Effective Data Graphics with SAS/Graph® Software

Lesley Harrschlitz, DOFASCO Inc.
Laurie Abell, DOFASCO Inc.

Introduction

Today's working world is focusing on information delivery. It has long been known that pictures are a powerful tool for aiding in communication and understanding of information. Graphics, and their proper use, therefore, play a key role in the delivery of information.

There are, however, many pitfalls in producing these designs that may lead to incorrect interpretation or incomplete understanding. It is important for designers to make their graphics effective communication tools.

This paper concentrates on the basic principles that are used to create an effective data driven graphic (text slides and diagrams are not included). These principles will apply to any manual or computerized approach to producing a picture. Following this general discussion, the paper continues with a discussion of how SAS/Graph Software can be used to produce a design which follows these principles, and concludes with a working example using SAS/Graph.

Why are pictures such an important communication tool? Why do we see them used in every visual information format? Television, magazines, reports, presentations; the best of all of these rely on exceptional visual aids. The reason for this is audience retention. We only retain 10% of what we read, 20% of what we hear, 30% of what we see; but 70% of what we see and hear. This is why we should attend a presentation of interest, rather than just reading the paper that accompanies it.

Now that we understand why pictures are so essential to information delivery, what about effective graphics? These are designs are those in which care is taken to maximize audience understanding, with some underlying purpose in mind. Effective graphics are planned based on principles of design, tailored to the viewing audience, and executed to show the information required in the correct context.

Creating Effective Data Graphics

Creating effective graphics is an art. There is no set formula to follow, no recipe that will guarantee success every time. We use a series of rules of thumb, or design principles, to improve the chance of success. There are three major steps in the creation of an effective graphic: definition (data and audience), selection (and design), and execution.

Definition:

What data? We must have access to the data, know how it is organized, and be able to show in detail how it will answer the questions asked. The most important rule of thumb is the GIGO principle. As with any form of production, garbage in equals garbage out. Our first step in defining a data graphic is to examine the underlying data. Is it relevant, complete, and truthful? If so, we have a good chance of getting our point across. If not, we should not pursue the graphic.

Who is the audience? The second part of definition is to identify the audience.

What is their background? Do they come from management positions, or are they line workers, or technical staff? Each audience comes to us with different expectations. Management may be interested in the big picture, summaries and overviews. Line workers may want more specific and directed information. Technical staff may be different again.

What preferences do they have? Do they prefer colour or black and white? Do they prefer large print and graphics which are spread out; or do they like comparisons, and size doesn't matter?

How large is the audience? The size of the audience for a presentation will determine the type of format used. Large audiences will find slides easier to view. Small audiences will find overheads very adequate.

What is their level of understanding? Since our objective is communication, we must adjust our graphics to meet the level of understanding of the audience from both a technical and language standpoint. Presenting the mathematical theories behind the flight of an airplane would not communicate to a group of workers with high school education. Equally, presenting graphics with text in an unfamiliar language would also lack communication. For example, at Dofasco many employees speak English as a second language. The Safety manual, which communicates vital information, is printed in several languages so that most employees may have a copy in their first language.

What is the purpose? After locating the data and identifying the audience, we must identify the purpose of the graphic. Do we want to inform our audience of the facts, persuade them to adopt our point of view, or motivate them to change something. Each of these purposes is valid and will influence the type of graphic we produce.

Selection:

Data driven chart types tend to fall into easily defined categories with several choices within the category. In making our choice within the category, we take into account the preferences of our audience and their expectations, as well as our purpose in presenting the graphic.

Items which Change with Time

A common type of data is that which changes with time. This includes such things as production per shift, the Dow Jones Index, the daily high temperature. We may choose to present this data in the form of a line chart, area chart or bar chart. One thing all of these have in common is that a measure of time is along the horizontal axis.
Items which change with time

We generally make certain types of comparisons based on a single point in time or over a defined period of time. For example, we may look at sales for each market or region over the past quarter. For this type of data, we commonly use a horizontal bar chart, with the comparison being made vertically through the chart.

