APPLIED COGNITIVE PSYCHOLOGY, VOL. 12, 173–190 (1998)

Judging Proportion with Graphs: The Summation Model

J. G. HOLLANDS[1]* and IAN SPENCE[2]

[1]*University of Idaho*
[2]*University of Toronto*

SUMMARY

People take longer to judge part-to-whole relationships with bar graphs than with pie charts or divided bar graphs. Subjects may perform summation operations to establish the whole with bar graphs, which would be unnecessary for other graph types depicting the whole with a single object. To test this *summation model*, the number of components forming the whole was varied with bars, divided bars, reference bars, and pies in three experiments. Response time increased with the number of components for bar graphs but there was little increase for other graph types in Experiment 1. An accuracy emphasis in Experiment 2 produced generally longer response times, but had little effect on the time per summation. The summation operation was not used when graphs were displayed briefly in Experiment 3, although subjects still took longer with bars. The estimated time for a summation operation is consistent with estimates derived from other research. In general, the bar graph is not effective for proportion judgments, and its disadvantage becomes potentially greater as the number of components increases.

Appl. Cognit. Psychol. **12**: 173–190 (1998)

INTRODUCTION

Despite widespread use of statistical graphics and graphical displays, knowledge of how people read graphs is relatively meager. However, researchers have made progress in two related areas of graphical perception: how well people perform with different graph types and what cognitive processes or operations might be involved in reading graphs (e.g., Casner, 1991; Cleveland, 1985, 1990; Gillan and Lewis, 1994; Hollands and Spence, 1992; Lohse, 1993; Pinker, 1990; Shah and Carpenter, 1995; Simkin and Hastie, 1987; Tan and Benbasat, 1990, 1993). A common finding is that different graphs are better for different judgment tasks. To account for this result, some researchers (e.g., Casner, Hollands and Spence, Pinker) have distinguished between relatively direct perceptual judgments (occurring when the demands of a task

*Correspondence to: Justin G. Hollands, Department of Psychology, University of Idaho, Moscow, ID 83844-3043, U.S.A. E-mail: hollands@uidaho.edu

The authors thank Pat Bennett, Brian Dyre, Sallie Gordon, and Mike Kahana for helpful comments, and Faith Lai for help with the data analysis. The first author additionally thanks Lochlan Magee and the Defence and Civil Institute of Environmental Medicine (Canada) for providing writing time and computer equipment during his stay there as a Visiting Fellow.

Contract grant sponsors: Natural Sciences and Engineering Research Council of Canada; Contract grant number: A8351.

CCC 0888–4080/98/020173–18 $17.50

Received 11 September 1996
Accepted 24 June 1997

correspond to the arrangement of data in a graph) and judgments involving a longer sequence of cognitive steps or mental operations (occurring when there is little task-graph correspondence). The anchoring framework of Tan and Benbasat is similar in that better performance occurs when the location of cognitive anchors in the graph corresponds to task demands.

Mental operations in judgments of proportion

Judging part-whole relationships is considered fundamentally important by mathematics educators (Leinhardt, Zaslavsky and Stein, 1990), and is a common task in everyday life. It is not surprising, therefore, that a common graph-reading task requires subjects to estimate proportion (e.g., what percentage is quantity A of the whole?). For example, Hollands and Spence (1992) required adult subjects to make judgments of proportion with bar graphs, divided bar graphs, and pie charts (Figure 1 shows examples of these graph types). In their experiments, subjects shown bar graphs (without a graduated scale) required more time and made larger errors than they did with divided bar graphs and pie charts. Although most bar graphs have a scale, it is often not graduated in percentage units; the individual attempting to estimate a proportion cannot do so by reading a value from the scale.

To account for their results, Hollands and Spence (1992) suggested that subjects sum heights or areas with bar graphs to establish the whole prior to estimating the part-to-whole ratio. Let us consider each of these operations in more detail. The input

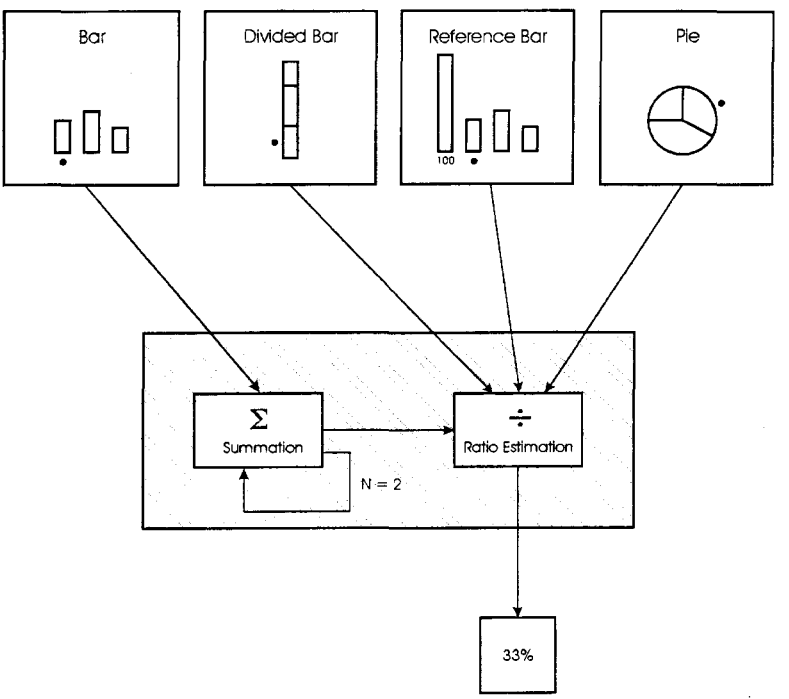

Figure 1. Hypothesized sequences of operations for judging proportions. Each graph type used in the experiments is shown portraying the same data. Since the number of components is 3, the number of proposed summations for the bar graph is $N = 2$

for a *summation operation* might be defined as two perceived or imaged areas (or lengths), and its output should consist of the sum of the areas or lengths. Other researchers have proposed similar summation processes with graphs: Casner (1991) proposed a 'stack heights' operation where the observer sums the heights of two areas placed side by side, and Gillan and Lewis (1994) described an addition process that could be used to establish the sum of values shown in a line graph or scatterplot. The input for a *ratio estimation* operation should consist of two components (imaged areas, lengths, angles, or arc lengths), at least one of which might be produced by a prior summation. The output would consist of the ratio of the smaller to the larger (e.g., 33%). Cleveland (1990), Gillan and Lewis (1994), and Simkin and Hastie (1987) have described similar ratioing operations.

