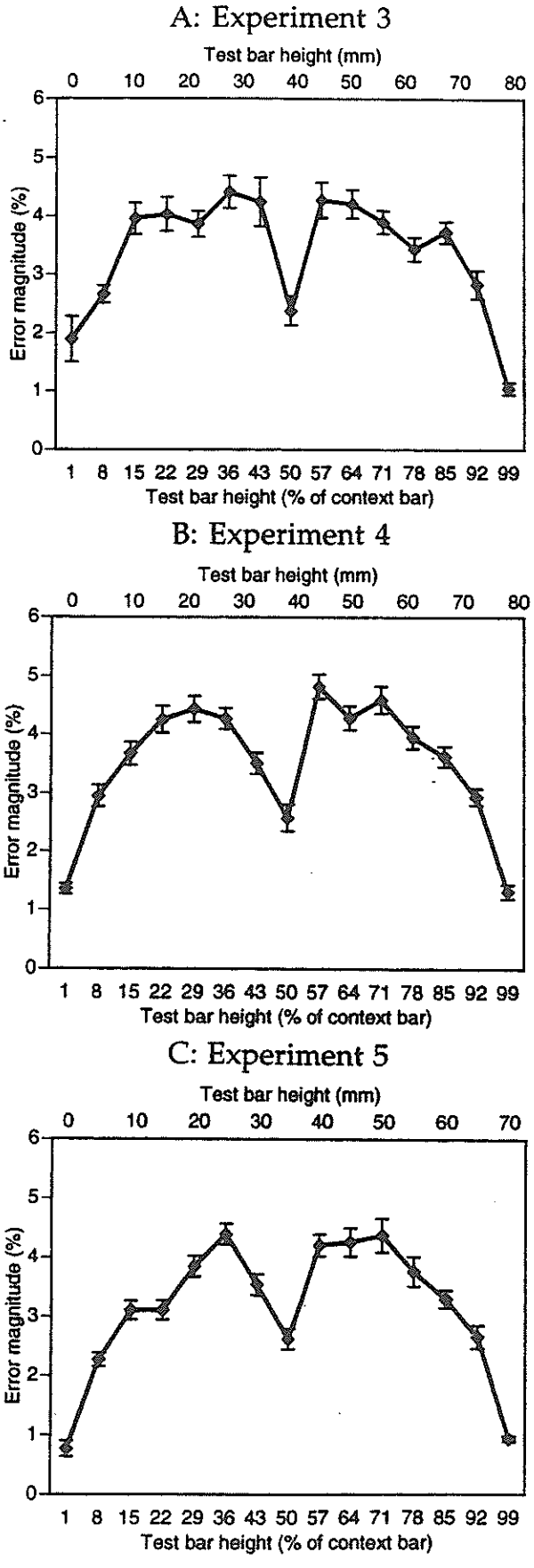
Discussion

Results from the magnitude estimation procedure used here were quite similar to those using the perceptual match procedure. In both Experiment 1 and Experiment 3, adding pictorial depth cues had a small but significant detrimental effect on accuracy. In both cases, the effect of the height of the bar relative to the surrounding graphical elements was about 10 times as powerful as the effect of rendering style. As in Experiment 1, adding depth cues had no systematic effect of lowering or raising participants' estimates of the height of the bars.

The effect of bar height on the raw errors probably reflects assimilation toward the context bar. This is similar to the assimilation seen in Experiment 2. (Note that, unlike the perceptual match procedure, the magnitude estimation procedure produced marked assimilation with no delay.)

Experiment 4: Magnitude Estimation With Parametric Depth Cue Manipulation

If the lower accuracy caused by extraneous depth cues observed in Experiments 1–3 is due to a roughly linear cue-combination process, then



parametrically manipulating the discrepancy between the perspective/shading cues and other depth information should give rise to a roughly linear change in the height judgment error magnitude. Experiment 4 was designed to test this possibility and provide more information about the effects of neighboring elements on height judgments.

Method

Participants. Forty Stanford undergraduates participated in partial fulfillment of a course requirement.

Stimuli and procedure. As in Experiment 3, the stimuli consisted of bar graphs with two elements. The same heights and ratios were used. Booklets were assembled and administered in the same manner.

