Rethinking "quantitative" methods and the development of new researchers

Stephen Gorard
School of Education
Durham University
s.a.c.gorard@durham.ac.uk

Abstract

The argument in this paper, supported by literature, evidence, illustrations and simulations, is as follows. The general standard of social science research still needs improvement, even after a number of attempts have been made to improve it. The most recent attempt in the UK is the Quantitative Methods Initiative, based on widening the appeal of social science that involves numbers. However, before any initiative like this could be successful what is envisaged as 'quantitative' methods needs to be radically altered to make it simpler and more logical. All work based on standard errors – significance testing, confidence intervals, power calculations, multilevel modelling and so on – requires complete random samples/groups as a mathematical necessity. Since complete random samples are so rare in social science, work based on standard errors is largely irrelevant and should not define what 'quantitative' methods are. The paper then shows that the logic of significance testing, power, and confidence intervals does not work anyway (even with a perfect random sample). Also, estimated standard errors are often inaccurate, and that inaccuracy propagates into the calculation of test statistics. The inaccuracy is worsened by missing data, and sometimes even more so by attempting to compensate for that missing data. The ensuing results cannot be trusted. Once released from consideration of probability distributions and the like, 'quantitative' methods can be further simplified. The consequence would be a social science in which the logic of working with numbers would be the same as the logic of working with any other kind of data. This would then have formative implications for so-called 'qualitative' methods, and leads to the realisation that there are, in fact, just methods. The things that make good research or trustworthy results are the same, and independent of the methods used – rigorous design, pre-specification as far as possible, full reporting, full consideration of missing data and lost cases, and setting out the complete logical argument from research findings to implications.

Capacity-building and the QM Initiative

In the UK, there is a long-standing concern about the quality and utility of social science research (McIntyre and McIntyre 2000, Gorard 2004, Platt 2012). The same concerns appear in many other countries. In education, for example, US journal editors and others have long reported that very few published papers were really worthy of acceptance. Wandt et al. (1965) deemed most published studies to be trivial or invalid, and often both. The key issues were lack of appropriate study design, incorrect methods of analysis, and lack of reporting assumptions or limitations. Forty years later, Ioannidis (2005, p.1) reported concerns that most current published research findings in medicine, science and social science are simply false. 'For many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias'.

A number of attempts have been made to improve research-capacity, and so raise the standard of research. A range of funding partners - including the Higher Education Funding Council for England, the British Academy, the Nuffield Foundation, and the Economic and Social Research Council - have decided that a key problem is the lack of 'quantitative' research. They have launched a joint national Quantitative Methods Initiative for social science (e.g. http://www.esrc.ac.uk/research/skills-training-development/qmi/index.aspx). The focus is on the development of new methods courses, teaching resources, new training opportunities, work placements, scholarships, and the embedding of 'quantitative' evidence in substantive courses for all social scientists. The model underlying all of this activity and expenditure is that 'quantitative' work in social science is good, especially in economics and psychology, but that it is not always well taught, and there is not enough of it outside those two disciplines.

Some of this diagnosis may be correct, although it is possible that the problems of social science are as much to do with lack of design, genuine curiosity and research integrity as to do with lack of number. Some of the solutions may also be necessary but they will probably not be sufficient. This paper shows that what is being taught as 'quantitative' methods is currently too often wrong, and even more often needlessly complex. It is these confusing approaches more than anything that are driving researchers away from the relatively simple use of measurements and frequencies in their work (Cooper and Shore 2008, Hampel 1997, Murtonen and Lehtinen 2003, Watts 1991). Traditional 'quantitative' methods are more part of the problem than any promise of solution, and 'the research enterprise may be distorted by statistical technique, not helped' by it (Berk and Freedman 2001, p.2). In psychology, an area of social science exempted from the QM initiative as already having enough sound QM work, it seems that data fiddling, publication bias and false positives abound (Simmons et al. 2011, p.1360). It is 'unacceptably easy… to accumulate (and report) statistically significant evidence for a false hypothesis'. And, of course, non-significant results are routinely unpublished, and so deleted from the record, across all areas of research (Pigott et al. 2013).

An alternative approach is to simplify what it means to use numbers in research, to stress that the underlying logic is the same for the analysis of any kind of data whether numeric or not, and only then try to improve uptake through the pedagogy and persuasion of something like the Quantitative Methods Initiative (but with a less divisive name). This paper is based on experiences of leading the ESRC Research Capacity-building Network, and subsequent activities within the ESRC Researcher Development Initiative, culminating in a recent project for the Quantitative Methods Initiative. Funders have repeatedly assumed that improvement in research will come about mainly through innovation in, or the wider utilisation of, complex or highly technical methods. However, there is very little evidence that the existing problems diagnosed by stakeholders and the literature critical of existing research that does not use numbers (above) are susceptible to solution merely by a change in the fashion of methods. The paper proposes instead the simplification (and correction) of methods of data collection and analysis as far as possible, much greater clarification of research as an argument based on genuine attempts to communicate each step, greater emphasis on prior study design, and research integrity. Only then does the creation of new teaching resources and approaches make sense. Teaching new researchers things that most of them would never need, or that mostly do not work as intended anyway, or that have traditionally been made overly complex to disguise limitations and flaws, is absurd. And this is so however innovative or charismatic the teaching is made to be. The paper explains and illustrates each of these points, and concludes with a proposed way forward to enhance research quality in which the red-herrings of 'qualitative' and 'quantitative' should not feature.

The real SE

A typical course in, or resource for, learning 'quantitative' methods will be based on random sampling theory and its derivatives. These techniques for conventional statistical inferences (based on formulas for the standard error of the mean, t-tests and so on) depend upon the strict assumption of random sampling (Berk and Freedman 2001). What does this mean, in theory and in practice?

The term 'standard error' is, like many things in statistics, ambiguous (see below). In theory, the standard error of the mean refers to the standard deviation of the distribution of the means of a specific measurement from a huge number of repeated random samples of the same size from a known finite population (Peers 1996). The specific measurement can be any value or characteristic associated with each member of the population (such as their height for a population of people, or the numbers of rooms in each house for a population of houses). Put another way, if we know the values (or scores) for a fixed population then we can estimate the distribution of means that we would obtain from many random samples of size N. The standard deviation of that sampling distribution, for any real positive value of N, is given the special name 'standard error'. Standard errors can be calculated for values and parameters other than the mean, but for simplicity this paper focuses on the mean (and the points made would be true about the standard errors for other parameters as well).

Altman and Bland (2005, p.903) explain it in this way:

> When we calculate the sample mean we are usually interested not in the mean of this particular sample, but in the mean for individuals of this type—in statistical terms, of the population from which the sample comes. We usually collect data in order to generalise from them and so use the sample mean as an estimate of the mean for the whole population. Now the sample mean will vary from sample to sample; the way this variation occurs is described by the "sampling distribution" of the mean. We can estimate how much sample means will vary from the standard deviation of this sampling distribution, which we call the standard error (SE) of the estimate of the mean.

The calculation of this true standard error (SE) thus requires considerable prior knowledge. It requires knowledge of the identity of every member of the population. In fact, this is the definition of a 'population' in this context. The 'population' refers to all members (cases) that are independent of each other (such as people or houses), each of which has an equal chance of being picked for a random sample during random selection. The required knowledge of the population also includes the measurement or value that is to be sampled, for every case (such as the person's height or the number of rooms in each house). A case selected at random that has no relevant measurement or value cannot be used to help compute the SE, and so must be discarded. Otherwise, not every case has a chance to be included when calculating the mean, and this will bias the sample. It would be like having a standard pack of 52 playing cards with a few cards so disfigured that it is not possible to say what they are. Drawing any valid card from this deck is no longer a 1/52 chance. Either, the unreadable cards must be discarded and the deck (population) redefined to be only those cards (cases) for which we have perfect knowledge, or else we must accept that any sample will not be random, and so a SE cannot be calculated and would not make sense anyway (see below).

