Why can't measurements based on mathematical models be more user-friendly? Problems, causes and suggestions

11 January 2012

Michael Wood

University of Portsmouth Business School

SBS Department, Richmond Building

Portland Street, Portsmouth

PO1 3DE, UK

+44(0)23 9284 4168

michael.wood@port.ac.uk .

Abstract

The outputs of mathematical models are often measurements on scales which may not be properly understood by some users – this includes both lay people and specialists in other areas. This may lead to these outputs, and the underlying models, being ignored, or misunderstood, or used or interpreted in ways that misrepresent the assumptions in the model. The practical consequences of these problems range from the enormous amount of time and energy devoted to trying to educate users, to the consequences of the misinterpretation of some financial valuation models which contributed to the recent financial crash. This paper analyses some examples of these problems with short case studies of *p* values in statistics, financial valuation models, university league tables and sigma measures, suggests that it is often possible to redesign measurements to make them easier to interpret, and proposes some principles for doing this. In the long term, the redesign of measurements from the perspective of potential users is likely to be an important facilitator of the growth and use of knowledge. However, in the short term there are powerful inhibiting factors.

Keywords: Mathematical models, Measurement, Public understanding of statistics, User-friendliness .

Introduction

Measurements are important in the modern world. We measure, for example, intelligence, the quality of university courses and industrial processes, the value of investments, the certainty of scientific conclusions, risk, and so on. The importance of these measurements stems from the fact that they are increasingly involved in decision making.

The word "measurement" refers to the process and result of measuring – which I will take to be the process of assigning a number to some aspect of the world in order to convey information about it. So we can measure lengths with rulers, temperatures with thermometers and the prices of goods in a shop by reading the label.

The measurements listed in the first paragraph above, however, are more complex: intelligence is measured by means of a test combined with some rules about how to use the results to compute intelligence scores, and the certainty of scientific conclusions may be measured by using the appropriate statistical models to compute *p* values.

These more complex measurements are typically related to mathematical or computer models of varying degrees of complexity. Ideally, perhaps, all users of a measurement would fully understand the model underlying it. In practice, however, the number of measurements in use, and the sophistication of some of the underlying models, means this is unlikely, and users' understanding may be seriously limited. This limited understanding of measurements has potentially serious consequences ranging from a failure to understand the significance levels used to quantify the certainty with which scientific hypotheses have been established, to a hazy understanding of the basis for quality scores for universities, to the potentially very harmful misunderstanding of the basis of the valuations of certain financial instruments which is widely viewed as a contributing factor to the recent financial crisis.

The users we are talking about here are not just lay people. They may include professional social scientists whose understanding of statistical concepts may be limited, but who need to make use of these concepts, and financial experts who are not familiar with the mathematics used in models to value certain financial instruments. Experts in one area frequently make use of these models and measurements without a full understanding of how they work and the assumptions they entail because "as specialized knowledge domains are becoming ever more onerous, experts are becoming more ill informed about related knowledge domains" (Ungar, 2000, p. 299).

The obvious remedy for these problems is simply to recommend, or to try to insist on, adequate education. In practice, however, this may not be practical given the extent of these measurements and the time necessary for users to become sufficiently expert (Ungar, 2000; Simon, 1996: 90-93). The alternative approach to the problem is simply to try to redesign the measurements, or the underlying models, to make them easier to understand.

In an earlier paper I discussed the idea of redesigning models and theories to make them easier to understand (Wood, 2002). In the present paper I focus on a more limited approach – redesigning the measurements which are produced by the models (although may involve some change to the models as well). These measurements form the interface between the mathematics and the user. My argument is that there are opportunities to make enormous improvements with very little cost. Besides enhancing understanding of these measurements, there is also the possibility of simplifying the education process – and the implications of this may be at least as important.

Attempts to popularize science and make it more widely understood are, of course, widespread. These range from popular books and articles on the latest science, or some issue in the news, to text books which aim to explain the concepts to a novice audience. Siemsen (2009), in a review of the views of Ernst Mach, distinguishes between "Pop" and "popular" science, the latter aiming to "teach fundamental scientific understanding to many people … Then we are also close to Popper's (1959) claim that scientific theories should not exist in a remote ivory tower, but that they should be made questionable by as many people as possible." If this is achieved, it is likely to encourage a more realistic trust in mathematical models.

