Cognitive Testing of Statistical Graphs: Methodology and Results

Colleen Blessing, Howard Bradsher-Fredrick, Herb Miller, Renee Miller, Robert Rutchik
Energy Information Administration
1000 Independence Ave., S.W., Washington, D.C., 20085, colleen.blessing@eia.doe.gov, howard.bradsher-fredrick@eia.doe.gov, herbert.miller@eia.doe.gov, renee.miller@eia.doe.gov, robert.rutchik@eia.doe.gov

Introduction

The evolution of graphical and statistical computer packages over the past 25 or more years has enabled analysts to devise increasingly complex graphical displays. More data can be depicted on a single graph in an increasingly elaborate way, which allows analysts to make increasingly complex statements about the data in this visual format. But can the reader of the report (or user of the electronic product) discern the overall messages being conveyed by the graphs? The Energy Information Administration (EIA) attempted to determine whether or not some formats are more effective than others in conveying a message, discerning trends, determining relative quantities, etc. In short, EIA has become interested in identifying a set of "best practices" to be employed by its analysts in creating graphs.

This paper describes progress toward improving EIA's graphical displays in publications and on its web site. The results of cognitive interviews conducted by the authors will be described. These tests were devised to determine the ability of diverse, high-level users to interpret and employ government statistics, as depicted in EIA graphs. While the graphics used by EIA have become increasingly complex, the ability of users to interpret these graphs has not been tested until this research work.

Methodology

To determine whether EIA graphs are understandable and useful, a team of EIA staff members, trained in cognitive testing and experienced in conducting usability testing of the EIA web site and cognitive testing of EIA survey forms, decided to cognitively test a variety of EIA graphs. EIA disseminates graphical information in a wide variety of formats, including histograms, bar charts, single or dual axis line graphs (with single or multiple lines), stacked bar charts, two- and three-dimensional pie charts, scatter plots, and others. Thus, it was important that the research tested a variety of different commonly used formats.

The principal objective of the research team was to determine how to best convey a variety of graphical messages. While this was only a subset of the wide variety of graphical display formats used by EIA, the research team focused on the following:

- Two variables having a relationship over time (sometimes shown as dual axis graphs),
- Multiple variables summing to a total variable showing changes over time for all of the variables (sometimes shown as "stacked bar charts").

There is a wide literature on the subject of graphical displays and their proper usage. Cleveland[1] states that graphs are to be used to show overall trends, patterns, or relationships in the data; compare two or more factors in a general or quantitative fashion; present large data sets in a comprehensive way; and analyze data. Along these same lines, Tufte[2] states that graphs should show the data (without distortion); serve a clear purpose; induce the viewer to think about the substance rather than about methodology, graphic design, and the technology of graphic production; present many numbers in a small space; encourage the eye to compare different pieces of data; reveal the data at several levels of detail; make large data sets coherent; and be closely integrated with the statistical and verbal descriptions of a data set. Tufte succinctly states the overall objective of graphical design as follows: "An excellent graph is one that gives

[1] William S. Cleveland, *Visualizing Data* (Summit, N.J.: Hobart Press), 1993.
[2] Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, CT: Graphics press), 1983.

the viewer the greatest number of ideas in the shortest time with the least amount of ink in the smallest space."[3]

The research design we employed called for testing the messages conveyed by the graph as follows:

- The "big" message,
- Approximate data point values,
- If a time series graph, ability to determine trends (increasing, decreasing, or no trend),
- Ability to identify the variables on the axis and the units employed,
- For a dual axis graph, ability to ascertain the relationship between the two (or more) lines.

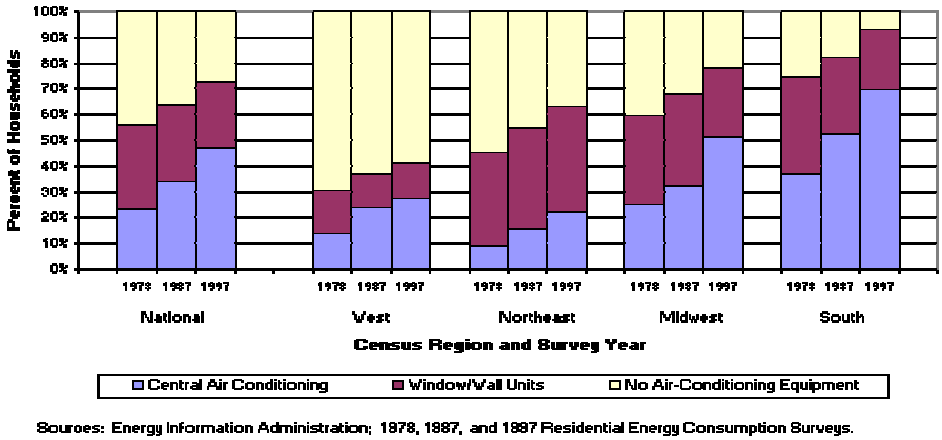
The research team selected five graphs that have been questioned as being fully comprehensible to EIA data users. Four of these graphs could be classified as dual axis graphs with a variety of formats, while the fifth was two stacked bar charts. For each of these graphs, an alternative single graph or a set of alternative graphs were designed which avoided the use of the dual axis or the stacked bars. For example, for a dual axis graph (two Y-axes) with two lines shown, two graphs were devised with a single Y-axis for each. An EIA staff member would show the original to participants and then the alternatives that EIA designed in order to determine which graph or set of graphs they could more accurately read the data and/or determine the message.

