# A RESPONSE TO WHITE AND GORARD: AGAINST INFERENTIAL STATISTICS: HOW AND WHY CURRENT STATISTICS TEACHING GETS IT WRONG

JAMES NICHOLSON
*Durham University*
*j.r.nicholson@durham.ac.uk*

JIM RIDGWAY
*Durham University*
*jim.ridgway@durham.ac.uk*

**ABSTRACT**

*White and Gorard make important and relevant criticisms of some of the methods commonly used in social science research, but go further by criticising the logical basis for inferential statistical tests. This paper comments briefly on matters we broadly agree on with them and more fully on matters where we disagree. We agree that too little attention is paid to the assumptions underlying inferential statistical tests, to the design of studies, and that* p*-values are often misinterpreted. We show why we believe their argument concerning the logic of inferential statistical tests is flawed, and how White and Gorard misrepresent the protocols of inferential statistical tests, and make brief suggestions for rebalancing the statistics curriculum.*

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

## 1. INTRODUCTION

White and Gorard (WG) present a number of interlinked arguments. We summarise the arguments in their paper briefly.

- Social scientists are using the wrong methods to understand social phenomena;
  - Too little attention is paid to the assumptions underlying inferential statistical tests (IST) (notably representativeness), so conclusions are often invalid;
  - IST has led to intellectual laziness where a set of mathematical techniques is used as a substitute for thinking about the phenomena being studied;
  - *p*-values are often misinterpreted;
- There are better things to do with curriculum time than teach IST;
- There is a long standing consensus in the statistical community that the logic underpinning IST is wrong;
  - These logical flaws extend to a family of techniques such as *p*-values, confidence intervals (CI), and standard errors;
- Social scientists who use IST are aware of these faults (poor methodology, illogical reasoning, flawed mathematics) but persist because these faults are deeply enculturated, and it is easier to conform to cultural norms (and have their work published) than to kick against the pricks.

*Statistics Education Research Journal, 16(1), 66-73, http://iase-web.org/Publications.php?p=SERJ*

## 2. AREAS OF BROAD AGREEMENT

We fundamentally agree with White and Gorard about their clarion call for a sea change in the ways that social science is conducted. We refer to two further papers that illustrate some of the maladies of social science research, and show that the community is aware of current bad practice. The Open Science Collaboration (2015) – actually Nosek and 269 co-authors – set out to replicate the results of 98 papers published in 3 well-regarded psychology journals. Only 39 out of 100 replications (two studies were replicated twice) were successful. "Low power research designs combined with publication bias favouring positive results together produce a literature with upwardly biased effect sizes'' (p. 6). The instability of *p*-values for *t*-tests on samples of around the size typically reported in psychology journals, drawn repeatedly from two populations where the effect size has been chosen to be typical of psychology experiments is remarkable (see Cumming 2012, or Cumming, n.d.); CIs are more stable. We return to these issues later.

Ioannidis (2005), in his much-cited paper *Why most published research findings are false* offers a critique of methodology in social science, and asserts:

> *"*A research finding is less likely to be true when the studies conducted in the field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytic modes; where there is greater financial and other interest or prejudice; and where more teams are involved in a scientific field."

Statistical inference is often employed to use data from a sample to estimate some parameter in a population, and to test an hypothesis about one or more parameters in a population. Statistical inference is only appropriate when one is confident that the sample is representative of the population (so it is important to avoid sample bias). WG argue that this is usually impossible in social science research. We broadly agree with the observations that WG have to make about the cavalier approach to the use of significant tests and confidence intervals etc., in contexts where the requirements for their use are not met. We also broadly agree with the observations they make about the misstatement of conclusions – there is a widespread sloppiness which encourages fallacious misinterpretation of probabilistic reasoning (including 'bright line' rules around the sanctity of $p < 0.05$). These are extremely important issues and deserve more attention than they currently attract. For us, these are amongst the most important arguments made in the WG paper; far too little of the resource available for social science research is devoted to large scale studies, or replication. There is too little of what WG call 'sophisticated description' and 'judgement based analysis'. Some social scientists act as if the use of a particular methodology – Null Hypothesis Significance Testing (NHST) – will overcome problems associated with poor design; the rationale for NHST is often misunderstood and misused.

We also agree that most textbooks on statistics place too much emphasis on NHST (without specifying the situations where NHST is and is not appropriate), and too little emphasis on issues of design (such as sampling) and how the interpretation of outcomes of NHST should be expressed. *Mind on Statistics* (Utts and Heckard, 2015) is an example of a textbook where key issues in NHST are addressed properly.

