Significance testing is still wrong, and damages real lives: a brief reply to Spreckelsen and van der Horst, and Nicholson and McCusker

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

Abstract

This paper is a brief reply to two responses to a paper I published previously in this journal. In that first paper I presented a summary of part of the long-standing literature critical of the use of significance testing in real-life research, and reported again on how significance testing is abused, leading to invalid and therefore potentially damaging research outcomes. I illustrated and explained the inverse logic error that is routinely used in significance testing, and argued that all of this should now cease. Although clearly disagreeing with me, neither of the responses to my paper addressed these issues head on. One focussed mainly on arguing with things I had not said (such as that there are no other problems in social science). The other tried to argue either that the inverse logic error is not prevalent, or that there is some other unspecified way of presenting the results of significance testing that does not involve this error. This reply paper summarises my original points, deals with each response paper in turn, and then turns to an examination of how the responders use significance testing in practice in their own studies. All of them use significance testing exactly as I described in the original paper – with non-random cases, and using the probability of the observed data erroneously as though it were the probability of the hypothesis assumed in order to calculate the probability of the observed data. It really is time for this absurd practice to cease.

Introduction and reprise

I recently published a paper summarising the evidence why we should no longer use, or accept the use of, significance testing in social science research (Gorard 2016). I used as support the writings of experts in the field such as Berk, Cohen, Freedman, Glass, Jeffreys, Meehl, Rozeboom and Tukey. In fact, I presented a summary of a copious and long-established literature that is rarely cited in methods resources – which has been critical of significance testing since its inception. It is intriguing that methods resources often portray method choices as schismic and disputed, sometimes to an absurd extent. But they hardly ever raise the disputed validity of significance tests (as opposed to rejecting all use of number in an unjustified way). Therefore, I deliberately sought to remedy this defect and show that experts have long been opposed to the use of significance tests, and that learned bodies and journals are increasingly spurning them or at least offering cautions about them. I also reminded readers of the damage created by publication bias, non-replicable results, vanishing breakthroughs and the like – all of which are associated with significance testing. In a very real sense, the argument against significance testing is no longer technical. The abuses and the inverse logic error are real, and it is simply a matter of waiting for more people to realise this. The argument now is largely an ethical one. How sure are defenders of the inverse logic error (and even defenders of 'abusing' significance tests) that they are right, and how willing are they for science and people in real-life to suffer for the sake of it?

The literature I cited, from a much larger pool, clearly shows two things. Significance tests are widely abused in the literature, and even if they were to be used as intended they cannot provide a useful result for analysis (without then abusing the result).

Significance tests *are* widely abused in the literature (and some examples are provided in the last section of this paper). Their computation is based on an assumed hypothesis, usually the nil-null hypothesis of no difference between groups (or no correlation, trend, or pattern). Given this assumption, and a complete set of randomised cases, and accurate measures (or counts) for each case, it is possible to assess how unlikely any observed difference between groups is. This is the p-value computed in significance tests. It is the probability of observing randomised data as extreme as actually observed in one study, given that the assumed hypothesis is true (or pD|H). However, if the measures are not accurate, or any data or cases are missing, or if the cases had not been randomised in the first place, then a significance test cannot work as intended. A computer or calculator will still provide an answer even where data is incomplete or otherwise non-random, but the algorithm will assume complete randomisation. Therefore, the answer will be incorrect, even in its own terms.

> Of course the number will be right only if all the assumptions made by the test were true. Note that the assumptions include the proviso that subjects were assigned randomly to one or other of the two groups that are being compared. This assumption alone means that significance tests are invalid in a large proportion of cases in which they are used (Colquoun 2014, p.3).