Parts of the whole

Parts of the whole exist when a total available supply is allocated into areas of demand. Budgets or expenses by category are examples of this type of data. The most commonly used chart for this is the pie chart, as it is the only chart available with one dimension. However, in some cases 100% horizontal bar charts or 100% area charts can be used.

Frequency distributions

These are generally used in production where multiple values of the same thing are checked for their distribution over a period of time. An example of this type of data is a single car's gas mileage over its last 200 tanks. This type of data is presented as a vertical bar chart with or without a superimposed distribution line, as can be done with SAS/QC® Proc CAPABILITY.

Relationships between variables

This type of data compares the value of one variable to that of the other. An example of this may be to compare production and price for a commodity. Types of charts used for this type of data include 2 and 3 dimensional scatter plots, bubble plots, and paired bar charts.
Execution:

Principles of Design

Now that the graphic has gone through the definition and selection stages, we are ready to proceed to the execution stage. The first step in this stage is to rough out the chart on paper using the basic design principles: unity, emphasis, balance, proportion, movement, and simplicity.

Unity

Unity is the presentation of a consistent image throughout a report or a presentation. Some of the points to remember are as follows:

- Use a consistent page orientation. Graphics should be all horizontal (landscape) or vertical (portrait). This prevents the audience from having to refocus from one slide to another.
- When using colour, we use only 2 or 3 throughout our graphics, and keep background and border colours consistent. We make selections from across the colour wheel for contrast, and adjacent on the colour wheel to emphasize. Shades of colours can also be used for emphasis.
- Fonts should be kept within the same family, i.e., Swiss, Zapf, Times, Helvetica, etc., with no more than 3 sizes used. Use a proportional and sans serif font for cleaner presentation.
- Use borders and repetition, and keep them consistent from one graphic to another.

Emphasis

Highlight the major point of the graphic. This helps the audience focus on the part of the graphic that we most want them to see. We can use a change in colour, font, size or pattern to highlight for emphasis.

Balance

Balance is based on the visual weight of the objects in a graphic. Dense objects "weigh" more than open objects, and large objects "weigh" more than a small objects. We place different objects on our graphic so that the overall weight is the same throughout the page. We can use symmetrical placement, which is more formal, or asymmetrical placement, which is more informal.

Proportion

Proportion is the relative size of objects within the graphic. Sometimes, as in cartoons, proportion is deliberately distorted in order to achieve an effect. For data graphics we use proportion to ensure that the data is not distorted, and therefore does not give a false impression, and also to give a clean professional appearance.

Movement

Movement is the flow through a graphic that gives the eye a natural path to follow. For most data driven graphics, movement is provided by the data itself - no special consideration needs to be made for this design element.

Simplicity

Simplicity is probably the most important part of graphic design. Most people expect to spend very little time looking at a graphic before achieving understanding. Therefore, we need to communicate quickly - simplicity gives us this quick communication. To achieve simplicity use lots of space around objects, connect similar elements, make the graphic legible, and use the KISS principle. (Keep It Simple, Sweetie)

An Example:

Figure 6 shows an example where these principles have been applied. Notice the consistency of fonts, the framing of the graphic image and legend, and the consistency of size which promotes unity. Emphasis is placed on the titles and axes labels using size and a swiss font instead of a swiss font (same family). Balance is symmetrical, which is very easy to achieve for data graphics. Axes text is carefully sized so that it remains in proportion to the graph, and movement occurs by naturally following the graph left to right, and then moving down to the legend.

Using SAS/Graph to Create Effective Graphics

SAS/Graph, with all the defaults in effect, rarely produces an optimal effective graph. However, there are a vast number of options that can be used with SAS/Graph to produce the very best of effective graphics.

The statements that allow for the use of these options are goptions, title, footnote, note, axis, legend, symbol, and pattern. As well, certain options on the proc and chart/plot statements can be used to good effect.