The calculation of the true standard error (SE) is based on complete knowledge of all relevant values for the population, and the generation of large numbers of complete random samples. These samples *must* be random, and the selection probabilities must remain the same from sample to sample in the repeated re-sampling that is needed to create the SE. This is part of how the SE is defined, and random sampling is inherent in its mathematical computation, and in the algorithm in any software used to calculate the true SE. As Berk and Freedman (2001, p.2) state "This is not a matter of debate or opinion; it is a matter of mathematical necessity". This means it is important to understand what a random sample is.

What is a random sample?

A 'random sample' is a subset of cases from the known complete population (see above), selected by chance in such a way as to be completely unpredictable. The chance element can be produced by observing radioactive decay, using specialist software or even Microsoft Office, via a random number table, or a mechanical process like a card shuffling machine. Strictly speaking, these all produce pseudo-random numbers because they are generated by a process of some kind, but they are all that is possible in reality. A true random sample should also permit the occurrence of the same case more than once (like drawing a card from a pack of 52, replacing the card and drawing another at random). The chance of such repetition occurring depends upon what proportion of the population is in the sample. If the population is large in relation to the sample this issue of permitting repetition within a sample may not matter much in practice, but it is important to remember that sampling without the possibility of repetition (of the card or whatever) is not really random sampling. It is important because compromises like this, however small, mount up. So far, we must accept that a 'random' sample is really pseudo-random, and that if it was drawn without the possibility of repetition then it is slightly less random again (if there is such a thing).

More importantly, a random sample must be complete in two senses. It must include every case that was selected from the population by chance. There can be no non-response, refusal or dropout in a random sample (by definition – again this is not a matter of opinion or belief). And every case selected must have a known measurement or value of the characteristic of interest (height, number of rooms or whatever). There must be no missing values. As before, these issues are assumed in the definition and computation of the SE, and are a matter of mathematical necessity.

Imagine the process of dealing a hand of 13 playing cards. If the cards are properly shuffled, we can say that the hand is random. If the deal involved picking 15 cards and then returning two low cards to the pack, the resulting hand is not random even though it has 13 randomly dealt cards. Similarly, dealing 13 cards and then replacing two low cards by two others drawn randomly from the pack would not yield a random hand. This is so, even though all 13 cards in the eventual hand had been drawn randomly. A random sample, despite being an ideal, is a simple thing and any deviation from random sampling must inevitably lead to a sample that is not random.

Problems of non-response

A situation where cases are selected for a random sample but some do not actually participate or do not provide responses for a key variable, also leads to a non-random sample. There may be non-response, refusal, untraceable cases, errors in the sampling frame, and unreadable or undecipherable data. If cases from a random sample drop out or do not provide responses for a key variable after a research study has started, the resulting sample is also non-random. Again, this attrition may be due to refusal, mortality, mobility, or undecipherable data at a later stage. In general, the larger the study, the more data it collects and the longer it lasts, the more likely it is to have substantial missing data. Many commentators suggest that 100% participation, of the kind required by statistical theory (to calculate the SE), is unheard of in practice (Cuddeback et al. 2004). For example, Lindner et al. (2001) examined all of the articles published in the Journal of Agricultural Education 1990-1999. All empirical reports had issues of non-response (even though most made no mention of the threat that this caused).

All non-response, of all of the kinds outlined above, creates a potential for bias in the results from the remaining cases (Peress 2010). It is most unlikely that the cases dropping out from a random sample are randomly distributed, and that is part of the reason why it is no longer a 'random' sample (Hansen and Hurwitz 1946, Sheikh and Mattingly 1981). Data based only on the achieved respondents is almost certainly biased, as can be demonstrated when population or administrative data are compared to (necessarily) incomplete surveys of the same cases, or where different studies have selected exactly the same 'sample' but then have different response rates (Dolton et al. 2000). Non-response can have a 'dramatic impact' on the data, leads to 'puzzingly different results' (Behaghel et al. 2009, p.1) and any bias in the substantive results caused by missing data generally cannot be corrected by any technical means (Cuddeback et al. 2004). There is no clear evidence that an estimate of propensity to respond (based on how many requests need to be made, or how late the response is) can be used to generate valid data similar to the cases that do not respond. As is shown later, neither can missing data be accurately replaced by using the evidence that has been collected. In fact, attempts at such replacement often make the bias worse. Weights can only be used *post hoc* to correct for variables for which all true population values are known, making weighting pointless, and weighting a sample in this way clearly cannot correct the values of other variables for which the true population value is not known (Peress 2010). It is crucial to recall that missing data of any kind must be assumed to be biased, as has been demonstrated repeatedly in practice, rather than occurring at random.

The impact of such bias can be illustrated via a simulation. When calculating an effect size, for example, it is important to consider how likely it is that any 'effect' could appear due to chance, and the volatility of small numbers. A simulation used 1,000 samples, each of 100 cases. Each case is a random integer in the range 0 to 9 (treating 0 as an integer). The first sample of 100 cases can be envisaged as the treatment group in an imaginary experiment, and the second sample of 100 cases as the control. The third sample can be a treatment group and the fourth another control in a second imaginary experiment. In this way, there are 500 pairs of samples representing the results of 500 experiments (each with N=200). The true effect size (calculated as the difference in means between the two groups in each pair, divided by the standard deviation for the control group) should be 0. The true mean for the population from which the samples were taken is 4.5, and both samples in each pair should be equivalent, because no experimental treatment has taken place. The simulation actually generated means from 3.74 to 5.3, with standard deviations from 2.59 to 3.24. The 'effect' sizes generated range from -0.39 to +0.35 despite all of the figures involved being random

numbers. This is a caution against accepting relatively small effect sizes from non-replicated studies.

If (biased) attrition is then simulated by deleting the highest 10% of scores in each sample of 100 cases, the sample means now range from 2.93 to 4.82. As should be obvious, missing data can distort findings, in this case by leading to a likely under-estimate of the population mean when using data from any one sample. The 'effect' sizes now range from -0.47 to +0.55. Attrition is capable of creating or strengthening entirely spurious effects as well as obscuring genuine ones. Given that the What Works Clearinghouse guidelines (http://goo.gl/8o7GqX) state that even 29% attrition is a 'low' level as long as the dropout is equal between groups (as it is precisely here), 10% is rather low in practice. Of course, the bias may not be as extreme as missing all scores of a similar type, but this simulation shows the considerable danger from even what WWC would, ludicrously, consider a very low level of attrition.

True probabilities and permutations

When we know the relevant values for the complete population, as would be required to calculate the true standard error, then we have no need of the standard error anyway. We could calculate the exact probability of obtaining any specified sample of values through a consideration of combinations and permutations. This would involve working out every sample of size N that could be drawn from the known population, and counting how many of these had the same value (such as the mean) as the specific sample we were interested in. This forms the basis for a 'permutation' test, originally suggested by Fisher, far superior to its alternatives because it always provides the correct answer rather than an estimate (Ludbrook and Dudley 1998, Ernst 2004). And it has one important area of possible use in practice.

Although t-tests, ANOVA and other statistical approaches are routinely used both to help decide on differences between groups and to decide whether these differences are generalisable to the wider population, these two situations are actually very different. It is strange that most analysts do not distinguish between them more clearly, and that they feel able to use the same approach such as an independent samples t-test to address both issues. The permutation test can certainly only be used in one of these situations.

Imagine a random sample of cases from a real survey that asked people about themselves and their attitudes. An analyst might divide the sample into those reporting themselves as male and female, and then compare an attitude score between these two groups. In this example, the analyst is using a t-test to try to estimate whether the difference in attitudes between the groups also appears in the population from which the sample was drawn. This is a question of generalisability. Without knowing the attitudes of the complete population, the analyst cannot conduct a combination or permutation calculation to see how often such a difference would arise. And whether they should even be using a t-test in this situation is addressed below.

Imagine instead, a researcher conducting a randomised controlled trial. They ask for participants in the trial and then randomly allocate all participants to either the intervention group or a control group. If they used a t-test at the end of the trial, it would be used for a very different purpose than above. It would be to try and estimate how likely it was that the observed difference between the groups would be observed merely because the two groups were randomised. There is no statistical way of generalising from the volunteers in the trial to

any wider population. For the purposes of the trial, the population is all of the volunteers from whom two random 'samples' have been drawn. But here the two random 'samples' put together constitute the complete population (in a way that does not happen when a sample is divided *post hoc* into two groups like male and female). Therefore, as long as no participant drops out, the analyst can use the perfect permutation test to compute exactly how unlikely it is that the groups eventually differ by as much as they do. In fact, it seems possible that the whole panoply of statistics was originally devised to deal with the second kind of situation (RCTs and similar where the complete 'population is known') and was simply abused when applied to the first situation (generalisation to an unknown population).

