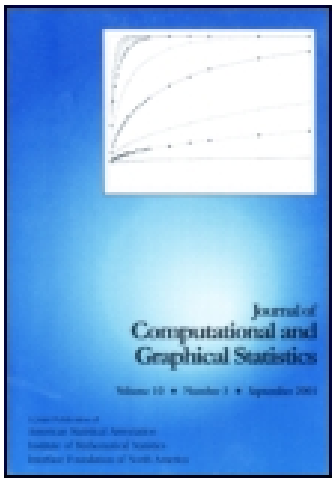

Journal of Computational and Graphical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

Statistician heal thyself: have we lost the plot?

Ian Gordon[a] & Sue Finch[a]
[a] Statistical Consulting Centre, The University of Melbourne, Victoria 3010, Australia. email:
Accepted author version posted online: 20 Dec 2014.

STATISTICIAN HEAL THYSELF: HAVE WE LOST THE PLOT?

IAN GORDON AND SUE FINCH*

Statistical Consulting Centre, The University of Melbourne

Abstract

In 1984, Cleveland suggested that statisticians have an important role in changing the use of graphics in science for the better. Thirty years later, we compared graphs published in top-rating applied science and statistics journals, evaluated for overall quality and against five principles of graphical excellence. Nearly 40% of the 97 graphs we sampled were rated as poor, with no striking differences between the applied science and statistics graphs. Better use of graphs requires better definition of variables, units of measurement, scales, groups and other graphical elements, and more routine use of grid lines on a 'standard' set of graphical forms. Progress over the next 30 years needs to be supported by changes in software defaults.

Keywords: Chartjunk, Statistical communication, Statistical graphics

*Statistical Consulting Centre, The University of Melbourne, Victoria 3010, Australia.
email: sfinch@unimelb.edu.au

THE REVOLUTION IN GRAPHICS?

> Those of us who believe graphing to be important and even essential to research would be well advised to think hard about why visual displays are not used more extensively in serious applied research.
>
> Gelman (2011, p.4)

The art and science of making good graphs in statistics was elucidated by John Tukey (1972, 1977). The ideas of Tukey's colleagues and successors including key figures Bill Cleveland (1985) and Edward Tufte (1983) have influenced the development of graphs in statistical software, making powerful tools for graphing data widely available today. In 1984 Howard Wainer showed us the "dirty dozen" for displaying data badly (Wainer, 1984), but it remains very easy to find poorly constructed and badly documented graphs in both lay and academic publications and presentations. In the same year, Cleveland concluded that "statisticians can play ... the leading role, in effecting an improvement of graphical communication in science" (Cleveland, 1984, p.265).

Other important reviews and commentary on graphics include the historical archive of Friendly and Denis (2014) and Wickham's (2013) survey of the literature on the design of statistical graphics. Over 60 years ago, a remarkable series of papers by Haemer in *The American Statistician* made numerous cogent points that have been re-emphasized in the modern era (Haemer, 1947a,b, 1948a,b,c, 1949a,b,c, 1950a,b, 1951), including problems with double scales, attempts at using a third dimension, and the unhelpful use of colour.

It is true that graphs in different contexts serve different purposes, and sometimes graphs unsuitable for scientific purposes may, in non-science publications, attract people to a topic (Gelman and Unwin, 2013). However, the purpose of statistical graphics, especially in scientific discourse, remains vitally important.

Every budding statistician is routinely taught statistical graphics, and the reforms begun by Tukey

are generally applauded. There is however a contradiction. In his tongue-in-check old-school defence of tables over graphs, Gelman (2011) suggested that there are many applied scientists who remain unconvinced about the value of graphs. Feinberg and Wainer (2011) found a strong preference for tables rather than graphs as a display format in three high ranking journals from applied disciplines as well as in the *Journal of Computational and Graphical Statistics*. Gelman and others (Cleveland, 1984; Gelman et al., 2002; Kastellec and Leoni, 2007) have identified a need for greater use of graphs in statistics and allied disciplines literature. Cook and Teo (2011) recommended the use of graphs over tables based on the speed and accuracy with which information about statistical simulation studies could be decoded. However, the skeptics or "table people" (Friendly and Kwan, 2011) are unlikely to be convinced of the power of graphics without examples of graphical excellence.

This raised questions for us. Have statisticians taken up the leading role Cleveland advocated? Are statisticians any good at making graphs? In particular, are graphs published in statistics journals consistent with principles of good graphics? Is the quality of graphs in academic journals in statistics any better than in allied applied science disciplines?

PRINCIPLES OF GRAPHICAL EXCELLENCE

Cleveland (1984), for example, provided guidelines for authors after reviewing hundreds of graphs in scientific publications. We draw on the work of Cleveland and Tufte to provide a framework for evaluating the quality of statistical graphics, and for educating students. This framework integrates the principled insights of Cleveland and Tufte with practical rules of thumb. A good graph will be consistent with all the principles, a poor graph will fail in several ways. The same rule of thumb can follow from more than one principle.

While graphs are perhaps underutilized, there is a strong convention dictating the production of *some* graph in many contexts. Conference presentations of applied science research, for example, frequently include graphics. A picture may be worth a thousand words; but a picture is not necessarily worth one or two words. We should first ask if the space required for a graph is justified. If the answer is yes, then an overall standard should guide the production — does this graph stand alone? While aspects of the context (the text of an article, the verbal presentation) can clarify features of the graph, a high quality graph should be able to be interpreted without elaboration. This should be the goal.

Principle 1 is *Show the data clearly.* This is Tufte's maxim: "Above all else show the data." (Tufte, 1983, p.92)

Detection refers to the fundamental question of whether the important properties of the data can be detected by the visual system (Cleveland and McGill, 1985, 1987). In print, the space devoted to a graph might simply be inadequate. Cleveland (1984) advises assessing how well a graph can stand reduction. A graph might be otherwise well-constructed but if many of the elements of the graph overlap, individual elements may not be able to be detected. The data will not be shown clearly, and the graph fails.

Figure 1 is an example taken from a recent issue of the *International Journal of Epidemiology* (van Raalte et al., 2011). It shows the lifespan variation (standard deviation) versus average lifespan

in 10 European countries. For each country separate points are provided according to gender and education level (high, medium or low). The authors had the challenge of representing five variables and distinguishing six groups (gender by education level) without using colour. They chose to label points to draw attention to countries of interest. The data are however not shown clearly because clutter has not been avoided and the main patterns are not clear.

Figure 2 shows these data more clearly; our thinking in the construction of Figure 2 was as follows. We considered redesign of the original figure, keeping in mind the original authors goal of illustrating the relationship between average lifespan and lifespan variation for different genders and education level. Hence a scatterplot is an appropriate choice. We also chose to restrict our figure to gray scale, as the original figure has no color. The authors followed a natural approach in trying to represent groups in terms of symbols (and in other contexts, colors). However this approach fails to show the data clearly, and in order to allow easy identification of different countries, we introduced country panels. This choice means that the data are no longer aligned on single common x and y axes, but rather on multiple identical x and y axes. This is an example where the third principle discussed below—alignment on common scales—is not entirely maintained, in order to achieve clarity. The choice of symbols and gray scale colors to represent gender and education groups is challenging. Symbol intensity is increased to show higher levels of education, and two different symbols types are used to identify gender. This is a relatively complex encoding of two factors, so faint lines are added to join education levels for each gender; the purpose of this is to make the encoding of factors more transparent.