For divided bar graphs and pie charts, we propose that a part may be directly compared to the whole using a ratio estimation operation because a physical object represents the whole in each case. The same would hold true for the *reference bar* graph shown in Figure 1, which is similar to the grouped dot chart devised by Cleveland (1985, p. 261). With conventional bar graphs, however, we propose that subjects perform a series of summation operations to form an estimate of the whole prior to the ratio estimation operation. If summation operations are performed sequentially with bar graphs, it follows that the more bars there are, the longer it should take to sum them. There is evidence for serial assembly of an image: Kosslyn, Reiser, Farah, and Fliegel (1983) showed that subjects took more time to assemble an image as the number of components increased. With the other graph types, there should be no increase in judgment time as the number of components increases.

More formally, total response time can be expressed by the linear equation:

$$T = \tau_S N + \tau_O + e,$$

where τ_S is the time for one summation operation, τ_O is the time for fixed overhead operations (e.g., perceiving the stimuli, estimating the ratio, executing the response), and e is random error. N is the number of summations, which is one less than the number of components in the graph, and thus T should increase linearly with the number of summations for bar graphs. For divided bar graphs, reference bar graphs, and pie charts, N is 0 regardless of the number of components, since no summation operations are required, and thus total time T should equal the time for the overhead operations, τ_O, plus random error. In other words, manipulating the number of components should affect the number of summation operations for bars; with the other graph types, the manipulation should have little effect. Figure 1 illustrates the model. In terms of other more general models (e.g., Casner, 1991; Pinker, 1990), estimating a proportion involves a simple perceptual judgment with divided bars, reference bars, or pies; in contrast, a proportion judgment with bar graphs requires complex cognitive inference. In terms of Tan and Benbasat's (1993) anchoring framework, divided bars, reference bars, and pies have an anchor or reference point representing the whole, whereas bars do not (i.e., the anchor must be constructed mentally).

As noted above, some graphical perception researchers (e.g., Cleveland, 1990; Gillan and Lewis, 1994; Lohse, 1993; Pinker, 1990) have discussed possible operations for graph reading, but did not treat proportion judgments in detail. Gillan and Lewis manipulated the number of additions required in a task, and asked their subjects to estimate the sum, but did not require their subjects to make proportion judgments.

Simkin and Hastie (1987) described several operations that could be used when judging proportion. However, they did not vary the number of components shown in their graphs, and so the effect of that manipulation on the operations used in proportion judgments has not yet been established.

Description of experiments

The goal of this study was to investigate the predictions of the summation model. In Experiment 1, the number of components was varied in four graph types to test the model. In Experiments 2 and 3, factors thought to affect the fixed overhead and summation operations were explored. In Experiment 2, accuracy was emphasized more than in Experiment 1 to determine if the summation operation would be affected by changes in instruction. In addition, the range of the number of components was increased and the size of the graphs was varied randomly across trials. The purpose of varying size was to prevent subjects from using the remembered size of the whole on earlier trials to assist their judgments. In Experiment 3, the display duration was shortened to determine if the summation operation would be disrupted in a speeded context.

EXPERIMENT 1

In Experiment 1, subjects judged proportions with the four different graph types and the number of components was varied. The summation model predicts a linear increase in response time with the number of summations (number of components minus one) for the bar graph, but not for the other graph types. Absolute error scores (the absolute value of the difference between the judged and true proportions) were also obtained.

Method

Subjects

Twenty-four students enrolled in an introductory psychology course at the University of Toronto participated and received course credit for participating.

Materials and apparatus

Ten sets of integers were separately generated for two-, three-, and five-component conditions. Each data set consisted of a selected integer (which would be plotted as the selected or to-be-judged component in the graph) and one or more remainder integers (which would be plotted as other components in the graph). Each selected integer was randomly generated, and was between 5 and 50 with no two selected integers closer than 2. Remainder integers were generated by iteratively obtaining the difference between 100 and the sum of already-generated integers, and then randomly choosing an integer less than this difference, until the appropriate number of integers was generated for the set. No remainder integer was less than 5. The sum of any set of integers was 100.

Each data set was plotted using each graph type shown in Figure 1, producing 120 graphs. As shown in Figure 1, a small circle (colored red on the display) indicated

the selected component. The ordinal position of the selected component was determined by selecting a random number between one and the total number of components. The prompt, 'What proportion is the dotted quantity of the whole?' was displayed below each graph.

The graphs were shown on a 30-cm display (measured diagonally) at 640 × 480 (horizontal by vertical) resolution, controlled by a program running on an IBM personal computer, which also measured response times. All graphs were drawn as white lines on a black background. The different graph types were drawn approximately the same size. The width of each bar was 10 mm, and a given vertical distance represented the same percentage for bars, divided bars, and reference bars (7 mm = 10 percent). The arc length for a pie segment of 10 percent was 19 mm. The radius of all pies was 30 mm. The viewing distance was approximately 50 cm.

Design and procedure

A 4 × 3 (Graph Type by Number of Components) within-subjects factorial design was used. The graph types were bar, divided bar, reference bar, and pie; the number of components was two, three, or five. The 10 trials in each of the 12 conditions were ordered randomly for each subject, forming a block. Each subject performed 12 blocks for a total of 120 trials and the order of blocks was counterbalanced using two randomly selected 12 × 12 Latin squares.

Each subject sat in front of the computer and read the instructions, which stated that the subject should judge the proportion indicated by the red circle and respond by typing an integer percentage on a numeric keypad. The instructions specified that the subject should be as quick as possible without sacrificing accuracy, that a backspace key could be used to correct an entry, and that the enter key should be pressed immediately after the percentage was typed. The experimenter told the subject to judge each graph independently. Because subjects may have been unfamiliar with the reference bar graph, they were instructed to compare the selected component to the one-hundred percent bar and to ignore other bars with that graph type.

On each trial the graph remained visible until the subject pressed the enter key. Subjects had an unlimited time to view the graph. Response time was the time from display onset until the enter key was pressed. After the enter key was pressed, a fixation point (an asterisk) was displayed at screen center for 2 s, and the next trial began. Subjects had 12 practice trials (one per condition) before the experiment. After the experiment, which required approximately an hour to complete, the subject was debriefed.