The purpose of this paper is to suggest how this aim can be assisted by bringing the languages of science and the layman closer together. In many cases the user-friendly options to be discussed below are as useful and rigorous as the specialists' original version, so we do not need an alternative "Pop" version because everyone would use the same terms.

In a sense, my aim is to encourage the translation of jargon into the language of the lay person. There is a widespread view that jargon in some fields of academic discourse is extensive, unnecessary, and just serves to exclude outsiders and reinforce the apparent expertise of the expert. By and large, mathematical jargon has escaped this criticism because it is seen as necessary. Sometimes this is doubtless true, but my case in this paper is that much mathematical jargon is unnecessary and could be replaced by more accessible terminology.

There are many examples of attempts to make mathematical measurements more user-friendly. To take a few more or less at random: Spiegelhalter (2005) advocates "funnel plots for comparing institutional performance … [because they] are very attractive to consumers of data". Wood et al (1998) make various suggestions for making the measurements used in statistical process control more user-friendly. MacKay (2008) in his book on sustainable energy tries to make numbers "accessible by expressing them all in everyday personal units". These are, however, just isolated examples; there seems to be no systematic philosophy which makes user-friendliness a requirement, not just an optional extra.

Measurements, of course, raise several other issues. There are different types or levels of scale (ratio, interval and ordinal being one common categorization), if users stand to gain or lose depending on the value of a measurement there may be various pressures to design measures more likely to give the desired result, or to bias the inputs in various ways, and there

are obvious issues of validity and reliability. These are not my primary concern here: my focus will be the user-friendliness of measurements although this factor undoubtedly interacts with some of these other factors.

I will start with some specific case studies to illustrate some of the problems and what it may be possible to do to solve them. I will then review some relevant background ideas, suggest some principles for redesigning measurements, and draw out some of the implications for science, public understanding of science, and the growth of knowledge.

Some case studies of problematic measurements

Statistical significance tests and *p* values

"Life expectancy was 3.9 years longer for Academy Award [Oscar] winners than for other, less recognized performers (79.7 vs 75.8 years; *P* = 0.003)" (Redelmeier and Singh, 2001: 955). This conclusion appeared to demonstrate that winning an Oscar gives actors an edge over their less successful peers in terms of life expectancy. The *p* value is an estimate of the probability of a difference of 3.9 years or more arising *if* there were in fact no difference in the life expectancies of the underlying "populations". However, the meaning of the *p* value is likely to be obscure to anyone without a good grounding in statistics (even the meaning of the word "population" may not be clear to the uninitiated) as commentators have been pointing out for at least 50 years (see, for example, Morrison and Henkel, 1970; Nickerson, 2000).

The statement that *p* = 0.003 is roughly equivalent to the statement that the data suggests that Oscar winners have a greater life expectancy with a confidence level of 99.85% (see Wood, 2012). The slight uncertainty takes account of the fact that the conclusion is based on the sample of Oscar winners and peers who have already lived: even if it were true that winning an Oscar does have a tendency to extend life expectancy, the randomness of all the other influences on life expectancy may mean that the result is just an accident of the particular sample, and not a universal truth. The confidence level formulation seems far easier to understand, although it is almost never (as far as I am aware) used in this way. On the other hand, confidence *intervals* are widely used in some fields including in a revised analysis of the Oscars data (Sylvestre et al, 2006).

Coulson et al, in a survey of 330 authors of published articles, tested authors' understanding of *p* values and confidence intervals. They concluded that "interpretation was generally poor", although slightly better for confidence intervals than *p* values. However, there was very clear evidence that many authors interpreted confidence intervals in terms of *p* values; those who interpreted confidence intervals without reference to null hypothesis tests gave a far better interpretation of the results than those who thought in terms of null hypothesis tests. These results are particularly interesting because the subjects were not naïve readers but authors of published articles. The comparison was with confidence *intervals*; there is no

systematic evidence about the interpretability of confidence levels for hypotheses other than intervals.

The result about Oscar winners and life expectancy has been challenged on the grounds that the statistical methods used were inappropriate (Sylvestre et al, 2006). Their revised estimate of the additional life expectancy due to winning an Oscar is given in the form of a 95% confidence interval extending from -0.3 years to +1.7 years (equivalent to a *p* value of 0.15, or a confidence level of a positive extension to life expectancy of about 93% – see Wood, 2012). This suggests a reasonably strong possibility that the extension to life expectancy might be negative – winning an Oscar might actually reduce life expectancy. This confidence based formulation of the problem seems far clearer: the 95% or 93% figures give a clear measure of how likely it is that extension to life expectancy will lie in a given range without the distraction of the "null" hypothesis of no difference.

