Journal of
Micro/Nanolithography, MEMS, and MOEMS

SPIEDigitalLibrary.org/jm3

How to Write a Good Scientific Paper: Figures, Part 2

Chris Mack

SPIE

How to Write a Good Scientific Paper: Figures, Part 2

This is the seventh in a series of editorials covering all aspects of good science writing.

The great statistician and graphical expert John Tukey said, "The greatest value of a picture is when it forces us to notice what we never expected to see."[1] While many graphic forms can help us accomplish this goal, the most useful for science has proven to be the x-y scatterplot. In 2012, about 1/3 of all figures in JM^3, and about 70% of all data plots, were x-y scatterplots.[2] The first modern scatterplot is attributed to John Herschel (1792–1871), son of William Herschel, the discoverer of Uranus and infrared light.[3] In 1833, John Herschel used a scatterplot of noisy binary star measurements to extract a trend "by bringing in the aid of the eye and hand to guide the judgment,"[4] thus fulfilling Tukey's goal. The scatterplot allows the viewer to visualize the important trends the data suggests, and possibly offer a theory to explain them, by imagining a line that passes "not through, but among them," as Herschel so aptly said.[4] By 1920, the scatterplot had come into widespread use as the tool of science we know it now to be.

The x-y scatterplot is "a diagram having two variates plotted along its two axes and in which points are placed to show the values of these variates for each of a number of subjects, so that the form of the association between the variates can be seen."[5] If the *x*-axis plots time, we generally call the graph a time-series plot and often use unique analysis or interpretive frameworks for the data due to the unique role of time in causality. Here I'll talk only to the more general x-y scatterplot and not to time-series plots specifically. I'll also (mostly) ignore the role of x-y scatterplots as a projection of multivariate data (three or more variables), as interesting and important as that role is, and instead concentrate on the basics of this most popular of science graphs.

What makes for a good x-y scatterplot? As for all graphs, the goal should be to allow the data to tell its story efficiently and effectively. The first rule of a graph is that it must help to reveal the truth.[2] The design and execution of an x-y scatterplot can either help or hinder this goal. And while graphs can aid both in data exploration and data presentation, I'll focus only on the latter here. Since I gave general advice on good graphics in part 1 of this editorial, here I will strive to be more specific through the use of examples.

Though I have only anecdotal evidence, I am quite confident that most JM^3 authors use Microsoft Excel to create their x-y plots (as well as most other graphs in their papers). Thus, my first example will explain how to turn the seriously awful default scatterplot of Excel into an acceptable graph for submission to JM^3, or any other scientific journal.

My example will be simple: a plot of (made-up) experimental data along with an equation that models that data. The before and after plots are shown in Fig. 1. Here is the sequence of steps I went through in Excel to move from the default to the final graph. I'm assuming that the final graph will fit within a single column in a two-column-per-page format. For journals with other page formats, some adjustments to these directions may be required.

1. Set chart area size to be 5 in. tall by 6.75 in. wide (this is 2× the final size required by JM^3, but it will shrink 50% when published since most scatterplots will fit in one column). The chart area height can be adjusted as needed, if the data suggests a better shape, but the 4:3 aspect ratio used here is a good default.
2. Set the chart font size to be 14 points (they'll end up being 7 pt after shrinking the graph 50%).
3. Remove the legend if not needed (try to put labels inside the graph if they fit rather than using a legend). If using a legend, see if there is room within the plot area to put it. In the example above, using the convention of symbols for data and a line for the theoretical equation means that the legend can be embedded in the caption.
4. Remove all gridlines.
5. Change axes line color from gray (the Excel default) to black and set to 1 pt thick.
6. Change major tick mark to "cross" and minor tick mark to "outside."
7. Format the chart area to have no border.
8. Format the plot area to have a solid black border (1 pt thick) and no fill.
9. Set the "axis crosses" point so that the two axes meet at the lower left corner.
10. Adjust the axes label numbers so that they have the proper number of decimal points.
11. If necessary, adjust the axes min and max values (Excel defaults are often poor). Remember that the goal is to use up almost all of the graph space with data, but try to keep the data points from overlapping onto the solid border surrounding the plot area.
12. Add axis titles, set to 18 point (less if titles are too long), no bold, and use a rotated vertical title.
13. Format the "data series" to have the preferred color and symbol or line type/style for maximum readability and differentiation between data series. I typically use a weight of 1.5 pt for my lines (the default of 2.25 is too heavy), and my preferred symbol is the open circle when more than one thing is being plotted at a time.
14. If using line segments to connect data points, never turn on the "smoothed line" feature.