Results

Response time

For each subject, a mean response time was computed for each of the 12 conditions. For each graph type, the subject means were submitted to a repeated-measures (within-subjects) regression analysis, with response time regressed on number of components. The regression model is similar to one described by Neter, Kutner, Nachtsheim, and Wasserman (1996, ch. 29), except that the independent variable is quantitative, rather than categorical. The upper part of Figure 2 shows the means of the subject means and associated standard errors for each condition. (The standard errors shown in all graphs were computed using within-subjects variation as recommended by Loftus and

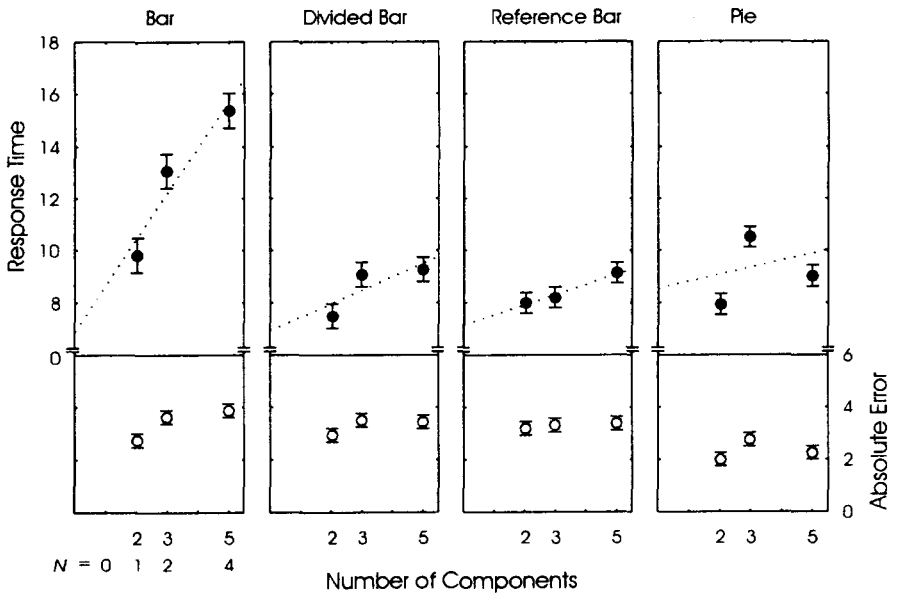

Figure 2. Experiment 1: response time (in seconds) and absolute error as a function of number of components and graph type. N = number of summations for the bar graph. Each error bar indicates plus or minus one standard error of the mean

Masson, 1994.) Response time increased with the number of components for bar graphs, $F(1,47) = 31.0$, $p < 0.001$. Response time also increased with number of components for divided bars, $F(1,47) = 5.3$, $p < 0.05$, and reference bars, $F(1,47) = 4.2$, $p < 0.05$, but not for pies, $F(1,47) = 0.95$. Table 1 shows the proportion of variance (total and within-subjects variance) accounted for by number of components: The proportions were higher for bar graphs than for any other graph type.

To compare slope and intercept values for the different graph types, an individual regression equations procedure for analyzing repeated measures was used (Lorch and Myers, 1990). Response time was regressed on the number of possible summations

Table 1. R^2-values (proportion of total variance and proportion of within-subjects variance accounted for by the summation model) by graph type for each experiment

	Graph type			
Type of variance	Bar	Divided bar	Reference bar	Pie
Experiment 1				
Total	0.14	0.03	0.02	<0.01
Within-subjects	0.40	0.10	0.08	0.02
Experiment 2				
Total	0.10	0.02	0.01	0.02
Within-subjects	0.21	0.06	0.05	0.05
Experiment 3				
Total	<0.01	<0.01	<0.01	<0.01
Within-subjects	0.02	<0.01	0.01	<0.01

(the number of components minus one) using the mean values for each subject, producing a slope and intercept value for each subject (the same slopes and intercepts would be obtained if the regressions were performed on each subject's raw data). This was done separately for each of the four graph types. Number of summations was used instead of number of components because the intercept then represents an estimate of the time for fixed overhead operations. (Using number of summations rather than number of components has no effect on the slopes.) Hence, a slope and intercept value were obtained for every subject in each graph type condition. The slopes obtained for each graph type were then compared in a one-way within-subjects ANOVA (as were the intercepts). The mean slope for bar graphs (1.8 s, which is an estimate of τ_S, the time for a summation operation) was greater than for any other condition (0.5, 0.4, and 0.2 for divided bar, reference bar, and pie, respectively), $F(3,69) = 8.7$, $p < 0.0001$, Student-Newman-Keuls (SNK), $p < 0.05$, and there were no differences in slope values among the other conditions, SNK, $ps > 0.5$. There were no differences among the intercepts, $F(3,69) = 1.9$, $p > 0.14$ (8.6, 7.4, 7.5, and 8.7 s, for bars, divided bars, reference bars, and pies, respectively). Figure 3 shows the mean slope and intercept values and associated standard errors for each graph type.

Absolute error

During debriefing, some subjects commented that with bar graphs they occasionally compared the selected component to the remainder instead of the whole, especially