Strangely, despite their obscurity and "near universal misinterpretation" (Cohen, 1994, p. 997) *p* values are the standard way of expressing this type of uncertainty in the social sciences. Statistical analyses of experiments, regression models, and so on, are all conventionally qualified by *p* values, although it would be very easy to use confidence intervals or levels.

Valuation models and the recent financial crisis

An important contributory factor in the 2008 financial crash was the acceptance by financial institutions and ratings agencies of valuation models for financial products which did not realistically reflect the risks involved. There were two key, faulty assumptions in these models. The first is an over-reliance on the idea that the future will resemble the past – data from the past was built into models which thus failed to anticipate the new circumstances as the crisis took hold. The second point relates to the statistical dependence of events. Suppose we think, based on past data, that the probability of one borrower defaulting on a loan is 10%, and the same applies to a second borrower. Then standard probability theory says that if the two events are statistically independent, the probability of both defaulting will be 10% of 10% or 1%. The difficulty arises if the two events are statistically dependent: if, for example the probability of the first defaulting if the second has already defaulted is 80% – perhaps because they are both subject to the same economic environment – the correct assessment of the probability of both defaulting is then 80% of 10% or 8% which is eight times the estimate based on statistical independence. Most financial models do take some account of statistical dependence, but the difficulty is that, in practice, events like this are usually dependent in complex ways which are extremely difficult to anticipate and model.

Over-reliance on the theory of the normal (Gaussian, or bell curve) distribution is essentially the same point because this is based on the assumption that variables can be viewed as the sum of a large number of small, independent factors. Commentators have pointed out since the 1960s (Buckley, 2011: 140-142) that stock markets suffer large losses far more frequently than the theory based on the normal distribution would suggest because these

influences are sometimes large or dependent on each other. If these models are accepted by decision makers, as seems often to be the case, the result is an under-estimation of the likelihood of large losses.

One such model is the "copular function approach" to default correlation – the relationship between the likelihoods of a number of individual credits defaulting – proposed by Li (2000). This model has become notorious because it was used to produce valuations for portfolios of credits which turned out to be wildly over-optimistic (see, for example, Salmon, 2009). The key points to note about Li's model are

1 It used market information because this "reflects the market agreed perception about the evolution of the market in the future" (Li, 2000, p. 48), which, Li argues, is the best information available. The difficulty is, of course, that the although the market arguably provided the best information available, it was still woefully inadequate.

2 It used normal distribution functions, which, as we have seen, are likely to be fallible.

In effect, Li's model took the current market traders' assumptions about default probabilities, used an elegant, but flawed, mathematical model to combine them and extrapolate them into the future, and then fed these predictions back to the traders. The result is that traders' very fallible assumptions seemed to be legitimized by the elegance of the mathematics.

To prevent this, the assumptions need to be seen, not as a technical detail for the eyes of statistical experts only, but as an essential part of the answer: "if these conditions are assumed, then …." If the conditions are not satisfied, then the model has no reliable predictions.

University league tables

University league tables are meant to summarize the quality of the universities, and of particular subjects within universities. There are now many such league tables (Usher & Medow, 2009). They generally take a number of inputs such as measures of research and teaching quality, student satisfaction, employment levels among graduates, and so on. These inputs are then processed by a mathematical algorithm to yield an aggregate score which is used to rank universities.

Clearly the results depend on the inputs, their relative weighting, and the precise method of calculation used for the aggregation. Different ranking systems use different inputs and aggregate the results in different ways, so not surprisingly the results differ to some extent. For example the Guardian league table for individual subjects (Hiely-Rayner, 2011) uses eight performance measures including, for example, results of a teaching quality survey and an assessment of career prospects, but not research quality. The sources of data and the definition of the measurements for each performance criterion are described, although not always in full detail – the "value added" score is "based upon a sophisticated indexing methodology" which is

only roughly described. The scores are then standardized to similar scales using standard deviations, and a weighted average is then computed. For non-medical subjects five of the performance criteria are weighted at 15%, and the other three at 10% or 5%, although no reason is given for this. Other rankings use different sets of performance criteria (many include research quality, for example), and aggregate them in different ways, and the resulting rankings are, of course, different.

