



## Discriminating Strata in Scatterplots

Stephan Lewandowsky; Ian Spence

*Journal of the American Statistical Association*, Volume 84, Issue 407 (Sep., 1989),  
682-688.

Stable URL:

<http://links.jstor.org/sici?sici=0162-1459%28198909%2984%3A407%3C682%3ADSI%3E2.0.CO%3B2-U>

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

*Journal of the American Statistical Association* is published by American Statistical Association. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/astata.html>.

---

*Journal of the American Statistical Association*  
©1989 American Statistical Association

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact [jstor-info@umich.edu](mailto:jstor-info@umich.edu).

©2003 JSTOR

# Discriminating Strata in Scatterplots

STEPHAN LEWANDOWSKY and IAN SPENCE\*

---

When multiple groups are shown in a scatterplot each stratum is represented by a different symbol; for example, three strata might be coded using red, green, and yellow circles. Various symbol types were compared by behavioral experiment: Subjects were fastest when strata were coded using different colors and slowest when strata were coded with confusable letters—but there were no differences in accuracy. Accuracy differed only when processing time was restricted, again with different colors and confusable letters representing the two extremes. We conclude that color is the optimal symbol type and show that measuring response latency in addition to accuracy is essential in research on graphical perception.

KEY WORDS: Perception; Speed-accuracy-trade-off function; Statistical graphs.

---

## 1. INTRODUCTION

Statistical graphs have been in use for some two centuries, originating with the work of Playfair (1786). Despite this long history and the ubiquity of graphs, little is known about how people perceive and process statistical graphs, and numerous appeals for more empirical investigations have been made (e.g., Cleveland and McGill 1984a, 1985; Cox 1978; Jacob 1981; Kruskal 1975). This article addresses one aspect of the perception of statistical graphs, namely the discrimination of multiple strata in scatterplots. In two behavioral experiments, we examine the roles of two major determinants of perceptual performance: symbol type and the expertise of the observer.

### 1.1 Symbol Type

Often distinct groups, or *strata*, are compared on a set of common variables. For example, the relation between brain capacity and body weight for three different species of primate might be shown in the same scatterplot: Observations for chimpanzees could be plotted using the letter *C*, those for monkeys using the letter *M*, and those for gorillas using the letter *G*. The *display symbols* would be *C*, *M*, and *G*, and the *symbol type* would be *letters*. Partly because no special equipment is necessary for character graphics, letters have often been used to code strata. Other possibilities, shown in Figure 1, include different shapes (e.g., circles vs. squares vs. triangles), different amounts of fill (open circles vs. filled circles vs. half-filled circles), circles of different colors, or oriented lines. Readers are invited to use crayons to add color to the as-yet-undifferentiated points in the first panel.

Clearly, performance on a task that requires differentiation of the strata—for example, determining whether the correlation between brain capacity and body weight is greater for chimpanzees than for monkeys—will be affected by the perceptual discriminability of the symbols. Cleveland and McGill (1984b) proposed a tentative rank ordering of symbol types, with colors assumed to provide optimal discriminability, followed by amounts of fill, then

different shapes, and finally letters. Some support for their rank ordering may be found in psychophysical work (e.g., Chen 1982) involving single graphical elements; however, predictions based solely on psychophysical work with single stimuli can be at variance with results obtained in experiments that attempt to mimic more closely the kinds of judgments made in actual practice.

### 1.2 The Expertise of the Observer

Consumers of graphs encompass all levels of expertise, from the statistically untrained reader of a newspaper to the researcher examining scientific data. In one investigation of the elementary perceptual tasks involved in graph processing (Cleveland and McGill 1984a), experts with considerable technical training, as well as novices, had to judge relative magnitudes of angles and lengths of lines. Experts and novices did not differ in the accuracy of their discriminations, suggesting that expertise does not always influence graphical perception. Cleveland and McGill, however, measured response accuracy only, and (as we will show) it is not possible to make definitive statements about performance differences between groups of subjects unless *both* accuracy and speed of response are measured. It must therefore remain unclear whether experts and novices were indeed equally skillful in the Cleveland and McGill experiments.

