AGAINST INFERENTIAL STATISTICS: HOW AND WHY CURRENT STATISTICS TEACHING GETS IT WRONG

PATRICK WHITE
Department of Sociology, University of Leicester
pkw4@leicester.ac.uk

STEPHEN GORARD
School of Education, Durham University
s.a.c.gorard@durham.ac.uk

ABSTRACT

Recent concerns about a shortage of capacity for statistical and numerical analysis skills among social science students and researchers have prompted a range of initiatives aiming to improve teaching in this area. However, these projects have rarely re-evaluated the content of what is taught to students and have instead focussed primarily on delivery. The emphasis has generally been on increased use of complex techniques, specialist software and, most importantly in the context of this paper, a continued focus on inferential statistical tests, often at the expense of other types of analysis. We argue that this 'business as usual' approach to the content of statistics teaching is problematic for several reasons. First, the assumptions underlying inferential statistical tests are rarely met, meaning that students are being taught analyses that should only be used very rarely. Secondly, all of the most common outputs of inferential statistical tests – p-values, standard errors and confidence intervals – suffer from a similar logical problem that renders them at best useless and at worst misleading. Eliminating inferential statistical tests from statistics teaching (and practice) would avoid the creation of another generation of researchers who either do not understand, or knowingly misuse, these techniques. It would also have the benefit of removing one of the key barriers to students' understanding of statistical analysis

Keywords: *Teaching statistics; Abuse of statistics; Inferential statistics; Significance testing*

1. INTRODUCTION

Over the past two decades, there has been a growing concern that there is a deficit of quantitative skills within the social sciences and humanities in the United Kingdom. Several reports have concluded that there is insufficient proficiency in numeric and statistical analysis among students, teachers and researchers (British Academy, 2011, 2015; Nuffield Foundation, 2012; Wiles, Durant, De Broe, & Powell, 2009). In response to this concern a number of high profile multi-million pound initiatives have been funded by a range of bodies, such as the Economic and Social Research Council (ESRC), the Higher Education Funding Council (HEFCE), the Nuffield Foundation, and the British Academy (BA). The UK has been unfavourably compared to the United States, which

Statistics Education Research Journal, 16(1), 55-65, http://iase-web.org/Publications.php?p=SERJ
© International Association for Statistical Education (IASE/ISI), May, 2017 – Non peer-reviewed discussion paper

reportedly has much higher levels of statistical capacity, perhaps because of the different nature of both the undergraduate and postgraduate curricula in that country. In the UK, subjects such as psychology and economics have been viewed as exempt from this problem, and are instead seen as models for other social science disciplines to emulate.

In our view, however, the social sciences in countries such as the US, and economics and psychology in the UK, all suffer from a damaging imbalance in the content of their statistics teaching. Regardless of discipline or geography, most statistical teaching focuses heavily on the use of inferential statistical tests (ISTs). In our experience (a combined total of more than five decades teaching or being taught statistics) this is often to the detriment of the understanding of the descriptive statistics (on which ISTs must necessarily be based) and at the expense of a more general statistical literacy or 'number sense'. Despite a long and sustained history of criticism of ISTs beginning almost immediately after their first use, little has changed in terms of what is taught to students and, subsequently, practised by researchers. In the language of one of the recent UK-based initiatives to increase statistical skills, we are proposing a genuine 'step change' in the content of statistics teaching. This would involve an immediate elimination of ISTs from the generic curriculum and from publications in academic journals. In the following sections we show why this apparently radical proposal is, in fact, not radical at all, and that once ISTs are fully understood it is the only reasonable course of action to take. Before moving on to look at the technical problems with ISTs, we briefly examine current teaching practice to show that, for many, any change will necessarily involve not only a change in understanding but also a shift in very well-established and seemingly intractable cultural practices.

