Improving Data

In the transactions between scientists and the media, influence flows in both directions. About 25 years ago, I wrote an oft-cited article with the ironic title, "How To Display Data Badly." In it, I chose a dozen or so examples of flawed displays and suggested paths toward improvement. Two major newspapers, *The New York Times* and the *Washington Post,* were the source of most of my examples, which were drawn over a remarkably short period of time. It wasn't hard to find examples of bad graphs.

Happily, in the intervening years, those same papers have become increasingly aware of the canons of good practice and improved their data displays profoundly. Indeed, when one considers both the complexity of the data often displayed and the short time intervals permitted for their preparation, the results are frequently remarkable.

Recently, I picked out a few especially notable graphs from *The New York Times*. Over the same time period, I noticed graphs in the scientific literature whose data had the same features, but were decidedly inferior. At first, I thought it felt more comfortable in the 'good old days' when we did it right and the media's results were flawed. But, in fact, the old days were not so good. Graphical practices in scientific journals have not evolved as fast as those of the mass media. It is time we learned from their example.

Example 1: Pies

The U.S. federal government is fond of producing pie charts, and so it came as no surprise to see Figure 1, a pie chart of government receipts broken down by their source. Of course, the person who created the chart felt it necessary to 'enliven' the presentation with the addition of a specious extra dimension. Obviously, the point of presenting the results for both 2000 and 2007 together must have been to allow the viewer to see the changes that have taken place over that time period. The only feature of change I was able to discern was a shrinkage in the contribution of individual income taxes.

I replotted the data in a format that provides a clearer view (see Figure 2) and immediately saw the diminution of taxes was offset by an increase in social insurance payments. In other words, the cost of tax cuts aimed principally at the wealthy was paid for by increasing social security taxes, whose effect ends after the first hundred thousand dollars of earned income.

I wasn't surprised that such details were obscured in the original display, for my expectation of data displays constructed for broad consumption was not high. Hence, when I was told of a graph in an article in *The New York Times Magazine*

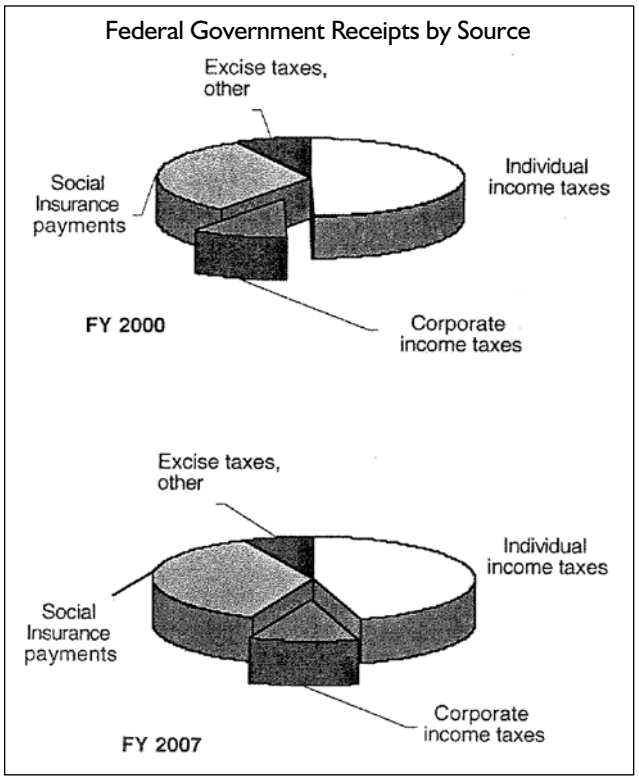


Figure 1. A typical three-dimensional pie chart. This one, no worse and no better than most of its ilk, is constructed from U.S. government budget data.

Graph courtesy of the Office of Management and Budget, *www.whitehouse.gov/omb/budget/fy2007*.