One problem that Fisher and others faced was that the permutation test was pragmatically impossible before the use of electronic calculators and computers. Even with a small population of 100 cases divided randomly into two groups of 50, there are $3x10^{93}$ permutations of the members of the two groups. Even with some short-cuts, this is a huge number, and real-life trials often involve many more than 100 cases. So, the significance tests and all that follow from them can be envisaged as an attempted heuristic (compared to the permutation algorithm) from the pre-computer era. They are meant to mimic the permutation test, but were still intended for use when all of the conditions met so far applied – so that the complete population was known, and the sample(s) was truly random. They were always intended for trials and similar research designs. Their problem, as a heuristic, is that they just do not work.

Problems with significance testing

In practice, the problems with significance testing should not matter because analysts so seldom work with complete random samples, and without this necessary element significance tests should not be conducted anyway.

> The fiction that probability statements are meaningful in the absence of random acts underlying them is preposterous (Glass 2014, p.12).

If a sample is not random or not complete, or the population from which it is drawn is not known or not complete, then there can be no standard error (see above). This, in turn, means that anything predicated on a standard error does not make sense – it cannot exist mathematically – unless it involves true random sampling. Significance tests, p-values, confidence intervals, power calculations and some complex statistical models such as multilevel modelling, are all clearly predicated on working with a random sample(s). This means that all such techniques are useless in practice for the simple reason that random samples generally do not exist in real research. Yet significance testing still abounds – both in stand-alone form, and even more frequently in the process of modelling such as regression or factor analysis. Analysts are either unaware of, or are ignoring, the mathematical assumptions.

The problematic logic of significance tests

The major problem with significance tests intended as heuristics for calculating an exact probability is that the probabilities they generate are the 'wrong' ones. In a two-sample t-test when the analyst is looking at a difference in means between two groups, for example, they ask whether this difference is 'significantly' different from zero (for the population from

which the samples were drawn). That is, they want to know given the two samples they actually achieved whether these were drawn randomly from the same population, and are therefore estimates of the same population mean. A significance test assumes from the outset that what is being 'tested' is true for the population, and so calculates the probability of obtaining a specific value from the random sample achieved (Siegel 1956). If we designate the null hypothesis that the two samples are from the same population as H_0, and the data obtained from the two samples as D, then what analysts want is the probability of H_0 given D. This is usually written as the conditional probability $p(H_0|D)$. And it is this probability that is implied when 'rejecting' a null hypothesis of no difference between the means. Analysts are saying, in effect, that the difference in the data is large enough (or sustained enough) to reject the idea that both samples come from the same population (i.e. that the observed difference between the two samples is solely due to the vagaries of random sampling).

Unfortunately, a significance test like a t-test does *not* calculate this probability or anything like it. Instead, it calculates $p(D|H_0)$ – the probability of obtaining the difference between the two samples assuming that they were both drawn from the same population. The conditional probabilities $p(D|H_0)$ and $p(H_0|D)$ both contain the same elements but they are very different. To assume that they are somehow the same would be like saying the probability of living in London if one owns a flat is the same as the probability of owning a flat if one is living in London. The two probabilities are not the same. Nor is either directly computable from the other. Knowing that 20% of London residents own a flat, for example, does not tell us, without a lot more information as well, what percentage of flat owners in the world live in London.

The p-value calculation in a significance test depends on the initial assumption of a null hypothesis about what is true for the population. As soon as it is allowed that the null hypothesis may not be true, the calculation goes wrong, and the p-value no longer exists. The actual computation for a significance test in practice involves no real information about the population, and this means that the same sample from two very different populations would yield the same p-values. A sample mean of 50 would, quite absurdly, produce the same p-value if the population mean were really 40, 50, 60 or 70 etc. This is because the population value is not known (else there would be no point in conducting the significance test), and the entire calculation is based only on the achieved sample value.

To illustrate the common misunderstanding of this, consider a simplified situation (Gorard 2014a). There is a bag, containing 100 well-shuffled balls of identical size, and the balls are known to be of only two colours. A sample of 10 balls is selected at random from the bag. This sample contains 7 red balls and 3 blue balls. The analytical question to be addressed is: how likely is it that this observed difference in the balance of the colours between the two samples is also true of the original 100 balls in each bag? The situation is clearly analogous to many analyses reported in social science research. The bag of balls is the population, from which a sample is selected randomly. A moment's thought shows that it is not possible to say anything very much about the other 90 balls in the bag. The remaining 90 might all be red or all blue, or any share of red and blue in between. There is no way any probabilistic calculation can be used to work it out. Yet the purpose of a significance test analysis here is to find out, via sampling, something about the balance of colours in the bag. Without knowing what is in the bag there is no way of assessing how improbable it is that the sample has ended up with 7 red balls. Once this impossibility is realised, the pointlessness of significance testing becomes clear.

What a significance test does instead is to make an artificial assumption about what is in the bag. Here the null hypothesis might be that the bag contains 50 balls of each colour at the outset. Knowing this, it becomes relatively easy to calculate the chances of picking 7 reds and 3 blue in a random sample of 10 balls. If this probability is small (traditionally less than 1 in 20, or 0.05) it is customary to claim this as evidence that the bag must have contained an unbalanced set of balls at the outset. This claim is obviously nonsense. The mere assumption of the null hypothesis tells us nothing about what is actually in the bag. For example, imagine that the bag started with 80 red balls and 20 blues. The sample is drawn as above, and contains 7 reds. The significance test approach assumes that there are 50 reds in the bag and calculates a probability of getting 7 in a sample of 10 balls. This probability will clearly be incorrect in reality because the balls are less balanced in fact than the null assumption requires. Now imagine that the sample is still the same but that the bag had 80 blue balls and only 20 red originally. The significance test approach again assumes that there are 50 reds in each bag and calculates the same probability of getting 7 red balls. This probability will also be clearly incorrect because the balls are less balanced than the null assumption requires, but now in the opposite direction. More absurdly, this second probability *must* be the same as the first one since they are both calculated in the same way on the same assumption. So the significance test would give exactly the same probability of having drawn 7 reds in a random sample from a bag of 80% reds as from a bag of 20% reds. This absurdity happens because the test takes no account of the actual proportion of each colour in the population. It cannot, since finding out that balance is supposed to be the purpose of the analysis.

Of course the probability of getting 7 reds from a bag containing 80 reds is different, *a priori*, to the probability of getting 7 reds from a bag containing 20 reds. But the significance test is conducted *post hoc*. There is no way of telling what the remaining population is from the sample alone.

For anyone who has spotted this misunderstanding, there is little doubt that their use of significance testing would cease (Falk and Greenbaum 1995). No one wants to know the probabilistic answer the tests actually provide (about the probability of the observed data given the assumption), and the test cannot provide the answer analysts really want (the probability of the assumption being true given the data observed). This conclusion is not new (Harlow et al. 1997). It has been known for a long time, perhaps since their earliest adoption, that significance tests do not work as hoped for, and may well be harmful because their results are so widely misinterpreted (Carver 1978). Significance testing and p-values are easily misunderstood, give misleading results about the substantive nature of results, and are ‘best avoided’ (Lipsey et al. 2012).

> Even if it can be demonstrated that data could or do meet the necessary requirements, the outputs of inferential statistics do not tell us anything we want to know (White 2014, p.27).

Standard errors in practice

In practice, rather than the theory of sampling outlined at the start, social scientists work with a very limited number of samples, often just one sample in each study. And, in practice, they do not know the relevant values or measurements for the complete population. Instead, they take a sample of cases precisely in order to gather the relevant values for the sample and so estimate the relevant values for the whole population. This is where ambiguity about the

‘standard error’ arises, because researchers cannot then compute the true SE for the population. If they had the relevant values for the complete population there would be no need for them to draw a sample to measure. A researcher already in possession of the relevant scores for a population would not usually want to take a sample from it. And if they did they would know its representativeness exactly, without any need for a SE.