In this experiment, the three graph types all made use of perspective depth cues, the apparent depth of which were parametrically varied by using diagonal lines of 4, 7, and 10 mm in length to connect the apparent front and back of each bar.

Results and Discussion

Parametric manipulation of extraneous depth cues had no statistically reliable effects on participants' raw errors or error magnitudes in height judgment. There was, however, a nonsignificant trend such that error magnitude went up as the perceived depth was exaggerated. The mean error magnitude for the smallest-depth-cue graph was 3.41% = 2.66 mm (*SEM* 0.088% = 0.069 mm); for the intermediate-depth-cue graph, it was 3.48% = 2.71 mm (*SEM* 0.090% = 0.070 mm);

Figure 9. For the magnitude-estimation task, error magnitude depended on bar height. Participants were more accurate for the tallest and shortest bars. (The unusually small error for the 50% bar may be due to a bias to choose round numbers.) Panel A shows data from Experiment 3, Panel B shows data from Experiment 4, and Panel C shows data from Experiment 5. Plotted points represent mean error magnitude for each test bar height, and error bars show 1 standard error of the mean.

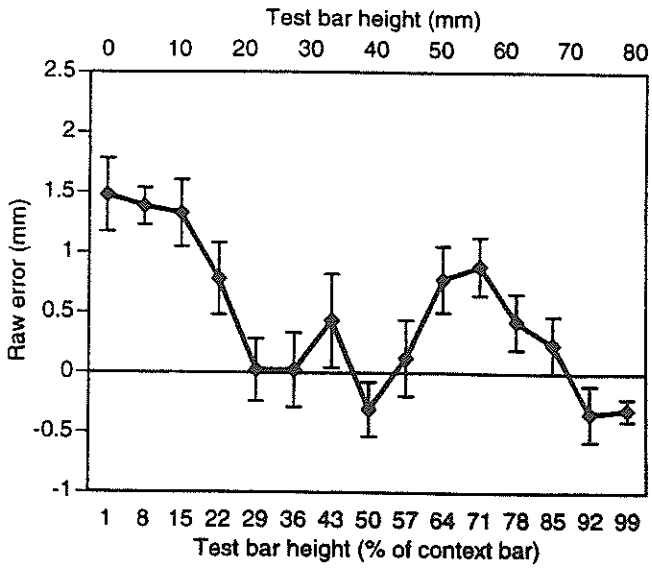


Figure 10. For the magnitude-estimation task, participants tended to overestimate the height of short bars and slightly underestimate the height of the tallest bars. The figure shows data from Experiment 3. Plotted points represent mean raw error for each test bar height, and error bars show 1 standard error of the mean.

and for the largest-depth-cue graph, it was 3.62% = 2.82 mm (*SEM* 0.092% = 0.719 mm).

The effect of bar height on both the raw errors and error magnitudes was quite similar to that seen in Experiment 3. Regarding the error magnitudes, as in Experiment 3, participants were most accurate for the tallest and shortest bars and less accurate for those of intermediate height. (Unlike in Experiment 3, there was no significant linear trend in this effect, $F[1, 3549] = .001, p = .973$.) As in Experiment 3, the 50% (39 mm) bar had an unusually low error magnitude, indicating that participants tended to report bar heights in round numbers. Participants tended to overestimate the height of the bars, mean raw error = .962% (.75 mm), $t(3599) = 12.6, p < .001$. This overestimation increased with the difference between the (taller) context bar and the test bar, $F(1, 3549) = 101, p < .001$. This pattern closely replicates that of Experiment 3 (see Figure 9B) and again suggests assimilation toward the context bar.

To summarize, this experiment replicated the effects seen previously of the test bar height on judgment accuracy and bias. It failed to show an effect of increasing the exaggeration of extraneous depth cues on accuracy, but this may have been due to insufficient power.

Experiment 5: Magnitude Estimation With Depth Cues and Other Extraneous Elements

Experiments 1 and 3 showed that for magnitude judgments at the time of viewing, adding extraneous depth cues lowers accuracy. How? One plausible explanation, described earlier, is that combining conflicting depth cues in particular is a source of error in magnitude judgments. A more parsimonious alternative is that adding any "junk" near the elements of interest induces error, and that this source of error is sufficient to explain the lower accuracy. Experiment 5 was designed to adjudicate between these two explanations, by comparing accuracy for volume graphs with accuracy for graphs that had extraneous elements similar to those in the volume graphs but no conflicting depth cues.