In a sense, it would be better if analytical software asked the user a few key questions about their datasets before running a significance test, and then 'refused' to do so unless the data were confirmed to be complete, accurate and randomised. In all examples of real social science that I have ever seen, that would mean that no one would be able to use significance tests, without lying explicitly to the software about the nature of their data. In a very real sense, all real-life users of significance tests, who know what a p-value is, are being similarly dishonest already – to the software, to others, and presumably even to themselves. That is part of the reason why my 2016 paper explained the whole thing as an ethical (integrity) issue rather than a technical one. Many people must be regularly flouting the basic assumptions of significance testing despite knowing better (and the examples in the last section of this paper are unremarkably illustrative of this pattern).

Incidentally, it might be helpful for those wondering what to do 'instead' of significance tests to continue this thought experiment. Imagine that analytical software such as SPSS had a binary switch, with perhaps a default setting that we are *not* using complete randomised data. The user would still be faced with almost all of the software that is useful. The only menu item affected would be for analysis, and even here almost every sub-item would remain. The user could still perform cluster and factor analyses, every kind of regression modelling available, correlation, and comparison of means, for example. In every method, 'effect' sizes would be used instead of p-values, such as in retaining or abandoning predictors in a regression model. SPSS might also add some useful basic stuff like a range of other 'effect' sizes. The biggest noticeable difference would be in the output. There would be no p-values or asterisks, so the output would be shorter, and easier to read and explain. But the substantive results would remain, and would simply have to be judged differently (Gorard 2006, 2015).

Significance tests anyway do *not* provide a useful answer for an analyst wanting to know how unlikely their assumed hypothesis is, and so by implication how plausible their alternative explanation is. This is so, even in the rare event that the analyst has complete and accurate data from fully randomised cases. They want to know pH|D (or even just pH) and this is a completely different probability to the one that significance tests provide (pD|H). As soon as the analyst tries to use pD|H to assess the probability of their assumed hypothesis they are committing the inverse logical error. The logic of *Modus Tollendo Tollens* does not work with probabilities. This is illustrated in my 2016 paper in a number of ways, but the key point is not mine. It has been made by many including Jeffreys (1937), and others long before.

However, all of this prior expertise is rejected by some commentators, perhaps including those whose main claim to expertise lies in teaching and using significance tests, or who want to reassure themselves and others about the damage caused by using an invalid approach, and so excuse their prior publishing record. Turkeys do not vote for Christmas, as the saying goes.

My 2016 paper has generated two critical responses in this journal - Spreckelsen and van der Horst (2016), and Nicholson and McCusker (2016). I deal with each briefly in turn.

Spreckelsen and van der Horst

Spreckelsen and van der Horst (2016) provide a reasonable summary of my argument in the 2016 paper. But in subsequent sections of their 'response' they try to cast doubt on the argument, and they do so largely by arguing against things I did not say in the paper and have not said elsewhere either. In fact, if they wanted to look they would see that almost everything they try to use to query my argument is something I agree with, and have written copiously about elsewhere and for a long time.

The title of their response asks 'Is banning significance testing the best way to improve applied social science research?' – to which the answer would have to be no. My paper does not say that this is the best or only way to improve the quality of social science. In fact, my paper suggests and refers to a number of ways in which handling and reporting numbers can be improved, although it also suggests that these more useful and valid approaches could be encouraged by ceasing to teach, conduct and report the misleading and confusing significance tests. For nearly two decades I have argued in favour of multiple methods approaches (e.g. Gorard with Taylor 2004), about the value of research design, and seeing each study as part of a larger research cycle (e.g. Gorard 2013), and the importance of clear reporting of things like attrition, data quality and threats to validity (e.g. Gorard 2015). Problems can occur at many stages in a research project, and this includes reviewing the literature, where I have proposed simply ignoring any p-values and using the effect sizes, if they have been reported or can be estimated (e.g. Gorard 2014). Reviews of literature reporting 'effect' sizes do show that positive results based on small samples are over-represented, suggesting some publication bias even in the absence of significance tests, as I have reported elsewhere (e.g. Gorard et al. 2016, Gorard et al. 2017).