In practice, we do not know the values for the population and have only the data for our one achieved sample. So, a convention has arisen that the ‘standard error’ of the entire population can be *estimated* from the standard deviation of the values in that one sample. The term ‘standard error’ is therefore also used, confusingly, to refer to what is only an estimate of the true standard error for the population from which that random sample is drawn. This estimated standard error for one sample is defined in statistical texts as the standard deviation of the sample, divided by the square root of the number of cases in the sample. It is important for what follows to recall that the estimated standard error for the population is therefore based solely on the values of one achieved sample, just like a significance test (above). The formula for the estimated standard error (se) makes no reference to the population itself.

Altman and Bland (2005, p.903) defined the standard error as the ‘standard deviation of the sampling distribution’ (see above), but they immediately move, in the next sentence and without justification, to stating that:

> Another way of considering the standard error is as a measure of the precision of the sample mean… The standard error of the sample mean depends on both the standard deviation and the sample size, by the simple relation SE = SD/√(sample size).

Cohen et al. (2011, p.97) put the necessary compromise more clearly:

> Strictly speaking, the formula for the standard error of the mean is:
>
> $S.E.=SD_{pop}/\sqrt{N}$
>
> However, as we are usually unable to ascertain the SD of the total population, the standard deviation of the sample is used instead.

This ambiguous nature of the term standard error, as demonstrated further below, encapsulates the problems with current use and training in statistics. A perfectly logical and mathematically correct construct like the standard error is created in theory based on a situation of perfect knowledge of the population, and the reality of being able to generate very large numbers of fixed-size complete samples at random from that population. It is logical, but fails in practice. Researchers either know the relevant population data, in which case sampling is not needed (in fact it would be absurd), or they do not know the population data in which case they do not know the true standard error. They also generally have only one sample. From this one sample they try to achieve the impossible, which is to estimate the true standard error of the population, even though they have no idea how representative of the population their one sample actually is. They then use this estimated standard error to help them try to decide how representative of the population their one sample is. It is like trying to measure the accuracy of a ruler, using only the same ruler. This sounds, and indeed is, invalid. The main problems are as follows.

First, the researcher interested in working with the standard error relevant to their one sample requires a random sample. As already shown, if the sample is not random there can be no standard error for it, even if the population values were known and many other repeated samples of the same size had been drawn from that population and even if *all* of these other samples were random. To be random, the sample must be complete in the sense that the values are known for each case in the sample, and there must be no missing cases (through non-response, dropout etc.). And the researcher must also know the identity of all other cases in the population, and all of these must have had a chance of being selected for the complete random sample. Most researchers have never been in such a situation and most are never likely to be. Therefore, no researcher should be working with the estimated standard error in practice.

Second, even if the researcher was in the ideal situation of having a true random sample, the 'logic' of standard errors does not work in practice. Like significance testing, it confuses two very different probabilities. The true standard error is based on knowing everything and then simply calculating the probability of selecting something (a sub-set of everything) – an easy calculation. The researcher estimating a 'standard error' from their one sample, on the other hand, is using the existence of something to try and estimate everything – an impossible calculation. And this confusion between two very different conditional probabilities then extends to the whole of sampling theory statistics as used by researchers (Gorard 2010). Significance tests provide a p-value which is the probability of the data in the achieved one sample, based on an assumption (or hypothesis) about the unknown data in the whole population. But the results are routinely treated as though they were the probability of the assumption being true, given the data in the achieved sample, even though these two probabilities are clearly different (see above). The same confusion occurs with confidence intervals and power calculations. The legitimate mathematical constructs of sampling theory based on the true standard error are not much help in real-life research, because in real-life the relevant values are only known for one sample and not for other members of the population, and because so few researchers have complete random samples anyway.

The dangers of estimating the standard error

The true standard error (SE) for a sampling distribution from a known population is very different to the estimated standard error (se) for the population based on one sample. Again, this can be illustrated via a simulation. Imagine that the population is the set of integers from 0 to 9 (treating 0 as an integer). From this population, 500 samples were drawn, each of 100 random cases. The true mean of the population is 4.5, and the SE of the sampling distribution is around 0.28 (the standard deviation of the 500 sample means). The achieved means of the 500 samples are normally distributed, with a lowest value of 3.64 and a highest value of 5.3. Each sample also leads to an estimate of the standard error (se), defined as the standard deviation of the sample divided by 10 (square root of N). These range from 0.25 to 0.33, and are uniformly distributed (meaning that the chances of a sample yielding 0.25 as the se is the same as the chances of it yielding the correct figure of 0.28). Note that where simulations were also run with a larger number of samples (1,000, for example), or a smaller number (50, for example), the sampling distribution SE remains the same, and the range of se also remains roughly the same. Naturally, changing the number of samples does not change the accuracy of any particular se, since each estimated standard error is based only on information from one sample. Each sample is independent, affecting the calculated SE but not the se (as is clear from their respective formulae).

In this simulation, using the range of sample means above, the estimates for the standard error (se) range from 11% or (0.25-0.28)/0.28 below the true figure, to 18% or (0.33-0.28)/0.28, above the true figure. These are very large maximum proportionate differences in estimating the standard error. They mean that any subsequent calculation using se will necessarily propagate any such differences, so affecting the eventual results. This is crucial because the standard error is the basis for so much of traditional statistics, underlying the calculation of power, test statistics in significance tests, confidence intervals, and the purported advantages of using technical approaches like multi-level modelling.

To illustrate the dangers, consider a one sample t-test. This simple test of significance is used for deciding whether one random sample mean comes from a population with a known mean (but unknown standard deviation). For example, we know that the population mean in our simulation with 500 samples of 100 cases each is 4.5. Suppose we consider one specific sample with a sample mean of 4. This is a reasonably close approximation to the population mean, and 100 cases is a reasonably large sample. The test is based on computing the test statistic t, and then looking up the critical value of t in a distribution table linked to degrees of freedom and the level of significance required (or the equivalent steps in software like SPSS). The statistic t is defined as the difference between the one-sample mean and the proposed null mean (from the null hypothesis), divided by the standard error. All parametric tests tend to use se or a combination of such estimated standard errors, to convert observed differences into multiples of standard errors (similar to computing standardised z-scores). To see whether any one of the 500 samples created for the simulation above was 'really' from the population of random integers, a statistician would take the null mean of 4.5. If the one sample mean were 4 then t should be calculated as (4.5-4)/0.28, using the true standard error SE. This would be 1.79. According to the table of t (e.g. http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf) the critical value for t, at p-value 0.05 (the standard) and 99 degrees for freedom (N-1), is just over 1.984 but below 1.99. Since 1.79 is below this critical (absolute) figure, tradition says the null hypothesis is retained, and the sample is assumed to be a fair one from the population of random integers (which it is).

Unfortunately, the analyst will not know the real standard error for the population, which is why the formula is always presented as the difference between the sample mean and the null mean (0.5 in the example), all divided by the estimated standard error. The simulation produced estimated standard errors as high as 0.33 or as low as 0.25. Thus, t could be calculated as anything from 1.5 or 0.5/0.33 to 2 or 0.5/0.25. This means that in practice the computed figure for t could be 16% or (1.5-1.79)/1.79 too small, or 12% or (2-1.79)/1.79 too high. A maximum proportionate difference of 16% in calculating a test statistic could make a difference to whether the sample mean was judged significantly different from the null mean or not. Here, since 2 is larger than the critical value for t, the null hypothesis would be rejected (incorrectly, as it happens) when the estimated standard error is as low as 0.25.

Note that this 28% (from -16 to +12) possible error range in the result for t is based on ideal textbook conditions, and a complete random sample, with a sample mean which is in fact a reasonable estimate. This 28% is not concerned with the sampling variation of the mean (which the t test is supposedly assessing). It is additional variation concerned with estimating the standard error. The simulation illustrates that the same mean from the same size sample can be judged either significantly or not significantly different to the population mean depending on the quality of the estimate of the true standard error. Note also that the 28% has

nothing to do with the alpha level, such as 5%, selected as a threshold for a significance test. Any calculation using the alpha level assumes, without justification as is shown here, that the standard error estimated from the one sample is correct. The p-value is a measure of uncertainty, but the measure itself is meant to be exact. Where the measure of uncertainty is itself uncertain (by up to 16% in this simple realistic example), the judgement of the actual uncertainty becomes very difficult indeed.