The participants for testing purposes were recruited as volunteers from a variety of different sources: the American Statistical Association Energy Committee members; attendees at the National Association of State Energy Officials Conference and the National Energy Modeling System Conference, and EIA staff members. These participants were generally expert energy data users. A total of approximately 26 participants were used in the testing of each of the graphical comparison sets. All of the testing was conducted between May, 2000 and August, 2001.

Stacked Bar Graphs

Two stacked bar charts were tested; Figure 1[4] shows one of these. The alternative was a set of five bar charts, one for each region. For the testing of the stacked bar charts versus the alternative set of five alternative bar charts, a participant was asked to read values from the charts and asked to assess whether the values are increasing, decreasing, or remaining the same. The experimental design provided for proper controls for learning and ordering effects.

Fig 1: Stacked Bar Chart: Type of Air Conditioning Equipment by Census Region: 1978, 1987 and 1997



Sources: Energy Information Administration; 1978, 1987, and 1997 Residential Energy Consumption Surveys.

[3] Ibid. p. 51.

[4] The actual graphs used for testing and appearing on the EIA web site were displayed in color.

Results: Stacked Bar Graph Versus Alternative Bar Graph
For the 26 participants a total of 62 exercise questions were posed, asking participants to read data from the stacked bar graph and 42 exercise questions were asked related to the alternative bar graphs. For the stacked bar, 6 of 62 (9.7%) responses were considered gross errors. For the alternative graphs, 2 of 42 4.8%) were considered gross errors, but 2 of 42 (4.8%) gave the "don't know" response. The range of correct answers was similar for both the stacked graphs and the alternative graphs.

Participants were given questions to test their ability to discern whether values are increasing or decreasing over time. For large increases or decreases, 1 of 46 questions resulted in a gross error for the stacked graph, while 0 of 32 questions resulted in a gross error for the alternative set of graphs. (1 of 32 resulted in a "Don't Know" response.) For small increases (changes over time from 36.4% to 38.8% to 41.0%) depicted on the graphs, for participants using the stacked bar chart, 11 of 15 (73.3%) characterized this change as "remaining the same," while 4 of 15 (26.7%) characterized this change as "increasing" or "slightly increasing." For these same small increases, 8 of 11 (72.7%) of the participants using the alternative bar charts characterized this change as "increasing," while 3 of 11 (27.3%) participants characterized the change as "slightly increasing." None of the participants using the alternative bar graphs characterized this change as "remaining the same."

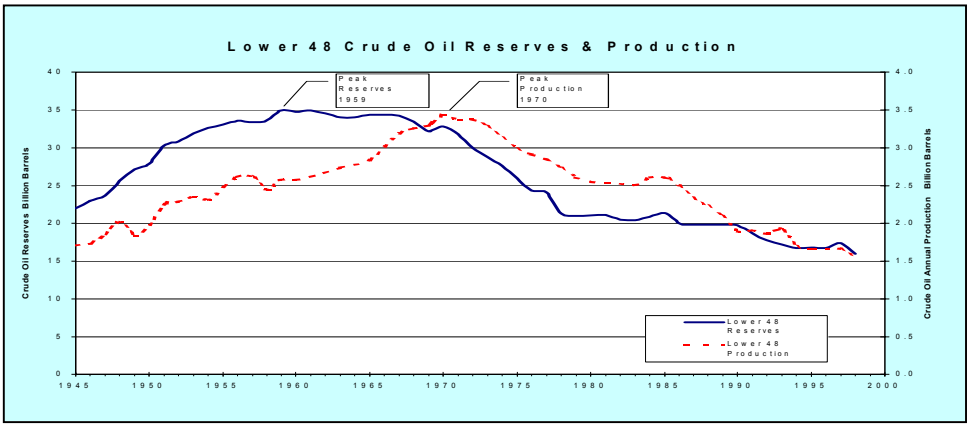
Summary and Recommendations: Stacked Bar Graphs
Participants made numerous mistakes when reading data from either the stacked bar graphs or the alternative bar graphs. With the stacked bar graphs, there was some tendency for participants to erroneously add the various bar segments rather then visualize and treat these quantities separately. However, discerning subtle trends is much more readily accomplished when using the alternative bar graphs in comparison to the stacked bar graphs.

