ASSESSING THE EFFICACY OF VISUAL DISPLAYS

Howard Wainer and Mark Reiser, The University of Chicago

Abstract

In an effort to measure the efficacy of several types of graphical displays, an experiment was performed in which a question was asked of the subject and his/her response time was measured. It was felt that any unambiguous display would allow the correct response eventually, and so response time seemed a natural dependent variable. The results indicated that for the type of question asked there was a definite order of preference with a tabular representation finishing last. This experiment is used as an exemplar of the difficulties involved in the empirical study of this problem area.

Introduction

On the surface the question, "Which of these two displays is better?" appears to be a perfectly reasonable one. Moreover, it initially appears that one could answer it using the traditional scaling methods that psychometricians have been using since the time of Thurstone. Sadly, a unique answer to this question is as difficult to determine as the answer to the question, "Which of two estimators is better?" In both cases there is no unique answer; it depends upon the situation. More precisely, one must specify the question in more detail before an answer can be found. Thus we must not ask which of two displays is better, but rather which is better for yielding the answer to a particular question. When we have reached the point where we know precisely what information we want to represent in a display we can easily test various candidates as to the clarity with which they convey that information. Previously we studied (Wainer, 1974) the efficacy of hanging histograms which have been proposed by John Tukey (Tukey, 1977) as an improvement on standard histograms. In that study we used a variant of Fechner's method of paired comparisons which is generally called (Bock & Jones, 1968) the constant method. In this task a series of stimuli are paired, one at a time, with a constant one, and the subjects are required to judge which member of the pair is x'er, where x is the underlying dimension of interest. The various stimuli are then arranged in the order of the general frequency in which they were preferred: the one which was preferred to the constant most frequently is considered the highest, and the one over which the constant was preferred most frequently is the lowest. The proportion of times a particular stimuli is preferred is converted to a scale value through an inverse normal transformation and the scale values are (through the intervention of a Thurstonian scaling model) considered to be intervally scaled. In the particular problem for which this technique was employed, the scale value of any of the various displays would reflect the extent to which that display emphasized the dimension x on which the subjects were judging. If that dimension was utterly lost in the display, then a plot of the physical values of the dimension against the subjective scale values would have a zero slope. If it was perfectly displayed the slope would be very steep indeed. Thus one could compare two (or more) display types on the basis of their slopes. Moreover, one could also parameterize each display type by its slope for that dimension. Thus a future data displayer could determine what display would best suit his needs by searching through a list which would detail the specifications of each display.

Two shortcomings of such an investigative method is that the experiment is rather tedious, and one may get slopes which are not uniform across the entire dimension. Furthermore, some questions which one might ask of a display do not lend themselves to this sort of scheme.

Another method for evaluating the efficacy of a display would be to determine the amount of time that it takes a person to extract a particular bit of information from the display. This seems like a reasonable way to go about this problem since any display should, at the very least, be unambiguous, allowing the average reader to extract the information contained in it. It seems reasonable to maintain that a better display will allow faster extraction.

Some months ago we had the good fortune to participate in a workshop on display techniques during which we looked carefully at the displays used in Social Indicators: 1973. After much complaining about some of the displays we tried to come up with alternatives which would be better. One display which suffered grievously at our hands was Chart 2/9 which is generally called a "bar chart," and which I leave to you to judge on the

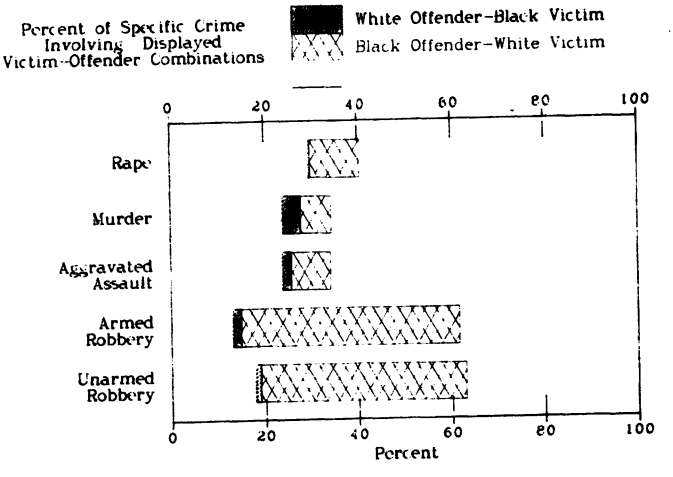

FIGURE 1. Bar chart showing race of victim and offender, by type of violent crime: 1967 (for 17 major cities).