Method

Participants. Forty Stanford University undergraduates participated in partial fulfillment of a course requirement.

Stimuli and procedure. The design of the stimuli and procedure followed that of Experiments 3 and 4.

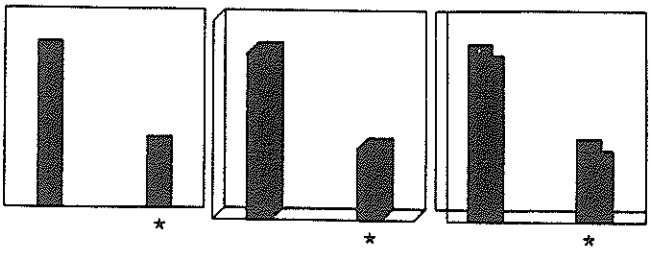


Figure 11. Examples of area (on the left), volume (in the middle), and foil (on the right) bar graphs used in Experiment 5. In each stimulus, an asterisk (*) marked the test bar; participants reported the height of the test bar as a percentage of the height of the context bar (which was a constant height).

This experiment contrasted three different bar types. Area and volume bars were the same as in the previous studies. Also, a new bar type was designed, which will be called a "foil." The foil graphs were created by taking volume graphs and straightening the parallelograms used to draw the sides and top of the box into rectangles. This left a figure with graphic elements similar to those of a volume graph, but with no illusion of depth (see Figure 11). Pre-testing showed that viewers did not see the foil graphs as having extension in depth.

The construction of the stimuli was basically the same as those used in Experiments 3 and 4; some small differences arose because these were prepared with different software (SuperPaint 3.0; Aldus Corporation, 1991). The height of the context bar was 71 mm rather than 78, and the bars were 11 mm wide rather than 19 mm. The frame was an 84 mm square rather than 92 mm. The same ratios of context bar to test bar were used. The length of the diagonals for creating a 3D effect and the spacing of the bars within the frame was the same as in Experiments 3 and 4.

Results and Discussion

Error magnitude varied significantly as a function of rendering type, $F(2, 4261) = 4.81$, $p < .001$. The mean error magnitude for area graphs was 2.93% = 2.08 mm (*SEM* 0.078% = .055 mm); for volume graphs it was 3.24% = 2.301 mm (*SEM* 0.086% = 0.061 mm); and for the foil graphs it was 3.27% = 2.32 mm (*SEM* 0.095% = 0.067 mm). Planned comparisons showed that errors for the foil and volume graphs were significantly higher than for the area graphs, $t(2878) = 2.72$ and 2.66, respectively, $p < .05$ (Bonferroni-corrected for three simultaneous comparisons). The volume and foil graphs did not differ, $t(2787) = .22$, uncorrected $p = .826$. As in Experiment 3, there was a slight but significant linear effect of bar height on accuracy, $F(1, 4261) = 8.44$, $p = .004$. Also, as in Experiments 3 and 4, the errors for the 50% (35.5 mm) bar were especially low (see Figure 9C).

Observers' raw errors depended on rendering style, $F(2, 4261) = 9.59$, $p < .001$. The mean raw error for the area graphs was 0.323% = 0.229 mm (*SEM* 0.109% = 0.078 mm); the mean for volume graphs was −0.056% = −0.040 mm (*SEM* = 0.121% = 0.086 mm); and the mean for foil graphs was −0.379% = −0.269 mm (*SEM* = 0.128% = 0.091 mm). However, only the area and foil graphs differed significantly, $t(2878) = 4.17$, corrected $p < .01$. The difference in raw error for volume and area graphs approached statistical reliability, $t(2878) = 2.32$, corrected $p = .062$, whereas the difference between volume and foil graphs did not, $t(2878) = 1.83$, corrected $p = .201$. Overall, there was no systematic bias in the raw errors, $t(4319) = .540$, $p = .589$.

These results argue against the proposition that depth-cue combination in particular is responsible for the error differences between area and volume graphs. Were that the case, one would expect that the pattern of results for the foil graphs would be quite different from that for the volume graphs. In fact, in both magnitude and bias of error, the volume and foil graphs were quite similar.

As in the previous experiments, the distortions due to rendering style were small compared with those due to the height of the test bar.