Howard Wainer,
Column Editor

Displays: Ours and the Media's

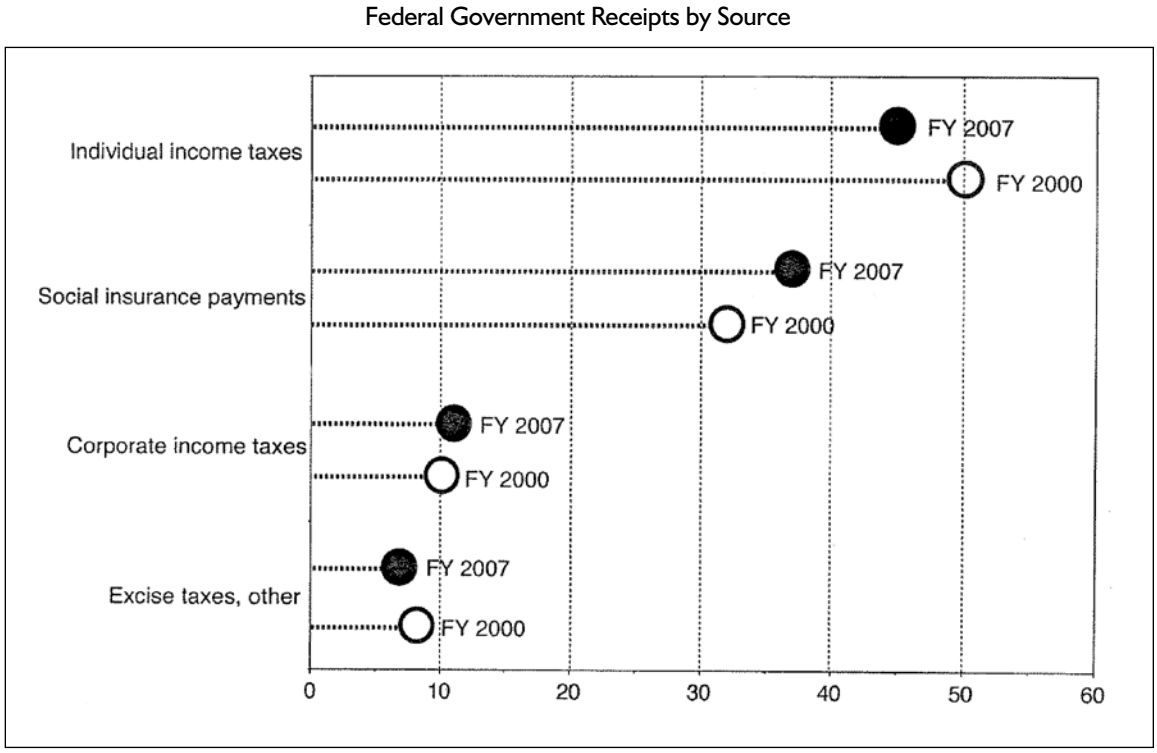


Figure 2. A restructuring of the same data shown in Figure 1, making clearer what changed in the sources of government receipts between 2000 and 2007

about the topics on which American clergy choose to speak out, I anticipated the worst. My imagination created, floating before my eyes, a pie chart displaying such data (see Figure 3). My imagination can only stretch so far, hence the pie I saw had but two dimensions and the categories were ordered by size.

What I found (Figure 4, from Rosen, 2007) was a pleasant surprise. The graph (produced by an organization named Catalogtree) was a variant of a pie chart in which the segments all subtend the same central angle, but their radii are proportional to the amount being displayed. This is a striking improvement on the usual pie because it facilitates comparisons among a set of such displays representing, for example, different years or different places. The results from each segment are always in the same place; whereas, the locations of all segments vary as the data change with pie charts. This plot also indicates enough historical consciousness to evoke Florence Nightingale's famous Rose of the Crimean War (see Figure 5).

Sadly, this elegant display contained one small flaw that distorts our perceptions. The length of the radius of each segment was proportional to the percentage depicted, but we are influenced by the area of the segment—not its radius. Thus, the radii need to be proportional to the square root of the percentage for the areas to be perceived correctly. An alternative figuration with this characteristic is shown in Figure 6.

Example 2: Line Labels

In 1973, Jacques Bertin—the maitre de graphique moderne—explained that when one produces a graph, it is best to label each of the elements in the graph directly. He proposed this as the preferred alternative to appending some sort of legend that defines each element. His point was that when the two are connected, you can comprehend the graph in a single moment of perception, as opposed to having to first look at the lines, then read the legend, and then match the legend to the lines.

Despite the authoritative source, this advice is too rarely followed. For example, Mark Reckase, in a rejoinder he wrote for publication in *Educational Measurement: Issues and Practice,* chose not to label the lines directly in a simple plot of two lines (see Figure 7)—even though there was plenty of room to do so—but instead chose to put in a legend. And the legend reverses the order of the lines, so the top line in the graph becomes the bottom line in the legend, thus increasing the opportunities for reader error.

Can the practice of labeling be made still worse? In Figure 8, from a *Journal of the American Statistical Association* article written by Danny Pfeffermann and Richard Tiller, comes a valiant effort to do so. Here, the legend is hidden in the figure caption, and again its order is unrelated to the order of the lines in the graph. Moreover, the only way to distinguish BMK from UnBMK is to notice a small dot. The only way I could think of to make the connection between the various graphical elements and their identifiers worse was to move the latter to an appendix.

How does *The New York Times* fare on this aspect of effective display? Very well, indeed. Two plots describing 100 years of employment in New Haven County can be seen in Figure 9.

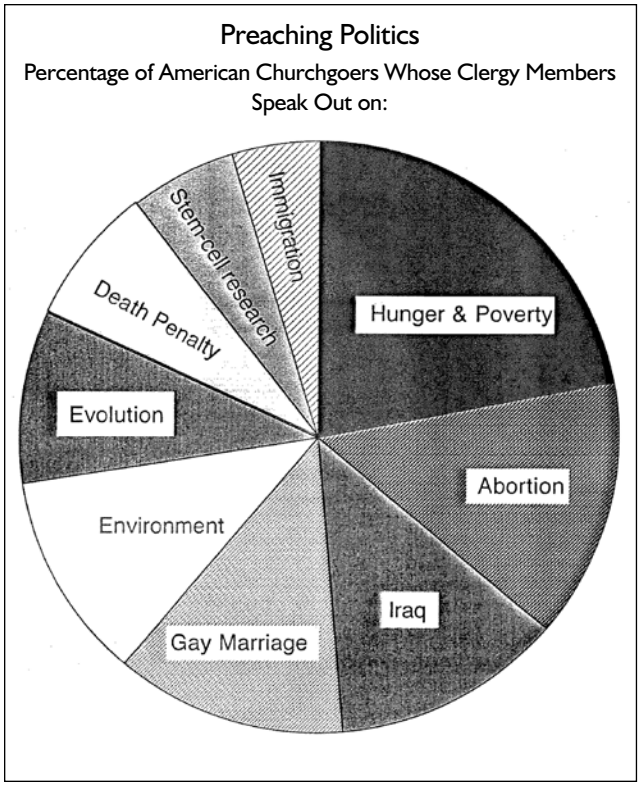


Figure 3. A typical pie chart representation of the relative popularity of various topics among the U.S. clergy