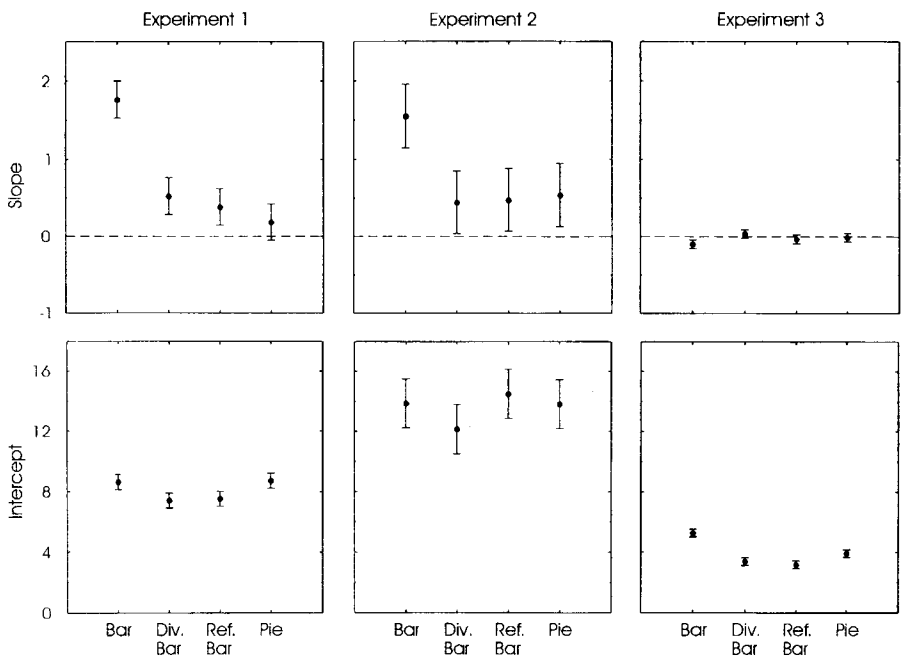

Figure 3. Experiments 1, 2, and 3: mean slopes and intercepts (in seconds) for response time regressed on number of summations for each graph type condition. Each error bar indicates plus or minus one standard error of the mean. The dashed lines on the upper graphs indicate slope values of zero. Div. Bar = Divided Bar; Ref. Bar = Reference Bar

with two components. To correct for this problem, a procedure described in the Appendix was used to adjust subjects' estimates. The procedure was performed on the bar graph data for all three experiments.

Absolute error was computed by taking the absolute value of the difference between the true proportion and the subject's estimate (adjusted estimate for bars), expressed as percentages. A mean absolute error value was computed for each subject, and the bottom part of Figure 2 shows the means of the subject means and associated standard errors. Comparison of response time and absolute error in Figure 2 does not suggest subjects were trading off speed for accuracy. The results of a within-subjects ANOVA (Graph Type by Number of Components) performed on the subject means showed smaller error for pies than other graph types, $F(3,69) = 6.3$, $p < 0.001$, SNK, $p < 0.05$, and smaller error with two components than any other number, $F(2,46) = 4.6$, $p < 0.05$, SNK, $p < 0.05$.

Discussion

The response time results were generally consistent with the predictions of the summation model. Response times increased with the number of components in the bar graph condition, at the rate of 1.8 s per component. Hence, the estimated duration of the summation operation ($\hat{\tau}_S$) was 1.8 s. The estimated duration of the fixed overhead operations ($\hat{\tau}_O$) was 8.6 s; this was not different from the mean values for the other graph types, consistent with the model's prediction that $T = \tau_O$ for graph types where the number of summations should be zero. The increase for the bar graph was greater than for the other graph types, although they also showed small increases (approximately 0.4 s per proportion). With more components, the stimulus is more complex, and may require more time to apprehend. Alternatively, more fixations might be required with a greater number of components. Casner (1991) estimated that the time for a fixation on a graphical item was 0.33 s in his experiments, which is consistent with our results.

EXPERIMENT 2

If the duration of the summation operation is to be a useful statistic, it should be relatively unaffected by changes in instruction. In Experiment 2, the instructions stressed accuracy and subjects were provided with feedback during practice trials. In Experiment 1 subjects were instructed to judge each graph independently. To further ensure that subjects made an independent judgment on each trial, the size of the graphs was varied randomly over trials in Experiment 2. Thus, with bar graphs, they could not use the remembered size of the whole on earlier trials to assist on the current trial. Finally, an eight-component condition was added to test the model's predictions over a wider range.

Method

Subjects

Sixteen students enrolled in an introductory psychology course at the University of Toronto participated in the experiment. They received course credit for participating. No subject in Experiment 2 served in Experiment 1.

Materials and apparatus

Materials and apparatus were similar to those in Experiment 1. Six sets of integers summing to 100 were independently generated for the two-, three-, five-, and eight-component conditions. These data were plotted for each graph type, producing 96 graphs, 6 per condition. The size representing 100% was randomly varied within a plus or minus 15% range. For the pie chart, this represented a change in radius; for the other graph types, the height of the bars was varied.

Design and procedure

A 4 × 4 (Graph Type by Number of Components) within-subjects design was used. The four graph types were bar, divided bar, reference bar, and pie; the number of components shown in each graph was two, three, five, or eight. The six trials in each condition were ordered randomly for each subject, forming a block. Each subject performed 16 blocks for a total of 96 trials and the order of blocks was counterbalanced using a randomly selected 16 × 16 Latin square.

The procedure was generally the same as in Experiment 1. However, in Experiment 2, the instructions stressed accuracy (i.e., 'Be as accurate as you can'), and subjects were provided with feedback during practice trials. For example, if the subject entered 44, and the proportion was 43, the computer would display, 'The correct response is 43. You were off by 1'. Subjects completed 16 practice trials (1 per condition) before starting the experiment.

Results

Response time

Mean response times for each of the 16 conditions were computed for each subject. These data were submitted to the same repeated-measures regression analysis used in Experiment 1, with response time regressed on number of components separately for each graph type. The upper part of Figure 4 shows the mean of the subject means and associated standard error for each condition. Response time increased with the number of components for bar graphs, $F(1,47) = 12.7$, $p < 0.001$. Increases in response time for the other graph types approached, but did not reach, conventional significance levels (divided bars, $F(1,47) = 3.0$, $p < 0.10$; reference bars, $F(1,47) = 2.5$, $p < 0.15$; pies, $F(1,47) = 2.27$, $p < 0.15$). Table 1 shows the proportion of variance (total and within-subjects variance) accounted for by number of components: As in Experiment 1, the proportions were higher for bar graphs than for any other graph type.

An individual regression equations procedure for analyzing repeated measures was used to compare slope and intercept values across graph types, as in Experiment 1. For each subject, response time was regressed on the number of possible summations separately for each graph type, producing a slope and an intercept value. Figure 3 shows the means of the slope and intercept values and associated standard errors for each graph type. The slope for bar graphs (1.6 s) was larger than the slopes for the other graphs (0.4, 0.5, and 0.5 for divided bars, reference bars, and pies, respectively), but the difference only approached conventional significance levels, $F(3,45) = 1.7$, $p < 0.18$. There were no differences among the intercepts, $F(3,69) = 0.37$, $p > 0.75$ (13.9, 12.1, 14.5, and 13.8 s, for bars, divided bars, reference bars, and pies, respectively).