The literature on multi criteria decision analysis is extensive and largely ignored by the compilers of these ranking tables (see, for example, Belton and Stewart, 2002) despite its relevance to the ranking task. Weighted average schemes are just one possibility and suffer from well-known pitfalls. A good performance on any criterion can compensate for a bad performance on another: this means that a very good score on student-staff ratio might compensate for a bad score on the student satisfactions surveys. An alternative approach might be to use thresholds on important criteria. The weights are also less straightforward than they may seem. The three criteria relating to the student feedback survey have a total weight of 25% in the Guardian's ranking, and "expenditure per student" has a 15% weighting. This may seem reasonable, but if it then turned out that the student survey results in all universities were fairly similar, whereas there were very big variations in expenditure it might seem more reasonable to give more weighting to the latter. (The standardization of the scores means that the spread of each set of scores will be similar.) There are a number of suggestions in the literature for dealing with this problem, but the point worth stressing here is that the weightings chosen by the ranking schemes are essentially arbitrary and in no sense "objective" even though a list of largely equal rankings may give this impression.

Despite their shaky foundations, these league tables undoubtedly have tremendous power. Customers use them to assess the value of a course at specific universities, and the universities themselves devote a lot of effort to managing their scores on key criteria in order to achieve a higher ranking next time.

Despite all this, most people who use these rankings to make decisions probably have little understanding of what the measurements mean, or of the assumptions behind them. In order to come to a clear decision about whether a particular league table is appropriate the user clearly needs to understand the basis of the scores on which the ranks are based, which may be difficult because the algorithm may be complicated, or because it may not be published. Given that this is unlikely, reliance on these league tables seems to be giving an unreasonable amount of power to the organizations behind these measurements.

One way round this problem would be *not* to aggregate the scores on different criteria so that users can balance the various criteria from their own – different – perspectives. This approach has been used by a number of organizations in various countries including the centre for Higher Education Development in Germany (Usher & Medow, 2009). Alternatively it would be possible to design a system to elicit information about preferences from individual students

and then use these to derive a ranking list customized to the particular individual (Giannoulis and Ishizaka, 2010).

Similar ranking schemes are used to measure the quality of many other goods and services – schools, hospitals, cars, even whole countries in the sense of their quality of life. Many of the same issue will apply to these.

The process sigma measure of process quality

This is a measure of the quality of a business or industrial process linked to the "Six Sigma" approach to management (Schroeder, Linderman, Liedtke, & Choo, 2008; Zu, Fredendall, & Douglas, 2008). For example, 430 defects per million opportunities (DPMU) corresponds to a process sigma of 3.33, and 6 sigma itself corresponds to 0 DPMU (using the calculator at http://world-class-manufacturing.com/Sigma/level.html accessed on 19 December 2011).

However, it is unclear what the purpose of the sigma scale is; defects per million opportunities (or as a percentage) is a straightforward and easily understood measure, whereas the process sigma measurement is seemingly devoid of intuitive meaning. It has an interpretation in terms of mathematical statistics – the sigma level is the value of the standard normal variable for which the single tail probability is the defects per million opportunities (although this is further complicated by the possibility of incorporating a "1.5 sigma shift" – the results cited above do not incorporate this). Usually mathematics is used to translate something which is difficult to understand or evaluate, to something that is easy to understand: e.g. mathematics can be used to show that $e^{\pi i}$ is the same as –1, the latter being far easier to understand. The process sigma measurement seems to involve exactly the opposite procedure: translating something that is easy to understand into something which is very obscure.

The obvious question is: why bother with the process sigma measurement? Why not stick with defects per million opportunities? One explanation is the persistence of conventional ways of speaking in the statistics where the normal distribution is very widely used: sigma values are a common (but by no means universal) route to estimating probabilities, so the idea of describing levels of uncertainty in terms of sigma values may seem natural. This is reflected in the use of the traditional measures of process capability (c_p and c_{pk}) whose definition also depends on sigma values: just the same question applies to these indices – why not replace them by a measure such as defects per million opportunities (Wood et al, 1998)?