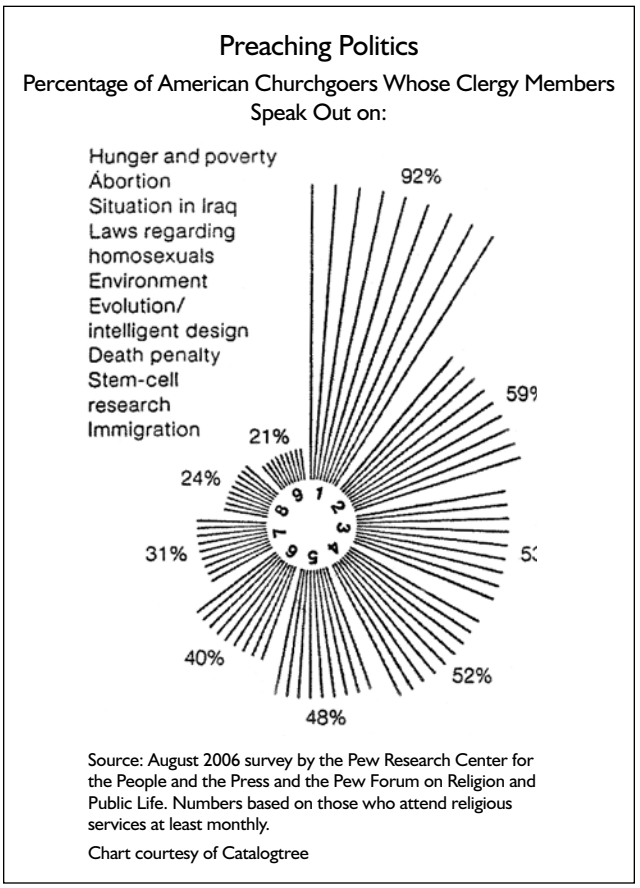


Figure 4. A display from the February 18, 2007, *The New York Times Magazine* (Page 11) showing the same data depicted in Figure 3 as a Nightingale Rose

In each panel, the lines are labeled directly, making the decline of manufacturing jobs clear. In the following week (Figure 10), another graph appeared showing five time series over three decades. In this plot, the long lines and their crossing patterns made it possible for the viewer to confuse one line with another. This possibility was ameliorated by labeling both ends of each line—a fine idea worthy of being copied by those of us whose data share the same characteristics.

Example 3. Channeling Playfair

On the business pages of the March 13, 2007, *Times* was a graph used to support the principal thesis of an article on how the growth of the Chinese economy has yielded, as a concomitant, an enormous expansion of acquisitions. The accompanying graph, shown here as Figure 11, has two data series. The first shows the amount of money spent by China on external acquisitions since 1990. The second time series shows the number of such acquisitions. The display format chosen, while reasonably original in its totality, borrows heavily from William Playfair, who many call the inventor of statistical graphics. First, the idea of including two quite different data series in the same chart is reminiscent of Playfair's 1821 chart comparing the cost of wheat with the salary of a mechanic (Figure 12). However, in plotting China's expenditures, those creating the graph had to confront the vast increases over the time period shown; a linear scale would have obscured the changes early on. The solution they