Goptions Statement

Use a Goptions statement to change the default graphics options. We can define default color, text size, and text font using the Colors, Htext, and Ftext options. For unity, use the Border option to place a border around the edge of the page (the default is Noborder). Also for unity and for proportion, use Rotate=Landscape or Rotate=Portrait to specify the orientation (the default orientation depends on the graphics device being used).

eg: GOPTIONS dev=vga rotate=landscape border ftext:awi88li

Title/Footnote/Note Statement

These statements can be used to add explanatory text to the graphic. A number of options are available to enhance these statements. The Colour, Font, and Height options directly affect the appearance of the text. The Angle, Line, and Rotate options
affect the baseline, and can be used to add emphasis. The Justify, Laplace, and Move(x,y) options control the placement of the text. Further emphasis can also be achieved by using the Box, Boolean, Blank, Draw, and Underline options. These are all placed before the text (encased in quotes) to which they apply. Note that this allows options to be changed mid-text to allow for full control over each line of text.

eg: TITLE1 cured h=4 l=aweslab 'Title Text';

Axis Statement

The axis statement is used to control the appearance of an axis for most SAS procedures. (One exception to this is G3D.) Axis statements are numbered(i.e. AXIS1), and then tied to the procedure with an option on the plot/chart statement. eg: PLOT Y=X VAXIS=AXIS1. The text associated with axes (values and labels), are controlled in much the same manner as the text statements use. Title, Font, Color, Angle, Justify, Rotation, and Height. The axis placement is controlled with the Origin and Offset options. The axis line is detailed with Color, Style, Width, and Length options. Tick marks are specified by the Order option, which determines a start point, finish point, and major tick mark values. For individual tick marks, Color, Style, Width, and value can be specified.

eg: AXIS1 label=(blue font=swiss h=1.5 as=90 'Label Text') values=(blue f=swiss h=73)

Legend Statement

The legend statement is used to control a legend which is automatically generated by SAS when a chart is created using the Subgroup=GroupVar option, or when a Y=X=Z plot is specified. Like Axes, Legends are numbered and tied to the procedure using the Legend option on the chart/plot statement. (eg. PLOT Y=X=Z/ LEGEND=LEGEND1.)

Label and Value text on a legend is controlled in the same manner as for Axes; using the Angle, Color, Font, Height, Justify, and Rotate options. The Subgroup option is useful in the Value options to specify which category the text is for: eg Legend1 Label=('ANSWER:') Value=(tick=1 'Yes' tick=2 'No'); Legend placement is controlled by the Position, Mode, Offset, and Origin options. The appearance of the legend is controlled with the Acros, Down, Shape, Frame, Cube, Cblock, Cborder, and Cshadow options.

eg: LEGEND1 label=(red font=swiss h=1.25 'Label Text') position=(bottom inside right);

Symbol/Pattern Statements

Symbol and Pattern statements are used to control the appearance of lines and areas within plots and charts. They are numbered and are connected through the chart/plot statements, sometimes automatically (PLOT Y=X=Z), and sometimes specifically (PLOT Y=X=1).

Patterns can be specified as solid, empty, lines rising right, lines rising left, or crosshatched. The line patterns can be spaced from 1 to 5, with 1 being the most open, 5 being the tightest. Pattern colors can also be specified.

Symbol statements are similar, with options split into those for the symbol, and those for the interpolation between symbols. Symbol options include Colour, Value, Font, Height and Width. Interpolation options include Colour, Style, Type and Width. Interpolation types include everything from simple lines to linear regression and stepwise bars.

eg. SYMBOL1 value=empty h=join l=5;
PATTERN 1=3 cured;

An Example:

Figures 7 and 8, are used to illustrate the use of many of these SAS/Graph options. We have chosen a comparative capability study. Capability studies are a measure of the ability of a process to produce an item which meets the customer's requirements. They have become a widely used statistical tool in many businesses, and Doefasco is no exception. In this case our comparative study separates data into that which comes from the Lead (front) end of the steel coil, and that which comes from the Trail (tail) end of the coil, and compares it to the overall results. Figure 7 shows the most basic SAS/Graph output which can be achieved using the graphic defaults. Notice that there are several problems. First, the text for the lead and trail diagrams is impossible to read. The default colors are masking the histograms, and in general, proportions are not very good.