Sigma levels are also part of the language used in the very different context of particle physics: the recent press coverage of the possible discovery of the Higgs boson made extensive use of sigma levels as a means of describing uncertainty levels, which prompted the *Times* newspaper to explain that 5 sigma means a chance of less than one in a million (14 December, 2011, page 6). However, in all these cases, there would be a good case for abandoning the sigma measures and returning to the core concept which is a probability expressed as defects per million opportunities or as a percentage.

The persistence of statistical convention is doubtless one factor encouraging the use of sigma measurements. This pressure might be supported by the desire of consultants and others to have a well defined product to sell. Six sigma is a nice slogan, with the Greek possibly suggesting profundity, and the added benefit that a lack of user understanding might encourage the uncritical acceptance of the wares of the consultant.

Further examples of problematic measurements

There are many further examples of measurements which could easily be improved from the users' perspective. Statistics is a particularly problematic area. We've discussed *p* values and sigma levels above; another example is provided by the habit of giving the results of regression models as a list of unexplained coefficients which are probably meaningless to many readers. For example Glebbeek and Bax (2004) give a standardized regression coefficient of –0.23 for one of their models estimating the impact of staff turnover in an organization on the financial performance of the organization. An alternative way of presenting the same result would be to say that this impact is –1778 units of currency (Dutch Guilders) per employee per annum, per 1% rise in staff turnover (Wood, 2010). This gives the reader an idea of the magnitude of the effect on a scale which would make sense to managers in the organizations involved. This is a fairly trivial, but typical, example. There are many more similar possibilities.

Problems

It is helpful to divide the problems found with these, and many other measurements, into the following four categories. These are not mutually exclusive: two or more may apply simultaneously. They are also inevitably fuzzy categories, and may depend on a simplistic idea of the "correct" understanding. But they are useful for my purposes here.

1. Failure to understand

Someone may not understand what a measurement means, and so may ignore it completely, or fail to appreciate the subtleties of the measurement. For example, the reader of the research on Oscars above may not understand what the *p* value means and so may ignore it, or just realise that it is some measure of research quality but without any clear idea of its meaning.

2. Misunderstanding of the basic concept

Other readers may have an idea of what they think *p* values, for example, are, but this idea might be wrong. One common misconception is that a very significant result, as indicated by a low *p* value, indicates a strong or important effect (Crettaz Von Roten, 2006); another is that the *p* value represents the probability of the null hypothesis being true. The first of these, in particular, is a serious misconception because it may lead to results of no practical significance being taken far too seriously, and conversely a lack of statistical significance does not mean that the effect does not exist and can safely be ignored, but rather that there is insufficient evidence to be sure.

3. Misunderstanding or ignoring the assumptions on which the result depends

The valuation models referred to above depend on various assumptions that may not be met in practice. The consequences, as we have seen, can be disastrous. *All* mathematical models depend on assumptions of varying kinds; sometimes these assumptions are innocuous, but at other times they need checking very carefully.

4. Unnecessary time and effort expended

The usual recommendation for people having difficulty understanding something is to spend more time thinking about it, read up the background, attend a course on the subject, etc. However, the possibility of redesigning measurements so that the time and effort spent on these activities could be reduced is potentially a far more powerful strategy. It is difficult to estimate the magnitude of these savings, but my estimate is they could be very substantial – I would say at least 50% of the time needed to learn aspects of statistics could be saved by the design of more appropriate measurements. This has obvious implications for education – either less time is needed, or we can achieve more.

Obviously the examples above are selected to demonstrate various specific points: as such they doubtless tend towards the extreme end of the spectrum. Getting a rigorous assessment of how widespread these problems are would be extremely difficult as each measurement would need to be analyzed and alternatives found or created, and there would never be a last word on how much improvement was possible. However, this is not necessary for my argument: all I am claiming is that these problems exist and improvements are possible, both in the examples considered and, almost certainly, in many other examples.

Why are these problematic measurements used?

My argument in this paper is a difficult one to communicate. There is a general assumption that knowledge, including how measurements are defined, is in some sense fixed, and that moulding it to make it more appropriate for the audience is dumbing down which, by implication, will entail a much restricted understanding. I hope I have shown in the above examples, at least, that this is not necessarily so. In many cases the problem is that the standard measurement system is designed by the experts who developed the field. The measurements are then part of the established paradigm: they are reinforced by the language used, tacit assumptions made and so on, and it may be difficult to take seriously the idea that the measurements could be redesigned. And the vested interests of the experts and the educational system can only reinforce this attitude. The result is that all users, including the uninitiated have to retrace the steps taken by the experts.