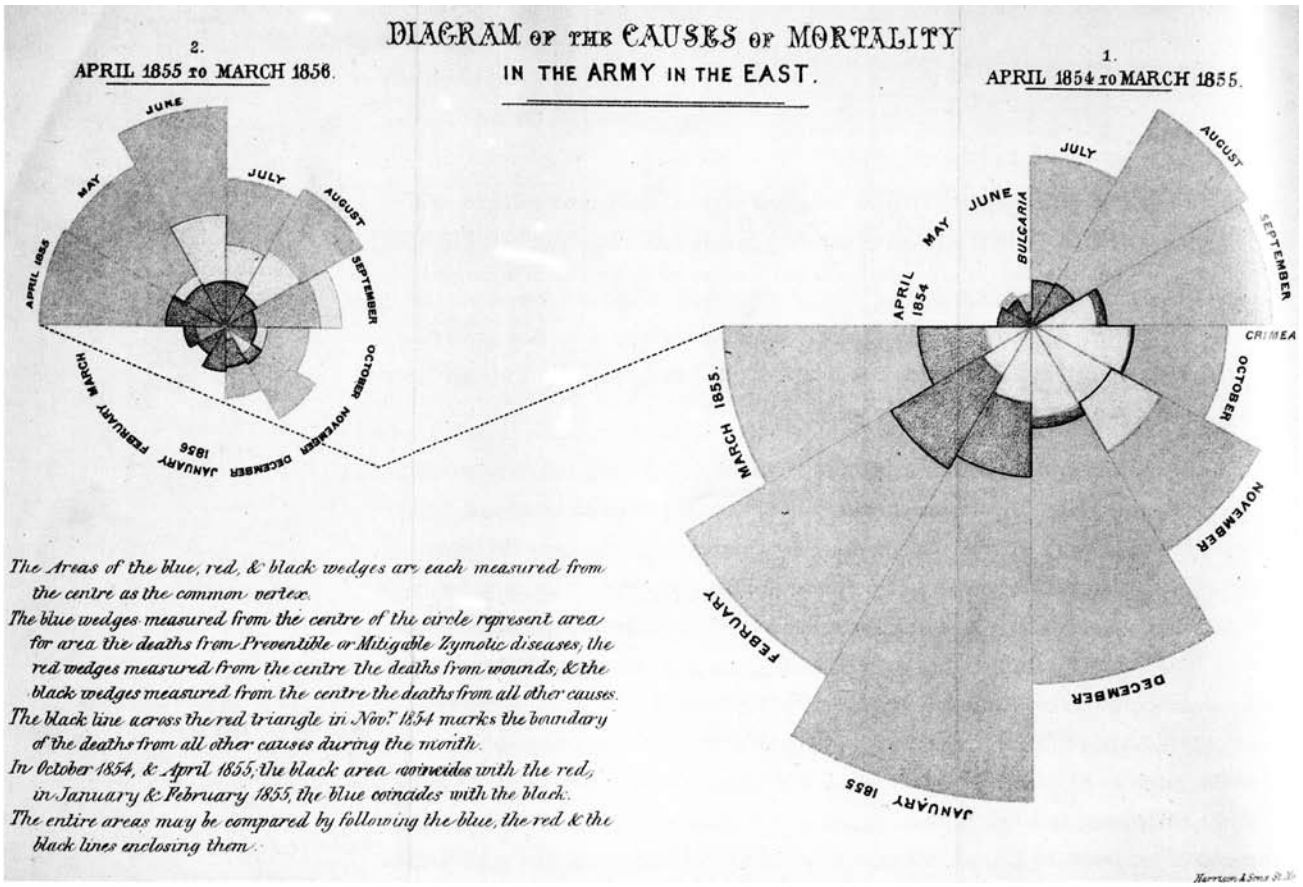


Figure 5. A redrafting of Florence Nightingale's famous 'coxcomb' display (what has since become known as a Nightingale Rose), showing the variation in mortality over the months of the year

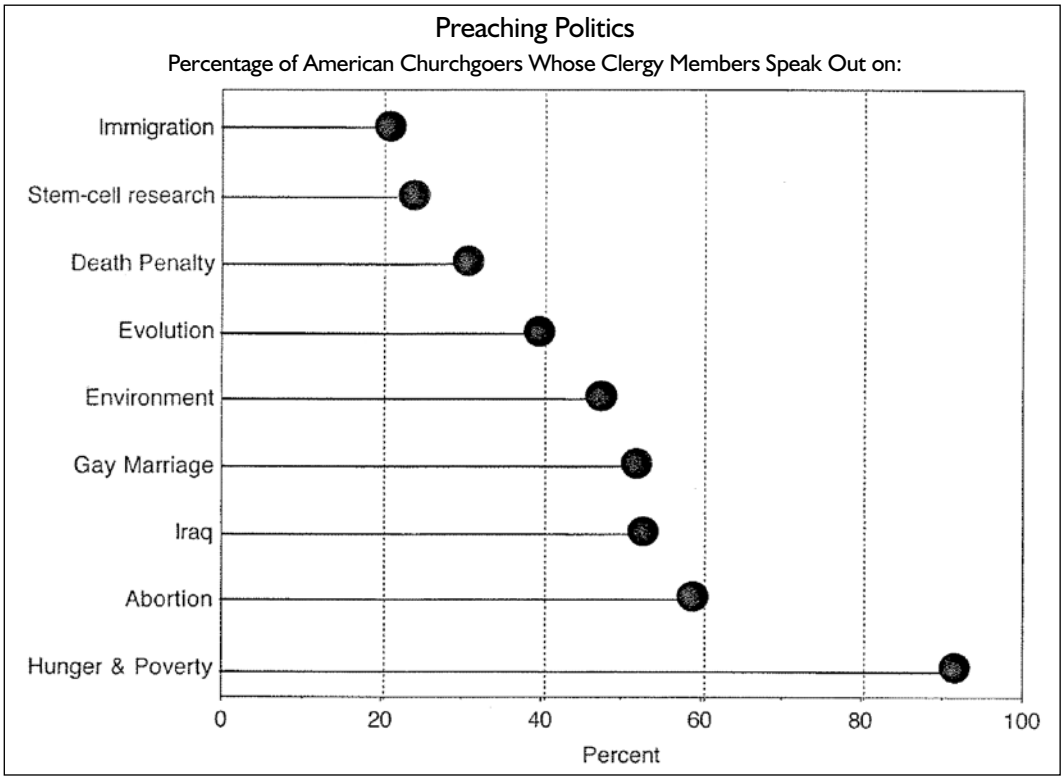


Figure 6. The same data previously shown in Figures 3 and 4, recast as a line-and-dot plot

chose also was borrowed from Playfair's plot of Hindoostan in his 1801 *The Statistical Breviary* (Figure 13). Playfair showed the areas of various parts of Hindoostan as circles. The areas of the circles are proportional to the areas of the segments of the country, but the radii are proportional to the square root of the areas. Thus, by lining up the circles on a common line, we can see the differences of the heights of the circles, which is, in effect, a square root transformation of the areas. This visual transformation helps place diverse data points onto a more reasonable scale.

The *Times'* plot of China's increasing acquisitiveness has two things going for it. It contains 34 data points, which by mass media standards is data rich, showing vividly the concomitant

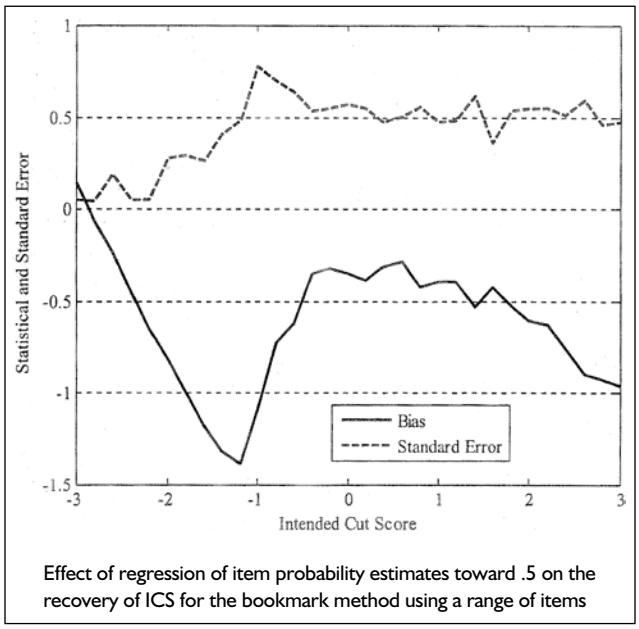


Figure 7. A graph taken from a 2006 article in *Educational Measurement: Issues and Practice* in which lines are identified through a legend—indeed, a legend whose order does not match the data—instead of identifying the graph lines directly

increases in both data series over a 17-year period. This is a long way from Playfair's penchant for showing a century or more, but in the modern world—where changes occur at a less leisurely pace than in the 18th century—17 years is often enough. And second, using Playfair's circle representation allows the visibility of expenditures over a wide scale.