## 2. OVERVIEW OF THE EXPERIMENTS

Both experiments employed similar stimuli, tasks, and experimental designs. Briefly, three strata were presented in a scatterplot and subjects had to compare two of the three strata and say which had the higher apparent correlation. This task requires that the subject be able to differentiate the strata in order to estimate the magnitudes of correlation. The primary experimental variable was the symbol type, and the response measures were accuracy and latency. The common elements of the experiments are summarized in the remainder of this section. Readers who wish to skip these details may proceed directly to Section 3.

### 2.1 Stimuli and Apparatus

The experiments were controlled by an IBM PC with a 50 × 40 cm Sony color monitor, and the stimuli, which

---

\* Stephan Lewandowsky is Research Associate and Ian Spence is Professor, Department of Psychology, University of Toronto, Toronto, Ontario M5S 1A1, Canada. This research was supported by Natural Sciences and Engineering Research Council of Canada Grants A8351 to I. Spence and G1779 to R. Baecker, A. Fournier, P. Muter, and I. Spence. The authors thank Jeff Bowers and Linda Tilley for testing the subjects, and two anonymous referees and an associate editor for their helpful comments on a previous draft.

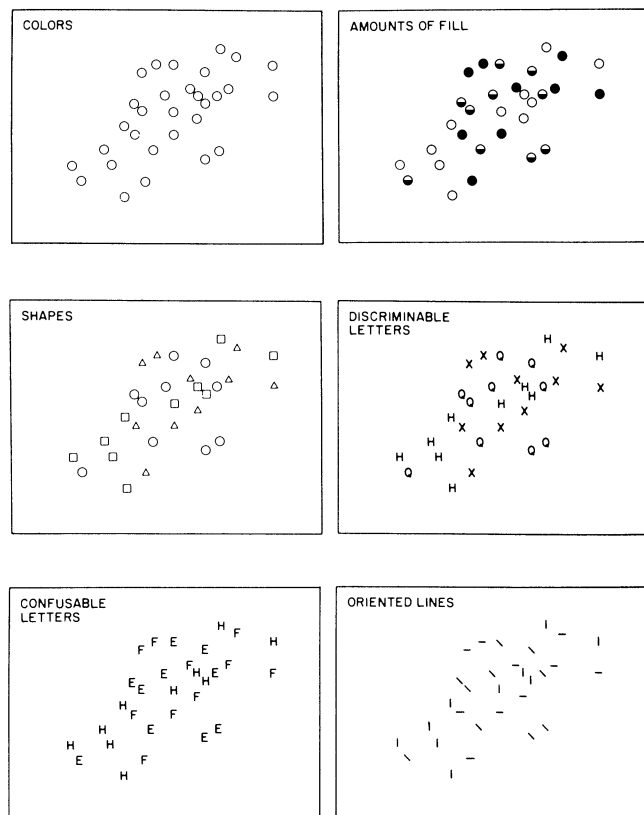


Figure 1. Several Examples of Different Symbol Types With Identical Strata.

consisted of a scatterplot of three strata of 10 points each, subtended a visual angle of approximately 15 degrees. The six panels of Figure 1 show examples of the symbol types used in the present experiments (the three colors were red, green, and yellow). For a given scatterplot and symbol type, the assignment of symbols to strata was random.

Experiments 1 and 2 employed 10 and 80 different scatterplots, respectively. The axes were scaled from 0 to 1, and the points represented samples from bivariate normal populations with correlations of .2, .5, and .8 for strata *A*, *B*, and *C*, respectively. The population means and standard deviations of the bivariate normals were identically  $\frac{1}{2}$  and  $\frac{1}{8}$  for all strata. During sampling, points were discarded if they fell outside the range of the axes or if there was overlap with an existing point. Three additional conditions were imposed: The sample Fisher *z* for a stratum could not be less than 0, the smallest of the three pairwise differences in *z* between strata could not be less than .1, and the largest pairwise difference in *z* could not exceed unity.

For Experiment 1, the average empirical correlations, over 10 replications, were .32, .50, and .76, and the Fisher *z* values were .36, .59, and 1.03 for strata *A*, *B*, and *C*. Thus the mean differences in *z* between strata were .23 for the *AB* comparison, .44 for *BC*, and .67 for *AC*. Expressed in terms of *r*, the differences were .18 for *AB*, .26 for *BC*, and .44 for *AC*.