There is often more than one useful way to represent data. To illustrate this, Figure 3 shows the data of Figures 1 and 2 from a different perspective. This is a panel plot with the panels defined by gender and education level. In this plot, countries are not identified. Figure 3 shows the negative relationship between average lifespan and the standard deviation of lifespan within each combination of gender and education level. This plot clearly shows that average lifespan

increases as education level increases, lifespan variation decreases as education level increases, average lifespan is longer for females than for males, and lifespan variation is greater for males than for females.

Principle 2 is *Use simplicity in design.* Good graphics have a high data-ink ratio. Tufte argues that "Every bit of ink on a graphic requires a reason. And nearly always that reason should be that the ink presents new information" (Tufte, 1983, p.96). This is one aspect of creating a simple design.

The burgeoning field of infographics is replete with offerings for the bad graphics hobbyist where the temptation to embellish graphs is all too strong. Pictograms are also making a comeback, particularly on the internet. Simplicity in design does not imply simplicity in data represented or that a graph should be limited to small data sets. Tufte's (1983) work on small multiples, and recent developments in dynamic graphics (Rosling, 2008) show rich data with simple designs.

Principle 3 is *Use good alignment on a common scale for quantities to be compared.* Our third and fourth principles arise from the framework developed by Cleveland and McGill (1985, 1987) for statistical graphics that integrates thinking about human perception. A graph *encodes* quantitative and categorical information using symbols, geometry and color. Graphical perception is the *visual decoding* of the encoded information. An empirical study of elementary graphical-perception tasks (Cleveland and McGill, 1985, 1987) showed that the most accurate decoding arose when graphical elements to be judged or compared were positioned along a common scale; next was position along identical, non-aligned scales. Figure 1 represents all data points with reference to a common scale (one for each axis) whereas Figure 2 uses position along identical, non-aligned scales. Good graphs maintain constant measurement scales, within reason. This is usually possible.

The ubiquitous pie chart does not conform to principle 3; the human eye is bad at comparing angles. Pie charts are often presented with the percentages made explicit; this acknowledges the failure of the graphic to accurately communicate quantitative information. Stacked bar charts fail in the same way; only one category is aligned against the common scale and often the percentages

are printed on the graph for clarity.

In our view, accurate decoding should support accurate estimation of the quantities represented. We do not agree with Ehrenberg's (1978) assertion that "Graphs usually fail if they do not have a simple story-line to tell. This restricts them to communicating qualitative aspects of the data."(p.87) Nor with a more recent expression of the same view:

> Unlike data tables, graphs are not meant to provide precise quantitative values. Graphs reveal patterns, trends, relationships and exceptions via the shape of the data that would be difficult to discern from a table of values. Grid lines are rarely needed in graphs to help readers assign accurate numeric values to the data; the approximate values that can be perceived without the aid of grid lines are almost always adequate. (Few, 2005)

Following this recommendation, echoed by Bigwood and Spore (2003, p.44), entails an unnecessary loss. A good graph can communicate the broad pattern *and* some details of the data. Often a scatterplot with suitable grid lines can actually represent the individual observations quite precisely. Light grid lines can help guide the comparisons, and can assist with accurate estimation of the values represented by the data points.

It is perhaps unfortunate that Tufte (1983) discussed the use of grid lines in his chapter on chartjunk, as his discussion is sometimes taken as supporting the removal of grid lines. Consider this advice on graph construction: "Chartjunk consists of unnecessary and distracting elements. The most common of these are grid lines ..." (Bigwood and Spore, 2003, p.43), and "Grid lines are commonplace in business graphs today but they are almost always chartjunk — visual content that adds no value, serves no purpose and distracts from the real data" (Few, 2005).

Here are Tufte's (1983) thoughts: "A gray grid works well and, with a delicate line, may promote some accurate data reconstruction than a dark grid." [p.116]. To Tufte, grid lines are not chartjunk;

they can facilitate estimation of *quantity*. Note how the grid lines facilitate comparisons of interest in Figure 2.

Principle 4 is *Keep the visual encoding transparent.* Building on the insight of Cleveland and McGill's (1987) regarding visual encoding and decoding, we have suggested the principle of "transparent visual encoding" (Finch and Gordon, 2014): the creator of a graph should aim to make the viewer's task — visual decoding — as simple as possible. If possible, the decoding necessary should be transparent: the viewer should be barely aware of doing it. If it is hard work to understand and explain a graph, the visual encoding is not transparent.

There are many ways in which it can be difficult for the reader to decode a graph. Again, simplicity in design and clear labelling of the graph contribute to easier decoding. All elements of the graph must be defined; for example, bars around point estimates might be used to show a standard deviation or a standard error, or a confidence interval. This should be stated explicitly.

The practical use of the graph must be considered: colour is only useful if the reader will see the graph in colour. Some form of grey scale tonal variation is an alternative to colour. Another option is to use different shaped symbols.

Poorly designed plots of time series data that are difficult to decode are easy to find, such as two time series shown as bars, super-imposed, leading to masking: a detection problem. Generally, the use of bars in situations where points might be used creates unhelpful decoding complexity.

When modern graphs use two scales on the same graph, even when these are clearly identified, there is scope for confusion, or for being misled, particularly if the reader looks at the plot casually.

Cleveland and McGill's (1985; 1987) notion of detection is important in ensuring transparent visual encoding, and in some cases ensuring that the data are detected will mean that data may need to be presented on common but non-aligned scales. In effect, this means using a panel plot. Panel plots violate Principle 3, since some of the data that we wish to compare are no longer aligned on

a common linear scale, although the second best option is used, and they are sometimes the best design. This is illustrated in Figure 2.

Cleveland (1985) also discussed the role of distance in visual perception. Distance refers to the proximity of data items to be compared on the graph (other than on the measurement scale). As the distance between items to be compared increases, the accuracy with which comparisons are made decreases. The idea can be taken further: often we want to compare quantities in different graphs, and the ease with which this can be done is certainly affected by how close the two graphs are together, even if they are not lined up. On an individual graph, faint grid lines can help reduce the problem caused by distance. Ordering the quantities by magnitude, if appropriate, can also help in graphs, as it can in tables (Feinberg and Wainer, 2011).

Principle 5 is *Use graphical forms consistent with Principles 1 to 4.* A set of standard and effective graphical forms can cover the majority of graphing needs, and are consistent our first four principles. These are: histograms, dotplots, boxplots, line plots, bar or dot charts, and scatter plots. The addition of panels to extend these is often powerful and appropriate. These are the types of graphs modelled, for example, by Cleveland (1984) and Tufte (1983).

Consider the scatter plot. A symbol represents a pair of observations simultaneously, one on the *x*-axis and one on the *y*-axis. This is an example of encoding: it is a long-established convention that the *x* and *y* values can be read off the graph by projecting onto the relevant axis: this is the decoding process. It is reasonably obvious, making it relatively transparent.

On the other hand, the rules for the boxplot are quite detailed and complicated. Users come to learn them and internalize the encoding involved. But it is not at all uncommon for a boxplot viewer to be uncertain about the way the whiskers are constructed. This is associated with wrongly constructed boxplots, or unhelpful variations.

The imperative to be creative in producing a good graph does not mean non-standard graphical forms should be used. Accurate decoding relies on standard forms.

The five principles provide a framework for evaluation of the graphs in our study. For each principle, a set of questions about features of a graph was articulated.