Based on this limited testing, a tentative recommendation is that graphical presentations need to be as simple as possible in order to avoid errors. Although stacked bar graphs save space, they should be avoided in favor of other graphical presentations. Moreover, considering the complexity of stacked bar graphs, if such figures are used it appears necessary to show the data in tabular form in addition to graphical form.

Dual Axis Graphs

Methodology: 1st Set
The dual axis graph shown in Figure 2 was tested on a total of 17 participants. This graph showed two lines of different colors without the axes being color-coded. Also, one of the lines is solid while the other is a dashed line. The alternative graphs (not shown) have two single-axis graphs one above another.

Figure 2: Dual Axis Graph



Participants were asked questions regarding the graphs. Below are the questions with a summary of the results.

1.What is the main message of this graph?
Most participants ascertained a message from the graph. The message according to the author of the graph is that "production follows reserves." A number of participants were distracted by the labels pointing to the peaks in reserves and production and decided that the peaks were the story. Comments included, "The message is when the peak years were," "Reserves and production peaked and then declined," "Reserves peaked first and production peaked later." Another participant stated the message as, "There is less oil in the market than in earlier years."

2. During the years 1985-1990, what was the amount of Lower 48 reserves?
This question required the participants to read a value from the left-side (more traditional) Y-axis. All but one participant answered correctly (16 of 17, 93%). Due to the fact that the value being requested was nearer to the right-side axis, one participant (incorrectly) read the answer from that axis.

3. About how much crude oil was produced in the peak production year?
This question required the participant to read a value from the right-side (less traditional) Y-axis. For their first answer 4 of 17 (24%) again used the left-side axis and obtained an incorrect value. After further thought, 2 of these 4 participants recognized their mistake and provided a correct answer.

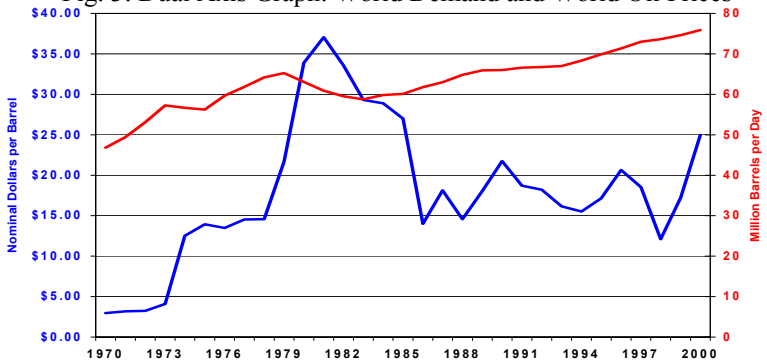
4. In what year did reserves fall below production?
This was intended to be a tricky question, since the order of magnitude of the values on the two Y-axes was different. Almost all participants quickly realized that there were two axes, but many thought they just repeated the same scale on both sides. Others realized that the scale for oil production was much lower than that for reserves, but they still proceeded to get this question wrong. The reserves and production lines intersect in the center of the graph. A total of 16 of 17 (94%) of participants said that reserves fell below production around 1970. This is a gross misunderstanding of the data. It appears as though crossing lines are more compelling visually than the two axes of different scales.

It was further observed in the testing that without the labels on the peaks (there are no labels on the lines themselves), you would not know which line was associated with which axis. An oil expert noted that s/he thinks in units of barrels per day, rather than billions of barrels. A number of participants turned the graph from side to side trying to decide what the two axes were.

Conclusions
Dual axis graphs with the same units on both axes are confusing. Although participants recognized that the scales on the axes are different, an intersecting line appears to confuse them. It was difficult for them to visualize the graphs, taking into account the change in scale, to perceive one line at the top and the other line far below. It is good practice to have a descriptive title that includes the message of the graph. It is also important that the graph does not get too cluttered with bubble boxes and descriptive phrases, although these can be helpful in conveying your message.

Methodology – 2nd Set
The dual axis graph shown in Figure 3 was used as the test vehicle for this experiment. This graph was color-coded with the line colors matching the associated axis, scale and labeling information. The alternative set of graphs included a design that had two graphs, one single Y-axis graph for each of the two lines shown.

Fig. 3: Dual Axis Graph: World Demand and World Oil Prices



The testing involved asking participants about the author's message in both the disseminated graph and the alternative graph. Participants were asked how they would statistically compute this data and graph it. Participants were also asked if they found anything confusing about either of the graphs.