The average empirical correlations, over 80 replications, for the three strata in Experiment 2 were .25, .54, and .76, and the Fisher *z* values were .26, .63, and 1.03. Thus

the mean differences in *z* between strata were .37 for the *AB* comparison and .77 for *AC*. Expressed in terms of *r*, the differences were .29 for *AB* and .51 for *AC*.

## 2.2 The Task

The two-alternative forced-choice task required subjects to decide which of two strata had the higher apparent correlation. The two strata to be compared on a given trial were indicated by symbols displayed below the scatterplot, and each symbol was associated—by its location on the screen—with a response key. If a square was shown at the lower left and a triangle at the lower right of the display, squares and triangles were to be compared. The left response key would be pressed to indicate that squares showed the greater correlation, and the right key would be pressed to indicate the opposite. All keys other than the two response keys, at opposite corners of the keyboard, were covered by an opaque mask. The third stratum was to be ignored and was included only because pilot testing showed that the task was too easy with only two strata present.

Although other experimental tasks could have been used, the two-alternative forced-choice task permitted particularly effective control of processing time and facilitated measuring response latencies. In addition, the judgment involved a cognitive perceptual component (assessment of degree of correlation) as well as a more fundamental perceptual component (discrimination of strata). Since the cognitive component was constant across experimental conditions, observed differences in performance must result from differential efficiency in perceiving and discriminating the strata; however, a cautious interpretation of the present results must consider the possibility that different performance patterns might obtain with other experimental tasks.

## 2.3 The Subjects

The novice subjects were University of Toronto undergraduates with no formal training in statistics who participated voluntarily for a fee of \$5 per hour. The expert subjects were senior graduate students or faculty in the Department of Psychology; all had received at least two years of training in statistics and experimental design, and all were regular users of statistical techniques. Expert subjects either participated voluntarily without pay (Experiment 1) or were paid at the same rate as the novices (Experiment 2). Subjects were screened for color deficiency (Hardy, Rand, and Rittler 1957).

## 2.4 Orientation and Practice

At the beginning of an experiment, all subjects received six orientation and four practice trials. The orientation trials introduced the novices to the concept of correlation by presenting single strata of 20 points, with the empirical correlation coefficient displayed below each plot. Subjects were told that the displayed number was a measure of “correlation” and that they should associate its magnitude with the appearance of the pattern of points. The first orientation trial displayed data with zero correlation, and

subsequent trials used progressively higher correlations (by increments of .2).

Four practice trials, involving the discrimination task, followed orientation. These were identical to experimental trials, except the comparisons were a little easier and feedback was given by printing the word *correct* or *wrong* below the graph after each response. All symbol types used in the experiment were presented during practice.

### 3. EXPERIMENT 1

Three groups of 14 subjects each participated. Two of the groups responded to the same symbol types and differed only with regard to the subjects' level of expertise. These groups were labeled E and N-1 (for experts and novices-1). Another group of novices, N-2, received a different selection of symbol types.

#### 3.1 Method

Altogether, five symbol types were compared: colors, amounts of fill, shapes, discriminable letters, and confusable letters. Each group of subjects saw only three different symbol types: Groups E and N-1 saw colors, shapes, and amounts of fill, whereas group N-2 saw shapes and the two sets of letters. The set of discriminable letters (little feature overlap) consisted of *H*, *Q*, and *X*, and the confusable letters (many features in common) were *H*, *E*, and *F*.

Each of the three possible comparisons for any given scatterplot was made by all subjects, resulting in a  $3 \times 3$  within-subjects design for each of the groups (E, N-1, and N-2). This design was replicated across 10 different scatterplots for each subject. Hence, an experimental session consisted of 90 trials presented in a unique random order. Each scatterplot remained on the screen until a response was made, at which point it disappeared, leaving a blank screen for 500 ms before the next trial began. Subjects were instructed in writing, and both accuracy and speed of response was emphasized.

#### 3.2 Results

Table 1 shows the means for accuracy (percentage correct) and latency (average of correct responses in seconds) in all conditions. For all groups, the particular pair of strata

compared affected both accuracy and response time. As expected, accuracy was highest when the most extreme strata (*AC*) were compared; it was lowest for the comparison involving the smallest average difference in correlation (*AB*). Correspondingly, response times were fastest for the *AC* comparison and slowest for *AB*. This variable did not interact with any of the other factors either in this or in the second experiment.