None of this is an argument for retaining significance testing in the way that Spreckelsen and van der Horst (2016) want to imply. For example, they say of the proposed ban on significance testing:

> since such a ban only aims at one aspect of the research process, we argue that it is too simplistic…. (p.7)

This is a non-sequitur. If the technique of significance testing were valid then they could simply show that it is. What might go wrong in other phases of a study is then not relevant to that key point. But we should not retain a bad method simply because there are other problems elsewhere or with other methods. That is akin to a medic refusing to treat a disease because it is not the only disease in the world. Similarly, Spreckelsen and van der Horst (2016) claim that having more tools available for analysis is better than having fewer (p.3). And again this is logically irrelevant. If significance tests worked then how many other tools there are does not matter. But as they do not work we should throw them away, even if there were no other tools. What Spreckelsen and van der Horst (2016) seem to be saying is that more tools are better even if those tools do not work – which is clearly not true. They also seem to be defending the use of significance tests with non-randomised cases here.

They cite Field (4th edition, 2013) approvingly as a widely used text for new researchers (para 4.3). But they seem to be unaware that since the first edition of this book, Field has realised that significance testing does not work and is dangerously misleading. Teaching of statistics needs to be radically altered.

> the standard approach in teaching, of stressing the formal definition of a p-value while warning against its misinterpretation, has simply been an abysmal failure (Selke et al. 2001, p.71)

Using null hypothesis significance testing is now described by Field (2011) as the "number one statistical *faux pas"*, and he goes on to say:

> No-one understands what a p-value is, not even research professors or people teaching statistics (Haller & Kraus 2002)

And he sums up that:

> a p-value is the probability of something given that something [else] that is never true is true, which of course it isn't, which means that you can't really get anything useful from a p-value other than a publication in a journal

This means that Field is one of those who has not been 'obstinate'. Instead, he has looked at the arguments and evidence, and had the courage to change his mind publicly since 2010. He even admits to previously making the inverse logic error:

> Many textbooks (including the first edition of my own Discovering Statistics Using SPSS) give completely incorrect, but commonly reproduced, explanations of what a CI [confidence interval] means.

In fact CIs are just p-values written in a different format (Colquoun 2014, p.12) with almost all of the same problems and a few added ones (such as their incomprehensibility).

I hope that Spreckelsen and van der Horst (2016) will now do something similar to Field. It is still possible for them to get published, write methods resources, and even teach their Q Step courses successfully without referring to significance tests.

Interestingly, in their response to my paper, Spreckelsen and van der Horst (2016) refer to themselves directly or indirectly as 'applied social science researchers' at least 10 times (the paper is very repetitive in several respects). I am not sure why they do this or what it might mean, over and above just being an actual researcher, but I am sure that I am also what they would term an applied researcher. Of the 1,000 or more books, chapters, papers and other pieces I have published, as an academic, the vast majority report substantive social science research, and none of them uses significance testing. This is relevant because it shows that whatever issues Spreckelsen and van der Horst might face in conducting or publishing research I will probably have met and overcome them before. It is certainly not necessary to use invalid procedures to get published or cited, or to have their work used in policy or practice.

Spreckelsen and van der Horst (2016) repeatedly cite another more technical response to my original paper - Nicholson and McCusker (2016) – in a way that suggests they agree with it but without saying why.

Nicholson and McCusker

A lot of the paper by Nicholson and McCusker (2016) is simply irrelevant – consisting of basic traditional statistics text book stuff, red herrings about insurance premiums, and merely restating their faith in significance testing. They suggest some errors in my paper that simply do not exist. For example, they state in para 4.7:

> which he erroneously refers to as 5.5% before later in the paragraph referring correctly to approximately 12%

They are suggesting that there is a typographical error or worse in my paper. Of course, none of us are immune to typos, but that sentence does not appear anywhere in my paper. The number 5.5 appears once only as a paragraph heading inserted by the editors, but not with a '%', and nowhere else in the text (as any electronic search will attest). They accuse me of creating a 'straw man' but I do not misquote and then argue with things that have not been said. I take care when disagreeing with a statement in the research of others to actually read their paper and quote directly and accurately (as I have advised in numerous methods publications for 20 years).