face of it whether it deserved our derision. Stephen Fienberg conceived of one alternative which he calls a "Floating Four-Fold Circular Display" (FCD), a sample of which is displayed in Figure 2. I contributed my own version of 2/9

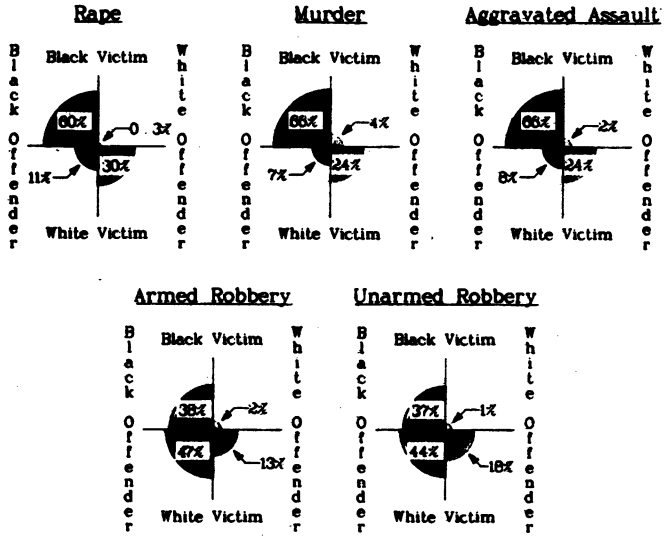

FIGURE 2. Floating four-fold contingency display showing race of victim and offender, by type of violent crime: 1967 (for 17 major cities).

which, for lack of a better name, I have denoted "cartesian rectangles." All three displays contain the same information, but each emphasizes a

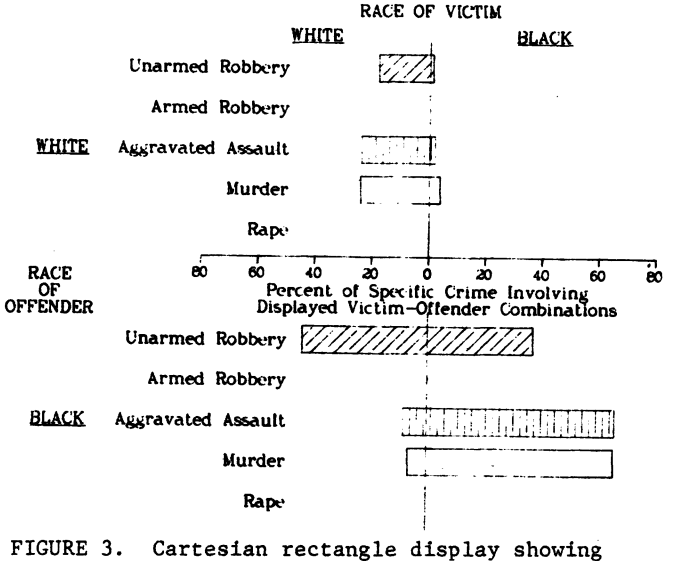

FIGURE 3. Cartesian rectangle display showing race of victim and offender, by type of violent crime: 1967 (for 17 major cities).

particular aspect of it somewhat differently. Clearly, to thoroughly test the efficacy of all these displays one would have to try a variety of different questions. Resource limitations prevented this, so the experiment which we shall describe shortly is limited in its implications, yet the methodology is generalizable. In addition to these three displays, we used as a standard of comparison a tabular display of the same data. It was felt that any saltworthy display should do better than a table of numbers.