Symbol type, on the other hand, had little effect on accuracy for any of the groups, but it did have a large effect on speed of response. For groups E and N-1, response latencies with colors were much shorter than with either shapes or amounts of fill. For group N-2, confusable letters produced much longer latencies—by about two seconds—than shapes and discriminable letters, which, in turn, did not differ. The corresponding analyses of variance are shown in Table 2. (Although we do not report the details here, a further experiment included oriented lines—panel six in Figure 1. Performance of the subjects was essentially the same as with shapes or amounts of fill.)

A comparison of group E with groups N-1 and N-2 showed the experts to be slightly more accurate than the novices. Unexpectedly, the experts took more time—about one second more—to make their judgments.

#### 3.3 Varying the Number of Points

To rule out the possibility that these findings were idiosyncratic to the particular choice of stimuli, another experiment was conducted, which varied the number of plotted points. Only three symbol types were used: colors, discriminable letters (*H*, *T*, and *X*), and confusable letters (*H*, *E*, and *F*). The size of the character matrix was reduced to accommodate a larger number of points on the screen, and the letter *T* was substituted for *Q*, which could not be constructed satisfactorily in the reduced matrix. One group of nine novice subjects examined scatterplots with 10 points per stratum, and another group of nine novice subjects examined scatterplots with 30 points per stratum. Thus the number of data points in each stratum was tripled for one group but, as before, no overlap between individual points was allowed.

The general pattern of results is summarized in Table 3. As before, and for both groups, there is little difference in accuracy associated with symbol type, although color

Table 1. Mean Accuracy (percent correct) and Latency (seconds) in Experiment 1 for the Three Groups and Various Symbol Types

Comparison	Experts			Novices-1			Novices-2		
	Shapes	Fill	Colors	Shapes	Fill	Colors	Shapes	HEF	HQX
Accuracy									
AB	73	69	69	66	61	66	57	63	62
BC	81	74	85	75	79	80	75	67	76
AC	89	89	91	84	86	88	79	78	77
Latency									
AB	10.5	10.8	9.5	8.8	9.3	7.4	9.2	10.6	9.2
BC	9.4	9.7	7.8	8.3	8.8	6.7	8.2	9.9	8.8
AC	9.2	9.4	7.1	7.9	9.0	6.5	7.7	10.6	7.5

Table 2. Analyses of Variance for Accuracy (percent correct) and Latency (seconds) in Experiment 1

Source of variation	df	MS for accuracy			MS for latency		
		E	N-1	N-2	E	N-1	N-2
Subject (S)	13	483	529	482	37.7	48.0	99.4
Symbol type (T)	2	296	117	60	40.9	50.8	53.4
S × T	26	165	102	100	2.2	2.6	5.0
Comparison (C)	2	4,003	5,017	3,317	33.9	6.1	11.9
S × C	26	108	105	231	2.1	1.3	3.0
T × C	4	138	62	205	1.3	.4	4.9
S × T × C	52	118	127	106	1.2	1.2	2.3

NOTE: MS is mean squares.

shows a slight advantage. But there are large differences in latency with colors leading to the fastest response times, followed by discriminable letters, and trailed by confusable letters. Increasing the number of points in each stratum had no effect on latency but produced a uniform improvement in accuracy. Thus the only noteworthy effect of increasing the number of points is that subjects seem better able to assess the degree of correlation, without any change in their ability to discriminate the strata.

Because the response latencies for the two groups are virtually identical, it seems that subjects probably attend to a stratum in a global fashion rather than selectively attending to each of the points in turn. If the latter were the case, one would expect longer latencies for the condition with the larger number of points. This supports our contention that psychophysical data on the discriminability of *single* graphical elements need not necessarily apply to the perception of more complex statistical displays.

#### 4. THE RELATION BETWEEN SPEED AND ACCURACY

The results of Experiment 1 are somewhat counterintuitive. First, one would probably anticipate that, in addition to being more accurate, experts would also respond more quickly than novices. After all, the label *expert* implies the ability to solve a certain class of problems particularly well, and problems that are solved well are usually also solved quickly. Instead, in the first experiment, experts were slower than novices. Second, the results did not correspond to the subjective impressions reported by subjects, who felt certain that they had been most accurate with colors; however, the actual level of accuracy was about the same for all symbol types, and only response latency varied considerably. We must therefore conclude that an examination of accuracy alone is insufficient to characterize performance. Had we measured accuracy only, we would have been led to believe that symbol type had little or no effect on perceptual performance.