OUR STUDY

Are statisticians leaders in graphical excellence as Cleveland suggested we should be? We examined published graphs in the scientific literature drawing from highly-regarded academic journals. Our focus is on this source of static graphs, arguably more important in practice than dynamic graphics. The form and structure of static graphs provide a basis for excellence in dynamic graphs where the design can be more challenging.

The Australian Research Council's 2010 "Excellence in Research Australia" evaluation provided ordinal rankings of academic journals across all disciplines, with the best journals receiving an A* rating. In all 1030 academic journals received this rating; this was the top 5% of journals. We expect papers published in statistics, applied science and social science journals rated A* to be a source of quality graphics, given competition to publish in these journals and the rigorous review processes.

Sampling and evaluation

The Australian Research Council classifies journals according to one or more field of research (FOR) codes. We identified A* journals publishing work in statistics and allied applied science disciplines by considering the first and second FOR codes.

There were 327 A* journals with FOR codes in the environmental sciences, agricultural and veterinary sciences, medical and health sciences, education, economics, psychology, but not statistics. We took our applied science sample from these journals. There were 17 A* journals with a FOR code in statistics; we took our statistics sample from these journals.

We sampled from all 17 statistics A* journals. We chose a random starting page within the most recently available issue online and sampled the first page with at least one graph starting from the randomly selected page. If necessary, we randomly sampled one of the *k* graphs on a page. We worked through from the random starting point to the end of the issue and then started at the first page and continued back to the start point, if required. We examined the three most recent issues of each journal in this way. One journal had no graphs in the three issues we examined, and one issue of another journal had no graphs. We therefore found 47 statistics graphs.

We took a simple random sample of journals from among the 327 applied science A* journals. A graph from a selected journal was sampled by finding the most recently available issue online. Within the chosen issue, the process was as for the statistics graphs. We sampled 55 applied science journals, and found 50 graphs; five journals had no graphs in the issue sampled. A list of the A* applied science and statistics journals from which we sampled graphs is in Appendix A.

Over 60 different features of each graph were coded; both authors reviewed all the graphs. The coded features related to the five principles of good graphics. For the principle *Show the data clearly*, we recorded, for example: Did the graph have detection problems? Are there undefined graphical elements? Are the axes labelled appropriately? Without direct reference to these features, each graph was also assigned an overall quality rating: poor, adequate, good or exemplary. Any

disagreements between the authors were resolved by discussion. Examples of some of the coding for two graphics in our study are given in Appendix B.

OUR FINDINGS

Overall quality

No graphs in our sample were rated as exemplary; 39% overall were poor, with a higher proportion in the applied science graphs than in the statistics graphs (Figure 4). Many of the features of graphs we coded were undesirable. We counted the number of undesirable features identified for each graph. Figure 5 shows dotplots of the number of undesirable features by the overall quality rating in each group. This illustrates that graphs with better overall quality ratings had fewer poor features, on average. This provides evidence of the coherence of the subjective quality ratings.

The overall quality was slightly lower for the applied science graphs than for the statistics ones (Figure 5). However, fewer of the statistics graphs (13%) than applied science graphs (26%) stood alone. We also coded if the graph had an obvious statistical problem; by this we mean, something which was a violation of standard theory and practice, wrong, internally inconsistent, or impossible, from a statistical point of view. This was true for 10% of the applied science graphs and 4% of the statistics graphs. For example, one graph showed a two dimensional 'scatterplot' where the x-axis was unlabelled, had no tick marks, and no apparent meaning. Another graph claimed to present quarterly data; the x-axis showed yearly labels but there appeared to be variable numbers of quarters (sometimes more than four) in each year. Captions of these graphs provided no clarification.

Representing data and inference

Most graphs represented data and/or estimates; see Figure 6. Only 4% of the applied science graphs represented something other than data and/or estimates, compared with one quarter of the statistics

sample. Statistics graphs included more examples of fitted models and theoretical distributions.

The principles of simplicity and transparency imply that point estimates should be represented graphically as points (rather than bars). Twelve graphs, all from the applied science sample, used bars instead. This was 41% of the applied science graphs that represented estimates.

Inferential results can be encoded graphically as point estimates with 'error' bars representing uncertainty; preferably the 'error bars' are confidence intervals rather than standard errors (Cleveland, 1985). Cleveland saw that "the difficulty . . . is that we are visually locked into what is shown by the error bars; it is hard to multiply the bars visually by some constant to get a desired visual confidence interval on the graph. Another difficulty, of course, is that confidence intervals are not always based on standard errors" (p.219).

In 13/29 (45%) of the applied science graphs with estimates, a representation of uncertainty was included; this compares with 4/20 (20%) in the statistics sample.

Only five out of the 17 graphs with 'error bars' showed confidence intervals; another five did not specify the meaning of bars. Visual decoding can be challenging with bars on bars — point estimates shown as bars with 'error' bars; half of the 12 applied science graphs with point estimates as bars had this feature.

In our view, inference is best represented graphically by plotting confidence intervals. There is a tradition in some disciplines of providing information relating to *P*-values on a graph; this includes exact *P*-values, relative *P*-values (e.g. $P < 0.05$) and star ratings (e.g. *, **, ***) corresponding to relative *P*-values. Relative *P*-values are relatively uninformative. *P*-values were reported in some form in 10/50 (20%) applied science graphs and one statistics graph; only three gave *P*-values to some decimal places.

What was done well?

Most graphs conformed to the principle of aligning elements to be compared along a common

scale (78%). In 19% of graphs we judged that the data could be shown more clearly if panels were used — in some cases the use of a single common scale reduced the clarity of the data.

Cross hatching, a source of visual noise, is now out of vogue (2% of graphs) presumably because of improvements in printing, and the use of colour. Similarly use of the principle of "Anaheim first" (Feinberg and Wainer, 2011) (alphabetical order) was rare; we judged only 4% of graphs could be improved with a reordering of elements.

What needs improving?

Many problems with the graphs related to detection — and just under half (45%) of all the graphs had detection problems. More than half (53%) the graphs has undefined abbreviations and 23% had undefined graphical elements. One example of an undefined graphical element was an interaction plot where the groups (lines) were not labelled; the reader had to infer them from the information in the body of the paper. Another example is undefined 'error' bars; they could be standard errors, standard deviations or confidence intervals.

Only around one third (35%) of the graphs provided suitable axis labels and in 14% of all graphs at least one of the axis labels referred to the variable measured rather than a point estimate relating to the variable measured. This can be ambiguous when, for example, the proportion of correct responses to a set of questions is measured for individuals and the average proportion is plotted but not clearly labelled. Ideally the tick mark labels on all axes should be horizontal — a simple instantiation of the principle of showing the data clearly. In 30% of all graphs, labels for at least one of the axes were not horizontal, a likely consequence of software defaults.

There were a number of ways in which Tufte's principle of simplification could be adopted. One quarter of the graphs used colour, but in almost half of these the use of colour was redundant. A legend was provided in 42% of the graphs but in one third of these we judged that direct labelling would have been a better option as it can reduce the amount of decoding the reader needs to do.