2. CURRENT TEACHING

Of those aimed at undergraduate level study and above, Field's (2009, 2013) series of textbooks on statistics are probably the best selling in the UK. At the time of writing, the latest edition of Field's (2013) text was rated as the third highest selling book in both 'Social Sciences Statistics and Research' and 'Probability and Statistics' section on Amazon. None of the better selling books were statistics texts aimed at undergraduate students. In the third edition (2009, p. 49) the author states that "most of this book deals with inferential statistics", a characteristic that Field's texts share with the vast majority of other books and resources in this area. Unlike many statistics textbooks, however, both the third and fourth (2013) editions contain sections warning of the problems with null-hypothesis significance testing (NHST) and *p*-values, with Field concluding that NHST "is flawed because the significance of the test tells us nothing about the null hypothesis" (2013, p. 76). This is clearly a welcome development in a field where many textbooks either ignore or downplay the assumptions underlying ISTs (e.g., Connolly, 2007) and incorrectly define their key outputs such as *p*-values, standard errors and confidence intervals (e.g., Somekh & Lewin, 2005, p. 224). But while Field does promote confidence intervals as alternatives to NHST and *p*-values, much of what follows in the book is 'business as usual'. Readers are warned of the flaws underlying these techniques and their outputs, but ISTs are then included routinely in the rest of the text.

Field's texts have not been chosen as an example because they represent the worst practice in the area. In fact, in some respects they are better than the majority of statistics textbooks available for undergraduates. The author explains statistical concepts clearly and uses an informal style to engage readers. Unlike some authors, Field is correct in several of his definitions of the outputs of ISTs, and honest about having previously defined CIs incorrectly in an earlier edition. And unlike most texts aimed at undergraduates, the problems with NHST and *p*-values are explained clearly. These texts were chosen, rather, for the following reasons. First, as this is a short article that is not intended to be a comprehensive review of the literature, there is only space to focus on a few texts. As Field's textbooks are widely spread, they will be familiar and accessible to the readership. Because of their market share, they also have the potential to be very influential. And lastly, as they are popular and generally well-respected, they cannot be dismissed as an easy or 'straw' target.

What Field's texts illustrates, however, is what appears to be the main obstacle to a much needed change in direction for statistical analysis – the acknowledgement of the flaws in NHST and ISTs perversely combined with a very much 'business as usual' approach to their use. Many commentators now generally agree that ISTs do not work as intended, and could not be used appropriately in practice anyway (see the references in Gorard, 2016, and in Siegfried, 2012, 2015). However, debate in this area – even among those with some knowledge of the problems with ISTs – seems to have stalled with a recognition of these problems, but a reluctance to make any fundamental changes to either research practice or the content of teaching. It is interesting to see that although Field (2011) listed NHST as one of his "top 5 statistical faux pas", he continues to use *p*-values and terms such as 'statistically significant' and 'not significant' in the interpretation of results in his actual reports of empirical work published years after he raises these concerns in his textbooks (e.g., Lester et al., 2015; Reynolds, Field, & Askew, 2014, 2015). It would be easy to explain this practice in terms of the requirements of editors, reviewers or co-authors. However, given Field's status as a leader in the teaching of statistics this defence would appear weak. Both White and Gorard have faced similar problems with reviews on occasion, but have managed to get all of their work published without reporting the results of ISTs.

This kind of position, of continuing to use a known flawed approach, seems to us to characterise, with some rare exceptions, the stance taken by all but the most 'radical' of commentators in this area. In a recent discussion of these issues in *The Psychology of Education Review*, one of our positions on this topic (Gorard, 2014 a) was characterised as "intentionally iconoclastic" (Howe, 2014, p. 14) and "purist" (Putwain, 2014, p. 17). Out of six responses, two were positive (Glass, 2014; White, 2014) and four negative. We are acutely aware that, after over 70 years of debate, ours is still very much a minority position. One of the defences offered by those who use and promote the use of ISTs is that they are being 'pragmatic'. In his response, Putwain (2014, p. 19) argues that pragmatists will not "be convinced solely by a technical argument that they may already be aware of". So what exactly are the problems with ISTs? And to what extent is ignoring these 'technical' problems a pragmatic response?