We agree that *p*-values are often misinterpreted. The American Statistical Association (ASA) statement on *p*-values (2016) makes the following 6 points:

1. *p*-values can indicate how incompatible the data are with a specified statistical model.
2. *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

## 3. AREAS OF DISAGREEMENT

We now address some points of disagreement with WG, starting with the assertion that there is a general agreement that the logic underpinning IST is flawed. On page 53 WG state "Many commentators now generally agree that ISTs do not work as intended*"*. In contrast to this assertion, the ASA statement on *p*-values is just concerned about improving the practice of NHST and communicating its proper meaning; it does not say ISTs do not work as intended.The ASA will host a symposium on statistical inference entitled *Scientific Method for the 21st Century: A World Beyond p<0.05* in the fall of 2017.

> "We're advancing statistical inference to advance research: Building on the American Statistical Association's statement on *p*-values, this symposium will focus attention on specific approaches for improving practice across three broad sets of activities:
> - Conducting research in the 21st century
> - Using research in the 21st century
> - Sponsoring, disseminating, reproducing, and replicating research in the 21st century" ASA (2017).

Next, we will assert that WG misrepresent NHST protocols, and that abandoning NHST amounts to throwing the baby out with the bathwater. We will:

- set out the logical basis of hypothesis testing and CI;
- set out some statistics fundamentals that relate to NHST
- make some observations about NHST and experimental design related to sample size;
- explain why we think CI and effect sizes are useful.

### 3.1. THE LOGICAL BASIS OF HYPOTHESIS TESTING

We begin by pointing to an inconsistency in WG. WG suggest (p. 59) "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*"*. This is problematic from their point of view; permutation analysis (PA) *is* IST. Consider the simple case of judging whether girls are better than boys on a shoot-em-up computer game. Performance data are gathered from girls and boys. PA works by pooling all the scores and partitioning them into two groups via some method of random allocation. An appropriate statistic (say the difference between the means) is then calculated for the two groups of scores, and is recorded. This is repeated a large number of times. The resulting distribution is then examined: how often does a value as large or larger than the one actually observed occur when the individual scores have been partitioned via a random process? If it occurs about half the time, one is likely to conclude that there is nothing to explain – these girls and boys perform at about the same level. If the observed mean difference in the girls' scores over the boys' scores actually occurred only once in 10 000 trials when scores are allocated to groups via a random process, one is likely to conclude that there is something

to explain. Notice that the PA allows us to draw conclusions only about observed scores – is there something to explain? It tells us nothing about either girls and boys, or shoot-em-up games. Any inference about matters of substance depends on judgements about the quality of the data in the context of the design of the study.

Tools for teaching with and conducting PA are readily available (e.g., Lock, Lock, Lock-Morgan, Lock, & Lock, 2013). Personal experience of teaching social science MA students – who need to read and understand articles in academic journals, but have no prior knowledge of statistics – via PA (starting with hands-on simulations of the technique) shows that this is an effective way to teach about the proper interpretation of '$p < 0.05$'. We note in passing that PA makes no assumptions about distributions, so can be applied to any data sets. PA is useful, but not as useful as conventional NHST based on assumptions of normality, when these assumptions apply, or where samples are large enough for the Central Limit Theorem to apply to the sampling distribution of the mean.

In the following sections, we move from PA to NHST as it is usually applied, and contrast our ideas about the logic of NHST with those of WG.

Informally, the logical stages of NHST in the example we have just given are:

- identify a question of interest (are girls better at shoot-em-up games than are boys?);
- gather relevant evidence via a well-designed study;
- analyse the data – are any differences observed consistent with the behaviour expected or observed with a plausible random mechanism? (either a parametric model or PA). If not, look for another explanation, i.e., the alternative hypothesis.

More formally, a NHST requires both a null and an alternative hypothesis to be stated before any analysis is done (this should be done before data is collected). The null hypothesis ($H_0$) must provide a sampling distribution for the statistic of interest (for example the mean of a distribution), and the alternative hypothesis ($H_1$) must allow you to identify what will constitute the most unusual outcomes if the null hypothesis were true.

We largely agree with the first part of WG (p. 55)

> "This approach [NHST] 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: $p(D \mid H_0)$. To create this probability at all it has to be assumed that the null hypothesis is true."

We disagree with the second sentence – this probability does not require an assumption that $H_0$ is true – it is the probability of obtaining these data, or data less likely than the data actually observed, *if* $H_0$ is true. There is an important distinction here in the logic: it does not create the circular argument WG claim. In addition, we disagree that this is *all* that NHST can tell us – we return to CI later.