Many psychological tasks involve a trading relation between speed and accuracy: Typically, a judgment is more likely to be correct if the subject responds slowly rather than quickly. Two subjects may differ in accuracy on the same task, one scoring 80% correct and the other scoring 90% correct, say, simply because the latter has chosen to wait longer before responding, and not because of any

inherent superiority. The difference lies exclusively in the subject's choice of response criterion and, in many cases, may be reversed by appropriate manipulation of payoffs or deadlines. It follows that if differences in accuracy are observed between two groups of subjects, in conjunction with a speed difference in the opposite direction, it is unclear whether the accuracy difference is due to a difference in response criterion or there is a more fundamental difference in capability. Experiment 1 presents this ambiguity: Accuracy differed between experts and novices, as did speed, but the speed difference was in the opposite direction.

This phenomenon may be examined more closely by forcing the observer to trade off accuracy for speed, by requiring that responses be made immediately after a signal rather than after unrestricted processing time. By varying the timing of the signal, the function relating speed and accuracy of response may be observed. This function is known as the speed-accuracy trade-off (SAT) function, where, usually, an observer's accuracy is an S-shaped function of processing time. Note that in Experiment 1, accuracy could not be observed as a function of processing time because both variables were under the subjects' control.

We offer three possible explanations of the differences between experts and novices. The first, a *criterion* explanation, assumes that experts were more cautious than novices and responded only when they felt certain—resulting in the observed increase in accuracy and a concomitant decrease in speed. This explanation postulates no inherent difference in skill between experts and novices and is not implausible, given Cleveland and McGill's (1984a) findings that experts were no more accurate than novices on basic perceptual tasks. The second explanation, a *skill* explanation, assumes that experts were indeed more skillful than novices but also happened to be more cautious. Finally, the *rate* explanation assumes that experts process graphical information in a fundamentally different way, eventually resulting in more skillful performance but requiring more processing time. Here the difference in response time is not associated with a criterion difference, but rather is a necessary consequence of the differential rate of processing.

#### 5. EXPERIMENT 2

In this experiment, processing time was manipulated by the experimenter in order to examine the three possible explanations of expert–novice differences. Experts and novices were tested using four different symbol types: colors, shapes, amounts of fill, and confusable letters.

Table 3. Mean Accuracy (percent correct) and Latency (seconds) for Varying Numbers of Points per Stratum

Symbol type	Accuracy		Latency	
	10 points	30 points	10 points	30 points
Colors	77	93	9.1	8.9
HTX	75	89	11.4	11.8
HEF	71	88	13.2	14.8

5.1 Method

Processing time can be controlled by means of a response-signal method, in which the stimulus is shown for a predetermined amount of time and subjects must respond immediately after—but not before—a response signal, regardless of how confident they feel about their decision. We refer to this procedure as *restricted*, in contrast to the *unrestricted* method of Experiment 1.

In Experiment 2, scatterplots were shown for a brief time and the response signal was the disappearance of the stimulus. Responses that trailed stimulus disappearance by more than 1.5 seconds were considered errors. Three expert and three novice subjects participated in 10 sessions of 160 trials each. The first and last sessions were unrestricted and thus provided a partial replication of the previous experiment. Sessions contained a subject-paced break after 80 trials and were variously scheduled between 6 and 48 hours apart.

Orthogonal to the four symbol types, two (*AB* and *AC*) of the possible three comparisons were used in all sessions. The two unrestricted sessions thus formed a  $2 \times 4$  arrangement that was replicated across 20 different displays per session. For the restricted sessions (2 through 9), the design was a  $2 \times 4 \times 4$  arrangement, where the third factor was the duration of stimulus presentation (1, 2, 4, or 8 seconds). Since a single session involved only 160 trials, the 32 cells of the design were replicated across pairs of sessions, and 10 different scatterplots were used in each pair. There were four pairs of restricted sessions (sessions 2 and 3, 4 and 5, 6 and 7, and 8 and 9) altogether, yielding a total of 40 replications per subject per cell.