3. WHAT ARE THE PROBLEMS?

Some of the problems with ISTs are more well-known than others. In our view, the problems that are least well-known – or at least most rarely aired – are the most important, as once these are understood they expose all ISTs as fundamentally problematic, and render the other issues moot. However, it is worth outlining all relevant issues with ISTs before drawing some conclusions about what should be taught in future methods courses. Directly below, we show why some of the commonly used arguments to defend the use of ISTs are flawed.

3.1. ABUSE IS THE ONLY PROBLEM

Many commentators have argued that there is nothing inherently wrong with inferential statistics and that problems only arise when they are used inappropriately or interpreted incorrectly. The main problem with this counterargument is that it is difficult – perhaps even impossible – to find examples of published research where these two criteria have actually been met. The authors have repeatedly asked those defending the use of ISTs to provide examples of their own work where ISTs have been used and interpreted correctly, and have never had more than an evasive response.

The calculations involved in ISTs assume either random sampling or random allocation to groups, and any non-response, drop-out or missing data would violate these assumptions. These issues affect most social scientific research to a greater or lesser extent, resulting in a situation where samples that remain truly random, or groups that remain properly randomised, are very rare or even non-existent (Cuddeback, Wilson, Orme, & Combs-Orme, 2004). And as drop-out and non-response have been demonstrated to be non-random in nature, this results in non-random samples, and groups that are not truly randomised (Sheikh & Mattingly, 1981). In short, the assumptions underlying ISTs are rarely – if ever – met. And these assumptions are, in the words of Berk and Freedman (2003), “empirical commitments” that affect the kinds of conclusions we can draw.

Abuse of ISTs – in terms of flouting the required assumptions – is clearly a widespread, almost universal, problem. It is, in fact, so common that it is reasonable to characterise it as ‘standard practice’. Many texts mention the key underlying assumptions only in passing, and some even encourage students to ignore them completely (e.g., Connolly, 2007). The fact that they are so rarely met is not an issue that is seriously discussed in any texts that we have encountered, other than our own. Even so-called experts in the field routinely use ISTs in their own work to analyse convenience samples or random samples with substantial drop-out or missing data (e.g., Sturgis, Brunton-Smith, Kuha, & Jackson, 2013). Bizarrely, they are even commonly used on population data when generalisation is not even an issue (e.g., Bukodi, Goldthorpe, Waller, & Kuha, 2014).

However, even in the unlikely situation that the required assumptions are met, it is unclear how interpreting the outputs of ISTs could be helpful to the researcher. As we show below, when these outputs are interpreted correctly, they produce information that is at best irrelevant and at worst misleading. Perhaps this is why, despite several requests on discussion forums, no advocates of ISTs have been able to produce published research reports containing these ‘correct’ interpretations. However, the burden of proof should be on those who advocate and use these techniques to show that they are both properly used and useful. Unfortunately, few of them seem to see it this way.

3.2. IT IS ONLY *P*-VALUES THAT ARE PROBLEMATIC

The second defence of ISTs is that *p*-values are the problem. Many texts will remind readers that statistical significance is not the same as substantive significance, for example. Much is made of the different interpretations of *p*-values by Fisher and Neyman and of the arbitrary ‘industry standard’ levels of 0.05, 0.01 and so on. Others will point out that with a sufficiently large sample any finding will be ‘significant’. Some even list and then debunk common misinterpretations of *p*-values (e.g., Mulaik, Raju, & Harshman, 1997).

These are all relatively common issues raised in debates about significance testing. However, there has been increasing recognition that one of the problems with *p*-values is what they tell us and, as importantly, what they do not. These problems are increasingly well-known but increased awareness of these issues has not been matched by any noticeable change in real-life practice.

As researchers, what we want to know is the probability of the null hypothesis being true (or false) given the data obtained, or $p(H_0|D)$. We want to know the probability that the difference or relationship that we observed in the sample (or experimental data) is due to the vagaries of random sampling (or allocation) rather than being a true reflection of data in our population. Unfortunately, NHST cannot tell us this. This approach can only tell us, via *p*-values, the probability of a difference or relationship at least as large or strong as that observed in our sample if our hypothesis was true in the first place: $p(D|H_0)$. To create this probability at all it has to be assumed that the null hypothesis is true.