Only 31% of graphs used grid lines. This is likely to be partly due to software defaults. In 30% (9/30) of the graphs with grid lines, the grid lines were too heavy. We judged that some/additional grid lines could be used in 84% of all the graphs. Some examples from our sample illustrate the importance of grid lines. One graph showed a time series of estimates of proportions (on a percentage scale); the final point estimate appeared to be plotted higher on the *y*-scale than the tick mark corresponding to 100%, and the upper bound of the confidence interval plotted around the point estimate was clearly over 100%. If the authors had added a grid to their graph, these problems would have become transparent. Another graph compared performance under two different conditions over time; under one condition the outcome was initially poor but improved over time whereas in the other condition initial performance was good but declined over time. An important question in this context is about the point of intersection of the two lines on this graph: at what time is the performance equivalent? Without grid lines, this was quite difficult to approximate.

There were 18 graphs that did not correspond to one of the recommended graphical forms; all but one came from an applied science journal. Many of the non-standard forms plotted point estimates as bars, sometimes with error bars. We found 'innovation' at its worst in the representation of a simple distribution. We have already mentioned an example where a simple distribution was represented in two dimensions — the second dimension was meaningless and unexplained. Another example showed an "empirical distribution"; it was described as a histogram but the bars plotted were not contiguous and the tick mark labels were not evenly spaced. A footnoted explanation suggested that the values on the *x*-axis corresponded to various percentiles of the distribution. This confusing representation of a simple distribution may have been produced with naive use of an Excel histogram function.

Applied science disciplines versus Statistics

The differences between the applied science and statistics graphs based on overall quality ratings and the number of undesirable features were small. In a number of, but not all, aspects of detail, the

statistics graphs appeared to be better. Figure 7 provides a caricature of graphs of estimates from an applied scientist (left panel) and a statistician (right panel), based on differences we observed between the applied science and statistics samples. The estimates plotted are the mean number of undesirable features in our study — broken down by the overall quality rating of the graph and the discipline (applied science or statistics).

The statistician sticks to a standard form (98% of statistics (S) graphs compared with 66% of applied science (AS) graphs) and avoids redundant use of colour (S: 6%, AS: 16%). When graphing estimates he labels his axes appropriately (axis labelled as variable rather than parameter estimate: S: 5%, AS: 41%), but generally uses non-horizontal tick marks (S: 55%, AS: 6%). The statistician's graph of estimates uses points rather than bars (S: 100%, AS: 59%). The applied scientist presents estimates as bars and includes a representation of the uncertainty of the estimates (S: 20%, AS:45%). In graphing estimates as bars, the applied scientists also represents uncertainty as a bar (50% AS graphs with estimates as bars).

CONCLUSIONS

In the best academic journals in 2012, we failed to find exemplars of graphical excellence. Graphs produced by statisticians and applied scientists left much to be desired and there is no clear evidence that the graphs in (high quality) statistics journals are better than the graphs in (high quality) applied science journals.

Generally, the cost of producing excellent graphics is small and the potential benefit of statisticians providing models for their applied colleagues to follow is large. Why is there no strong connection between valued principles (Tufte's work, for example, has been cited thousands of times) and actual practice? Is there a failure of understanding, of practice, or of both?

In our introduction we argued that an excellent graph stands alone — it can be accurately decoded without reference to text or verbal explanations. In many of the graphs we examined, there was a communication gap as the reader is challenged to unambiguously decode details, perhaps because the person encoding the graph was too familiar with the data and its context. This may be a failure of practice and attention to detail. Here we provide a checklist to elucidate all the important points of detail. While other checklists have been produced (e.g. Duke (2014), Government Digital Service (no date)), ours is explicitly related to the goal of graphical excellence, and developed from the findings of this study. Editors and reviewers, as well as authors, could make good use of this checklist. The graphs we sampled had survived the editing and review process.

This study examined graphs that authors have included in their articles. Often, however, graphs are simply omitted. This can be an important obstacle to communication, when insights of inferences or other results could be shown in a visually incisive way. In a sense, these omitted graphs are "missing values" in our study; examining this issue systematically would be a worthwhile sequel to this work.

Authors of journal articles should not regard the process of graph creation as straightforward and benign. They should understand that skill and expertise are needed, and that these can be acquired.

At the very least, those producing graphs should familiarize themselves with the thinking and insights of the key works of Cleveland (1985) and Tufte (1983, 1997); the simple books by Robbins (2005) and Evergreen (2014) are useful sources for novices. Wong's (2010) book is more suitable for those in the information graphics world, including the media.

The ease of producing a graph consistent with our principles from a software default varies substantially from package to package. Many of the graphs we reviewed were plausibly produced with little effort by the encoder to change software defaults. Examples of such defaults that are not conducive to excellent graphics are the use of grey backgrounds in Minitab, the absence of grid lines in several packages (*R*, Minitab, SPSS), and the production of tick labels on the y-axis that are not horizontal (*R*). There was an absence of grid lines in many graphs, for example, that appear to have been produced using standard *R* functions. In many software packages, it is straightforward for the user to make changes to the default graphical style. This does not however appear to be the statistician's or the applied scientist's routine practice.

Many graphical faults would disappear if software packages came with defaults that were consistent with principles of good graphics. *ggplot2* (Wickham, 2009, 2011; Chang, 2012), provides an excellent model of this; for example, a background grid is the default. Statisticians who care about quality graphics should focus some of their lobbying efforts on software producers. Software tools can make the production of good graphics hard work. If our defaults look more Tufte-esque, perhaps the steps to excellence would be easier to take.

A checklist for good graphical practice

How clear is your purpose in communication?

- ☐ What relationships or patterns can you identify in the graph?
- ☐ Are these the relationships or patterns you intended to represent?
- ☐ Can the viewer identify the patterns you wish to illustrate?
- ☐ Are the important comparisons you wish to show salient?

Make clarity a high priority.

- ☐ Does the graph have a clear title?
- ☐ Are the axes labelled?
- ☐ Are the units of the variables measured defined?
- ☐ Are the units of observation clear?
- ☐ Is the graph large enough?
- ☐ Would ordering groups or variables plotted improve the graph?
- ☐ Are all the graph labels horizontal?
- ☐ Use points to plot estimates (e.g. means, proportions) rather than bars.

Choose standard forms fit for your purpose.

- ☐ Use bars around points to indicate the precision of the estimates.
- ☐ Plot the estimates of interest (e.g. mean differences with confidence intervals) rather than standard summary statistics (group means).
- ☐ Plot inferences to support stories about models.
- ☐ Plot data to support stories about distributions and variation.

Consider detection issues.

- ☐ Can all the data points be seen?
- ☐ Are patterns in the data clear?

- Are the fonts large enough?
- Would it help to use jittering, or another form of representing multiple, identical values?

Would panels help?

- Are there grouping variables that can be used to panel the graph?
- Do the grouping variables correspond to the variation of interest?
- Would additional panels help?

Align quantities to be compared on a common scale.

- Has distortion of the data been avoided by using the same scales for the same measurement?
- Are measurements made on the same scale plotted on the same scale?
- Would transposition improve the graph?

Does the graph have grid lines?

- Light grey grid lines will help with accurate interpretation.

Are all the elements of the graph defined?

- What do points on the graph correspond to?
- Are estimates (e.g. means, proportions) plotted on the graph clearly defined?
- Are bars around points on the graph clearly defined?

How much decoding work does the viewer have to do?

- Is it easy for someone unfamiliar with your data to interpret your graph?
- Does the graph stand alone?
- Try it on a friend!