The simulation showed that in practice the one sample mean could vary from 3.64 to 5.3 (see above). Using the smallest estimate of the standard error and 3.64 as the smallest sample mean, this means that t could be calculated as large as 0.86/0.25 or 3.44. This means that the computed figure for t could be out by 92% or (3.44-1.79)/1.79. A proportionate error of anything like 92% in the calculation of t, based on a perfectly good and complete random sample of 100 cases, means that t is dangerous to use in practice. Even when used under perfect conditions and as intended, the propagation of the initial difference in estimating the se from one sample is too much. The results could be very misleading (but of course never noticed because the true SE would not be known). Here, both extreme values and many in between would lead to the 'incorrect' rejection of the null hypothesis. Using a poor estimate of the SE from one sample to judge the quality of the estimate of the mean from one sample does not work. Working with one sample, *post hoc*, it is not possible to know whether that sample is any good or not.

If we then accept that, in practice, a complete random sample is a rare phenomenon and we allow for some dropout or non-response the situation becomes worse again. In the simulation, the highest 10% of cases in each sample were deleted, leaving a sample size of 90 for each of the 500 samples. This represents bias in dropout or non-response, since evidence shows that such missing values are not random in nature (see above). Of course, the real population mean (4.5) and the SE of the sampling distribution (0.28) remain the same. But the samples now lead to worse representations of these true figures (considerable under-estimates). One common approach to handling such missing data is to compute a replacement value based on the data that is available. In the simulation, the missing cases were replaced by the mean of their samples (each of 90 cases now). The true population values still remain the same. However, the estimated standard errors become lower, now ranging from 0.21 to 0.27, respectively 25% and 4% below 0.28. Every single sample now yields an estimate of the standard error that is too low.

The simulation illustrates that trying to replace missing data often leads to more problems than it solves (Gorard 2014b). 'The standard error measures sampling variability; it does not take bias into account' (Berk and Freedman 2001, p.3). And adjustments made *post hoc* where the population values are not known are unlikely to work, and can make many samples worse. The difference from the true SE will carry forward into any calculation of significance such as those based on t or F, and the impact will be even more dangerous than when the sample is complete (above). Estimating the se as anything like 25% too low will lead to clearly erroneous claims of significance. In this example, again with a null mean of 4.5, t could be calculated as 4.1 or 0.86/0.21 (an extreme, based on the largest difference between the means divided by the smallest estimated se). This value of t would be 129% or (4.1-1.79)/1.79 larger than the true value of t for these data. This would mean that our sample mean would be judged significantly different from the null mean at a p-value of considerably less than 0.001. In fact, any sample with a mean outside a range of 4.2 to 4.8 would be judged, via a t test, to be significantly different from the population with a mean of 4.5, if using 0.25 as the estimated standard error.

Further issues with significance testing

So far, the main problems described with traditional approaches to QM, in the form of inferential statistics, are that:

- they are mostly irrelevant since researchers seldom work with complete random samples from a known population anyway;

- the logic of significance tests does not provide researchers will a useful answer;

- they are based on the standard error for the sampling distribution of the population, which will never be known in practice, and the inaccuracy in estimating it from one sample can make test results misleading (even in their own terms).

These should be enough to mean that QM moves away from sampling theory and its derivatives entirely. Yet, there are plenty more problems, some of which are outlined here.

There is widespread confusion about the level of the accepted p-value in a significance test (usually 5%) in relation to prediction and replication. Imagine sitting down to a game of Monopoly or something similar, and rolling two dice with a result of a 2 and a 3. Would you assume from that result that the dice were biased towards rolling 2s and 3s? I suspect not. I suspect you would not even notice the outcome, and yet it has an *a priori* probability of only 1/18 or 5.6%. I suspect you would not consider the dice biased even if your first roll yielded a 2 and another 2 (probability of only 1/36 or 2.8%). When social scientists conduct significance tests on their data, they use probabilities at about this level of 5% or less to help them decide that an outcome is so unlikely that they are justified in assuming that their starting assumption is wrong. In the Monopoly example, the probabilities are calculated assuming the dice are unbiased. If around 5% or less is evidence that our starting assumption is wrong then we should be suspicious of that first roll with a probability of only 1/18 or 1/36. But if 1/18 or 1/36 is not sufficiently unlikely to demand a new set of dice in Monopoly then how can it be small enough to make important decisions in social science?

The answer lies in prediction and replication. Predicting correctly (beforehand obviously) that the dice will roll a 2 and a 3 would be much more impressive. Strictly speaking, this is what the 5.6% probability refers to – while the *post hoc* probability is really 100%. Yet researchers and statistical 'experts' have largely forgotten this important aspect of probability. Analyses and predictions are meant to be part of the design of a study from the outset. Dredging the data after collection for possible patterns is something entirely different (Gorard 2013a). Also, the 5% for a significance test was always intended to be for a one-off analysis. Nowadays analysts run hundreds or even thousands of tests, often unknown to them as part of their regression or other modelling (where significance tests are conducted for individual independent variables and for the model as a whole). This is like rolling the dice hundreds of times, and then announcing the first 2 and a 3 result as remarkable. On the other hand, if the dice roll a 2 and a 3 repeatedly, that is also more impressive. All individual studies must be regarded as tentative, however strong their evidence, until independently replicated (Frank et al. 2013).

Any large real dataset divided into two groups by any criterion will yield differences between the two groups to some extent (unless by infinitesimal chance the two are identical to any number of decimal places). As Meehl (1967) has shown, any nil-null hypothesis is extremely unlikely to be true. This simply means that all analyses should lead to statistically significantly different results in any comparison between groups, as long as N is large enough. Viewed in this light, a non-significant result is an admission that the sample size is not sufficient to find the difference that must be there. The key issue for social science, not of whether there is a difference but whether it is large enough to pursue, is not addressed at all by significance tests. Since the nil-null hypothesis is never strictly true, some alternative hypothesis must be true instead. Given that there are an infinite number of alternate hypotheses, the results of a significance test do not help narrow down these alternatives. Genuine analysis requires a rather different approach (see below).

Having ignored all of these problems, it is no surprise that medicine, science and social science have experienced decades of 'vanishing breakthroughs' or findings that cannot be repeated and are never useful in practice (Matthews 1998). As explained at the start, many published results are false, and the chief culprit where numeric analysis is involved is the nonsense that is significance testing. QM, as currently envisaged, *is* the problem not the solution.

Confidence intervals and power

The problems that make significance tests unworkable have been dealt with at length. Once they fall, the rest of inferential statistics, as currently practiced, falls as well. With significance tests no longer used, power as currently defined becomes a red herring.

'Statistical power is the probability that if the population ES [effect size] is equal to delta, a target we specify, our planned experiment will achieve statistical significance at a stated value of alpha' (Cummings 2013, p.17). Therefore, power calculations would only make any sense if an analyst were planning to conduct a significance test. Since such tests do not work, it follows that power calculations do not work. We will not know the true value of delta (the effect size), else we would not need to conduct the study. But power calculations are circular and highly sensitive to minor variations in any estimated value such as delta (Gorard 2013).

Confidence intervals (CIs) are similarly disguised significance tests, based on the estimated standard error and with all of the dangers shown above. CIs and p-values are interchangeable and based on the same assumptions (Cummings 2013). Therefore, CIs do not work for the same reasons that significance tests do not work.

A 95% confidence interval, for example, is calculated as the sample statistic (such as the mean) plus or minus 1.96 times the standard error of the statistic (Peers 1996). The meaning is intended to be that if many (imaginary) random samples of the same size were taken from the same population, and the CI calculated for each, then the set of all CIs would contain the population mean 95% of the time. This does not imply that there is a 95% chance of a one sample CI containing the population mean. CIs therefore have the same backwards 'logic' as estimated standard errors and significance tests (*modus tollendo tollens*), worsened by being recursive in the sense that their definition incudes itself (Gorard 2014a).