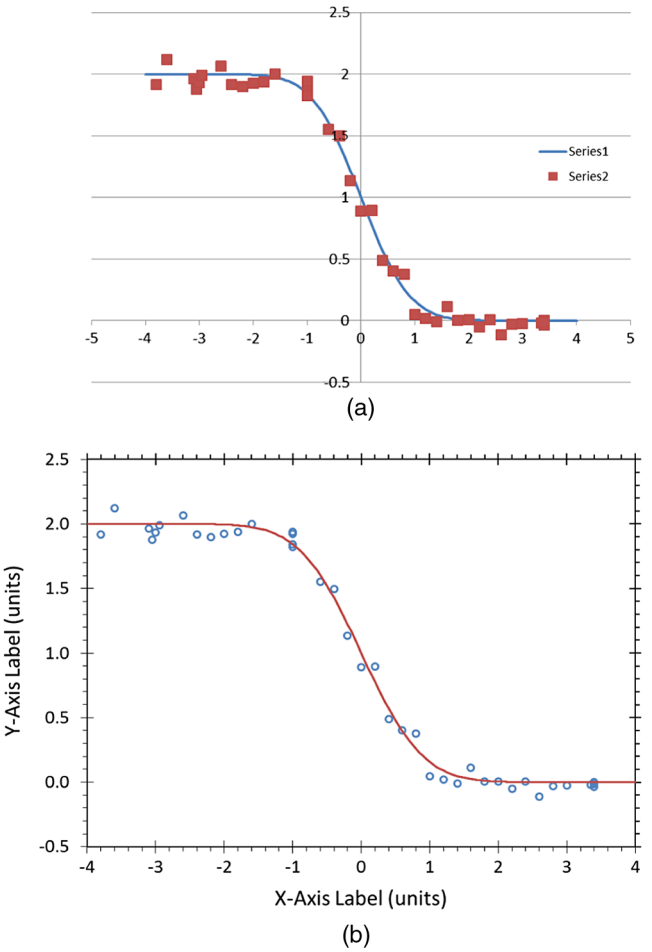

Fig. 1 Excel graphs of the same data: (a) default scatterplot settings, and (b) after proper formatting. Symbols show data, and the solid line shows the fitted equation.

15. Make sure there is no title.
16. Add a baseline in the graph if doing so is helpful for interpreting the data, but don't include a $y = 0$ line by default.
17. Preferred: put tick marks on the right and top of the plot area bounding box (this is tricky to do in Excel but can be done using a "secondary axis").

That's a lot of steps. But every step left out produces a less adequate graph. Note that some of these steps can be described as aesthetic, though making a graph more pleasing to the eye is generally synonymous with making it more readable. For example, the open circle data symbols enable one to see behind the symbol to the line and to other data points. In the original graph with the solid square symbols, can you tell how many data points are at $x = -1$ and $x = 3.4$? When using more than one symbol, be sure to consider the symbols' size and shape for maximum visibility when there is overlap.

The next example (Fig. 2) shows how labels can sometimes be fit into the graph to avoid the need to refer back and forth to a legend.