APPENDIX A: JOURNALS IN OUR SAMPLE

Applied science journals

Advances in Agronomy
American Journal of Transplantation
Annual Review of Neuroscience
Arthritis and Rheumatism
Behavioral and Brain Sciences
Brain
Cell Metabolism
Cochrane Database of Systematic Reviews
Cognition
Cognitive Science
Diabetes
Educational Administration Quarterly
Educational Researcher
Endocrine Reviews
Environment International
Fish and Fisheries
Health Psychology
IEEE Transactions on Evolutionary Computation
IEEE Transactions on Image Processing
IEEE Transactions on Information Technology in Biomedicine
Immunity
Indoor Air: international journal of indoor air quality and climate
International Endodontic Journal
International Journal for Parasitology
Investigative Ophthalmology and Visual Science
JAMA: Journal of the American Medical Association
Journal of Abnormal Psychology
Journal of Accounting and Economics
Journal of Biomechanics
Journal of Bone and Joint Surgery-American Volume
Journal of Cognitive Neuroscience
Journal of Educational Psychology
Journal of Endodontics
Journal of Experimental Psychology: Animal Behavior Processes
Journal of Experimental Psychology: General
Journal of Law Economics and Organization
Journal of Phonetics
Journal of Political Economy

Journal of Thrombosis and Haemostasis
Lancet Oncology
Midwifery
Molecular Nutrition and Food Research
Molecular Pharmacology
Morphology
Psychological Science
Sports Medicine
The Economic Journal
The Linguistic Review
The Review of Financial Studies
Trends in Ecology and Evolution

Statistics journals

Annals of Applied Probability
Annals of Applied Statistics
Annals of Probability
Annals of Statistics
Bioinformatics
Biometrics
Biometrika
Biostatistics
Epidemiology
International Journal of Epidemiology
Journal of Business and Economic Statistics
Journal of Computational and Graphical Statistics
Journal of the American Statistical Association
Journal of the Royal Statistical Society Series A (Statistics in Society)
Journal of the Royal Statistical Society Series B (Statistical Methodology)
Probability Theory and Related Fields*
Statistics in Medicine

*This journal was eligible according to the study protocol, but no graphs were found in the issues sampled.

An Excel file containing exact references to the 97 plots assessed in this study is available from the authors on request.

APPENDIX B: EXAMPLE OF CODING IN OUR STUDY

The coding of some of the features of graphs is illustrated in the table below, for two graphs in our study.

The first is Figure 2 from “More variation in lifespan in lower educated groups: evidence from 10 European countries” (*International Journal of Epidemiology* (2011), 40, 1703-1717, used with permission).

The second graph is Figure 1 from “A survival analysis approach to modeling human fecundity” (*Biostatistics* (2012), 13, 4-17, used with permission).

Example of some coding for the figures shown.

Scatterplot, first graph	Line plot, second graph
Adequate caption	Inadequate caption
Suitable axis label(s)	Poor axis label(s)
Detection problems	No detection problems
Undefined graphical elements	Undefined graphical elements
Graph does not stand alone	Graph does not stand alone
Elements to be compared are aligned	Elements to be compared are aligned
Gridlines could be added	Gridlines could be added
Graph is a standard form	Graph is a standard form
Panels could be used	No additional panels needed
Summary statistics/estimates shown	Summary statistics/estimates shown
Uncertainty on estimates not shown	Uncertainty on estimates not shown
No obvious statistical problem	No obvious statistical problem
Overall rating: Adequate	Overall rating: Poor

References

Bigwood, S. and Spore, M. (2003), *Presenting Numbers, Tables, and Charts*, New York, NY: Oxford University Press.

Chang, W. (2012), *R graphics cookbook*, Sebastopol, CA: O'Reilly Media, Inc.

Cleveland, W. (1984), "Graphs in scientific publications," *The American Statistician*, 38, 261–269.

— (1985), *The Elements of Graphing Data*, New York: Chapman and Hall.

Cleveland, W. and McGill, R. (1985), "Graphical perception and graphical methods for analyzing scientific data," *Science*, 229, 828–833.

— (1987), "Graphical perception: the visual decoding of quantitative information on graphical displays," *Journal of the Royal Statistical Society. Series A (General)*, 150, 192–229.

Cook, A. R. and Teo, S. W. (2011), "The communicability of graphical alternatives to tabular displays of statistical simulation studies," *PloS One*, 6, e27974.

Duke, S. (2014), "Best practices recommendations," https://www.ctspedia.org/do/view/CTSpedia/BestPractices.

Ehrenberg, A. S. C. (1978), "Graphs or tables?" *The Statistician*, 27, 87–96.

Evergreen, S. D. (2014), *Presenting Data Effectively: Communicating Your Findings for Maximum Impact*, CA: SAGE Publications.

Feinberg, R. A. and Wainer, H. (2011), "Extracting sunbeams from cucumbers," *Journal of Computational and Graphical Statistics*, 20, 793–810.

Few, S. (2005), "Grid lines in graphs are rarely useful," http://www.perceptualedge.com/articles/dmreview/grid_lines.pdf.

Finch, S. and Gordon, I. (2014), "The development of a first course in statistical literacy for undergraduates," in *Topics from Australian Conferences on Teaching Statistics*, eds. MacGillivray, H., Martin, M. A., and Phillips, B., New York: Springer.

Friendly, M. and Denis, D. J. (2014), "Milestones in the history of thematic cartography, statistical graphics, and data visualization," http://datavis.ca/milestones//.

Friendly, M. and Kwan, E. (2011), "Comment on 'Why tables are really much better than graphs'," *Journal of Computational and Graphical Statistics*, 20, 18–27.

Gelman, A. (2011), "Why tables are really much better than graphs," *Journal of Computational and Graphical Statistics*, 20, 3–7.

Gelman, A., Pasarica, C., and Dodhia, R. (2002), "Let's practice what we preach: turning tables into graphs," *American Statistician*, 56, 121–130.

Gelman, A. and Unwin, A. (2013), "Infovis and statistical graphics: different goals, different looks," *Journal of Computational and Graphical Statistics*, 22, 2–28.

Government Digital Service (no date), "Data visualisation: Creating valuable and meaningful graphics to help analyse data," http://www.gov.uk/service-manual/user-centred-design/data-visualisation.html.

Haemer, K. W. (1947a), "Hold that line," *The American Statistician*, 1, 25.

— (1947b), "The perils of perspective," *The American Statistician*, 1, 19.

— (1948a), "Double scales are dangerous," *The American Statistician*, 2, 24.

— (1948b), "Question 9: negative numbers and semilog paper," *The American Statistician*, 2, 18.

— (1948c), "Range-bar charts," *The American Statistician*, 2, 23.

— (1949a), "Presentation problems: area bias in map presentation," *The American Statistician*, 3, 19.

— (1949b), "Presentation problems: the supplementary-scale chart: two charts for the price of one," *The American Statistician*, 3, 11.

— (1949c), "Question 23: graphic presentation," *The American Statistician*, 3, 10.

— (1950a), "Presentation problems: a simplified ranking chart," *The American Statistician*, 4, 21.

— (1950b), "Presentation problems: color in chart presentation," *The American Statistician*, 4, 20.

— (1951), "The pseudo third dimension," *The American Statistician*, 5, 28.

Kastellec, J. P. and Leoni, E. L. (2007), "Using graphs instead of tables in political science," *Perspectives on Politics*, 5, 755–771.