Are there other alternatives that might perform better? Perhaps. In Figure 14 is a two-paneled display in which each panel carries one of the data series. Panel 14a is a straight-forward scatter plot showing the linear increases in the number of acquisitions China has made over the past 17 years. The slope of the fitted line tells us that over those 17 years, China has, on average, increased its acquisitions by 5.5/year. This crucial detail is missing from the sequence of bars, but is obvious from the fitted regression line in the scatter plot. Panel 14b shows the increase in money spent on acquisitions over those same 17 years. The plot is on a log scale, and its overall trend is well-described by a straight line. That line has a slope of 0.12 in the log scale that translates to an increase of about 51% per year. Thus, the trend established over these 17 years shows

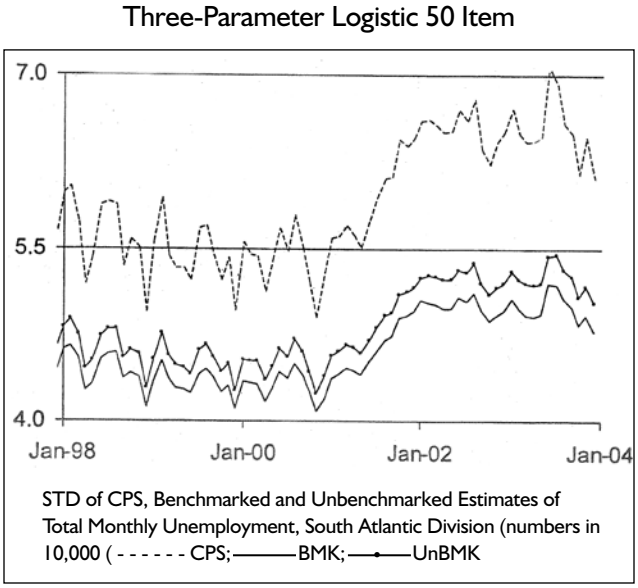


Figure 8. A graph taken from a 2006 article in the *Journal of the American Statistical Association* in which the three data series are identified in acronym form in the caption. There is plenty of room on the plot for them to be identified on the graph with their names written out in full.

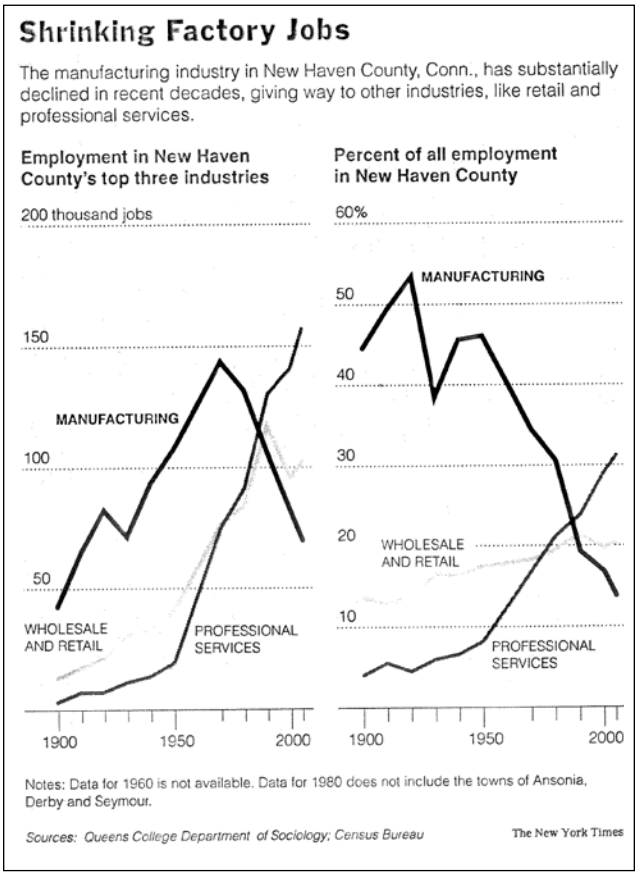


Figure 9. A graph taken from the Metro section of the February 18, 2007, edition of *The New York Times* showing two panels containing three lines each, in which each line is identified directly