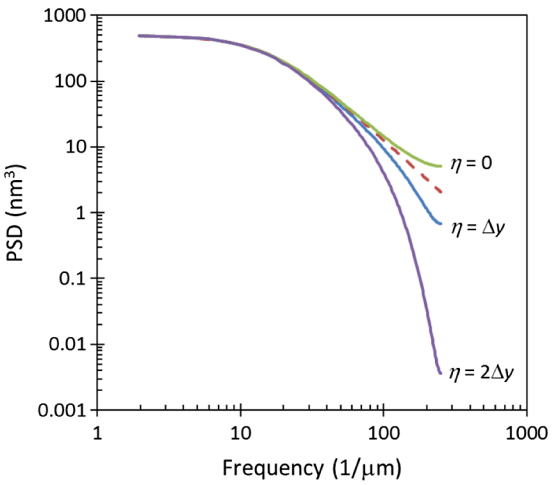

Fig. 2 Labels within the graph avoid the need for a legend. The color used here improves readability online but is not needed for comprehension when printed in black and white. The dotted line is explained as being the reference curve in the figure caption of the original.[6]

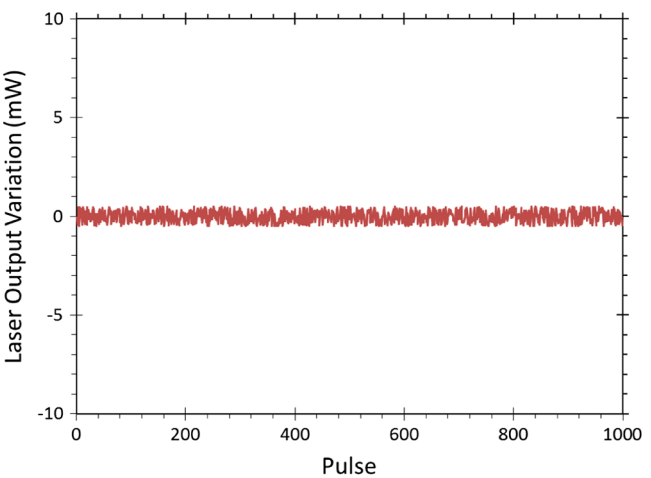

Fig. 3 A wasted graph. The *y*-axis is chosen to give the impression that the there is little variation in the output, but if we can't see any variation in the data, why show the graph?

A regular problem I encounter is a graph with data that fails to use up the space in the plot area. In Fig. 3, the authors wish to show how stable their laser is, so they stretch the *y*-axis range to be ten times the data range. As result, we can't see the variation in the data. So why bother showing the graph? A similar effect can be obtained by including zero on the *y*-axis scale even though no data are near zero (imagine a plot of Earth's global surface temperature in Kelvin, then starting the *y*-axis at zero—global warming would disappear). This is an example of advocating rather than informing—using graphs to hide rather than reveal the truth. If there is nothing in the data worth seeing, the graph should be replaced with simple statistics: mean, standard deviation, min/max of the output, and maybe a statement that a linear regression gave a slope that was not statistically different from zero. If there is something worth seeing in the data, then adjust the *y*-axis scale so that it can be seen.

There are other ways to mislead with an x-y scatterplot, some not as subtle as the previous example. Unitless axes are a favorite of those who, at a minimum, do not wish to reveal the whole truth. An axis without unambiguous labeling should never be allowed. Using "arbitrary units" for a *y*-axis is a bit trickier, since there are some cases where such a label is appropriate (a relative measure, based on a local uncalibrated standard that can be used to compare similar measurements). A common example is the relative intensity used in spectral analysis. Arbitrary units are never preferred, but sometimes necessary. Arbitrary units should never be used to hide known units that the author does not want to reveal. Additionally, arbitrary units have an arbitrary scale, but not an arbitrary zero point. Thus, when arbitrary units are used the graph must mark the zero point on the scale.