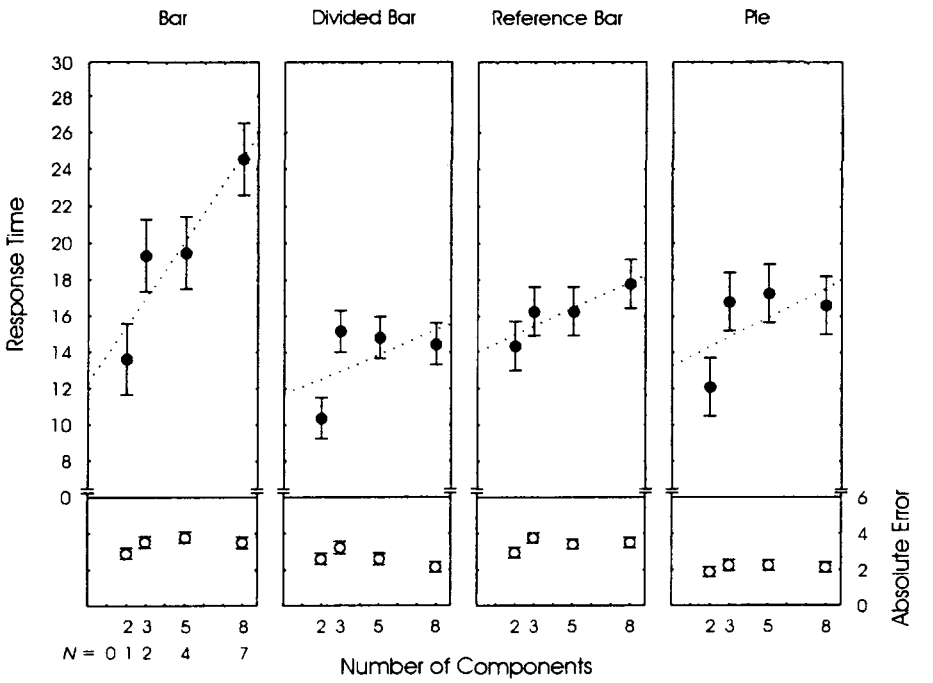

Figure 4. Experiment 2: response time (in seconds) and absolute error as a function of number of components and graph type. N = number of summations for the bar graph. Each error bar indicates plus or minus one standard error of the mean

Absolute error

Subjects' bar graph responses were adjusted as in Experiment 1 using the procedure described in the Appendix. A mean absolute error value was computed for each subject in each condition, and the lower part of Figure 4 shows the means of the subject means and associated standard errors for each condition. A comparison of response time and absolute error in Figure 4 suggests that subjects were not trading off speed for accuracy. A 4 × 4 within-subjects ANOVA showed a main effect for number of components, $F(3,45) = 4.4$, $p < 0.01$. The order of number-of-component conditions from smallest to largest error was: 2, 8, 3, 5. Subjects made smaller errors with two components than three or five, SNK, $p < 0.05$. Subjects produced smaller errors with divided bars and pies than other graph types, $F(3,45) = 8.3$, $p < 0.001$, SNK, $p < 0.05$. There was no interaction between graph type and number of components, $F < 1$.

Discussion

Consistent with the summation model, response times increased with the number of components in the bar graph condition, at the rate of 1.6 s per component. Hence, the estimated duration of the summation operation ($\hat{\tau}_S$) was 1.6 s, close to the value of 1.8 s obtained in Experiment 1. Hence, design differences between Experiments 1 and 2 had little effect on τ_S. The estimated duration of the fixed overhead operations ($\hat{\tau}_o$) was 13.9 s in Experiment 2, several seconds longer than the value of 7.6 s obtained in Experiment 1. The intercepts for the other graph types and for non-increasing bar

graph subjects were similarly larger in Experiment 2. Emphasizing accuracy (and possibly varying the displayed size) apparently increased the time for the fixed overhead operations, but had little effect on summation, suggesting that the summation rate remained reasonably consistent despite different instructions. The other graph types showed no increase with the number of components (although the increases approached conventional significance levels), and the difference in slopes failed to reach conventional significance levels.

Only sixteen subjects were run in this experiment, and therefore the power was low for detecting anything but large effects. The increase in variability evident in Experiment 2 (compare the error bars in Figures 2 and 4, or in Figure 3) is partly due to the smaller sample size. Also, since subjects emphasized accuracy over speed, response times were longer than in Experiment 1 and variances consequently larger.

EXPERIMENT 3

In the previous experiments, subjects were given unlimited time to respond and accuracy was stressed (especially in Experiment 2). If the graph is displayed for only a short time, however, subjects should not have sufficient time to perform the sequence of summation operations with bar graphs, and therefore response time should not increase with the number of components. If subjects cannot perform summation operations with bar graphs, how could they make proportion judgments? An individual bar could be compared to the irregularly-shaped area of the entire bar graph. Estimating the size of the whole should require more time for bar graphs than for other graph types. The whole would be irregularly shaped for bars (regardless of the number of bars). In contrast, the physical object that represents the whole has a simple shape (a rectangle or a circle) with other graph types. Therefore intercept values for bar graphs should be higher than for the other graph types in Experiment 3, but slope values should not differ.

Two display durations were used: 1 s and 3 s. We did not anticipate that subjects could make effective use of summation operations in as short an interval as 3 s, but to further ensure that subjects could not use a summation strategy, we included the shorter duration of 1 s.

Method

Thirty-two students enrolled in an introductory psychology course at the University of Toronto participated and received course credit for participating. No subject in Experiment 3 served in Experiment 1 or 2. Materials, apparatus, and design were generally the same as in Experiment 2, except that half the subjects saw each graph for 1 s, and the other half for 3 s. Thus, a $2 \times 4 \times 4$ (Display Duration by Graph Type by Number of Components) between-within design was used with display duration serving as the between-subjects variable.