(Percent)	White Victim		Black Victim	
Crime	White Offender	Black Offender	White Offender	Black Offender
Murder and nonnegligent manslaughter	24.0	6.5	3.8	65.7
Aggravated Assault	23.9	8.4	1.8	65.9
Forcible Rape	29.6	10.5	0.3	59.6
Armed Robbery	13.2	46.7	1.7	38.4
Unarmed Robbery	17.9	43.9	1.1	37.1

TABLE I. Race of victim and offender, by type of violent crime: 1967 (for 17 major cities).

The experiment

Sixteen right-handed students at the University of Chicago participated in the first phase of the experiment. An assertive statement was presented to the subject, and then followed by one of the displays. Each statement took the form: "In the crime of armed robbery (rape, aggravated assault), white (black) criminals victimize whites (blacks) more often than they victimize blacks (whites)." The subject's task was to decide whether the statement was true or false, based on the information in the display. Subjects were to indicate their response by pressing one button for true and another one for false.

Before the experiment, subjects were told that some of the displays represented fictitious data, so they would not be able to respond based on any previous knowledge. In fact, two sets of data were used, one real and one fictitious. Two displays of each type were made, one from each set of data. On the odd numbered trials the subjects were presented with a statement and then a display from the veritical data set. On the even numbered trials, subjects were presented with the same statement and display type as in the preceding odd numbered trial, but the display portrayed the fictitious data instead. Thus, the odd numbered trials could be considered practice or training trials, although the subjects were told that there were no practice trials. Four statements were used; thus eight trials were required for each subject. Across subjects, each question was paired with each display, and question-display pairs were balanced for order of presentation. All statements and displays were presented tachistoscopically, and responses were timed electronically. The dependent variable was response time, although response speed (1/time) was also calculated. In addition, when the presentation trials were completed, each subject was asked to order

the displays from the one which he thought was easiest to use to the one that he thought was hardest.

The results

Table II represents the results of the experiment. Note that the means yield a very peculiar artifact in that on the second trial the

TABLE II

Summary of Results

Untransformed Data

	Mean	Midmean	Standard Error
First Trial			
Rectangles	24.46	23.12	3.55
Bar chart	20.96	18.79	2.42
FCD	31.82	25.40	5.97
Table	28.74	22.20	5.99
Second Trial			
Rectangles	11.28	10.51	1.08
Bar chart	11.59	11.01	1.53
FCD	17.37	14.35	3.80
Table	16.55	16.61	1.61

Neg. Inverse Transformed

First Trial			
Rectangles	-.0512	-.0444	.006
Bar chart	-.0553	-.0537	.005
FCD	-.0450	-.0409	.006
Table	-.0491	-.0460	.006
Second Trial			
Rectangles	-.1025	-.0971	.01
Bar chart	-.1093	-.0915	.015
FCD	-.1084	-.0732	.022
Table	-.0696	-.0612	.007

	Judged Preference (S.E.)
Rectangles	2.1 (.95)
Bar chart	2.2 (.83)
FCD	3.1 (1.1)
Table	2.5 (1.1)

mean response time for the table is 16.55 seconds and for the Four-fold Contingency Display 17.37, thus indicating that the FCD is worse than the table of numbers. However, after an inverse transformation to speed we see that the table's mean speed is .0696 sec^{-1} and the FCD's is .1093 sec^{-1}. A reversal! This is caused by some outliers in the FCD. A truer picture of what is happening is seen in the mid-means (25% trimmed means) which still shows the cartesian rectangles as the display of choice (for this kind of question) followed closely by the bar charts. Next we have the FCD, while the table of numbers brings up the rear. Although the order of displays for the second trial is not the same as in the first trial, the display occasion effect was not significant as can be seen in Table III. It seems that the data from the second trial should be the more reasonable one from which to draw conclusions, since it represents subjects' performance after some practice.

The judged preferences in Table II were obtained by averaging the subjects' orderings of the graphs, where 1 indicated that the subject thought that the graph was easiest to use and 4 indicated that the subject thought the graph was hardest to use. The subjects' preferences seem to agree, in spirit, with the results from the analysis of time. The FCD was judged hardest to use, but this judgment may reflect unfamiliarity more than difficulty of use.