One common and important application of the x-y scatterplot is to compare different graphs (thus adding a third variable, sometimes more). Figure 4 shows a 2×3 array of

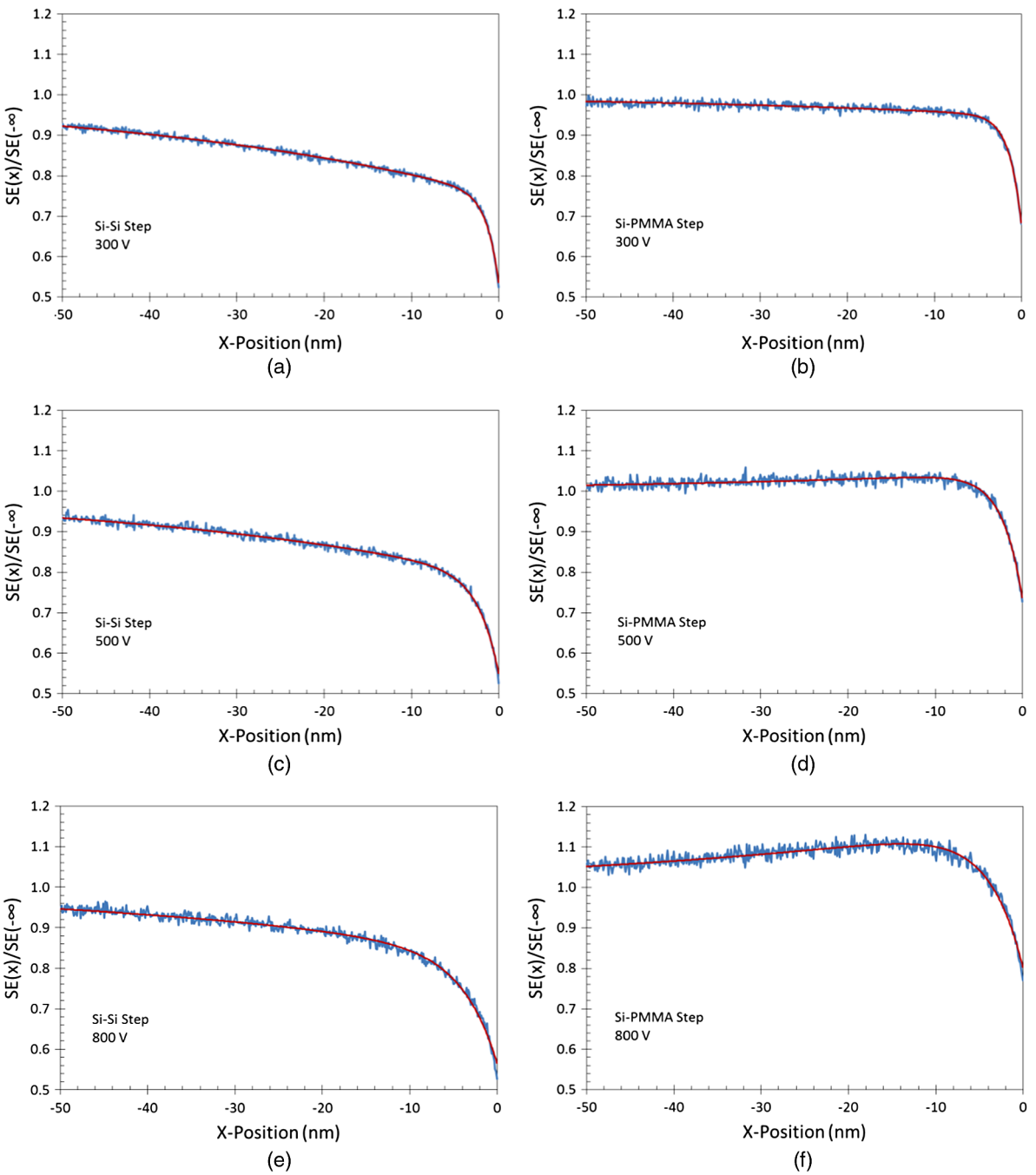

Fig. 4 Comparison of Monte Carlo simulations to an analytical model (ongoing work by the author, soon to be published). The smooth (red) line is the equation and the jagged (blue) line is the Monte Carlo simulation results. Both vertical and horizontal comparisons between graphs are enabled by matching the *x*-axis and *y*-axis scales of every graph. Note that in this case redundant axes labels could be removed.

graph multiples, matching x-axis and y-axis scales to allow easy comparison. With small multiples, many more graphs can be compared.