At first sight the differences between these two probabilities may seem slight or unimportant. However, as Gigerenzer (2003) and many others have clearly demonstrated, these probabilities are not interchangeable, and nor can one be calculated or estimated from the other without the kind of information we do not usually have. If a diagnostic test is 90% accurate in both directions in detecting or missing a symptom that occurs in 1% of the population, then the probability of testing positive or p(Pos|Sym) if one has the symptom (or negative of one does not) is 90%. But the probability of having the symptom if tested as positive or p(Sym|Pos) is only about 8%. To calculate one and then use it to mean the other (as is routine practice in ISTs) is clearly wrong and would lead to completely misleading results (Cohen, 2004). One can convert from one probability to the other using Bayes' theorem (see below) but even this requires the analyst to know the unconditional probability (such as the 1% in this example) which they would never know in practice.

Cohen (1994, p. 998) illustrates the absurdity of ignoring these facts when interpreting NHST and *p*-values. The logic underlying their interpretation is as follows:

- If H_0 is true, then the data obtained would probably not occur.
  But this result has occurred.
  Therefore, H_0 is probably not true.

The problem with this line of reasoning is that it is formally identical to:

- If a person is an American, then he is probably not a member of Congress.
  But this person is a member of Congress.
  Therefore, he is probably not an American.

There is not space here to repeat the fully worked logic of this example but sceptical readers are encouraged to read it in full, and can also refer to Field's (2013) fully worked explanation using a different example. Some commentators still maintain that the examples given by Cohen and Field are not convincing. They agree that *p*-values can only provide p(D|H) – it is impossible for them to try and argue otherwise – but they argue that this is still useful information. We have not yet, however, seen a convincing argument outlining exactly how they could be useful or, perhaps more importantly, a published research report where *p*-values are interpreted correctly in terms of p(D|H) and then shown to reveal useful information.

While, as shown earlier in relation to Field (2009, 2013), some texts do acknowledge this issue, even then it makes little difference to the general approach in the book as a whole. In many texts, however, the interpretation of *p*-values advocated by the author is simply incorrect (see the examples in Gorard, 2015 b).

3.3. THE 'NEW STATISTICS' SOLVES THESE PROBLEMS

In recent years, much has been made of what has been called the 'new statistics' (see, e.g., Cumming, 2012). Ironically, little of what is presented under this heading is new at all and, as we show in this section, much of it is problematic. However, one theme that appears to run through writers advocating this approach is less reliance on *p*-values. This can only be regarded as a positive development but, unfortunately, the alternatives that are proposed are not much better. Attention to the size of effects is clearly important, and something that was often obscured by the greater attention paid to *p*-values than to the size of any differences or relationships. But conventional 'effect sizes' such as Cohen's d and Hedge's g still have their limitations. However, what is more important here is the move away from the use of *p*-values towards confidence intervals (CIs) and standard errors (SEs).

Advocates of using CIs suggest that they are superior to *p*-values, as they specify a range of possible outcomes rather than relying on the kind of arbitrary 'cut off' used with *p*-values (e.g., $p < 0.05$ or $p < 0.01$). This is not quite true as, just like with *p*-values, the 'acceptable' level for a CI is reported with exactly corresponding arbitrary cut-offs (e.g., 95%, 99%). However, a more fundamental problem with CIs is a logical one, similar to the one that affects NHST and *p*-values.

Before examining this problem in detail, it is worth looking at the definitions of both SEs and CIs. The SE is the standard deviation of a sampling distribution of a very large number of equally sized random samples drawn from the same population. It is important to point out that this figure is hypothetical in almost all research as only one sample is actually selected. The SE is anyway not a particularly intuitive measure, but small SEs are preferred, as they suggest that if we selected lots of identical samples in the future they would have similar characteristics to the one we actually have drawn. The SE can then be used to calculate CIs. However, as Gorard (2014 a, p. 6) shows, what CIs actually mean is often misrepresented in methods texts. For example a 95% CI is routinely mis-described as the range within the true population mean is likely to fall. In fact, its meaning is much more complex and less useful. It is hypothetical, recursive and inverse logical in nature – making it very difficult for readers to comprehend. It means that if a very large number of samples of the same size had been drawn from the population, and each of their CIs computed, then 95% of these CIs are likely to contain the true population mean. The CI an analyst is faced with in practice can be envisaged as one of these CIs.