Robbins, N. B. (2005), *Creating more effective graphs*, Hoboken, NJ: Wiley-Interscience.

Rosling, H. (2008), "Gapminder," `http://www.gapminder.org/`.

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

— (1997), *Visual Explanations*, Cheshire, CT, USA: Graphics Press.

Tukey, J. W. (1972), "Some graphic and semigraphic displays," in *Statistical Papers in Honor of George W. Snedecor*, ed. Bancroft, T., Iowa State University Press, pp. 293–316.

— (1977), *Exploratory Data Analysis*, Reading, Mass: Addison-Wesley.

van Raalte, A., Kunst, A., Deboosere, P., Leinsalu, M., Lundberg, O., Martikainen, P., Strand, B., Artnik, B., Wojtyniak, B., and Mackenbach, J. (2011), "More variation in lifespan in lower educated groups: evidence from 10 European countries," *International Journal of Epidemiology*, 40, 1703–1714.

Wainer, H. (1984), "How to display data badly," *American Statistician*, 38, 137–147.

Wickham, H. (2009), *ggplot2: Elegant graphics for data analysis*, New York: Springer.

— (2011), "ggplot2," *Wiley Interdisciplinary Reviews: Computational Statistics*, 3, 180–185.

— (2013), "Graphical criticism: some historical notes," *Journal of Computational and Graphical Statistics*, 22, 38–44.

Wong, D. (2010), *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts and Figures*, New York: Norton and Company.

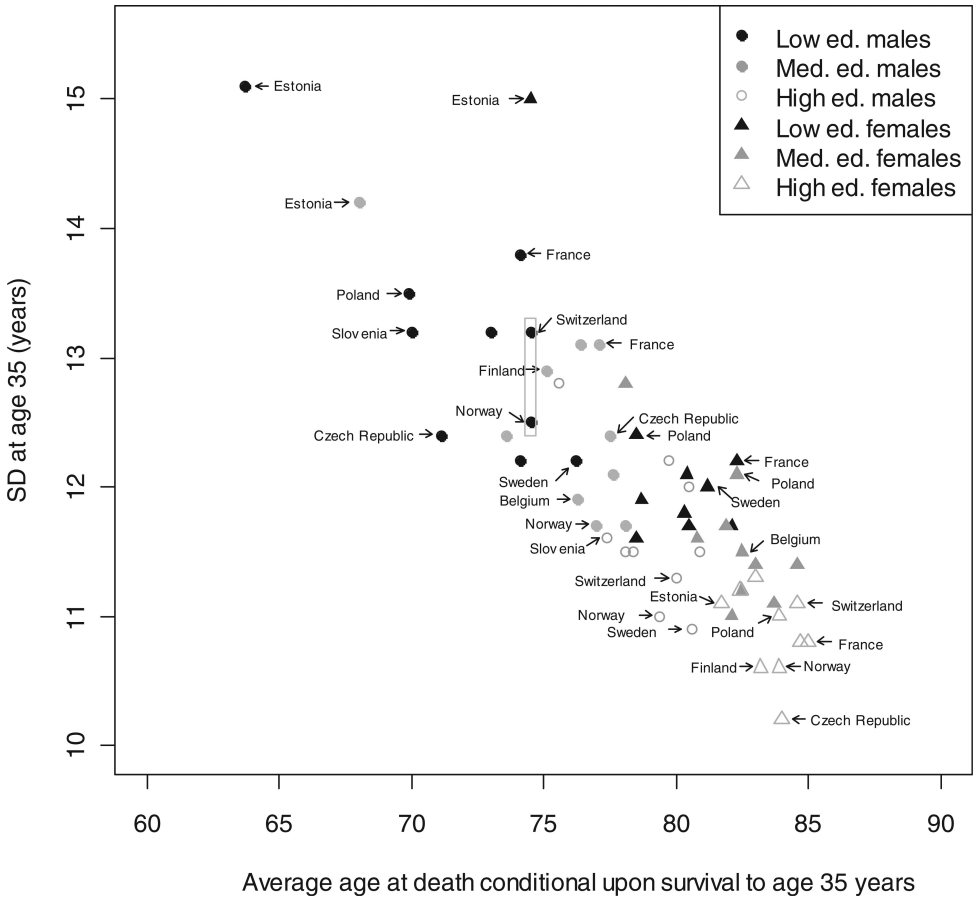

Figure 1: Scatterplot of lifespan variation versus average lifespan [vanRaalte et al., 2011]. The original caption reads: *Relationship between lifespan variation (SD at age 35 years) and average lifespan (conditional upon survival to age 35 years) by sex and education. All data points in Table 1 and 2 are plotted, but some are not labelled to avoid clutter.*

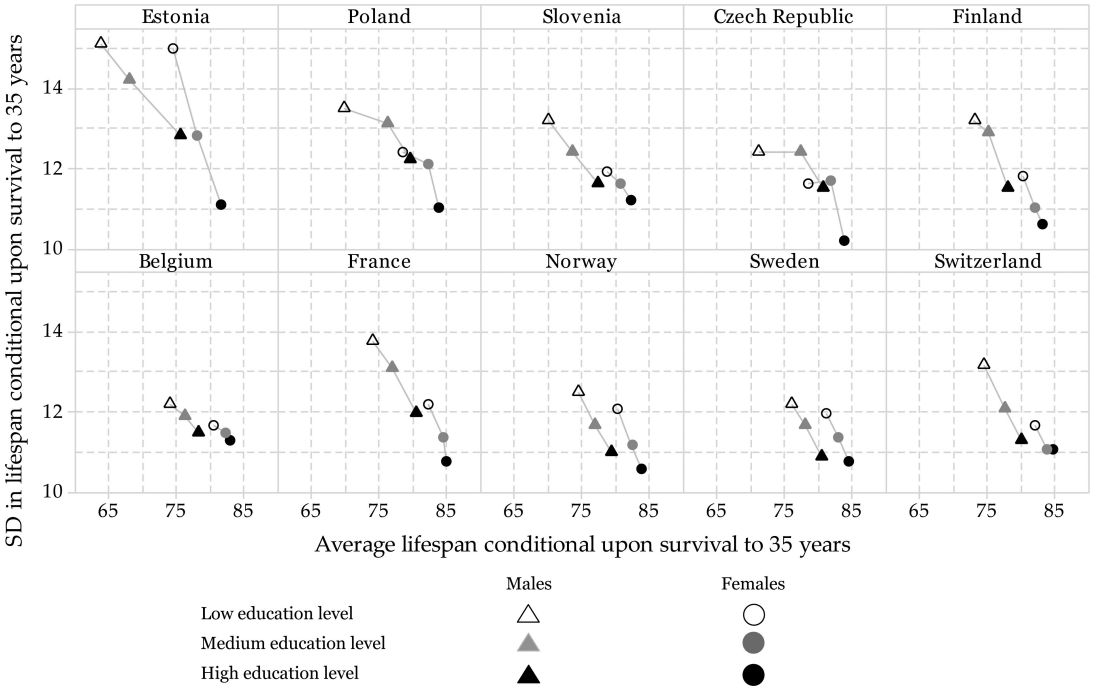

Figure 2: Scatterplot with panels that shows the data from Figure 1 clearly.