The following is the program that is run to produce the enhanced output in Figure 8. The bolded sections are those that were added to override defaults and produce the enhanced graph.

SAS/Graph Program:

gopt opt1=aweslab;html=2 rotate=devrot;
symbol value=none c=black;
symbol2 value=none c=blue h=4;
pattern 1 v=none c=black h=4;
pattern2 v=none c=blue h=4;
legend frame label=none shape=line(.5 in) down=3
value=(h=1.75 tw1 'Lower Spec Limit: 62' t=3.5 'Overall RockweU Hardness';
axist order=5 to 11 by 5
legend1 frame label=none shape=line(.5 in) down=3
value=(h=2.25 tw1 'Percent';
axist order=55 to 73 by 2
label=(h=2.25 'Hardness')
ttitle1 h=3 faxitals 'Rockwell Hardness for Lead End';
proc capability data=large.capable graphics;
spec 1sJ=62 usl=88;
histogram hardness/normal(fill)
where end='Lead';
nrun;
ttitle1 h=3 faxitals 'Rockwell Hardness for Trail End';
proc capability data=large.capable graphics;
spec 1sJ=62 usl=88;
histogram hardness/normal(fill)
where end='Trail';
nrun;
goptions norotate;
ttitle1 h=3 faxitals 'Overall RockweU Hardness';
proc capability data=large.capable graphics;
spec 1sJ=62 usl=88;
histogram hardness/normal(fill)
where end='Overall'
nrun;
goptions rotate;
...program continues
DOFASCO Tin Plated Steel Products: Process Capability

Data used to create this graph is fictitious.

Figure 7

DOFASCO Tin Plated Steel Products: Process Capability

Data used to create this graph is fictitious.

Figure 8
One thing that we should always keep in mind with data driven graphics is the possibility in inadvertently misrepresenting our data. As an example, usage data, in hours per month, is kept by work area for photocopiers. We are interested in those installations showing usage rates of over 80 hours per month, as these installations may require decisions regarding hardware upgrades, or additional equipment purchases. Figure 9 shows our initial graph.

On the surface this graph looks quite good, and certainly seems to illustrate the point of which work have the highest usage. However, we have fallen prey to what Edward Tufte calls the Lie Factor. Tufte defines the lie factor as follows:

\[
\text{Size of Effect in Graphic} \quad \frac{\text{Size of Effect in Data}}{\text{Size of Effect in Data}}
\]

As a rule of thumb the Lie Factor associated with any graphic should fall between .95 and 1.05.

Figure 10 shows the calculation of the lie factor for our current graphic. The final value of 15.05 lies far outside the acceptable limits. What this means is that a casual observer of the graph would tend to think that the difference between Area A and Area P is several times larger than it actually is.

There are several options that we could use to correct this situation. The simplest is to begin the horizontal axis at zero.

Summary

The steps to creating an effective graphic are as follows:

1. Define the Graphic:

What data are we using, for whom, and why? What is our purpose in creating the graphic? What response do we expect to get?

2. Select the Graphic:

Who is the audience, what are their preferences. What type of data do we have. If our data is parts of a whole, and our audience hates pie charts, let's make another selection from that group. Possibly a 100% horizontal bar chart. If there is no audience influence we chose the chart that best suits the purpose of the graphic.

3. Create the Graphic

Rough out the graphic manually using the basic design principles: unity, emphasis, balance, proportion, movement, and simplicity. Remember the GIGO and KISS principles here. Also pay close attention to the Lie Factor when dealing with data driven graphics.

When using SAS to generate the graph make use of the enhancing options and statements that are available. Use goptions, title, footnote, note, axis, legend, pattern, and symbol statements. Remember, all of these will default, use only the ones you require to get the desired effect.

The most important considerations are to keep the graphics simple, easy to understand, and to the point.


© SAS/Graph, and SAS/QC are registered trademarks of SAS Institute Inc., Cary, NC, USA.