Some commentators (including the editors of this issue, who are quoted directly below) may argue that:

> If the variance is known, then 95% of the sample means observed will lie in an interval centred on the true population mean (whatever it is) which is the same width as the confidence interval constructed around that sample mean. There is a 1-1 correspondence between the times the 95% CI captures the true value of the population mean and the times that the sample mean lies within 1.96 standard errors of the mean.

This may or may not be true but it is completely irrelevant. It is akin to someone saying in defence of significance tests, if we know the actual probability of H_0 then we can tell exactly how accurate our *p*-value for the null hypothesis is. But it is clear that if we know the true value of H_0 then we would not need to run a significance test to assess its probability. Similarly with CIs, if we know the variance of the sampling distribution then we must know its true mean as well, because the variance is computed by aggregating the deviation of each data item from that mean. Therefore, if we know the true variance there is no point in estimating an estimated range for that true mean. We can simply compute exactly how far any sample mean is from the actual mean (and even that sounds like a completely pointless activity).

CIs (and SEs) may appear at first sight to be less problematic than *p*-values but, as is shown above, they share the same fundamental flaw. All of these techniques rely on making an assumption about the very thing we are trying to find out. The assumption is then incorporated into the calculations that are intended to discover whether that assumption is correct or not (with *p*-values the assumption is that the null hypothesis is necessarily true, and with CIs it is that the sample variance is equal to, or is a good estimate of, the population variance). It is much easier simply to state what we know (such as N, the sample mean and standard deviation, and the level of attrition) and let the reader judge the academic and practical importance of these findings.

4. WHAT SHOULD WE DO INSTEAD?

One of the most common objections to abandoning the use of ISTs is that there is, as yet, nothing to replace them with. The absence of alternatives that achieve exactly the same outcome as ISTs strive for is understandable given the impossibility of the task that ISTs attempt to achieve. This argument, however, is the equivalent of recommending quack medicine to treat an incurable disease simply because an effective treatment has yet to be developed. This objection is clearly invalid. One of the further arguments for eliminating ISTs from normal practice is that while the use of ISTs continues to dominate statistical teaching and practice, it is unlikely that many alternative approaches to judging the representativeness of findings will be developed. However, it is not clear that an exact substitute for ISTs is possible or even required. We present several possibilities for future practise, below. All are feasible, and mostly compatible with each other.

4.1. DO NOTHING INSTEAD

One possibility is to do nothing, in terms of statistical practice. The flaws in the reasoning underlying ISTs demonstrate the impossibility of producing precise estimates of population values (or even ranges) from research based on a single sample. This does not mean that we should stop using large, high-quality random samples or striving for high response rates. Conversely, it implies exactly the opposite. It can be demonstrated that a large random sample will usually produce statistical estimates that are close to the population value. ISTs are often used to invoke a false sense of security in findings when samples are not very large or have other problems with their design or execution. Focusing on producing high-quality data to begin with, rather than on more and more complicated techniques to address inadequacies in data sets, will generally lead to results that are representative even if the representativeness of these findings cannot be summarised in a precise figure. But acknowledging a degree of uncertainty about representativeness – a known unknown – is surely much better than producing a ‘precise’ figure that is flawed.

4.2. TREAT THE DATA AS A LIMITED POPULATION

The use of ISTs with population data – and the accompanying appeal to ‘superpopulations’ or ‘hyperpopulations’ (see White, 2014; Gorard, 2015 a) – is surely an indicator of the ‘cult’ of IST use among researchers. It appears that many researchers cannot interpret data without the ‘crutch’ of ISTs, even when representativeness is not an issue because data on all of the population are available. Even the well-known advocate for statistical literacy Hans Rosling, in his famous Joy of Stats programmes (Rosling, n.d.), uses ISTs on what can only be viewed as either population data or a convenience sample. The rot has really set in when even supposed experts in the field cannot play by the rules of the techniques they are so keen to advocate.