5.2 Results

The first unrestricted session of Experiment 2 represents a partial replication of Experiment 1. The mean accuracy and latency scores for symbol type for both groups of subjects are presented in Table 4. With performance averaged across symbol type, experts were more accurate than novices (86% vs. 83%) and responded more slowly than novices (12.5 vs. 11.4 seconds). Thus the first unrestricted session replicated the expert–novice difference of Experiment 1. This pattern was not maintained in the last unrestricted session, where the novices were less accurate than they had been initially and during the restricted sessions. Presumably, the novices responded prematurely in the last session, when it was possible, because they were eager to finish the experiment after some 1,600 trials.

Figure 2 shows the empirical SAT functions, obtained

Table 4. Mean Accuracy (percent correct) and Latency (seconds) for the First Unrestricted Session in Experiment 2

Symbol type	Accuracy		Latency	
	Experts	Novices	Experts	Novices
Shapes	88	81	11.8	10.8
Fill	87	83	11.3	12.3
Colors	83	88	9.3	9.9
HEF	85	78	17.4	12.5

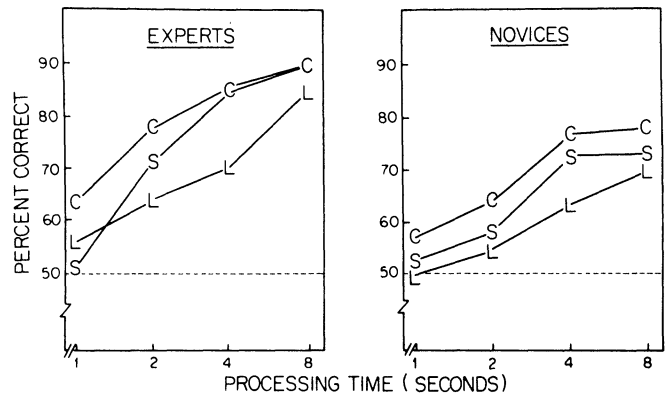


Figure 2. Accuracy as a Function of Processing Time for Experts and Novices: C denotes color; S denotes shapes and amounts of fill; L denotes confusable letters.

from the restricted sessions, for different symbol types for experts (left panel) and novices (right panel). Since performance hardly differed between shapes and amounts of fill, the averaged data for those two conditions are shown by the points labeled *S*. The empirical SAT functions show that experts were more accurate than novices at all processing times, for all symbol types. Accuracy differences were also observed between the different symbol types, with colors clearly superior to all others, and confusable letters producing the worst performance. Shapes and amounts of fill were associated with an intermediate degree of accuracy. Furthermore, in agreement with the results of Experiment 1, performance for all symbol types was at about the same level of accuracy at the longest processing time, and it was similar to that obtained in the unrestricted sessions.

6. VIEWING CONDITIONS

The foregoing experiments showed that, unless processing time is restricted, symbol type does not affect the accuracy of discrimination. At first glance this might appear to suggest that the choice of symbol type is unimportant when designing scatterplots with multiple strata for publication. Unrestricted processing time, however, may be an ideal that is rarely approached in practice: In many contexts, we read and study under the pressure of deadlines and devote minimal time to the examination of accompanying graphs. Moreover, in some conditions—for example, when looking at a slide during a lecture—viewing time is under the control of another. Also, in real life, graphs are often viewed under conditions where other, concurrent, demands are made on the perceptual and cognitive systems. For many reasons, viewing times in real life are seldom truly unrestricted, and therefore the restricted conditions of Experiment 2 may be more relevant to practical application than first appears.

7. GENERAL DISCUSSION

Two aspects of our results deserve discussion: the performance differences between experts and novices, and the differences in accuracy between symbol types under restricted processing time. Based on our results, we can also give some practical advice for choosing symbols.

## 7.1 Experts and Novices

We offered three possible explanations for the expert–novice difference observed in Experiment 1: Experts either were simply more careful than novices (criterion explanation), or were indeed more skillful (skill explanation), or differed in the rate of processing (rate explanation). Experiment 2 clearly ruled out the criterion or rate explanations and endorsed the skill explanation. Experts discriminate strata in scatterplots more accurately than novices, even if the experts are forced to respond before they feel confident of their judgment. If given the choice, however, experts will use a more conservative decision criterion than the novices. Thus the novices, already disadvantaged by their inferior skills, respond sooner than the experts and thereby further compromise their level of performance. It is also remarkable that novices cannot overcome their disadvantage even after some 1,600 experimental trials.