The ensuing difficulties of comprehension lead to the common errors of using CIs as though they could handle non-random cases, and of interpreting a CI as a range of likelihood within which a desired parameter will fall. Neither is true (Watts 1991). Confidence intervals clearly do not work with non-random cases and missing data (Gorard 2014b). Like significance tests their calculation makes no reference to the population, and yet they are routinely presented as being an assessment of how close a parameter (like a sample mean) is to the unknown population mean. CIs do not work for the same reasons that significance tests do not work – we do not know the true SE, the estimated se can be seriously inaccurate, and we cannot use this one-sample estimate to decide whether this one sample is a good representation of the population. CIs are just the estimated standard error in different clothing. Like significance tests they use an estimate of variation from one sample to estimate how close an estimated measure (like a mean) from exactly the same sample is to that measure in the population. And like significance tests, CI calculations can yield exactly the same values from completely different populations.

For example, imagine that a sample mean was 50, with a specific standard deviation, and that this was drawn from a population with mean 60. The CI would have a particular range centred around 50. Now imagine that all else remains the same but that the population mean was actually 70. The CI would remain the same because the CI is unrelated to the actual population mean. This shows that a CI based on an estimate of 50 for a real value of 60 would imply the same level of accuracy as for a real value of 70. In practice, and even when used as intended, CIs convey no useful information.

Cluster randomised samples

The final substantive issue from traditional statistics dealt with briefly in this paper concerns the notion of cluster randomised samples (Rhoads 2014). These are, like many things in statistics, ambiguous. If samples are drawn at random from a list of clusters (like institutions or areas), and then measures are taken from cases within each cluster, the sample is really a random sample of clusters (chapter 16, Cochrane Collaboration Handbook - http://handbook.cochrane.org/chapter_16/16_3_3_methods_of_analysis_for_cluster_randomized_trials.htm). N would be the number of clusters sampled, rather than the number of 'cases' contained within all of the clusters, if the sample is considered to be a random one. This creates no problem, and everything that applies to cases in random samples (above) also applies to clusters as cases. Cases can be anything, including observations, texts, people, or institutions. There are a range of other simple approaches that also work perfectly well for handling problems when dealing with clustered data (Bland 2003). Some authorities, including Cochrane, might express concern that treating clusters as cases reduces 'power'. But, as shown above, power as currently defined is not relevant (unless a significance test is planned).

Some commentators (below) advocate using the measures taken within each cluster, instead of the cluster, to define their N. They want a larger N. But to a great extent which N they use makes no difference in practice to their results. Imagine a randomised controlled trial involving two groups, each group consisting of 10 schools, with each school having 100 pupils. A test is given to the full 2,000 pupils. The average test score for the 1,000 pupils in the first group must be identical to the average test score for the 10 schools in that group. It does not matter whether N is treated as 10 or 1,000. The actual difference in mean test scores between the two groups (the headline result of the trial) will remain the same.

These writers then worry that using the measures taken within clusters for the calculation of test statistics may lead to an under-estimate of the standard error, because these measures may be more similar within clusters than would be expected by chance. Of course, as shown throughout the paper, they should not be using test statistics anyway. Their cases do not generally occur within clusters like schools, prisons, regions, or hospitals by chance. So the occurrence of the cases within clusters is *not* random, meaning not that the SE is too low but that there is no SE (by definition, see above). In this situation, a standard error could only exist at the cluster level, and only if the clusters used as cases in a sample had been selected at random with no refusal, dropout or missing values.

Yet these writers persist, perhaps so that they can use increasingly complex techniques such as multi-level modelling to overcome a problem that does not really exist. If cases within clusters are selected randomly (as well as the clusters themselves) there are now two levels of randomisation. But the randomisation within each cluster is constrained, so the writers present the concept of the intra-cluster correlation. This means that in power calculations to decide on the required size of a sample, a design effect is calculated from the intra-cluster correlation and then used to multiply the required number of cases by. The design effect is defined as 1 + (the number of cases per cluster -1) x the intra-cluster correlation (Kish 1965). It is not clear what to do if, as is likely, the number of cases and the intra-cluster correlations vary between clusters.

Aside from the fact that significance tests do not make sense and power calculations do not make sense without them, what is also strange about this design effect is that it takes no account of the sampling fraction within the clusters. If all of the cases within each cluster are involved then there is no standard error at the cluster level, but the design effect takes no notice of that. This means that the design effect will be considered larger (and so the sample worse in quality) if all cases are used in each cluster than if only a few are selected at random. Whereas, in truth and all other things being equal, the population data within clusters must be preferred to sampled data. Of course, if there is no sampling within each cluster, the design effect should not be calculated. But the advocates of multi-level modelling say that cluster effects are present even in such complete population data. Some extreme commentators, such as Goldstein (2008), therefore want analysts to provide sampling theory derivatives such as confidence intervals for population data. And to try and justify this, they have used the notion of an infinitely large super-population to suggest that even populations are merely random samples from some larger imaginary super-population. These are the lengths that such writers have to go to in order to defend their sampling theory practices.

For example, Camilli quotes Goldstein as arguing that statisticians are not really interested in generalising from a sample to a specified population but to an idealised super-population spanning space and time. Goldstein claims that 'social statisticians are pretty much forced to adopt the notion of a "superpopulation" when attempting to generalise the results of an analysis' (Camilli 1996, p,7). As Glass counters, (also in Camilli 1996), such imaginary populations are simply the evidence we have from the sample writ large, and necessarily having the same characteristics. This is the complete inverse of statistical inference, and makes any attempt to assess sampling errors erroneous. 'I think we're lost when we accept statistical inferences based on data that weren't observed, and moreover do not exist conceptually… [Goldstein is]… playing a game with the language' (Camilli 1996, p.11). These so-called superpopulations are imaginary, data has *not* been randomly sampled from them (Berk 2004, p.52), and the fact that analysts have regularly invoked them is no

justification for retaining them (Freedman 2004, p.989). 'At the risk of the obvious, inferences to imaginary populations are also imaginary' (Berk and Freedman 2001, pp.1-2). The kind of convoluted sophistry proposed by Goldstein and many others is retarding the progress of social science (Robinson 2004). Like everything discussed so far, it needs to be dropped to make QM both correct and incidentally easier for a wider audience to understand.

What are the alternatives?

Further simplifications

Once released from consideration of probability distributions and the like, 'quantitative methods' can then be further simplified. Without being as hard to understand as significance testing, one of the other things that newcomers report as strange in their QM training is the formula for the standard deviation. It comes from a time when it was felt that absolute values were hard to work with. So the standard deviation (SD) as a measure of dispersion tries to avoid absolute values by squaring and summing the deviations from the mean, and then taking the square root of the result. However, the standard deviation does not avoid the use of absolute values in practice. A square root (for a real number) has a negative and a positive form, yet software like SPSS, and practice in journal publishing, only reports the absolute value of the square root as *the* standard deviation.

Newcomers generally find the average absolute deviation from the mean easier to understand and to compute. This mean absolute deviation has many other practical advantages over the SD (Gorard 2005) and is finding increasing use in areas from dental age estimation (Ashith and Acharya 2014) to computer science (Anand and Narashimha 2013). Using the mean absolute deviation as a preferred measure of dispersion, it is then possible to modify other forms of analysis to make them easier to portray and more robust in face of extreme scores. There is a mean absolute deviation version of the effect size, correlation coefficient and regression model, for example (Gorard 2013b, 2015a, 2015b). However, useful though these may be for the future, not all are yet ready for widespread use, and none of them leads to the kind of simplification that follows from simply not using significance testing (and all its disguised forms).

Thinking about analysis more clearly

Working with numeric data in social science is at once less complex and more skilled than most commentators portray. It is less complex and less scary because the most difficult bits mathematically, that many novices find off-putting, are useless, irrelevant, illogical and misleading. Significance tests and everything associated with them like confidence intervals and the use of p-values in modelling do not work as intended. No one should use these techniques nor accept their use unproblematically in anything that they read. Avoiding their use is easy enough to do. When reading the work of others who have used such techniques, these elements can be ignored, and their evidence and argument judged on its own merits. For example, if a researcher has used a p-value from a significance test to decide whether to pursue a difference in scores between two groups, the p-value should be ignored (Gorard 2014a). Instead, the study design, scale, bias, data quality, accuracy of scoring, variability of response, and above all the meaning of the numbers involved can help decide whether this difference is noteworthy or not. There is no technical or statistical way to do this. It requires

judgement, making the task *at once less complex and more skilled than most commentators portray* (Gorard 2006).