Results

Response time

For each trial, response time was the time from display offset until the enter key was pressed. A between-within $2 \times 4 \times 4$ ANOVA showed no effect of display duration on

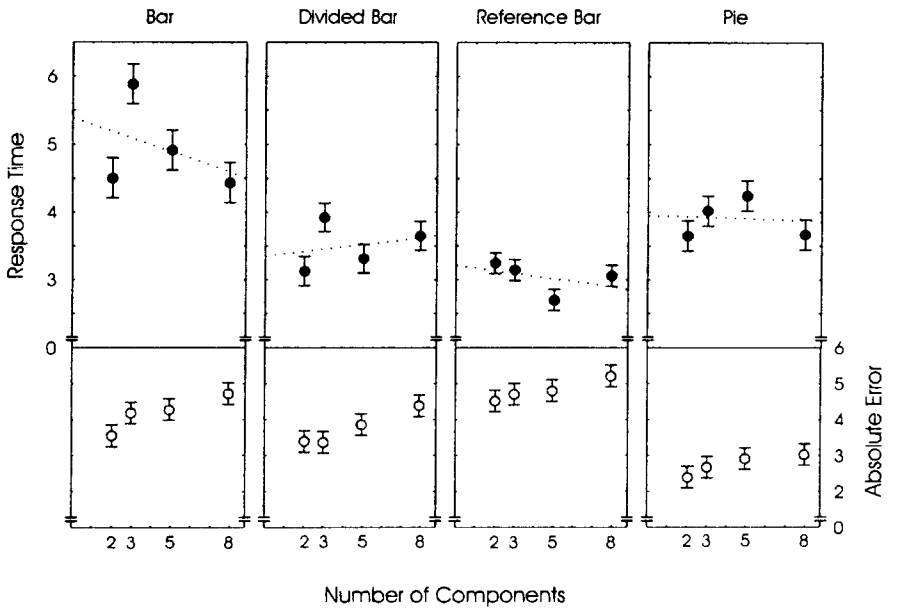

Figure 5. Experiment 3: response time (in seconds) and absolute error as a function of graph type and number of components. Each error bar indicates plus or minus one standard error of the mean

response time, nor did display duration interact with any other factor ($p > 0.05$ in all cases). Therefore subsequent analyses were collapsed across this factor. Mean response times for each of the remaining 16 conditions were computed as in Experiment 2 for each of the 32 subjects. These data were submitted to the same repeated-measures regression analysis used in Experiments 1 and 2. The upper part of Figure 5 shows mean response times (averaged across subjects) for each condition.

Response times did not increase (or decrease) with the number of components for the bar graph, or any other graph type (bars, $F(1,95) = 2.3$, $p > 0.13$; divided bars, $F(1,95) = 0.58$, $p > 0.44$; reference bars, $F(1,95) = 1.2$, $p > 0.27$; or pies, $F(1,95) = 0.03$, $p > 0.87$). Table 1 shows the proportion of variance (total and within-subjects variance) accounted for by number of components for each graph type. In contrast to the Experiment 1 and 2 results, there was little difference in proportion of variance accounted for by the model among the graph types. This is true regardless of whether one considers proportion of total variance or within-subjects variance.

As in the previous experiments, an individual regression equations procedure for analyzing repeated measures was used to compare slope and intercept values across graph types. In contrast to earlier results, there was no difference among the slopes for the four graph types (−0.10, 0.04, −0.04, and −0.01 for bars, divided bars, reference bars, and pies, respectively), $F(3,93) = 0.98$, $p > 0.40$. Intercepts for the divided bars, reference bars, and pies were 3.4, 3.2, and 3.9, respectively; all of these intercepts were lower than the bar graph intercept of 5.3 s, $F(3,93) = 12.3$, $p < 0.0001$, SNK, $p < 0.05$, but they did not differ among themselves, SNK, all ps > 0.12. Figure 5 shows the mean slope and intercept values and associated standard errors for each graph type.

Absolute error

Absolute error values were adjusted as in Experiments 1 and 2 and the lower part of Figure 5 shows mean values. A $2 \times 4 \times 4$ between-within ANOVA showed that there was a general increase in error as the number of components increased, but only the difference between two and eight proportions was significant, $F(3,90) = 6.7$, $p < 0.001$, SNK, $p < 0.05$. The smallest errors were made with pie charts, and the largest errors were made with reference bar graphs, $F(3,90) = 22.4$, $p < 0.001$, SNK, $p < 0.05$. There was no interaction between graph type and number of components, $F(9,270) < 1$. There was no effect of display duration, nor did duration interact with any other factor ($p > 0.05$ in all cases).

Discussion

A comparison of Figure 5 with Figures 2 and 4 shows that the results of Experiment 3 are different from earlier results. As predicted, response time did not increase with number of possible summations for bar graphs in Experiment 3, whereas there was an increase in Experiments 1 and 2. Presumably, response time did not increase for bar graphs in Experiment 3 because subjects did not have sufficient time to perform summation operations. Subjects generally required less time to judge proportions in Experiment 3 than in Experiments 1 and 2. There would be no advantage to delaying after display offset in Experiment 3, but in the earlier experiments the subject could take as long as necessary to reach an internal accuracy criterion since the graph remained visible until the subject responded. The short display durations of Experiment 3 appear to have forced a relatively quick strategy unaffected by the number of bars.

The intercept for bars was larger than for any other graph type, and the slopes for the other graph types did not differ from the slope for bars. Hence, response times were generally longer for bar graphs than for other graph types in Experiment 3, implying some extra processing for bars. Estimating the size of the whole may be more time consuming when the area of the whole is irregularly shaped, as with bar graphs.

There was some evidence for an increase in absolute error with number of components in Experiment 3. In Experiment 1, subjects could take longer when there were more components, and sometimes did. It appears that since subjects could not take more time in Experiment 3, an increase in error with more components resulted. The speeded task used in Experiment 3 also appears to have increased absolute error somewhat, especially compared to Experiment 2 where accuracy was emphasized.

It might be argued that Experiment 3 attempted to prove the null hypothesis (Frick, 1995). However, the power to detect differences was much higher in Experiment 3—variability was low and the sample size was larger. If an effect existed Experiment 3 possessed fairly high power to detect it. Nevertheless, the effect of number of components on response time in speeded proportion judgments was negligible.

GENERAL DISCUSSION

In Experiments 1 and 2, response times increased with the number of components for bar graphs, but not for other graph types. The slope and intercept of the bar graph

regression equation provided an estimate of the duration of the summation operation, and the duration of the fixed overhead operations, respectively. An emphasis on accuracy in the Experiment 2 instructions had little effect on the estimated time for the summation operation, but increased the time required for fixed overhead operations by several seconds. In Experiment 3, speeded presentation appeared to prohibit use of the summation operation with bars and subjects used a strategy unaffected by the number of bars.