Contrast this with the account given by WG (p. 55): *"*As researchers, what we want to know is the probability of the null hypothesis being true (or false) given the data obtained, or $p(H_0 \mid D)$. 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.*"*

Researchers might want to know the probability of the null hypothesis being true [$p(H_0 \mid D)$], but this is unknowable. The ASA statement on the interpretation of *p*-values makes it clear that NHST does not give $p(H_0 \mid D)$ – and does not suggest that this is problematic.

WG argue that the following examples have the same logical structure:

> "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.

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."

We disagree that this analysis sheds any light on the logic of the NHST: the fundamental problem is that one cannot use being a member of Congress as a test of being an American. NHST requires both a null hypothesis which generates a sampling distribution for the statistic of interest, and an alternative hypothesis which identifies what would be the most extreme observations. If the observation falls into that extreme category, then the conclusion would be to reject the null hypothesis and conclude there was evidence to suggest it is not true. However, being a member of Congress is proof absolute that the person *IS* an American, not evidence to suggest that they are not. Further, NHST is designed to facilitate statements about samples (e.g., is it plausible that two samples are drawn from the same population?) not about individuals.

NHST makes no claims about $p(H_0 \mid D)$, and so does not conflate conditional probabilities.

### 3.2. NHST: STATISTICS FUNDAMENTALS

Statistics is about making sense of incomplete information, to support decision making in the face of uncertainty. A key issue is to know something about the quality of the information collected. When dealing with incomplete knowledge, you don't know everything, *but you also don't know nothing* – knowing where you are on that continuum is valuable, and it is an integral part of how hypothesis tests can provide valuable information on which to base decisions. WG (p. 59) say "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."

We assume that this is the example used in Gorard (2014, p. 4): "Of course, the probability of getting seven reds from a bag containing 80 reds is different, a priori, to the probability of getting seven 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."

The example appears to conflate two different statistical tools; one where one has some *a priori* belief about the population ('the factory marble bagging machine should put 80% reds and 20% blues into each bag'), where NHST is appropriate, and the other where the content is unknown, i.e., one is using a sample to estimate proportions of reds and blues in the population, where CI is the appropriate approach.

For the NHST, one needs a null hypothesis, and one can make a judgement about its plausibility, based on evidence. One *can* estimate the composition of the marbles in the bag – this ability to make plausible predictions in the face of uncertainty is part of what makes statistics useful. On the evidence of 7 reds and three blues in a random sample of 10 marbles taken from a bag with 100 marbles in it, it is far less likely that no reds remain in the bag than that 63 reds remain in the bag; one would be foolish to gamble otherwise.

A key issue is to know something about the quality of the information collected. One of the strengths of using the NHST framework (assuming a well-designed study) is that the power function of the test, and effect size, taken alongside the *p*-value, give a good assessment of the likely robustness of the evidence collected – including showing when there is simply not enough data on which to make a decision with any degree of certainty.

### 3.3. NHST AND EXPERIMENTAL DESIGN: SAMPLE SIZE

The effect size is an important component in the design of NHST: the standard deviation of the underlying population effectively determines how easy, or otherwise, it will be to pick up any specified difference in means, and allows researchers to calculate, in the design phase, what sample size should be used.

The importance of this can be illustrated by the following example: the distribution of male adult heights in the UK can be modelled reasonably by a Normal distribution with mean 178 cm and standard deviation 10 cm. A 10% difference in the height of two groups would represent a very large effect size indeed (1.78) and one would expect to be able to identify this difference even on the basis of small samples. In contrast, a 10% difference in the means of two uniform distributions (the initial one on the unit interval) reflects a tiny effect size (0.173) and one would need a huge amount of data (a sample size more than 100 times greater than for the effect size of 1.78) to detect the difference, reliably.

The important lessons from the Open Science Collaboration (2015) and Ioannidis (2005) are that the mechanics of null-hypothesis significance testing (NHST) based on the Normal distribution can be used to explain why so many replications failed (Open Science Collaboration) and why 'most published research findings are false' (Ioannidis). We agree with WG that much social science research is badly done, but argue that some of the problems arise because the mathematical modelling underpinning NHST has been ignored, and not because of flaws in the logic of hypothesis testing *per se*.

The mathematics surrounding NHST provides some principles for research design. Small samples do not provide good evidence on which to make decisions, but how large should a sample be if it is to provide reasonably robust insight? If an effect size is chosen which the researcher feels constitutes an important difference, then the minimum sample size needed in order for the test to have a specified power in detecting that size of difference can be calculated. Note that the calculation of minimum sample size has three inputs – the significance level to be used in the test, the effect size which is felt to be important, and the desired power of the test. The smaller the effect size to be detected (i.e., to return an observed value in the critical region of the test, and hence a significant *p*-value) the larger the sample size needed; the smaller the significance level chosen, the larger the sample size needed; and the greater the power of the test required, the larger the sample size needed (for more details, see Nicholson and McCusker, 2016).