## 7.2 The Best Symbol Type

Since we have suggested that viewing time is often limited, even when nominally unrestricted, the results of Experiment 2 are particularly relevant to the choice of symbol type. Except possibly at the longest processing time, colors produced the best performance, followed by shapes, amounts of fill, and finally confusable letters. If technically possible, therefore, different strata in scatterplots should be represented by symbols of different colors.

If color is not available, the results suggest that either shapes, amounts of fill, or discriminable letters may be used with no great loss in accuracy. This is interesting in light of Cleveland and McGill's (1984b) conjecture that amounts of fill should yield better discrimination than different shapes, and that letters should not be used to code strata. We found that strata represented by the (confusable) letters *H*, *E*, and *F* were indeed difficult to discriminate, but when the letters *H*, *Q*, and *X* were used instead, performance was no worse than for shapes. Consequently, one must distinguish between confusable and discriminable sets of letters. The former should not be used, but the latter—provided that care is taken in their selection—are comparable to either shapes or amounts of fill.

Discriminable letters have one inherent advantage over all other symbol types because they provide an immediate mnemonic. For example, if *M* and *F* were used to code strata for males and females, respectively, no legend would be required to explain the symbols. If, on the other hand, squares and triangles were used, the observer would have to memorize two associations (triangles: males; squares: females) without any appreciable gain in discriminability. Therefore, if color cannot be used, and if discriminable letters for the different groups can be found, letters may be the best choice.

It is important to realize that performance differences between the different symbol types are not trivial. For example, if we compare the performance of experts and novices in Experiment 2, we see that the use of color boosts the novices to a level above that of the experts with con-

fusable letters. In that case, the deficit of being a novice may be partially removed by use of the proper symbol type.

## 7.3 Using Color to Code Strata

Several considerations should be borne in mind if color is to be used. Color is a *nonlinear* function of the wavelength or frequency of light so that although reds and blues are at opposite ends of the physical continuum, psychologically they are fairly similar. We may easily confuse a deep purple-red with a blue-violet. Perception of color turns out to be better described by a circle than by a line, which fits the observed data exceedingly well [see Sekuler and Blake (1985) for an elementary introduction]. It is not necessary to be terribly exact about describing the relative similarity of colors, and any color circle (e.g., in Sekuler and Blake 1985) should be sufficient for choosing colors that are fairly discriminable in practical situations. When choosing colors for a display, a data analyst should contrive to arrange that the chosen colors be fairly far apart in the two-dimensional representation. Of course, as more symbols are required, when there are many strata, this becomes increasingly difficult to arrange, but then one must question whether any form of scatterplot display will be useful.

Although we have advocated the use of color wherever possible, it is necessary to remember that more than 8% of males and fewer than 1% of females have some difficulty distinguishing color. The most common deficiency is a red–green deficiency, and in our experiments we encountered three males with this deficiency. Their mean accuracies (not included in the overall analysis) were 82% for colors (recall that we used red, green, and yellow), 86% for discriminable letters, and 78% for confusable letters. Thus these subjects are still able to perform the task quite well, but their accuracy with discriminable letters exceeds their accuracy with colors. Their mean latencies are similarly revealing: 30.8 seconds (colors); 15.0 seconds (discriminable letters); 17.6 seconds (confusable letters). These subjects take twice as long with colors as with letters, and they take much longer with colors than normal subjects. Thus it is clear that for subjects with color vision defects (approximately 5% of the population), the use of color to code strata may be far from optimal, especially if red and green are used.

## 7.4 Using Letters to Code Strata

The ability of human subjects to discriminate and identify individual alphabetic letters has been extensively studied by psychologists. It is clear that some letters are more easily confused than others—for example, did you notice the substitution in the fifth word of this sentence? Although these confusions depend, to some extent, on characteristics such as font size and typeface, there is a consistent pattern of confusability between distinct pairs of letters. Several information-processing models have been proposed to account for the differential confusability of letters, and although there is some disagreement on whether

Table 5. Features in Common for Letters of the Alphabet (after Geyer and DeWald 1973)