Fixed overhead operations

This paper has focused on one cognitive operation — summation — and has not yet discussed in great detail other operations that might be used in proportion judgments. With divided bars, reference bars, and pies, some ratio estimation process must occur in order for the subject to produce an accurate response. Ratio estimation should also occur after the summation sequence with bars. The intercept values described in our paper include the time for ratio estimation, but must also include other operations, which we have generally classified as fixed overhead operations. Given the large intercept values obtained, there is time for several additional operations to occur. Gillan and Lewis (1994) described several operations that might occur in ratio estimation for a divided bar graph. It is likely that similar operations are at work with reference bars and pie charts. In order, the following six processing steps seem likely: locating the part (Gillan and Lewis's 'search' operation); encoding the magnitude of the part; encoding the magnitude of the whole; comparing the magnitudes (ratio estimation); and response planning. In addition, our response times included the time the subject took to type the response on the keypad, so a response execution stage should also be included in this list, for a total of six processes. Using Gillan and Lewis's estimated average time-per-operation for arithmetic steps (1.8 s) produces a total estimated time of 10.8 s; non-arithmetic steps (0.8 s) produce a total response time of 4.8 s. In our task, which probably involved a mixture of arithmetic and non-arithmetic operations, the estimate for fixed overhead operations was 7.6 s in Experiment 1. Hence, the time for fixed overhead operations as estimated in Experiment 1 appears to fall within the boundaries predicted by the Gillan and Lewis model. In Experiment 2, accuracy was stressed, and the estimated time for fixed overhead operations was 13.9 s. Subjects may have repeated some operations as a check in order to keep error small in this experiment.

Reliability of summation rate estimate

The estimated time for the summation operation was approximately 1.7 s per summation (averaged across Experiments 1 and 2). The emphasis on accuracy in the Experiment 2 instructions had little effect on the estimate. Gillan and Lewis (1994, Experiment 2) required subjects to make (unspeeded) estimates of the sum of quantities shown in graphs (line graphs and scatterplots), and increased the number of additions required from two to five. By fitting a regression line to the data portrayed in their Figure 4, we have obtained an estimated summation rate of 1.6 s per summation (averaged across the two graph types). Since their task did not require a proportion judgment, and therefore no ratio estimation operation was necessary (nor locating or encoding the magnitude of the whole), one would expect that the

intercept for their experiment would be less than those obtained in our experiments. Indeed this was so; the estimated intercept for the Gillan and Lewis experiment was 2.6 s, less than any of the intercepts for our experiments.

Individual differences

A number of recent studies (e.g., Carswell and Emery, 1992; Lohse, 1993; Nowicki and Coury, 1993) have shown that individual differences can play a role in how people extract information from graphs. Hence, we checked for performance differences among subjects in all three experiments. A comparison of the R^2-values in Table 1 (proportion of total variance vs. proportion of within-subjects variance) suggests that between-subjects variability was high in Experiments 1 and 2. We found that some subjects in these experiments showed little increase in response time for bar graphs as the number of components increased. This might be interpreted to mean that some subjects did not perform summation operations with bars. The Experiment 3 results showed little variation among subjects, which suggests that individual differences were minimized in that experiment. In the speeded situation, all subjects performed a quick judgment, perhaps involving an area estimate of the irregularly-shaped whole. Subjects showing little increase in Experiments 1 and 2 might have used a similar strategy.

However, when individual slope values were ranked, there was no evidence for a bimodal distribution, with some subjects clearly showing an increase and other subjects not. The observed variation in slopes might reflect individual differences in summation rate rather than two qualitatively different cognitive strategies. Unfortunately, we have insufficient data to classify the subjects on an a priori basis (e.g., cognitive style, see Carswell and Emery, 1992; or expertise, Shah and Carpenter, 1995), and so we cannot draw any conclusions concerning individual differences in the current experiments. As other researchers have found, there are pronounced individual differences in graph reading; our data provide one more example.

Note that our estimates of summation rate are based on the assumption that all subjects performed the summation operation. If this is incorrect, and some subjects did not sum components, the estimated duration of the operation may be slightly longer than that reported here. However, it is encouraging that our estimate and the estimate obtained from analysis of the Gillan and Lewis (1994) data are similar.

Implications for design

The results from Experiments 1 and 2 (shown in Figures 2 and 4, respectively) indicate that the bar graph is not an effective display for judging proportion, and that its disadvantage becomes greater as the number of components increases. Even when the task was speeded, as in Experiment 3, subjects took longer with bar graphs, although the disadvantage did not increase with the number of components.

Divided bars, reference bars, and pies showed generally good performance across the three experiments. However, pie charts produced the smallest absolute error values in all three experiments. Despite the opinions of some critics (e.g., Cleveland, 1985; Macdonald-Ross, 1977; Tufte, 1983), the accumulated evidence suggests that the pie is an effective format for judgments of proportion and yields accuracies as good as or better than other graph types for that task (Eells, 1926; Hollands and

Spence, 1992; Simkin and Hastie, 1987; Spence, 1990; Spence and Lewandowsky, 1991). It appears that the pie chart is the graph of choice for communicating part-to-whole relationships, with divided bars and reference bars serving as reasonable alternatives. Bar graphs, in which no single object represents the whole, are not as effective for this task.

The summation model makes predictions pertinent to graphical display design. It predicts that any display not representing the whole by a single object will require extra time for proportion judgments, and that the extra time will increase with the number of components. For example, imagine a line graph showing several horizontal lines, each representing the number of sales of one product over several years. Suppose a graph reader wanted to estimate the proportion one product's sales were of all product sales in a given year. Since no single object in the graph represents the whole (i.e., all sales for a given year), the time for the proportion judgment should lengthen as the number of components increased. On the other hand, an improved bar or line graph might be constructed by surrounding the graph by a frame whose height was equal to 100%. Since the frame height represents the whole, no summation operation would be needed to establish its magnitude. Performance should be similar to that obtained with the reference bar or divided bars. Therefore the model could be used to predict performance with new or untested formats.