An obvious question then is what analysts can do instead, in terms of analysis or computation. There are a number of techniques that might help in portraying to others the grounds for any judgement. These include simple tables, graphs, percentages, proportions, averages, mean absolute deviations, effect sizes, odds ratios, attainment gaps, regression coefficients, correlation coefficients, and various indices of inequality. All should be used to clarify an important point for the reader (no undigested SPSS output please). Useful examples of this approach, using SPSS without significance tests, appear in the video tutorials by Patrick White (https://www.youtube.com/user/patrickkwhite).

Missing data could be handled using the permutation test (above) but replacing all the missing data with the 'least attractive' score in the same group/arm. The least attractive score could the one that is most counterfactual to the overall finding. If the result survives this hard test it must be robust. Put another way, robustness can be assessed as the proportion of cases that would have to be replaced with counterfactual data to invalidate the finding (Frank et al. 2013). This approach has been simplified, explained and illustrated as the number of counterfactual cases needed to disturb a finding (Gorard and Gorard 2015). It is also useful to separate thinking about the scale, strength or trustworthiness of the findings (an activity internal to the study) from whether the findings might be more generally true (which is always more speculative).

All of these approaches could be useful, and many involve nothing more than elementary arithmetic. They do not require sampling distributions, probability density functions, or standard errors (these can be left to mathematicians). The data obtained in any study can also be treated as a population in its own right, and inference discarded. However, these techniques do not provide a 'push-button' answer. There is no 'push-button' answer, any more than there is when analysing non-numeric data. What analysts should do instead is what they would have done if they had never heard of significance tests, confidence intervals and the like.

It is also worth reminding readers that even if they worked as intended, CIs and p-values could not address measurement error, missing data or bias. Using population data avoids the uncertainty of random sampling. Otherwise, the larger a sample (or number of cases) is, the more secure the findings will be. The better the sample is, in terms of random selection/allocation and full response, the more secure the findings will be. The more accurate the measurements are, the more secure the findings will be. And the larger the estimated mean is in proportion to the sample standard deviation (or the simpler mean absolute deviation) then the more secure the findings will be. There is no technical way of synthesising all of these factors to provide an overall estimate of quality. The only approach to numeric analysis of this sort is judgement. It requires considerable skill, and is best approached with great clarity and with any arguments laid out as simply as possible. Gorard (2014c) provides a 'sieve' that can be easily used to consider the study research design, scale, attrition, data quality, and other threats to the validity of a research result, synthesising these various factors into an overall judgement. The impact of using such an approach more generally would be a social science in which the logic of working with numbers would be the same as the logic of working with any other kind of data.

Conclusions

Much of the panoply of traditional statistics is inappropriate for real-life research where samples are so seldom truly random or normally distributed, where non-response and dropout always occur, and the use of population data is growing in prevalence. Even under ideal circumstances the logic of significance testing, p-values, standard errors, and confidence intervals does not provide analysts with the answers they seek. All such approaches are generally useless, and should be abandoned. This yields several immediate benefits. Learning about 'quantitative' research becomes much easier and less daunting for those who feel that it does not make sense but cannot express why, and so tend to avoid all use of numbers instead. And it becomes clearer that the underlying logic of analysing data is the same whatever format that data takes. This has many implications for those purporting to be 'qualitative' researchers, in areas such as scale, transparency, and warrant. Real analysis is harder than the traditional 'push-button' approach to any work with numbers in social science. It is simple conceptually but hard because it requires skill, judgement, clarity, and integrity.

There is a tradition of trying to draw a distinction between so-called 'quantitative' and 'qualitative' research when conducting and reporting on data analyses. This is scientifically misleading and confusing for new researchers, and the QM initiative is likely to make this worse, if only in terms of its exclusionary name. There is no particular difference between analysing data consisting of qualities and data consisting of measurements. A researcher reporting that 51% of people in a survey agreed with a statement is saying something very similar to a researcher reporting that many people in interviews made something like a particular quoted statement. Research measurements are generally of qualities, while analytical statements about qualities are generally numeric in form ('most', 'few', 'none'). The widespread use of random sampling theory derivatives such as standard errors, significance tests, and confidence intervals tends to make working with numbers look very different to working with other forms of data. But as soon as these are removed, as illogical and inappropriate in real-life research, that difference disappears.

All research designs, such as longitudinal, experimental or case study, are independent of the method of data collection used within them. A study that contacted the same people as research informants every year for ten years would be longitudinal. If it involved a survey of those people it would be longitudinal. If it involved interviews with those people it would be longitudinal. If it involved collecting existing data such as payslips from those people it would be longitudinal, or if the people were videoed or their height was measured. If the study did all of these things combined or different things every year it would still be longitudinal. There are no 'quantitative' or 'qualitative' designs, and since analysis is predicated in the design, there are no distinctly 'quantitative' or 'qualitative' analyses. There is no difference in analysing a set of data dependent upon whether it was gathered face-to-face, using IT, or on paper. Put another way there is no more difference between analysing text and analysing numbers than there is between analysing text and analysing photos or smells. Each may require some special skills but their overall logic remains the same.

The design elements of any research design, such as selecting the cases to be involved, allocating them to groups, or planning an intervention, are also all independent of the method of data collection used (Gorard 2013a). If a particular sample size is needed to make a believable finding about what people state about some phenomenon it does not matter whether those people make the statement to the researcher via a recording, a form, a computer, or conversationally in passing. All other things being equal, the larger the sample

the stronger and more convincing the research results will be. And all other things being equal, a random sample will provide a less-biased estimate of a more general population than any other kind of sample. Research craft knowledge like this remains valid whatever kind of data is then collected from that sample.

Analytical decisions, once all of the clutter is removed, then tend to become simpler and more similar across a range of situations. Most of them are of the form 'how big is this difference?', 'how strong is this pattern or trend?', or 'how valid is this exception?'. There is no technical way of answering any of these questions. For example, 'how strong is this pattern?' might depend on the estimated scale, the number of cases, and the variability of the phenomenon being investigated. Such factors can be summarised, such as by an 'effect' size, but this does not really answer the question. It simply converts the question to 'how important/stable is this effect size?'. There is no valid standard scale for judging the importance of an effect size. Its importance may depend also on the costs, benefits and unintended consequences of answering the question one way or another. Really all that an analyst can do is present the evidence, and explain clearly and fully why they believe that the evidence leads to the conclusion they make. They need to 'warrant' their conclusions, by marshalling logic and evidence to create a compelling argument that leads to a conclusion, and stands up to critical scrutiny (Gorard 2002, Phillips 2014). The software, graphs, frequency counts or effect sizes are useful tools for presenting the findings and sorting them out for analysis. The real analysis is the judgement that follows.

Acknowledgements

The ideas presented in this paper were sharpened by activities conducted as part of the ERSC Researcher Development Initiative, and by the very helpful comments of Patrick White.