Natural languages have an affinity for metaphors and analogies which undoubtedly assist communication in various subtle ways. When we say "it's not rocket science" we have the background to appreciate that rockets are complicated beasts requiring unimaginable expertise to design. Even that last sentence uses the word "beasts" in a non-literal sense to bring to the reader's mind the idea that there is some vaguely threatening and unpredicable at stake. The

same process occurs in expert communities who may refer, for example, to sigmas as a picturesque (perhaps conjuring up pictures of ancient Greek philosophers whom the experts would like to resemble) way of describing small probabilities, or to *p* values as a short hand for very particular probabilities. However, difficulties arise when these terms are exported out of the expert communities to people without the assumed background. That is when the experts need to be encouraged to use ordinary language in ordinary ways.

Some of the examples above come from the field of mathematical statistics. Many of the concepts employed in statistics are subtle ones, and their implementation often requires advanced mathematics, but despite this they are used by an increasing number of uninitiated users in an ever expanding variety of contexts – with some of the consequences we have discussed.

In most of these cases the problems are probably due to the inertia in any large system. However, in a few cases there may be a deliberate conspiracy to force untransparent measures on uninitiated users. It is difficult to see any other explanation for the six sigma measurement – the beneficiaries here would be the consultants and other experts peddling their supposed expertise. Similarly the university league tables are of obvious benefit to their publishers, and also to the universities which do well in the lists. These measurement systems are in effect brands: part of their strength, deliberately or not, is the relative obscurity of the mathematical models underpinning them which means that uninitiated users are unlikely to have the expertise and confidence to challenge them.

The prestige of mathematics is a factor which may be used to assist in the branding of measurement systems. It may also be an important factor without any sort of conspiracy. For many people, it is tempting to assume that if complicated mathematics is employed it must be employed for a good reason, and to accept the measurement on this basis alone. This temptation may be greater for those who do not understand the mathematical basis of the model and so the assumptions on which it depends. The effect of employing complex mathematics may be to encourage uncritical acceptance, whether by design or not.

The desire of ordinary users to understand measurements may be a comparatively weak counter to pressures like the prestige of mathematics and the expert community. It is possible that many people would prefer to see statistics as unintelligible rather than face the effort of trying to understand. And some measurements like the university quality measurements may derive their credibility largely from the fact that they are accepted, leading to the migration of the best students and faculties to the universities with high scores. In this case, the details of the derivation are in a sense irrelevant, but it does seem to give unjustified power to algorithms which may be arbitrary, or may be verging towards a conspiracy. The survival of "memes" like *p* values or university league tables may depend on factors far removed from naïve assumptions about their validity and user-friendliness.

To recap, the design of many measurement systems may depend on factors like the historical accident of how they were developed, or commercial pressures to brand a particular approach, or simply the unjustified prestige of complicated mathematics. Given this, it is likely that many measurements can be redesigned for make them more appropriate for current users and uses. The word here, of course, is "design", not discovery. What is needed, and is usually lacking, is a mindset which views this as a possibility worth pursuing.

Principles for redesigning measurements

The above examples are very varied, so it is difficult to generalize about strategies for redesigning measurements. However, the following principles are a tentative start.

1. Consider if the measurement should be scrapped or ignored

I have argued above that there is a strong argument for ignoring the university rankings at the aggregate level, and the sigma measure as used in Six Sigma. Measurement systems do tend to proliferate in the modern world, and it is always worth asking whether a new measurement adds any value to what is available already.

2. Use models and methods which are as transparent as possible

It is often possible to use computer simulation methods instead of mathematical probability theory for deriving statistics such as *p* values and confidence intervals (see for example Wood, 2005). The former have the advantage that users can almost literally see how they work, so their interpretation and limitations are likely to be far clearer than if results are derived from mathematical models that are not accessible to most users. Similarly the mathematical formulae for regression, or best fit, models can be replaced by trial and error methods which make the underlying rationale far clearer (Wood, 2001). There are likely to be similar opportunities in many other areas.

3. Output measurements should be as useful as possible (for the intended users)

This is obvious, but often forgotten. The typical audience for statistical results does not want *p* values; they want an estimate of the probability of, or confidence in, for example, the hypothesis that Oscar winners do live longer on average. And applicants for universities do not really need an overall quality measure for each university; rather they want measures of the particular aspects of the universities that are of interest to them as individual.