According to the author of the report which included the graph, the intended message here was fairly complex: "While demand growth may have been slowed by rising oil prices between 1970 and 1979, it took a large per barrel price increase (from $15 to $35 per barrel) over a relatively short time period to cause global demand to decline in demand. Between 1979 and 1983, world oil demand fell 6.5 million barrels per day. Since 1983, global oil demand has generally risen regardless of oil prices. During the last 17 years, world oil demand growth has slowed twice – once in the late 1980's and early 1990's as the former Soviet Union economy collapsed, and again in 1998 when the Asian economies were hit hard due to their financial crisis."

Results

Participants generally recognized that the dual axis graph and the alternative set of graphs were intended to show a relationship between price and demand. However, no one really noticed the complexity of the message intended by the author. There was confusion about the dual axis graph. One participant noted that it was not immediately obvious which line was associated with which axis. Another participant stated that s/he was only guessing about the lines and axes. Yet another participant noted that the scales of the axes could be manipulated to show various existing or non-existing relationships.

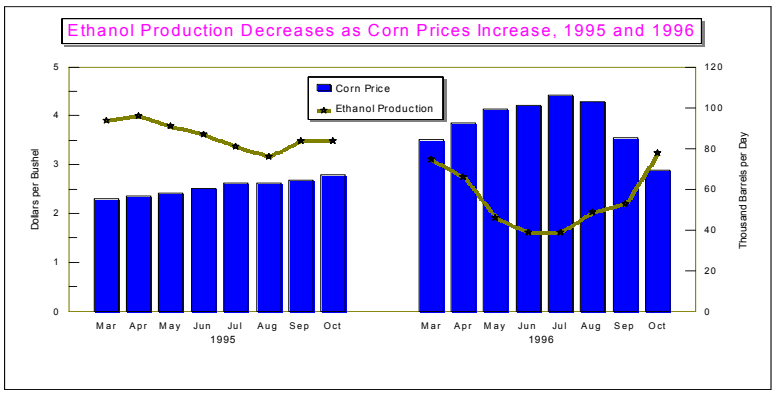
Conclusions

Complex messages are not easily conveyed through the use of dual axis graphs. In fact, participants can become distrustful of the intent of authors due to the author's ability to manipulate the scales of the axes. Moreover, participants have difficulty in easily perceiving the association between the lines and the axes.

Methodology: 3rd Set

The graph tested here (See Figure 4) can be characterized as a dual axis graph with both a line and a bar graph. During the course of the testing, two different alternative sets of graphs were tested. Initially, the sets of alternatives showed percentage changes in the values or indices. The final set of alternatives tested showed four graphs, two for 1995 and two for 1996, each having price and production on separate plots. The testing was conducted in such a way as to control for ordering and learning effects. The primary objective of the testing was to determine if participants understood the intended message of the graph.

Fig. 4: Dual Axis Graph with Line and Bar Plots



Results

All participants were able to arrive at the correct message after some period of time, whether using the test graph or either set of alternative graphs. One participant noted that if being able to read values was important, then the dual axis graph was preferable to the initial set of alternative graphs that showed

percentage changes or indices. The most prevalent comment was that the relationship between price and volume was easier to see on the dual axis graph.

Using a bar graph for the price and a line graph for production also received mixed results, with some preferring it and some finding it confusing. Other items of confusion were the use of barrels per day while using monthly data and showing partial years on the X-axis with space between the years.

Conclusions
Dual Y-axis graphs can be useful in conveying the relationship between price and volume when these are carefully labeled. Titles can be an aid in helping users get the overall message. Tables should be used along with graphical displays if users want to know exact numbers. Also, dual axis graphs can provide more information on data values than one-scale graphs that use percentage changes or computed indices.

Methodology: 4th Set

The dual axis graph to be tested here shows three lines, two of which are associated with one axis while the third line is associated with the other axis (See Fig. 5). The alternative graph removed one of the lines, changed "MBOPD" to "Thousand Barrels Per Day," and generally made the graph neater in appearance. Participants were asked to evaluate the test graph versus the alternative with respect to the messages being conveyed, ability to read data from the graph and ability to understand the abbreviations employed.

Results
Participants generally had "no clue" regarding the point of the graph. Participants were inclined to think that the title of the graph was the intended main message to be conveyed by the author. Participants could accurately read data from the graph, but had no understanding of the abbreviation, "MBOPD" or the jargon, "Texas First Purchase Price."

Conclusions
Dual Y-axis graphs are confusing and tic marks should be clearly marked on graphs. Also, to convey an accurate data point, a table is preferable to a graph. Graph designers should also be careful not to use acronyms or industry jargon.

Fig. 5: Dual Axis Graph with Three Lines