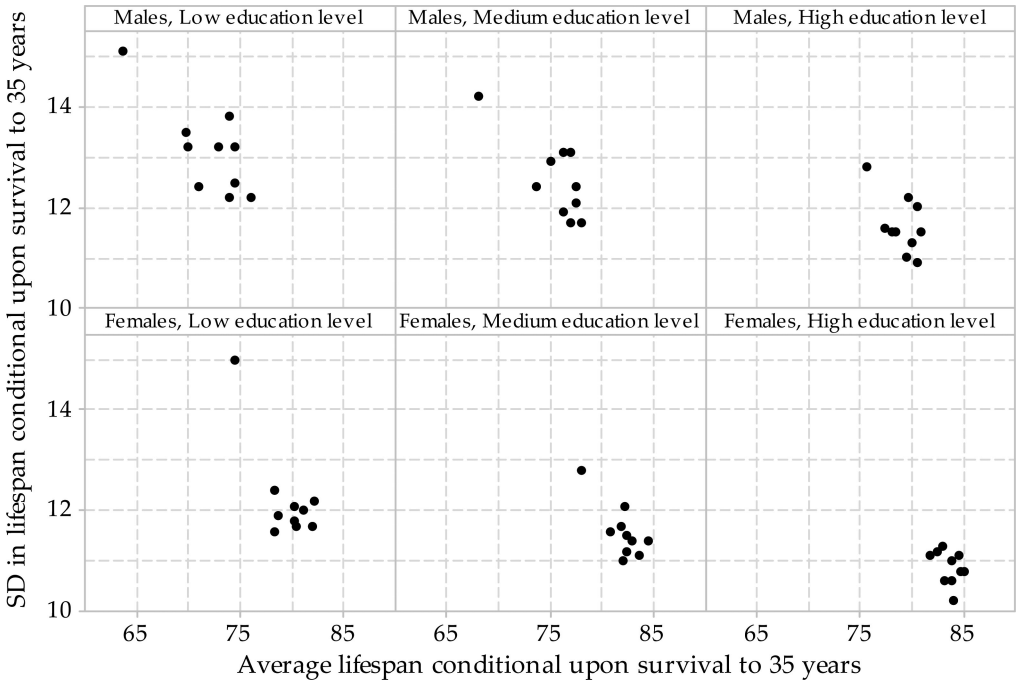

Figure 3: Alternative panel plot for the data from Figure 1. Each data point represents a country; the countries are not labelled.

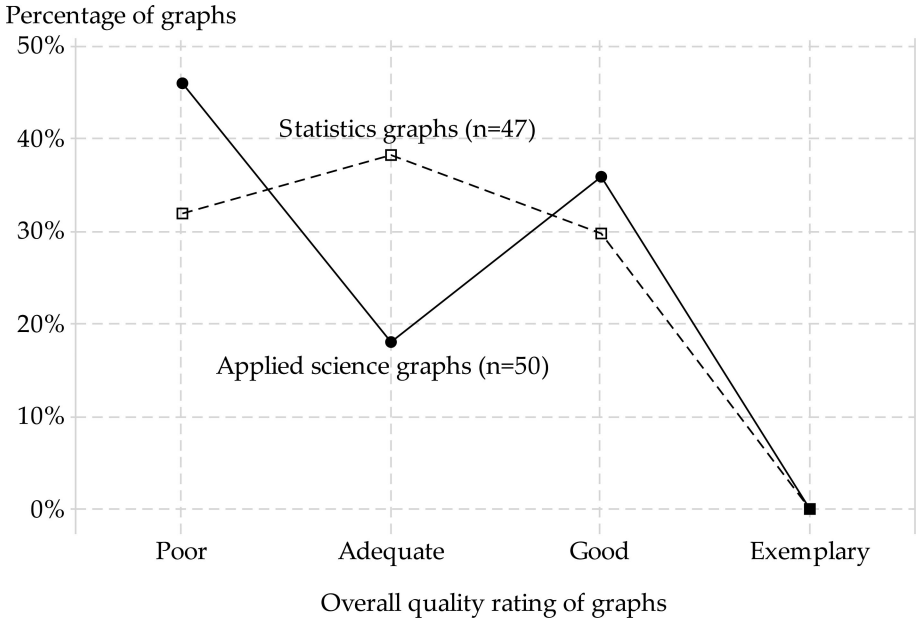

Figure 4: Quality rating of graphs sampled from A* journals in statistics (n = 47) and applied science (n = 50).

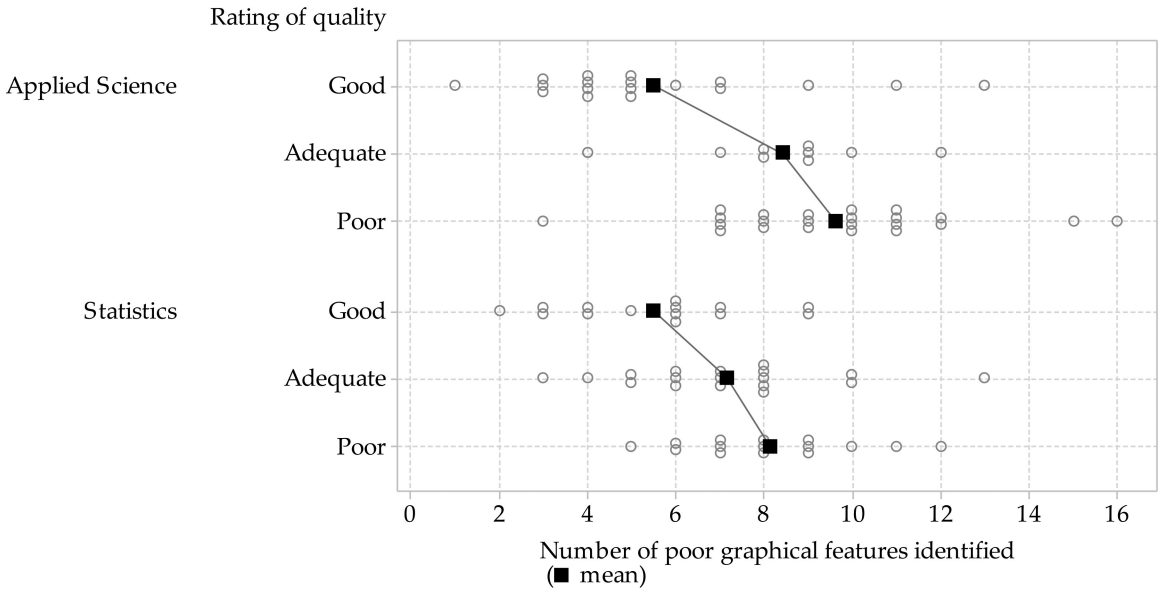

Figure 5: Number of poor features in graphs by overall quality rating and discipline.