Although this background does not bode well for our second proposal – treating each sample as a limited population – this approach has many advantages. It may encourage larger-scale research and, perhaps ironically, more care when designing samples and thinking about response rates. Data could be analysed without ISTs and the results could be applied relatively unproblematically to the sample at hand. In relation to generalisation, the question that should then be asked is: 'Do we have any reason to believe that these results would or would not apply to the wider population?' The answer to this question would depend on sampling strategies, response rates, the size of any relationships, and so on. This approach is in contrast to the common use of ISTs and associated techniques to try to 'mop up' problems with data; it would instead place emphasis on the generation of high-quality samples in the first place, as the (albeit flawed) safety net of supposedly 'fixing it later' would have been removed.

4.3. DEVELOP ALTERNATIVE TECHNIQUES

Considerable resources are used within the academic community in developing and refining ISTs. These efforts are often devoted to solving problems relating to data that are incompatible with the assumptions underlying these techniques. The majority of articles in methods and statistics journals tend to be of this nature and there is clearly a community of mathematically talented individuals with the skills to solve new problems as they arise. However, given the underlying flaws of ISTs – which cannot be overcome by any amount of mathematical manipulation because the problem is not a mathematical one – this talent pool is not being used as effectively as it could be.

Gorard and Gorard (2016 a) have made some progress in this area with their idea for, and technique for calculating, the number of counterfactual cases needed to 'disturb' a finding. In a response to the publication of this idea, Kuha and Sturgis (2016), and Gorard and Gorard (2016 b) refined this idea further. This example shows that progress can be made with the development of alternative techniques when only four individuals exchange ideas. At the moment this is one of the few examples of an alternative technique to ISTs but, perhaps because ISTs have become the accepted status quo, there is little evidence of the wider methods and statistics community working on such alternatives. Imagine the breakthroughs that could be made if a larger group of experts came together to work on these kinds of approaches.

5. WHAT SHOULD WE TEACH STUDENTS?

Both of the authors of this paper currently devote some of their teaching time to covering ISTs. This is necessary at present because of their widespread use. We make sure our students understand the assumptions required for these techniques, and teach them the correct interpretations of the associated outputs (such as what the *p*-values are actually the probabilities of). We hope that most students do not then have to be discouraged from using ISTs, as the problems with their use should become apparent during the course of their learning. It would be heartening to imagine a future in which we could ignore ISTs altogether because their use had been confined to the history books. Although we spend relatively little time covering these techniques we are aware that they are the principal focus on many statistics modules. Not only does this mean that students spend a great deal of time learning a technique that is both flawed and potentially misleading, but it also restricts the number of other procedures that they can learn in the allotted time. We would advocate a curriculum centred on the ideas of 'sophisticated description' and 'judgement-based analysis'.

Much can be learned from the basic analyses on which ISTs are currently intended to 'piggyback'. Almost all univariate, bivariate and multivariate techniques can be separated from the ISTs that they are often associated with. Those strategies – such as multi-level modelling – that are inseparable from inferential techniques become redundant once the flaws with ISTs and their associated outputs are recognised. Many of the most complex analyses merely represent attempts to ameliorate situations in which data have not met the assumptions underlying particular ISTs.

Distributions can still be summarised, differences measured and relationships between variables discovered without the need for ISTs. Furthermore, concentrating on 'sophisticated description' and ignoring ISTs would allow many more useful techniques to be covered in a similar timeframe. Another, incidental, benefit would be that many of the arbitrary cut-offs associated with ISTs (whether they be particular alpha levels or confidence intervals of certain widths) would be removed, leading to a situation whereby the judgement of analysts is pushed to the fore. While judgement is required at many stages of statistical analysis, reliance on ISTs has resulted in a 'deskilling' of analysis when it comes to the interpretation of results. It takes no skill at all merely to recognise that $p < 0.05$ or to report a 95% confidence interval. Without these 'crutches', the researcher has to think much more carefully about the robustness of their findings and the implications in substantive terms (see Gorard, 2006). A further benefit for researchers and students alike would be a disruption and hopefully a diminution of the 'file drawer' problem resulting from 'significant' results being published disproportionately.