They get to near the end of p.7 (in a 10 page 'response' paper) before then discussing a small part of my paper in some detail. They largely ignore the substantive points about abuses, my demonstration about conditional probabilities, the discussion of the sociology and psychology of the obstinacy, almost all of the literature and expertise cited, and the damage caused by significance tests in practice. If they do not like the demonstrations or examples in my paper (or as presented interactively in staff and public seminars to which they were invited but which they did not attend), then they could look at the simulations and analyses making the same substantive points in Hunter (1997), Trafimow and Rice (2009), Colquoun (2014), and others, as cited in my original paper.

The key to understanding such simulations is the realisation that in real-life not all interventions are effective, not all variables are correlated, and so on. In many studies, the null hypothesis is not just *assumed* to be true for the purposes of computing a significance

test, it will actually be true for all practical purposes. For example, the IES (US) and EEF (UK) have funded a large number of educational interventions evaluated via randomised controlled trials. Only around 12% have been deemed to be effective – and this may be an over-estimate because some of these will be false positives wherein the results occurred by chance. Using 12% of feasible funded interventions as a base figure it is possible to estimate how accurate a significance test using a 5% threshold (p-value) would be. If the results of 1,000 such interventions were assessed by use of significance testing with an 80% chance of detecting a true positive (the standard accepted rate, often termed 'power'), then only 12% (120 or fewer) would actually work, of which only 80% or 96 of them would produce a significant result on average (the other 24 would be erroneously treated as ineffective). Of the remaining 880 genuinely ineffective treatments, 5% or 44 of them would also produce 'significant' results just by chance, meaning that over 31% (44/140) of the positive results would be false ones, and the findings of these studies would be misleading or worse. The accuracy of a significance test is nothing like 95% (Colquoun 2016). Where the real-life likelihood of an effect is lower than 12% then the accuracy of significance tests will be even worse than 69%. Add in the false positives among the 12% deemed effective in practice, the 24 false negatives among the 120, the common practice of p-hacking and using many tests on the same data, and widespread publication bias, and it is clear that most results published on the basis of 'significant' results would be wrong even if they met the most basic assumptions of the test. But again, in reality, most real-life uses of significance tests are based on non-randomised cases or incomplete data anyway. Almost none of the published 'significance' results will be valid.

The main argumentative point made by Nicholson and McCusker (2016) is that the inverse logic error, described by Cohen (1994), Colquoun (2014) and many others, is irrelevant when significance tests are used properly. In fact they say that I, and presumably therefore Cohen and all of those others, are setting it up as a 'straw man' (para 7.4). They agree that significance tests generate the probability of the data observed given the assumed hypothesis (pD|H), and in their view it is this probability that can be used correctly in analysis. In their view, people do not confuse it with the probability of the assumed hypothesis being true given the data observed (pH|D). They agree with me, Cohen and all of those others that pD|H is not the same pH|D, and that the one can be large and the other small etc. for the same data. The two are completely different probabilities. And a p-value is a form of pD|H. In this we are in agreement. However, pD|H is of no use in analysis unless it is used to assess pH|D in some way. The standard and traditional way is to simply conflate the two probabilities so that if pD|H is small this is used somehow to reject the assumed hypothesis, exactly as though it were really pH|D. Altering the phrasing from 'rejecting a hypothesis' to 'assessing the fit' of the assumed model does nothing to change this basic error. I have never seen an analysis that presented and explained pD|H correctly and made no attempt to use it to assess the likelihood of the assumed hypothesis (H) somehow. This is the key point. Once the inverse logical error is understood, significance testing becomes meaningless in real-life. I hope that Nicholson and McCusker will soon come to this realisation, as Field did (above).