### 3.4. WHY CI AND EFFECT SIZES ARE USEFUL

With a random sample, the sample mean is the best available *point* estimate of the population mean. However, the CI uses the variability of the underlying population and the sample size to construct an *interval* estimate which improves the quality of information available from the estimation process by providing some idea of how precise the estimate is (in a non-technical sense of 'precise').

On page 7, WG reproduce a paragraph we wrote in responding to their expression of interest in submitting to this special issue, this paragraph was written to correct a definition offered by WG:

> "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."

But WG then say that whether our observation is true or not is irrelevant. They argue that to know the variance of a sampling distribution you need to know the population mean. If that were true there would be no need to construct a CI. They conclude the paragraph by saying:

> "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)."

NHST and CI protocols recognise that there is no way to determine the true population mean or variance through taking a sample. However, we can make estimates of both these parameters – and if a large representative sample is used, these can provide useful, bounded, approximations. The sample mean is the best point estimate and the CI gives valuable information about the precision of knowledge available from the sample.

WG (p. 54) rightly point out that "statistical significance is not the same as substantive significance", because of the conflating role of the sample size. Decisions on whether or not to implement an intervention are often made on the basis of anticipated effect sizes and estimates of the costs associated with obtaining those effect sizes, *if* there is evidence to suggest (via a statistically significant outcome with a robust sample size) that the intervention makes a substantive difference, but *p*-values on their own do not measure effect size, i.e., they do not give any insight into whether a significant difference is important.

## 4. CONCLUSIONS

WG make important and relevant criticisms of some of the methods commonly used in social science research, often reflected in social science courses on research methods: too little attention is paid to the assumptions underlying IST, which can invalidate the conclusions drawn, and *p*-values are often misinterpreted. Too much time is devoted in introductory statistics courses to NHST, too little attention is paid to the design of studies, and too little attention to 'sophisticated description' and 'judgement based analysis'. We are in broad agreement with these ideas; our own work focuses heavily on 'sophisticated description' of large scale data.

The important points of disagreement centre on the logic underpinning NHST, and on the usefulness of confidence intervals and effect sizes. WG seem (paradoxically) to accept the usefulness of PA, even though this is IST. For us, the mathematical apparatus surrounding NHST is useful both in explaining the phenomena of low rates of replication reported by the Open Science Foundation (2015), and in designing studies with appropriate statistical power. We would advocate approaches to teaching statistics that introduce IST via PA, and use simulations to demonstrate phenomena such as the instability of *p*-values, and the relationships between effect size to be detected, power, and sample size perhaps using the Open Science Foundation (2015) paper as a stimulus.

We agree with WG that researchers and students should be encouraged to make qualitative judgements about the likely robustness of results, and about the likelihood of replication. Publishers should ensure that textbooks are vetted carefully by both statisticians and social science researchers before publication.

## REFERENCES


ASA (2017). Notice of the ASA symposium on statistical inference. [Online: ww2.amstat.org/meetings/ssi/2017/]

ASA (2016). Statement on statistical significance and *p*-values. *The American Statistician, 70*(2), 131–133. [Online: amstat.tandfonline.com/toc/utas20/70/2?nav=tocList]

Cumming, G. (2012). Understanding the new statistics.New York: Routledge.

Cumming, G. (n.d.). Dance of the *p*-values. *Intro Statistics*. Melbourne: La Trobe University. [Online: www.youtube.com/watch?v=5OL1RqHrZQ8]

Gorard, S. (2014). The widespread abuse of statistics: What is the problem and what is the ethical way forward? *The Psychology of Education Review 38*(1), pages missing.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine* 2(8): e124. [Online: journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124]

Lock, R., Lock, P., Lock-Morgan, P., Lock, E., and Lock, D. (2013). *Statistics: unlocking the power of data.* New Jersey: Wiley.

Nicholson, J. & McCusker, S (2016). Damaging the case for improving social science methodology through misrepresentation: re-asserting confidence in hypothesis testing as a valid scientific process. *Sociological Research Online 21*(2)11. [Online: www.socresonline.org.uk/21/2/11.html]

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science 349*(6251). doi: 10.1126/science.aac4716.

Utts, J, and Heckard, R. (2015). *Mind on statistics*. 5th ed. Stamford, CT: Cengage Learning.

Wasserstein, R. L. & Lazar, N. A. (2016). The ASA's statement on *p*-values: context, process, and purpose. *The American Statistician, 70*(2), 129–131. [Online: amstat.tandfonline.com/toc/utas20/70/2?nav=tocList]