Feature	Letter																									
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
External																										
Horizontal					2	1					1									1						2
Vertical		1		1	1	1		2	1	1	1	1	2	2		1		1		1	2					
Slant (/)	1																					1	1		1	
Slant (\)	1																						1	1		1
Convex segment		2	3	2			3			1					4	1	4	1	2		1					
Open																										
Horizontal				1			1													2						
Vertical										1												1				
Horizontal wedge		1									1								1						2	1
Vertical wedge											2		1	2					1			1	1	2	1	2
Internal protrusion												1											1			
Internal intersection	2	1				1	1			2		1							1							
Horizontal bar	1					1	1	1	1																	
Crossing-bar slant																	1									
Vertical symmetry	1								1	1			1		1					1	1	1	1	1	1	1
Horizontal symmetry		1	1	1	1				1	1		1							1							1

humans use “features” (Geyer and DeWald 1973) or the spatial frequency content of letters (Harvey, Roberts, and Gervais 1983) as the basis for internal representations, models based on both approaches fit the data quite well. Feature models propose that the visual system detects discrete features (lines, edges, and contours in varying orientation) of the stimulus. Thus letters that have more features in common are more likely to be confused. Conversely, letters with few common features will rarely be mistaken for each other. We suggest that data analysts make use of Table 5 (adapted from Geyer and DeWald 1973), in which each letter is characterized by whether it possesses any of 15 different features. These features are classified into two groups labeled *external* and *open*. The former refers to features associated with the periphery of the letter—for example, an *A* has no external horizontal line, but a *T* does. Open features include both internal features of the letters and the orientations of concave features. For example, the letters *H* and *Q* share no features in common and neither do *Q* and *X*, whereas *H* and *X* share only two open features, namely vertical and horizontal symmetry. On the other hand, *H* and *E* share four common features, as do *F* and *E*. Thus it seems reasonable to assume that *H*, *Q*, and *X* form a relatively distinct set, compared to *H*, *F*, and *E*. Of course, the experiments in this article support that supposition.

## 8. CONCLUSION

The design of statistical graphs may be aided by behavioral research. When the performance of observers of statistical graphs is examined, it is desirable to measure not only accuracy but also response latency. Failure to do so

may result in ambiguous or uninterpretable results. When different strata are represented in a scatterplot, the use of color to code different groups is advised. If color is not available, shapes, amounts of fill, or letters—provided they are highly discriminable—may be used with little loss in accuracy. Letters have the advantage of providing a mnemonic for the strata in the plot.

[Received June 1987. Revised December 1988.]

## REFERENCES

- Chen, L. (1982), “Topological Structure in Visual Perception,” *Science*, 218, 699–700.
- Cleveland, W. S., and McGill, R. (1984a), “Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods,” *Journal of the American Statistical Association*, 79, 531–553.
- (1984b), “The Many Faces of a Scatterplot,” *Journal of the American Statistical Association*, 79, 807–822.
- (1985), “Graphical Perception and Graphical Methods for Analyzing Scientific Data,” *Science*, 229, 828–833.
- Cox, D. R. (1978), “Some Remarks on the Role in Statistics of Graphical Methods,” *Applied Statistics*, 27, 4–9.
- Geyer, L. H., and DeWald, C. G. (1973), “Feature Lists and Confusion Matrices,” *Perception and Psychophysics*, 14, 471–482.
- Hardy, L. H., Rand, G., and Rittler, M. C. (1957), *AO H-R-R Pseudoisochromatic Plates*, New York: American Optical Company.
- Harvey, L. O., Roberts, J. O., and Gervais, M. J. (1983), “The Spatial Frequency Basis of Internal Representations,” in *Modern Issues in Perception*, eds. H.-G. Geissler, H. F. M. Buffart, E. L. J. Leeuwenberg, and V. Sarris, Rotterdam: North-Holland, pp. 217–226.
- Jacob, R. J. K. (1981), “Comment on Trees and Castles,” *Journal of the American Statistical Association*, 76, 270–272.
- Kruskal, W. H. (1975), “Visions of Maps and Graphs,” in *Auto-Carto II: Proceedings of the International Symposium on Computer Assisted Cartography*, ed. J. Kavaliunas, Washington, DC: U.S. Bureau of the Census & American Congress on Survey and Mapping, pp. 27–36.
- Playfair, W. (1786), *The Commercial and Political Atlas*, London: Corry.
- Sekuler, R., and Blake, R. (1985), *Perception*, New York: Knopf.