At the end, Nicholson and McCusker seem to agree with me (para 7.5) that there is widespread abuse of significance testing, but if they genuinely agree with this then there is little or no use for significance tests in real-life research where the necessary conditions are usually not met in practice. An important way of discovering what they really believe about all of this is to examine what do they, and Spreckelsen and van der Horst, do in practice in their own research.

Do any of them really care?

At time of writing, Nicholson has no publications listed on his staff webpage, is not registered on Google Scholar, and has no empirical pieces listed in Research Gate. It is therefore not possible to say how he uses or abuses significance tests in practice. His response to my paper is based on writing about something he has never done, or at least not publicly reported (a real-life, empirical, social science project involving numbers).

McCusker has only one empirical piece involving numbers on his staff webpage (Botturi et al. 2012). This is clearly based on a sample that is only 47% complete (p.361), meaning that even if it had been randomised initially, and this is not stated, it can no longer be considered random in any way. Yet McCusker goes on to use significance tests in the standard style, exactly as described in my original paper, using the p-values so generated to decide on the 'significance' or otherwise of differences between groups (pp.365 and 374). Nicholson and McCusker (2016) say they are aware of this problem. In fact, they say "We agree wholeheartedly that testing is often abused….but we believe the problem is that testing is used in contexts it should not be'. This means that Nicholson and McCusker either do not follow their own 'wholehearted' advice (the context of 53% missing cases is clearly somewhere that significance tests should *not* be used). Or they think that using significance tests with 53% of cases missing is somehow justified statistically – i.e. they are plain wrong.

Spreckelsen has a few empirical pieces on his staff webpage, of which Montgomery et al. (2014) is used as an example here. This describes a trial in which cases were randomised to two groups, but not all cases had complete data and so fewer cases were analysed than randomised initially. Thus, it is not clear that the cases analysed were random any longer. Despite this, in Table 2 and elsewhere, the paper goes on to use significance tests in the standard way, confusing pD|H with pH|D in the usual way. But more absurdly, in Table 1 Spreckelsen presents data from the cases after randomisation but before the intervention, and uses significance tests to decide if the post-randomisation differences between groups are 'significant' or not. This is clearly nonsense. If the cases are truly randomised then any differences *must* have occurred by chance, and so no differences can be what they term 'significant' (i.e. judged not to have arisen by chance). All differences must have arisen by chance because that is what random means. If, on the other hand, the cases had *not* been randomised then Spreckelsen cannot even begin to justify using a significance test.

There are fewer empirical pieces for van der Horst on her staff webpage, of which van der Horst et al. (2016) is used here as an example. This study analyses 13,986 observations taken from 2,331 respondents, from a secondary data source that originally had 62,053 observations from 16,806 respondents. It is not clear whether the cases were originally randomised, but it *is* clear that with this level of attrition and case selection the resulting sample is nothing like random. But, again, the author goes on to conduct and report the p-values from significance tests, in Table 8 and 9 for example (here used as part of regression modelling).

None of these comments suggest that the three empirical pieces are worthless. In fact, the latter two are pretty good. In all three the substantive findings survive simply removing the references to significance testing and p-values, and relying on the data and 'effect' sizes reported instead. These findings are then easier to judge in terms of their trustworthiness (Gorard 2015), shorn of the irrelevant, misleading, and therefore potentially dangerous significance tests. But my simple summary of their prior work does show that those

attempting to defend significance tests are also those who abuse them, while often decrying exactly the same kind of abuse in methods articles, as a kind of meaningless litany or incantation. In the last five years I have conducted a large-number of systematic and scoping reviews covering perhaps 60,000 research reports, over and above my usual reading and reviewing (e.g. Gorard et al. 2011, Gorard and See 2013a, 2013b, See and Gorard 2015a, 2015b, Gorard et al. 2016). In real-life research, I have *never* seen a significance test used in a correct context (meeting the basic assumptions) and with the results described correctly. As my original paper said, it is time for everyone to accept this and move on. There *are* lots of other things to improve as well.