4. Use appropriate names, scales and units for measurements

The principle here is to ensure that the measurement fits in with the users' frame of reference by adjusting these relatively trivial aspects. *P* values, for example, have an uninformative name (significance level is little better), and the scale is a reverse one with low *p* values corresponding to high degrees of certainty for the hypothesis of interest (the alternative to the null). Contrast this with the idea of confidence levels which have an informative name (confidence), high values do correspond to high confidence, and the conventional use of percentages encourages users to see the cited confidence level as a proportion of total confidence.

The name of the measurement should reflect meaning of the result, not the method used to get there or some historical accident of where the original idea came from – so R^2, for example, should be described as the proportion of variation explained. And details such as the choice of 68%, 95% and 99.8% for funnel plots (Spiegelhalter, 2005), and conventional control lines, which are driven by the method of derivation from standard errors (or sigmas), should be replaced by more obvious choices such as 90%, 99%, 99.9%.

5. Include a list of assumptions

In essence mathematical modelling is conditional reasoning: if we make these assumptions, then these conclusions follow. Often, the assumptions are seen as a technical detail which is not important for informal work. The example of the financial valuation instruments above shows that this is serious error. Assumptions of statistical independence, and that the future can be predicted by extrapolating past trends, are vital and should be appreciated by users.

Conclusions: an agenda for research

Measurements based on mathematical models are widely misinterpreted or ignored by their intended users. This is a problem leading to the waste of the information that the measurements could have provided to users, and to the consequences of misunderstandings which are potentially serious, particularly in fields such as medicine and finance. My argument here is that a very powerful strategy for dealing with the problem is to redesign measurements to make them more appropriate for their intended users and uses, and I have given a few examples of what might be possible, and suggested some principles for achieving this. Another important byproduct of this approach would be a potentially enormous reduction in the time and energy needed for the often unsuccessful effort to educate users to understand the unnecessary complexities of many mathematical measurements.

The difficulty, of course, is the mindset that says that the experts' version cannot be altered. This is a powerful mindset: as far as I can see the ideas proposed here about, for example, even the possibility of changing relatively superficial aspects, such as the naming of measurements and the scales used, seems to be rarely, if ever, mentioned. History is often recommended as a good thing in so far as it stops us repeating the mistakes of the past, but if taken to extremes, it may make everyone repeat the tortuous route often taken to good ideas. The problem is not just inertia: the vested interests of universities and other commercial forces may be an important factor in the preservation of unnecessary complexity.

Markets are a powerful tool for encouraging sensible offerings to flourish. Software has become far easier to use over the years because user-friendliness is a necessary pre-requisite for software to be used. Training courses for software have probably declined as the idea that you can pick it up as you use it has spread. However, this process is far from perfect: to turn my computer (using Windows XP) *off* I need to click on the button labeled start! The market for ideas like mathematical measurements is much less developed and efficient: there is little confidence in the idea that ordinary people can pick up an acceptable understanding by practice

and trial and error, and in fact, as we have seen, this often leads to disaster. Part of the problem is the feedback mechanism: users of a software package know when they are not succeeding, but this is often not true of naïve users of mathematical measurements.

We need to challenge the culture that sometimes seems to assume that complex mathematical models are a good idea just because they are complex. Sometimes complexity may be necessary, but if a simplification is possible, which would bring the equivalent power "to the masses", we should surely take this route. The peer review system, which legitimizes academic knowledge, may be a hindrance here because new models and measurements are scrutinized by experts who, for obvious reasons, may not acknowledge the problem. Non-experts are not in a position to offer a critique because they do not understand, so the clique of experts may generate more and more complex ideas. In relation to statistics, to answer Crettaz Von Roten's (2006) question "Do we need a public understanding of statistics?", my answer would be yes, and furthermore experts in areas that make use of statistics (such as the authors surveyed by Coulson et al, 2010) are an important part of the public who need to understand.

These issues all need further research. There are several potentially useful avenues for research. At the conceptual level, devising new ways of measuring (e.g. the confidence levels suggested by Wood, 2012), and analyzing the conceptual background necessary for understanding various measurements with a view to simplifying this background as much as possible, are both important avenues to explore. On an empirical level, surveys of misunderstanding and experiments to see how these can be improved by changing the measurements used (e.g. Coulson et al, 2010) are of obvious importance. And there are also interesting issues about the extent of people's understanding of the conditional nature of mathematical reasoning – and the importance of the conditions or assumptions – and the extent to which mathematical models are, or are not, trusted.