TABLE III

Analysis of Variance

Analysis of Variance for Reaction Time

Source	Sum of Squares	DF	Probability
Grand Mean	52992	1	.0001
Display	1472	3	.0255
Occasion	4839	1	.0001
Subjects	10393	15	.0001
Display X Occasion	112	3	.8641
Error	15957	105	
Total	85767	128	

Analysis of Variance for Speed

Source	Sum of Squares	DF	Probability
Grand Mean	.6954	1	.0001
Display	.0096	3	.1179
Occasion	.0720	1	.0001
Subjects	.0780	15	.0002
Display X Occasion	.0084	3	.1597
Error	.17	105	
Total	1.03	128	

Table III gives a summary of the analysis. Although the effect of display is significant when reaction time is the dependent variable, unfortunately, it is not significant when the dependent variable is transformed to speed. The change is due, no doubt, to large outliers in the

data from the first trials. The most likely remedy is to obtain more data in order to yield more stable results, i.e., smaller error variance.

Discussion

The type of question was formulated to give an advantage to the bar chart, since a more quantitative question (e.g., "True or False--16.5% of all rapes are Whites on Blacks.") would be far easier to answer with the other displays wherein these numbers can be read directly rather than obtained through subtraction of two points on the x-axis. For this type of question the table might do better still.

Even with this edge the bar chart did not run away from the competition. We thus conclude that even for this situation (that is, the one in which it seems best suited) the bar chart is not the easiest to use, and other display types are to be preferred. It is interesting to note further that the bar chart was apparently the most familiar display to the subjects since in the training trial they responded most rapidly to it. Even with this advantage it still did not win. We feel that the more innovative display types would do even better with a more extensive training period.

It seems to us that a catalogue of display types could be prepared (much as Cal Schmid has done in his Handbook of Graphic Presentation) which would not only include categorizations of various displays but also some sort of parameterization indicating how good each display type is for each of a variety of purposes. The prospective user could then reach into this bag and pull out the one which most nearly fills all of his needs. This has two interesting sidelights. First, it implies that a great deal of empirical work of the sort we have tried to illustrate must be done, although one would hope that it would be done with greater experimental cleverness than we were able to muster. Second, it places an additional load on the prospective displayer to explicitly determine what particular aspect of his data he is most interested in emphasizing. The displayer's emphasis can be determined by his readers by checking back to the gedanken handbook mentioned previously to see what aspect of the data the display of choice was supposed to emphasize.

We believe that investigations using such methods as multi-dimentional scaling could be very useful in determining the perceptual dimensions that are involved in the perception of a display. This would facilitate the development of a theoretical structure of display construction. In addition, it may be that different kinds of audiences respond differentially to different kinds of displays; thus the individual differences models for scaling (both uni- and multi-dimensional sorts may be very helpful.

We are pursuing this currently, but we're insufficiently far along to be able to report on it at this meeting. Perhaps later. We should not end without a short comment on general graphics with no particular purpose in mind. A picture is probably the best way of finding something for which one is not explicitly looking, and so it may be that there are some general purpose display techniques which are not as good as a special purpose display for a particular question but yet are all around good performers--a graphical equivalent to the Jackknife. Whereas the all around good performers should be treasured for exploration, special purpose displays seem to be required for investigation in depth. Just as the Jackknife can drive screws (or test regression coefficients) a special tool can do each of these special jobs better. Thus, if a special job is at hand one should use the right tool, but if there is a family of jobs, one should use the most efficacious general tool. Just which is which can only be obtained through careful empirical work.

References

Bock, R. D., & Jones, L. V. The measurement and prediction of judgment and choice. San Francisco: Holden-Day, 1968.

Schmid, C. F. Handbook of Graphic Presentation. New York: The Ronald Press Company, 1954.

Tukey, J. W. Exploratory data analysis. Reading, Mass.: Addison-Wesley, 1977.

Wainer, H. The suspended rootogram and other visual displays: An empirical validation. The American Statistician, 1974, 28, 143-145.