As mentioned above, alternative techniques to ISTs have started to be developed and these can be taught to students instead of ISTs. In addition to Gorard and Gorard's (2016 a) 'number needed to disturb' sensitivity approach, Gorard (2014 b) has also developed a security approach to rating the trustworthiness of research findings. Older methods, such as permutation analysis, that was impractical when first developed by Fisher, is now possible because of increased computing power, and can be used as a superior alternative to ISTs in randomised controlled trials. There are also some very promising aspects of innovations in Bayesian analyses which allow the probability of the data obtained given a true null hypothesis to be converted into the much more useful probability of the null hypothesis being true given the data obtained (Gorard, Roberts, & Taylor, 2004).

6. SUMMARY: CAN WE STOP TEACHING ISTS COMPLETELY?

For the present, students are likely to come across ISTs in journals and other research reports, and so it is important that they are made aware of them. However, the problems with their use should be covered at the same time and material relating to ISTs should occupy the minimum space possible in the curriculum. We both use an exercise involving a bag filled with two different colours of marbles to demonstrate to students why probabilities can only be calculated if an assumption is first made about the contents of the bag. This helps them to understand how making such an assumption is necessarily incompatible with trying to find out whether that assumption is actually true.

Providing students with clear and accurate definitions of *p*-values, SEs and CIs (as the most common outputs of ISTs) should further persuade them that these measures are not useful when interpreted correctly. This knowledge will also provide them with a robust defence if they are challenged by those advocating the use of ISTs. We hope we would see this kind of teaching as a transitory stage, aiming towards a situation where these techniques are no longer taught because they are no longer used. Although we do not feel that this is a likely situation in the very near future, in light of the problems with ISTs outlined above, we do not see it as 'radical', 'purist' or 'iconoclastic'. Once the issues are fully understood, such an approach can only be seen as sensible. In teaching terms, it would also reduce an unnecessary learning load, simplify explanations, and open the use of numeric data to a far wider audience and range of users.

REFERENCES

Berk, R. A. & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In Blomberg, T. G. & Cohen, S. (Eds.), *Punishment and social control* (2nd ed.; pp. 235–254). New York: Aldine de Gruyter.

British Academy (2011). *Society counts: quantitative skills in the social science and humanities*. London: British Academy.

British Academy (2015). *Count us in: quantitative skills for a new generation*. London: British Academy.

Bukodi, E., Goldthorpe, J., Waller, L., & Kuha, J. (2014). The mobility problem in Britain: new findings from the analysis of birth cohort data. *The British Journal of Sociology, 66*(1), 93–117.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*(12), 997–1003.

Connolly, P. (2007). *Quantitative data analysis in education*. New York: Sage.

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

Cumming, G. (2012). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Field, A. (2011). Top 5 statistical faux pas. *Methodspace*. [Online: www.methodspace.com/profiles/blogs/top-5-statistical-fax-pas]

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London: Sage.

Gigerenzer, G. (2003). *Reckoning with risk*, London: Penguin.

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

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

Gorard, S. (2014 a). The widespread abuse of statistics by researchers: What is the problem and what is the ethical way forward? *The Psychology of Education Review*, *38*(1), 3–11.

Gorard, S. (2014 b). A proposal for judging the trustworthiness of research findings. *Radical Statistics, 110*, 47–59. [Online: www.radstats.org.uk/no110/Gorard110.pdf]

Gorard, S. (2015 a). Rethinking "quantitative" methods and the development of new researchers. *Review of Education, 3*(1), 72–96. [Online: doi.org/10.1002/rev3.3041]

Gorard, S. (2015 b). Context and implications document for: Rethinking "quantitative" methods and the development of new researchers. *Review of Education, 3*(1), 97–99. [Online: doi.org/10.1002/rev3.3042]

Gorard, S. (2016). Damaging real lives through obstinacy: re-emphasising why significance testing is wrong. *Sociological Research Online, 21*(1).