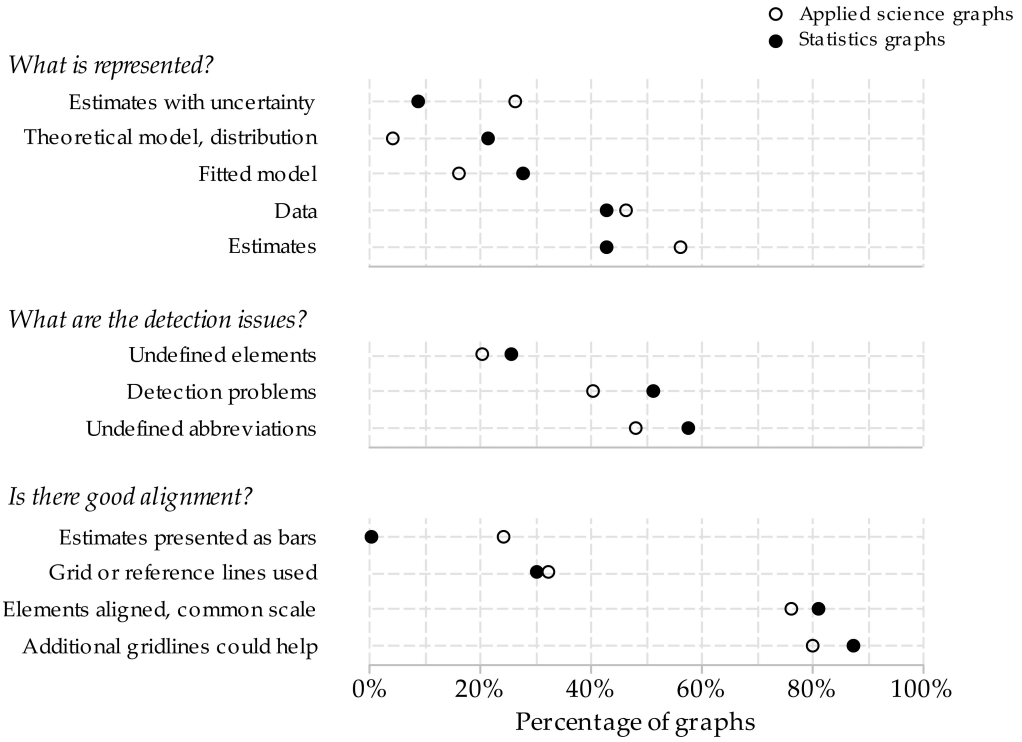

Figure 6: Percentage of graphs sampled from A* journals in statistics (n = 47) and applied science (n = 50) with various features.

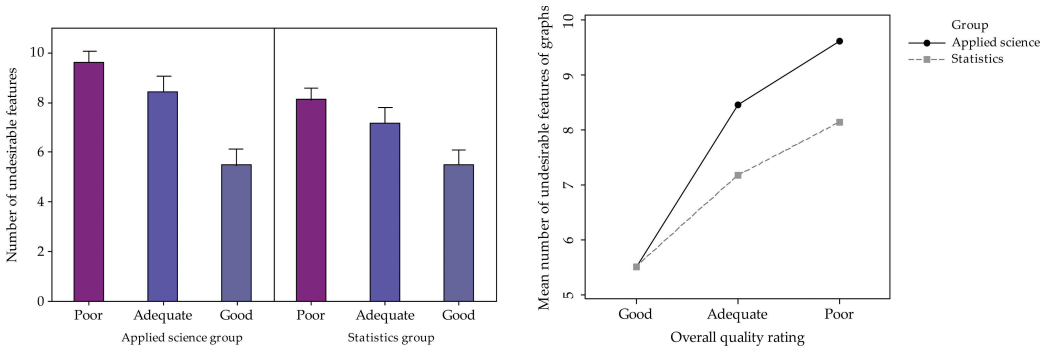

Figure 7: Caricature of graphs produced by an applied scientist (left panel) and a statistician (right panel); see text for details.

Fig. 1. Plot of $\lambda^{(k)}(j)$ for the New York State Angler Prospective Pregnancy Cohort Study for an average-age nonsmoking couple by parity for (a) cycle 1, (b) cycle 2, (c) cycle 3, and (d) cycle 4.

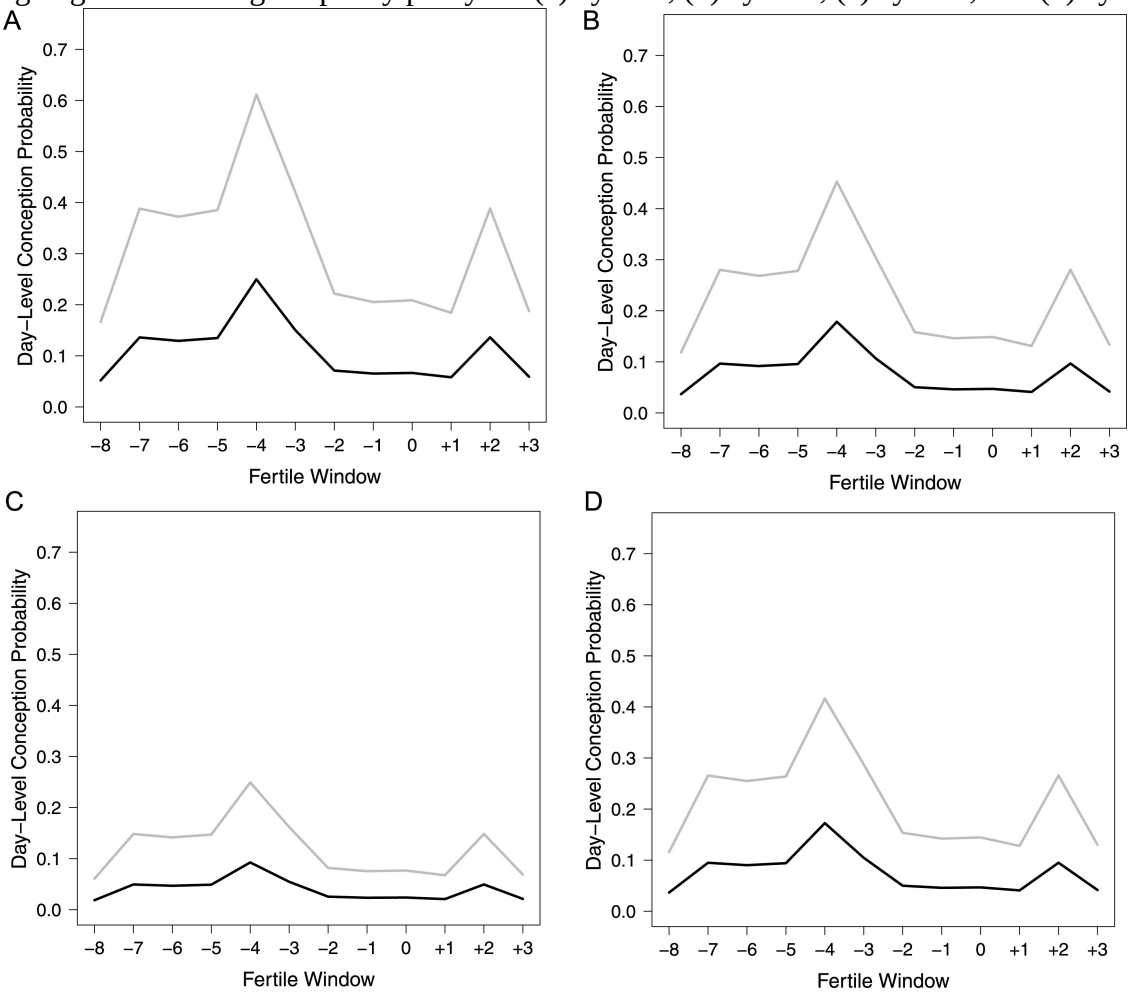

Figure 2 Relationship between lifespan variation (SD at age 35 years) and average lifespan (conditional upon survival to age 35 years) by sex and level of education. All data points in Tables 1 and 2 are plotted, but some are not labelled to avoid clutter