References

Botturi, L., Bramani, C. and McCusker, S. (2012) Boys are like Girls: Insights in the Gender Digital Divide in Higher Education in Switzerland and Europe, *Journal of Universal Computer Science*, 16, 3, 353-376

Cohen, J. (1994) 'The Earth is Round (p<.05)', *American Psychologist*, 49, 12, 997-1003

Colquoun, D. (2014) An investigation of the false discovery rate and the misinterpretation of p-values, Royal Society Open Science, http://rsos.royalsocietypublishing.org/content/1/3/140216

Colquoun, D. (2016) The problem with p-values, Aeon, https://aeon.co/essays/it-s-time-for-science-to-abandon-the-term-statistically-significant

Field, A. (2011) Top 5 statistical faux pas, http://www.methodspace.com/top-5-statistical-fax-pas/

Field, A. (2013) Discovering Statistics Using IBM SPSS Statistics: 4th edition, London: Sage

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

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

Gorard, S. (2014) An easy way to ignore significance testing, https://www.dur.ac.uk/resources/education/research/Findingswithandwithoutsigtests.pdf

Gorard, S. (2015) A proposal for judging the trustworthiness of research findings, ResearchED January 2015, http://www.workingoutwhatworks.com/en-GB/Magazine/2015/1/Trustworthiness_of_research

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

Gorard, S. and See BH (2013a) *Do parental involvement interventions increase attainment? A review of the evidence*, London, The Nuffield Foundation, http://www.nuffieldfoundation.org/sites/default/files/files/Do_parental_involvement_interventions_increase_attainment1.pdf

Gorard, S. and See, BH. (2013b) *Overcoming disadvantage in education*, London: Routledge

Gorard, S., See, BH and Davies, P. (2011) *Do attitudes and aspirations matter in education?: A review of the research evidence*, Saarbrucken: Lambert Academic Publishing

Gorard, S., See, BH and Morris, R. (2016) *Review of effective teaching approaches in primary schools: Main report of findings*, London; DfE

Gorard, S., See, BH and Siddiqui, N. (2017) *The trials of evidence-based education*, London: Routledge

Gorard, S., with Taylor, C. (2004) *Combining methods in educational and social research*, London: Open University Press

Haller and Krause (2002) Misinterpretations of significance: a problem students share with their teachers?, MPR-Online, 7, 1, 1-20

Hunter, J. (1997) Needed: A Ban on the Significance Test, *Psychological Science*, 8, 1, 3-7

Jeffreys, H. (1937) *Theory of probability*, Oxford: Oxford University Press

Montgomery, P., Burton, J.., Sewell, R., Spreckelsen, T. and Richardson, A. (2014) Fatty acids and sleep in UK children: subjective and pilot objective sleep results from the DOLAB study – a randomized controlled trial, *Journal of Sleep Research*, doi: 10.1111/jsr.1213

See, BH and Gorard, S. (2015a) Does intervening to enhance parental involvement in education lead to better academic results for children? An extended review, *Journal of Children's Services*, 10, 3, 252-264

See, BH and Gorard, S. (2015b) The role of parents in young people's education - a causal study, *Oxford Review of Education*, 41, 3, 346-366

Selke, T., Bayarri, M. and Berger, J. (2001) Calibration of p values for testing precise null hypotheses, The American Statistician, 55, 1, 62-71

Spreckelsen and van der Horst (2016) Is banning significance testing the best way to improve applied social science research?, *Sociological Research On-line*, 21, 3

Trafimow, D. and Rice, S. (2009) A test of the null hypothesis significance testing procedure correlation argument, *The Journal of General Psychology*, 136, 3, 261-269

van der Horst, M. *et al.* (2016) Pathways of paid work, care provision, and volunteering in later careers: Activity substitution or extension?, *Work, Aging and Retirement*, http://dx.doi.org/10.1093/workar/waw028