References

Belton, V., & Stewart, T. J. (2002). *Multiple criteria decision analysis: an integrated approach*. Boston: Kluwer.

Buckley, A. (2011). *Financial crisis*. Harlow: Pearson Education.

Cohen, J. (1994, December). The earth is round (p< .05). *American Psychologist*, 997-1003.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology , Article 26.*

Crettaz Von Roten, F. (2006). Do we need a public understanding of statistics? *Public Understanding of Science , 15*, 243-249.

Giannoulis, C.; Ishizaka, A. (2010). A Web-based decision support system with ELECTRE III for a personalised ranking of British universities. *Decision Support Systems, 48*(3), 488-497.

Glebbeek, A. C., & Bax, E. H. (2004). Is high employee turnover really harmful? An empirical test using company records. *Academy of Management Journal, 47*(2), 277-286.

Hiely-Rayner, M. (2011, May 17). *Guardian University Guide 2012: methodology.* Retrieved December 19, 2011, from http://www.guardian.co.uk/education/2011/may/17/guardian-university-league-table-methodology

Li, D. X. (2000, March). On default correlation: a copular function approach. The Journal of Fixed Income , 43-54.

MacKay, D. J. (2008). *Sustainable energy - without the hot air*. Cambridge: UIT.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy*. London: Butterworths.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Redelmeier, D. A., & Singh, S. M. (2001). Survival in Academy Award-winning actors and actresses. *Annals of Internal Medicine , 134*, 955-962.

Salmon, F. (2009, March 17). *Recipe for disaster: the formula that killed Wall Street*. Retrieved December 19, 2011, from Wired Magazine: http://www.wired.com/techbiz/it/magazine/17-03/wp_quant?currentPage=all

Schroeder, R. G., Linderman, K., Liedtke, C., & Choo, A. S. (2008). Six Sigma: definition and underlying theory. *Journal of Operations Management , 26*, 536-554.

Siemsen, H. (2009). The Mach-Planck debate revisited: democratization of science or elite knowledge? *Public Understanding of Science* , published online July 24, 2009.

Simon, H. A. (1996). *The sciences of the artificial.* Cambridge, Massachusetts: MIT Press.

Spiegelhalter, D. J. (2005). Funnel plots for comparing institutional performance. *Statistics in Medicine , 24*, 1185-1202.

Sylvestre, M.-P., Huszti, E., & Hanley, J. A. (2006). Do Oscar winners live longer than less successful peers? A reanalysis of the evidence. *Annals of Internal Medicine , 145*, 361-363.

Ungar, S. (2000). Knowledge, ignorance and the popular culture: climate change versus the ozone hole. *Public Understanding of Science , 9*, 297-312.

Usher, A., & Medow, J. (2009). A global survey of university rankings and league tables. In B. M. Kehm, & B. Stensaker, *University Rankings, Diversity and the New landscape of Higher Education* (pp. 3-18). Rotterdam: Sense Publishers.

Wood, M., Capon, N., & Kaye, M. (1998). User-friendly statistical concepts for process monitoring. *Journal of the Operational Research Society, 49*(9), 976-985.

Wood, M. (2001). The case for crunchy methods in practical mathematics. *Philosophy of Mathematics Education Journal, 14* (a web journal at http://www.ex.ac.uk/~PErnest).

Wood, M. (2002). Maths should not be hard: the case for making academic knowledge more palatable. *Higher Education Review, 34*(3), 3-19.

Wood, M. (2005). The role of simulation approaches in statistics. *Journal of Statistics Education, 13*(3), http://www.amstat.org/publications/jse/v13n3/wood.html .

Wood, M. (2010). The use of statistical methods in management research: a critique and some suggestions based on a case study. *http://arxiv.org/abs/0908.0067*.

Wood, M. (2012). Liberating research from null hypotheses: extending the idea of confidence instead of using *p* values. http://arxiv.org/abs/0912.3878.

Zu, X., Fredendall, L. D., & Douglas, T. J. (2008). The evolving theory of quality management: the role of Six Sigma. *Journal of Operations Management , 26*, 630-650.