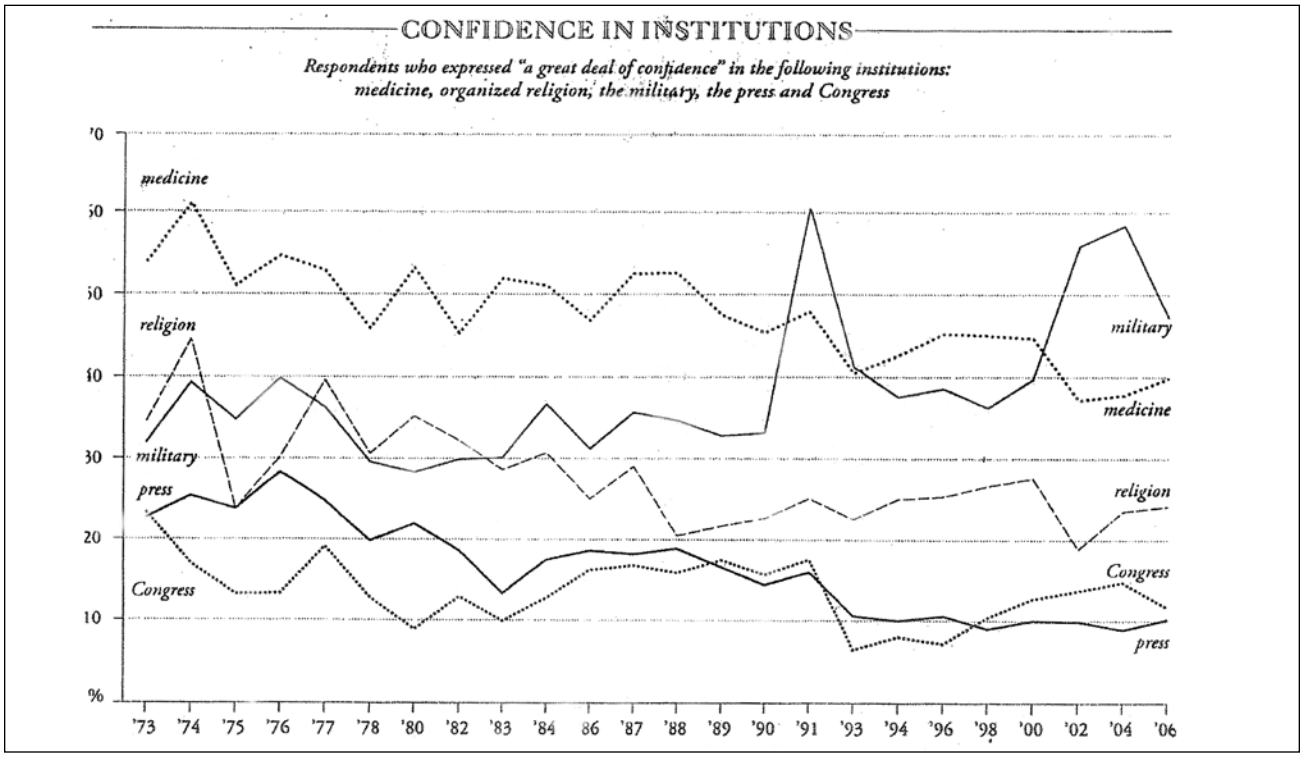


Figure 10. A graph taken from the Op-Ed section of the February 25, 2007, edition of *The New York Times* showing five long lines each, in which each line is identified directly at both of its extremes, thus making identification easy, even when the lines cross. Graph courtesy of Ben Schott; Data courtesy of NORC, General Social Survey (*www.norc.org/projects/General+Social+Survey.htm)*

that China has both increased the number of assets acquired each year and acquired increasingly expensive assets.

The key advantage of using paired scatter plots with linearizing transformations and fitted straight lines is that they provide a quantitative measure of how China's acquisitiveness has changed. This distinguishes it from the *Times* plot, whose primary message was qualitative, although it contained all the quantitative information necessary to do these calculations.

Persistent Practices

William Playfair set a high standard for effective data display more than 200 years ago. Since that time, rules have been codified (e.g., American Society of Mechanical Engineers standards in 1914, 1938, and 1960) and many books have been published that describe and exemplify good graphical practice. All these have had an effect on graphical practice. But it would appear from my sample of convenience, that the effect was larger on the mass media than the scientific literature. I don't know why, but I will postulate two possible reasons. First, because scientists will make graphs with the software they have available and would tend, more often than is proper, to accept whatever are the default options for that software. Producers of displays for large-market mass media have, I suspect, a greater budget and more flexibility. The second reason why poor graphical practices persist is akin to

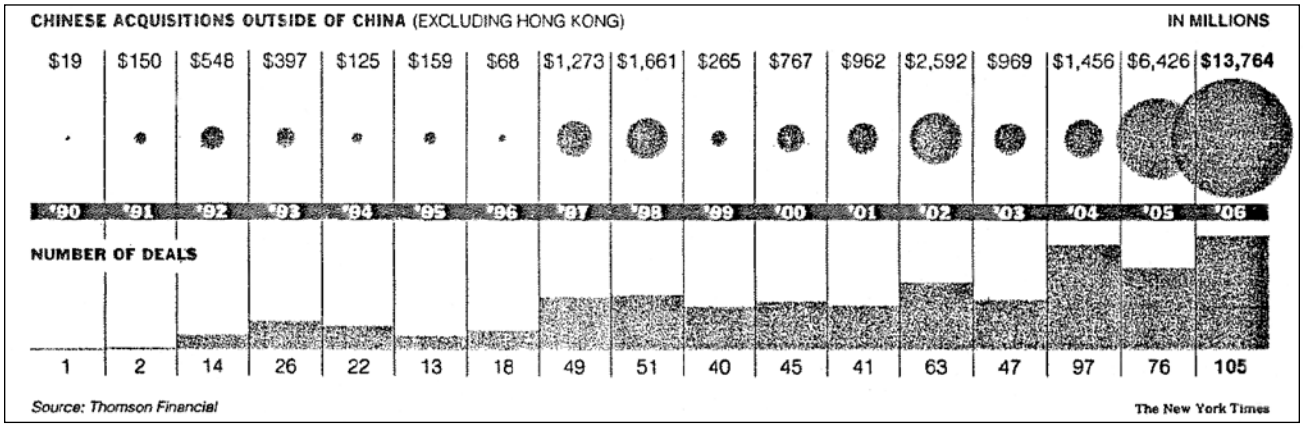


Figure 11. A graph taken from the Business section (Page C1) of the March 13, 2007, edition of *The New York Times* showing two data series. One is of a set of counts represented by bars; the second is money represented by the area of circles.