References

Altman, D. and Bland, M. (2005) Standard deviations and standard errors, *BMJ*, 331, p.903, doi: 10.1136/bmj.331.7521.903

Anand, M. and Narasimha, Y. (2013) Removal of salt and pepper noise from highly corrupted images using mean deviation statistical parameter, *International Journal on Computer Science and Engineering*, 5, 2: 113-119

Ashith B. Acharya (2014) Forensic dental age estimation by measuring root dentin translucency area using a new digital technique**,** *Journal of Forensic Sciences*, 59, 3, 763-768

Behaghel, L., Crepon, B., Gurgand, M. and Le Barbanchon, T. (2009) *Sample attrition bias in randomized surveys: a tale of two surveys*, IZA Discussion Paper 4162, http://ftp.iza.org/dp4162.pdf, accessed 060714

Berk, R. (2004) *Regression Analysis: A constructive critique.* Thousand Oaks, Ca: SAGE

Berk, R. and Freedman, D. (2001) *Statistical assumptions as empirical commitments*, http://www.stat.berkeley.edu/~census/berk2.pdf, accessed 030714

Bland, M. (2003) *Cluster randomised trials in the medical literature*, http://epi.klinikum.uni-muenster.de/StatMethMed/2003/Freiburg/Folien/MartinBland.pdf, accessed 25/4/06

Camilli, G. (1996) Standard errors in educational assessment: a policy analysis perspective, *Education Policy Analysis Archives*, 4, 4

Carver, R. (1978) The case against statistical significance testing, *Harvard Educational Review*, 48, 378-399

Cohen, L., Manion, L. and Morrison. K. (2011) *Research methods in education* (7th Edition), London: Routledge

Cooper L. and Shore F. (2008) Students' Misconceptions in Interpreting Center and Variability of Data Represented via Histograms and Stem-and-leaf Plots, *Journal of Statistics Education*, 16, 2, 13 pages

Cuddeback, G.. Wilson, E., Orme, J. and Combs-Orme, T. (2004) Detecting and statistically correcting sample selection bias, *Journal of Social Service Research*, 30, 3, 19-30

Cumming, G. (2013) The new statistics: why and how, *Psychological Science*, dio:10.1177/0956797613504966

Dolton, Lindeboom, M. and Van den Berg, G. (2000) *Survey Attrition: A taxonomy and the search for valid instruments to correct for biases*, http://www.fcsm.gov/99papers/berlin.html

Ernst, M. (2004) Permutation methods: a basis for exact inference, *Statistical Science*, 19, 4, 676-685

Falk, R. and Greenbaum, C. (1995) Significance tests die hard: the amazing persistence of a probabilistic misconception, *Theory and Psychology*, 5, 75-98

Frank, K., Maroulis, S., Doung, M. and Kelcey, B. (2013) What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences, *Educational Evaluation and Policy Analysis*, 35, 4, 437-460

Freedman, D. (2004) Sampling, in M. Lewis-Beck, A. Bryman and T. Liao (Eds.), *Sage encyclopaedia of social science research methods* (pp.987–991). Thousand Oaks, Ca: SAGE

Glass, G. (2014) Random selection, random assignment and Sir Ronald Fisher, *Psychology of Education Review*, 38, 1, 12-13

Goldstein, H. (2008) Evidence and education policy – some reflections and allegations, *Cambridge Journal of Education*, 38, 3, 393-400

Gorard, S. (2002) Fostering scepticism: the importance of warranting claims, *Evaluation and Research in Education*, 16, 3, 136-149

Gorard, S. (2004) The British Educational Research Association and the future of educational research, *Educational Studies*, 30, 1, 65-76

Gorard, S. (2005) Revisiting a 90-year-old debate: the advantages of the mean deviation, *The British Journal of Educational Studies*, 53, 4, 417-430

Gorard, S. (2006) Towards a judgement-based statistical analysis, *British Journal of Sociology of Education*, 27, 1, 67-80

Gorard, S. (2010) All evidence is equal: the flaw in statistical reasoning, *Oxford Review of Education*, 36, 1, 63-77

Gorard, S. (2013a) *Research Design: Robust approaches for the social sciences*, London: SAGE

Gorard, S. (2013b) The possible advantages of the mean absolute deviation 'effect' size, *Social Research Update*, 65, Winter 2013, pp.1-4, http://sru.soc.surrey.ac.uk/SRU64.pdf

Gorard, S. (2014a) The widespread abuse of statistics by researchers: what is the problem and what is the ethical way forward?, *Psychology of Education Review*, 38, 1, 3-10

Gorard, S. (2014b) Confidence intervals, missing data and imputation, *International Journal of Research in Educational Methodology*, 5, 3, 693-698, http://cirworld.org/journals/index.php/ijrem/article/view/2105/pdf_51

Gorard, S. (2014c) A proposal for judging the trustworthiness of research findings, *Radical Statistics*, 110, 47-60

Gorard, S. (2015a) Introducing the mean absolute deviation 'effect' size, *International Journal Research and Methods in Education*, 38, 2, 108-114
Gorard, S. (2015b) An absolute deviation approach to assessing correlation, *British Journal of Education, Society and Behavioural Sciences*, 5,1, 73-81
Gorard, S. and Gorard, J. (2015) The number of counterfactual cases needed to disturb a finding: a new approach to summarising scale, effect size and attrition, *International Journal of Social Research Methodology*, (under review)
Hansen, M. and Hurwitz, W. (1946) The problem of non-response in sample surveys, *Journal of the American Statistical Association*, *41*, 517–529
Hampel, F. (1997) *Is statistics too difficult?*, Research Report 81, Seminar fur Statistik, Eidgenossiche Technische Hochschule, Switzerland
Harlow, L., Mulaik, S. and Steiger, J. (1997) *What if there were no significance tests?*, Marwah, NJ: Lawrence Erlbaum
Ioannidis, J. (2005) Why Most Published Research Findings Are False, *PLoS Med*. Aug 2005; 2(8): e124, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/, accessed 060714
Kish, L. (1965) *Survey Sampling*, New York: John Wiley & Sons, Inc
Lindner, J., Murphy, T. and Briers, G. (2001) Handling non-response in social science research, *Journal of Agricultural Education*, 42, 3, 43-53
Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K. and Busick, M. (2012) *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*, Washington DC: Institute of Education Sciences, p.13
Ludbrook, J. and Dudley, H. (1998) Why permutation tests are superior to t and F tests in biomedical research, *The American Statistician*, 52, 2, 127-132
Matthews, R. (1998) Flukes and flaws, *Prospect*, November 20 1998, http://www.prospectmagazine.co.uk/features/flukesandflaws
McIntyre, D. and McIntyre, A. (2000) Capacity for research into teaching and learning, Swindon: Report to the ESRC Teaching and Learning Research Programme
Meehl, P. (1967) Theory - testing in psychology and physics: A methodological paradox, *Philosophy of Science*, 34, 103 – 115
Murtonen, M. and Lehtinen, E. (2003) Difficulties experienced by education and sociology students in quantitative methods courses, *Studies in Higher Education*, 28, 2, 171-185
Peers, I. (1996) *Statistical analysis for education and psychology researchers*, London: Falmer Press
Peress, M. (2010) *Correcting for Survey Nonresponse Using Variable Response Propensity*, Journal of the American Statistical Association, http://www.rochester.edu/College/faculty/mperess/Nonresponse.pdf
Phillips, D. (2014) Research in the grad sciences, and in the very hard 'softer' sciences, *Educational Researcher*, 43, 1, 9-11
Pigott, T., Valentince. J., Polanin, J., Williams, R. and Canada, D. (2013) Outcome-reporting bias in education research, *Educational Researcher*, 42, 8, 424-432
Platt, J. (2012) Making them count: how effective has official encouragement of quantitative methods been in British sociology?, *Current Sociology*, 60, 5, 690-704
Rhoads, C. (2014) Under what circumstances does external knowledge about the correlation structure improve power in cluster randomized designs?, *Journal of Research on Educational Effectiveness*, 7, 2, 205-224
Robinson, D. (2004) An interview with Gene Glass, *Educational Researcher*, 33, 3, 26-30
Sheikh, K. and Mattingly, S. (1981) Investigating nonresponse bias in mail surveys, *Journal of Epidemiology and Community Health*, *35*, 293–296

Siegel, S. (1956) *Nonparametric statistics for the behavioural sciences*, Tokyo: McGraw Hill
Simmons J., Nelson L. and Simonsohn U. (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant, *Psychological Science*, 11:1359-66. doi: 10.1177/0956797611417632
Wandt, Edwin, Adams, Georgia W., Collett, Dorothy M., Michael, William B., Ryans, David G., & Shay, Carleton B. (1965) *An evaluation of educational research published in journals*, Report of the Committee on Evaluation of Research, American Educational Research Association, unpublished report
Watts, D. (1991) Why is introductory statistics difficult to learn?, *The American Statistician*, 45, 4, 290-291
White, P. (2014) Against inferential statistics, *Psychology of Education Review*, 38, 1, 24-28