Figure Quality from a Production Standpoint

The final step in ensuring a good quality figure in your published paper is to make sure the submitted figure matches the production requirements of the journal. I'll talk here specifically about JM[3] requirements, but I don't think they are much different from most other journals. A few of the largest publications, such as *Nature* or *Science*, employ professional editors who can reset a graph to the standards of the journal. For most publications, however, it is up to the author to get the graph right. Below are some hints, given to me by the SPIE publications staff, that will make the production process go more smoothly, and the resulting graph higher quality.

- Submit high-resolution figures. The quality of the published figure is only as good as the original file—it cannot be improved by the typesetter. A resolution of 100 dpi (dots per inch) looks great on your computer screen but is inadequate for print. A minimum of 300 dpi is required, but 600 dpi is preferred. Thus, a one-column wide photograph must be at least 1000 pixels across.
- Submit full-size figures (7 in. wide) but remember that they will, in general, be reduced 50% to fit within one column. Make sure that the fonts, lines and other elements of the graph will hold up to this reduction (see my font-size suggestions in the Excel example above). Try shrinking the graph 50% and printing it out yourself as a test.
- High-contrast color graphics are great for online viewing, but the figures still need to be readable in grayscale for black-and-white printing (unless you pay for color printing). Colors such as red and blue, which are easy to distinguish online, are the same shade of gray when printed in black and white. If lines or symbols must be distinguished in a legend or caption, use different line styles and symbols instead of relying solely on color.
- Don't submit JPG files – the image compression often compromises the quality of the figure. TIF files have no compression, but if the file size is unmanageable try using "LZW" compression.
- If a figure contains multiple parts, they should all be laid out in one file, not submitted as individual files. This is important because it lets the author determine how a figure should be arranged for the reader (horizontal versus vertical, for example, for proper comparison). The parts should be clearly labeled with (a), (b), etc. (lowercase Roman text in parentheses). Put the labels outside the graph area, either centered below the figure part or on the left near the bottom. For example, Fig. 4 was submitted as a single one-page file.
- For an x-y scatter plot in Excel, I generally copy and paste the figure into Microsoft Word (full 7 in. width), print it to a PDF file (one figure per PDF file), and then submit the PDF.

Conclusions

When presenting results, a good graph is like a good scientific theory—once you see it, everything just makes sense. But arriving at such a point takes care and consideration. In part 1 of this series on figures, I talked at a high level about what makes for good graphics. Here, I provided more pragmatic advice geared toward a specific type of graph—the ubiquitous x-y scatterplot. Keeping in mind the advice from both parts of this pair of editorials will, I hope, lead to graphs that help you, the author, achieve your goal of effective and efficient communication.

Chris Mack
Editor-in-Chief

References

1. John W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, Reading, MA (1977).
2. Chris A. Mack, "How to write a good scientific paper: Figures, part 1," *J. Micro/Nanolith. MEMS MOEMS*, **12**(4), 040101 (Oct–Dec 2013).
3. Michael Friendly and Daniel Denis, "The Early Origins and Development of the Scatterplot," *Journal of the History of the Behavioral Sciences*, **41**(2), 103–130 (Spring 2005).
4. John F. W. Herschel, "On the investigation of the orbits of revolving double stars," *Memoirs of the Royal Astronomical Society*, **5**, 171–222 (1833).
5. *The Oxford English Dictionary*, 2nd ed., Oxford Univ. Press (1989).
6. Chris A. Mack, "Systematic Errors in the Measurement of Power Spectral Density," *Journal of Micro/Nanolithography, MEMS, and MOEMS*, **12**(3), 033016 (Jul–Sep, 2013).