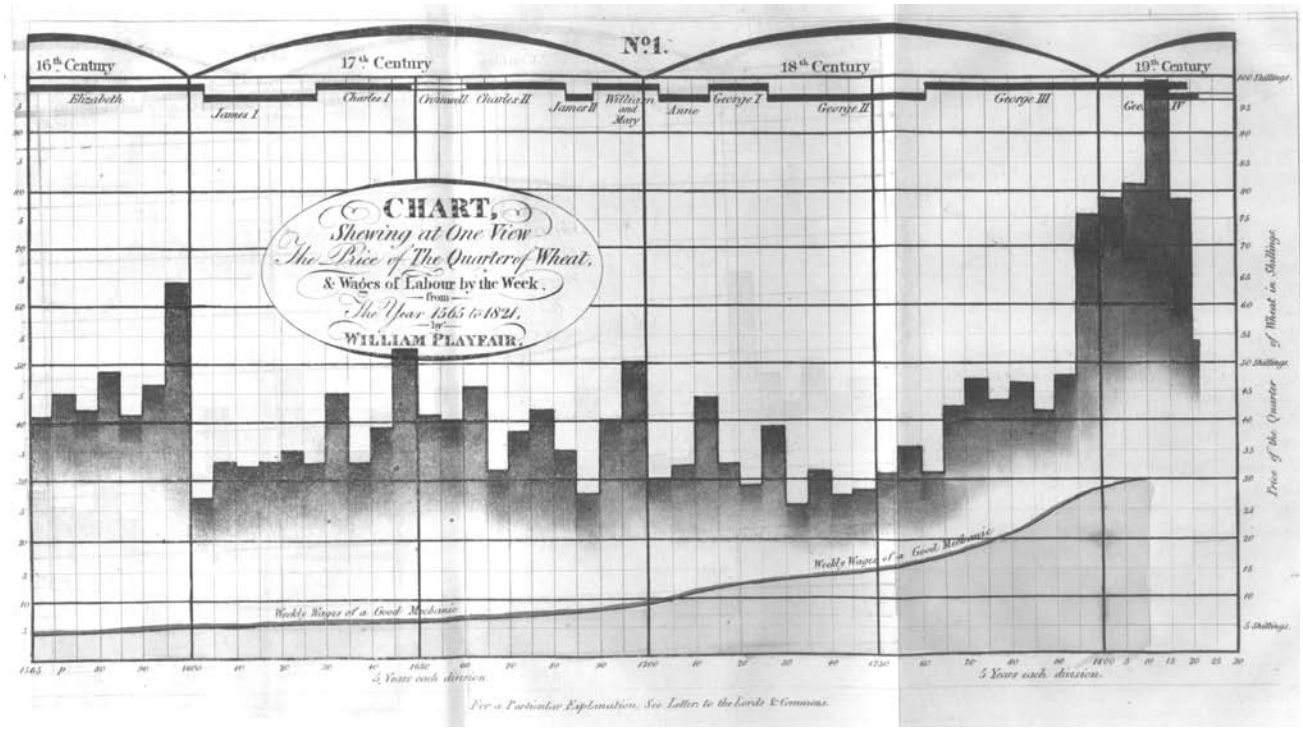


Figure 12. A graph by William Playfair containing two data series meant to be compared. The first is a line that represents the "weekly wages of a good mechanic," and the second is a set of bars that represents the "price of a quarter of wheat."

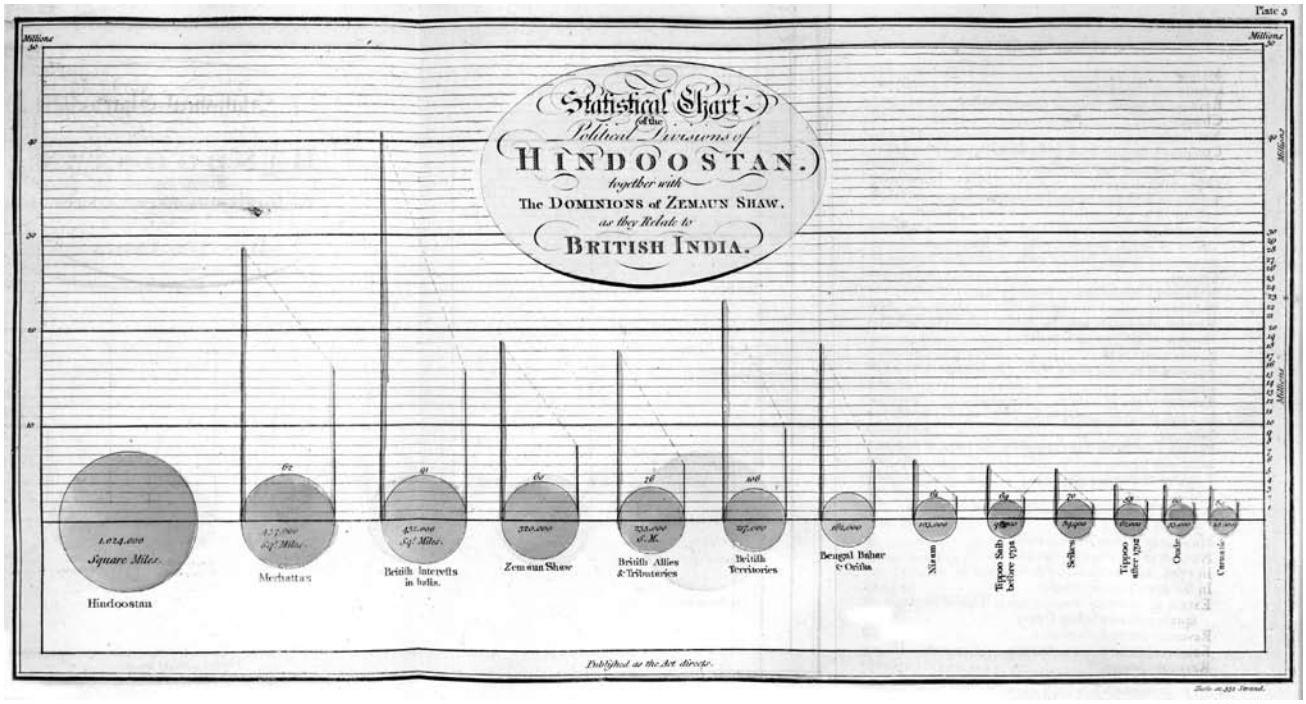


Figure 13. A graph by William Playfair containing three data series. The area of each circle is proportional to the area of the geographic location indicated. The vertical line to the left of each circle expresses the number of inhabitants, in millions. The vertical line to the right represents the revenue generated in that region in millions of pounds sterling.