CONCLUSIONS

That bar graphs might take longer than other graphs for judgments of proportion may not be surprising and has, in fact, been demonstrated previously (Hollands and Spence, 1992). However, by manipulating the number of components this study demonstrated a number of important results, both theoretical and practical. First, a serial summation process appears to be used to establish the whole with bars, and this is probably the source of their disadvantage for proportion judgments. Second, an estimate of the average rate at which the summation process occurs has been obtained — roughly 1.7 s per proportion — and this rate was unaffected by changes in instruction. Third, with more components, the bar graph becomes gradually less effective for depicting proportion. Finally, in speeded situations, the summation process appears to be abandoned, but subjects still take longer with bars, reflecting the increased difficulty of estimating an irregular area. In summary, the results show that when there is task-graph correspondence (e.g., proportion judgments performed with a pie chart), relatively direct perceptual judgment can occur. In contrast, direct perceptual judgment cannot occur with a bar graph, and a longer sequence of cognitive steps is required, requiring extra processing time.

REFERENCES

Carswell, C. M. and Emery, C. (1992). Effects of stimulus complexity and cognitive style on spontaneous interpretations of line graphs. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 1498–1502). Santa Monica, CA: Human Factors Society.

Casner, S. M. (1991). A task-analytic approach to the automated design of graphic presentations. *ACM Transactions on Graphics*, **10**, 111–151.

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

Cleveland, W. S. (1990). A model for graphical perception. In *Proceedings of the Section on Statistical Graphics* (pp. 1–32). Alexandria, VA: American Statistical Association.
Eells, W. C. (1926). The relative merits of circles and bars for representing component parts. *Journal of the American Statistical Association*, **21**, 119–132.
Frick, R. W. (1995). Accepting the null hypothesis. *Memory and Cognition*, **23**, 132–138.
Gillan, D. J. and Lewis, R. (1994). A componential model of human interaction with graphs: 1. Linear regression modeling. *Human Factors*, **36**, 419–440.
Hollands, J. G. and Spence, I. (1992). Judgments of change and proportion in graphical perception. *Human Factors*, **34**, 313–334.
Kosslyn, S. M., Reiser, B. J., Farah, M. J. and Fliegel, S. L. (1983). Generating visual images: units and relations. *Journal of Experimental Psychology: General*, **112**, 278–303.
Leinhardt, G., Zaslavsky, O. and Stein, M. K. (1990). Functions, graphs, and graphing: tasks, learning, and teaching. *Review of Educational Research*, **60**, 1–64.
Loftus, G. R. and Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin and Review*, **1**, 476–490.
Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human–Computer Interaction*, **8**, 353–388.
Lorch, R. F. Jr and Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16**, 149–157.
Macdonald-Ross, M. (1977). How numbers are shown: a review of research on the presentation of quantitative data in texts. *Audio-Visual Communication Review*, **25**, 359–407.
Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied linear statistical models* (4th Edn). Chicago: Irwin.
Nowicki, J. R. and Coury, B. G. (1993). Individual differences in processing strategy for a bar graph display. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* (pp. 1315–1319). Santa Monica, CA: Human Factors and Ergonomics Society.
Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73–126). Hillsdale, NJ: Erlbaum.
Shah, P. and Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General*, **124**, 43–61.
Simkin, D. and Hastie, R. (1987). An information processing analysis of graph perception. *Journal of the American Statistical Association*, **82**, 454–465.
Spence, I. (1990). Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance*, **16**, 683–692.
Spence, I. and Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, **5**, 61–77.
Tan, J. K. H. and Benbasat, I. (1990). Processing of graphical information: a decomposition taxonomy to match data extraction tasks and graphical representations. *Information Systems Research*, **1**, 416–439.
Tan, J. K. H. and Benbasat, I. (1993). The effectiveness of graphical presentation for information extraction: a cumulative experimental approach. *Decision Sciences*, **24**, 167–191.
Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

APPENDIX–ADJUSTMENT OF ESTIMATE PROCEDURE

In all experiments, some subjects commented that with bar graphs they occasionally compared the selected component to the remainder instead of to the whole, especially with two components. Suppose, for example, a subject was shown a graph containing two parts, the first forming 40 and the second 60 percent of the whole, and was asked to estimate the first part. If the subject erroneously compared 40 to 60, instead of 40 to 100 (40 + 60), then the approximate value of the response would be 67 percent. To correct for this problem, we can assume that the second part has a value of 100,

and the sum of the two parts should be $67 + 100 = 167$. Hence, the subject's response should be adjusted to $67/(67 + 100) = 0.4 = 40\%$. This adjusted estimate represents how the subject might have responded if the correct ratio had been employed.

However, only certain trials should be adjusted; we must guard against the possibility of adjusting the estimate associated with a proper proportion judgment. In the earlier example, let us suppose the subject's actual estimate was 64%. The absolute difference between the subject's estimate and the true proportion (absolute error) would be $|64 - 40| = 24$. But since the subject compared the component to the remainder, rather than to the whole, the error might be better expressed as $|64 - 67| = 3$. If the first error measure is larger than the second, as it is in this example, we might assume that the subject compared the component to the remainder, instead of to the whole, and that the response should be adjusted. However, we cannot distinguish a priori which trials were judged properly. Subjects make errors even if they use the whole for their judgment. Hence, we ought to ensure that the difference between the two error measures is larger than some threshold value, such as 10%.

These ideas can be expressed more formally:

$$\text{if } |P - \Pi| - \left|P - \left(\frac{\Pi}{1 - \Pi}\right)\right| > \delta, \quad \text{then } P^* = \frac{P}{P + 1}.$$

The value Π is the true proportion, P is the subject's estimate, P^* is the adjusted estimate, and δ is the threshold value. (Π, P, and P^* are here treated as decimal fractions, rather than percentages, to simplify the expression.)

Thus, if the response was closer to the ratio with the remainder denominator (plus δ) than the correct denominator, the adjusted estimate was set equal to $P/(P + 1)$. The threshold value was set to 0.10 (or 10%) because higher values did not substantially alter the number of trials adjusted, and lower values required adjustment of a high number of trials in the three- and five-component conditions. The procedure adjusted 79 estimates (out of 570, or 13.9%) in Experiment 1. Forty-three (out of 190, or 22.6%) were in the two-component condition, and 18 (out of 190, or 9.5%) occurred in each of the other component conditions. Similar percentages of trials were adjusted in Experiments 2 and 3. The procedure was applied to bar graph data in all three experiments.