Gorard, S. & Gorard, J. (2016a). What to do instead of significance testing? Calculating the 'number of counterfactual cases needed to disturb a finding'. *International Journal of Social Research Methodology, 19*(4), 481–490. [Online: doi.org/10.1080/13645579.2015.1091235]

Gorard, S. & Gorard, J. (2016 b). Explaining the number of counterfactual cases needed to disturb a finding: a reply to Kuha and Sturgis. *International Journal of Social Research Methodology, 19*(4), 497–499. [Online: doi.org/10.1080/13645579.2015.1126494]

Gorard, S., Roberts, K., & Taylor, C. (2004). What kind of creature is a design experiment? *British Educational Research Journal, 30*(4), 575–590.

Howe, C. (2014). A response to Gorard: The widespread abuse of statistics by researchers: What is the problem and what is the ethical way forward? *The Psychology of Education Review, 38*(1), 14–16.

Kuha, J. & Sturgis, P. (2016). Comment on 'What to do instead of significance testing? Calculating the "number of counterfactual cases needed to disturb a finding"' by Stephen Gorard and Jonathan Gorard. *International Journal of Social Research Methodology, 19*(4), 491–495. [Online: doi.org/10.1080/13645579.2015.1126495]

Lester, K. J., Lisk, S. C., Mikita, N., Mitchell, S., Huijding, J., Rinck, M., & Field, A. P. (2015). The effects of verbal information and approach-avoidance training on children's fear-related responses. *Journal of Behaviour Therapy and Experimental Psychiatry, 48*, 40–49.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In Harlow, L. L, Mulaik, S. A., & Steiger, J. H. (Eds.), *What if there were no significance tests?* (pp. 61–106). Brighton: Psychology Press.

Nuffield Foundation (2012). *Promoting a step-change in the quantitative skills of social science undergraduates*. [Online: goo.gl/7cYbVc]

Putwain, D. (2014). A response to Gorard: widespread abuse of statistics by researchers: What is the problem and what is the ethical way forward? *The Psychology of Education Review, 38*(1), 17–19.

Reynolds, G., Field, A. P., & Askew, C. (2014). Effect of vicarious fear learning on children's heart rate responses and attentional bias for novel animals. *Emotion, 14*(5), 995–1006.

Reynolds, G., Field, A. P., & Askew, C. (2015). Preventing the development of observationally learnt fears in children by devaluing the model's negative response. *Journal of Abnormal Child Psychology, 43*(7), 1355–1367.

Rosling, H. (n.d.). *The joy of stats*. Gapminder.

Sheikh, K. & Mattingly, S. (1981). Investigating non-response bias in mail surveys. *Journal of Epidemiology and Community Health, 35*, 293–296.

Siegfried, T. (2010). Odds are, it's wrong. *Science News, 177*(7), 26. [Online: www.sciencenews.org/article/odds-are-its-wrong]

Siegfried, T. (2015). P value ban: small step for a journal, giant leap for science. Science News. [Online: www.sciencenews.org/blog/context/p-value-ban-small-step-journal-giant-leap-science]

Somekh, B. & Lewin, C. (2005). *Research methods in the social sciences*. London: Sage.

Sturgis, P., Brunton-Smith, I., Kuha, J., & Jackson, J. (2013). Ethnic diversity, segregation and the social cohesion of neighbourhoods in London. *Ethnic and Racial Studies, 37*(8), 1286–1309.

White, P. (2014). A response to Gorard: The widespread abuse of statistics by researchers: What is the problem and what is the ethical way forward? *The Psychology of Education Review, 38*(1), 24–28.

Wiles, R., Durrant, G., De Broe, S., & Powell, J. (2009). Methodological approaches at PhD and skills sought for research posts in academia: a mismatch? *International Journal of Social Research Methodology, 12*(3), 257–269.

PATRICK WHITE
School of Media, Communication and Sociology
University of Leicester
University Road, Leicester, LE1 7RH,
UK