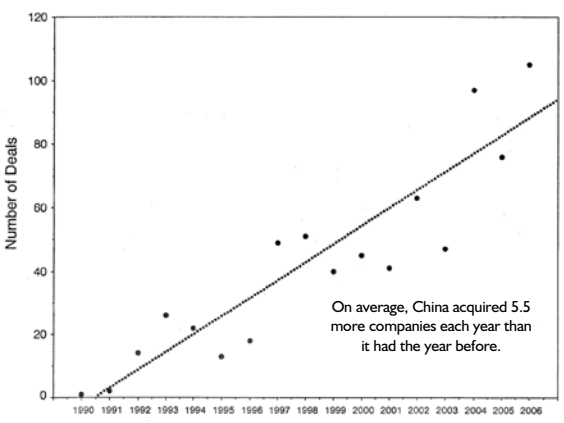


Figure 14. a.

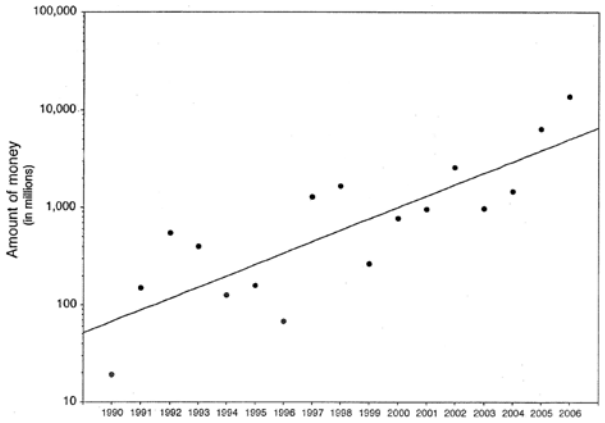


Figure 14. b.

Figure 14. The data from Figure 11 redrafted as two scatter plots. The plot of money is shown on a log scale, which linearizes the relationship between the amounts and the passing years. The superposition of regression lines on both panels allows the viewer to draw quantitative inferences about the rates of growth not possible with the depiction shown in Figure 11.

Albert Einstein's observation of the persistence of incorrect scientific theory: "Old theories never die, just the people who believe in them." C

Further Reading

Bertin, J. (1973). *Semiologie Graphique, 2nd ed.* The Hague: Mouton-Gautier. (English translation done by William Berg and Howard Wainer and published as *Semiology of Graphics,* Madison, Wisconsin: University of Wisconsin Press, 1983.)

Friendly, M. and Wainer, H. (2004). Nobody's Perfect. *CHANCE,* 17(2): 48–51.

Pfeffermann, D. and Tiller, R. (2006). Small-Area Estimation with State-Space Models Subject to Benchmark Constraints. *Journal of the American Statistical Association,* 101: 1387–1397.

Playfair, W. (1801/2005). *The Statistical Breviary; Shewing on a Principle Entirely New, the Resources of Every State and Kingdom in Europe; Illustrated with Stained Copper-Plate Charts, Representing the Physical Powers of Each Distinct Nation with Ease and Perspicuity.* Edited and introduced by Howard Wainer and Ian Spence. New York: Cambridge University Press.

Playfair, W. (1821). *A Letter on Our Agricultural Distresses, Their Causes, and Remedies.* London: W. Sams.

Reckase, M.D. (2006). Rejoinder: Evaluating Standard Setting Methods Using Error Models Proposed by Schulz. *Educational Measurement: Issues and Practice,* 25(3): 14–17.

Rosen, G. (2007). Narrowing the Religion Gap. *The New York Times Magazine,* February 18, Page 11.

Tufte, E.R. (1983/2000). The Visual Display of Quantitative Information. Cheshire, CT: Graphics Press.

Tufte, E.R. (1990). *Envisioning Information.* Cheshire, CT: Graphics Press.

"Magnum esse solem, philosophus probabit, quantus sitmathematicus."

— Seneca, Epistulae 88.27

Translation: "While philosophy labors to prove the obvious, mathematics takes its measure."

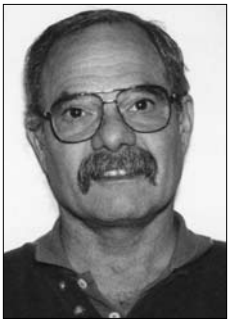
Tufte, E.R. (1996). *Visual Explanations.* Cheshire, CT: Graphics Press.

Tufte, E.R. (2006). *Beautiful Evidence.* Cheshire, CT: Graphics Press.

Wainer, H. (1984). How To Display Data Badly. *The American Statistician,* 38: 137–147.

Wainer, H. (2000). *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot, 2nd ed.* Hillsdale, NJ: Lawrence Erlbaum Associ¬ates.

Wainer, H. (2005). *Graphic Discovery: A Trout in the Milk and Other Visual Adventures.* Princeton, NJ: Princeton University Press.



Column Editor: Howard Wainer, Distinguished Research Scientist, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104; *hwainer@nbme.org*