Is banning significance testing the best way to improve applied social science research? – Questions on Gorard (2016)

Thees F Spreckelsen & Mariska van der Horst
Univeristy of Oxford; Univeristy of Kent

Abstract

Significance testing is widely used in social science research. It has long been criticised on statistical grounds and problems in the research practice. This paper is an applied researchers' response to Gorard's (2016) 'Damaging real lives through obstinacy: re-emphasising why significance testing is wrong' in Sociological Research Online 21(1). He participates in this debate concluding from the issues raised that the use and teaching of significance testing should cease immediately. In that, he goes beyond a mere ban of significance testing, but claims that researchers still doing this are being unethical. We argue that his attack on applied scientists is unlikely to improve social science research and we believe he does not sufficiently prove his claims. In particular we are concerned that with a narrow focus on statistical significance, Gorard misses alternative, if not more important, explanations for the often-lamented problems in social science research. Instead, we argue that it is important to take into account the full research process, not just the step of data analysis, to get a better idea of the best evidence regarding a hypothesis.

Keywords: Statistical significance, Transparency, Replication, Ban, Hypothesis testing, Controversy.
Word count (incl. endnotes/references): 6920

Introduction

In this paper we discuss the conclusions Stephen Gorard (2016) draws in his recent contribution to this journal in short: "significance tests just do not work, even when used as intended in statistical analyses, and […] their widespread use should cease immediately" (Gorard 2016: 1.1).[i] We maintain that how arguments in the paper are put forward is unlikely to persuade applied researchers to change how they work. In fact, the way the arguments are presented may make researchers defensive of their work and therefore might make them *less* likely to change their practices. The goal of this paper is not to get into a statistical discussion of whether there is value in the p-value (Nicholson and McCusker (2016) assess Gorard's paper on that dimension), but to evaluate this discussion – and the tone in which this discussion is held – from the point of view of applied social science researchers such as ourselves. We examine how Gorard's conclusions and recommendation are supposed to follow from his arguments, and contrast his paper with other efforts to improve social science research.

Our article starts out by revisiting the original article's arguments, followed by a critical reflection of these arguments, and alternatives to Gorard's conclusions. We also evaluate what this discussion means for teaching and reviewing of publications. This way we hope to provide an applied researchers' perspective on an important debate which Gorard's article claims is settled.

Damaging Real Lives (DRL) – revisited

Gorard argues in his article that *even when used correctly,* significance testing based on p-values do not provide the sought after information. In his words: "The purpose of this paper is to remind readers that significance tests just do not work, even when used as intended in statistical analyses, and to argue that their widespread use should cease immediately" (Gorard 2016: 1.1). The arguments are based on three general critiques: (1) a critique of the principles of significance testing based on p-values, (2) a critique of how researchers use significance testing, and (3) a critique of the wider culture of significance (in reference to the "P-value culture" (Gorard 2016: 2.3)) in all academic domains that perpetuates the use of significance testing. Gorard uses these points to conclude that the use of significance testing is ethically wrong. We will briefly summarize these arguments.

A large part of the arguments against significance testing in the work of Gorard is based on the often-stated "inverse probability"-fallacy (Gorard 2016: 3.7). This refers to the incorrect use of the conditional probability that the data is true given the null-hypothesis (D|H0) to make inferences about whether the null-hypothesis is true given the data (H0|D). One cannot reverse these probabilities as nicely illustrated by the example: "The probability that a hanged person will be dead will be very high, but the probability that any dead person had been hanged would be very low" (ibid.). However, according to Gorard, researchers are often interested in whether the null-hypothesis is true, and are thus not testing the question they are actually interested in, but still often interpret it as if it does give information about whether the null-hypothesis is true. Moreover, by testing null-hypotheses, researchers ignore any prior knowledge to the topic.

The second point made throughout the text relates to problems with the logic of the p-value. Specifically, Gorard maintains that there is a problem with the assumption behind the p-value, namely there "must be no bias in the study design, and no measurement error, non-response or sample dropout" (Gorard 2016: 2.1). When there is not a completely random sample, for example due to missing data or because population data is used, Gorard states that the p-value is meaningless as no standard error can be calculated in these cases (ibid.: 3.1). Only when we are talking about a completely random sample and all assumptions behind the p-value are met, the significance test can be interpreted as the likelihood of finding a difference/relationship/pattern at least as strong as observed in the data. However, since the assumptions behind the p-value are so unrealistic in real-life according to Gorard, the calculated p-value is usually meaningless. Moreover, and related, he points out that to test no difference or

relationship or pattern at all (nil-null hypothesis) in the data does not seem to be useful, as usually one would not expect it to be completely zero in the population.

Gorard adds a set of four examples so that readers could see these problems in practice. Thus Gorard makes his points using both logical and practical arguments in order to convince the reader. He encourages the reader to replicate, something that Nicholson and McCusker (2016) did and critique (and hence, we will not discuss these examples in this paper).

Next to arguments against significance testing, Gorard also reviews past arguments as to why research practices have not changed. These practices constitute a "culture of significance", in which publications hinge on "statistical significance" and significance testing is given the status of a "religious ritual" (Gorard 2016: 6.5). Gorard argues that this culture of significance ensures that individuals keep using significance testing while no-one should.

Crucially Gorard goes further than criticising what significance testing is, how it is used, and how it is seemingly embodied in research, reviewing and teaching. He argues that as a direct result of the flaws in the logic and use of significance tests, an unethical situation has arisen, referring to the argument as "The ethics of the situation". According to Gorard (2016) the issue on the merits of significance testing is settled and consequently it is ethically wrong to still perform or teach significance testing: "The paper ends by arguing that this is no longer a technical or scientific issue but chiefly an ethical one." (ibid.: 1.1). According to him, research that uses significance testing is not only poor value for money (ibid.: 7.1) but also stands in the way of genuine scientific progress. Based on these "practical and ethical" arguments Gorard concludes:

> *"anyone using significance tests, allowing them to pass peer-review for publication in their journals, teaching them to new researchers, or otherwise advocating them in any way, is part of a (hopefully) diminishing group causing untold real-life damage. Where they previously did so through ignorance, they should now cease. But anyone who continues with any of these actions despite reading the material in this paper (and others) is causing that damage deliberately."* (Gorard 2016: 8.4)

In short, Gorard asserts that the debate for or against significance testing is settled and has problematic, even harmful, consequences. Therefore it is unethical to continue using and teaching it.

Questioning DRL's conclusions

As applied researchers (not statisticians) we are not commenting on the accuracy of the statements – which are widely debated and do not seem as clear-cut as the author describes them. For readers interested in this, please read Nicholson and McCusker (2016). Instead, we argue that the conclusions and subsequent recommendations do not follow from the arguments presented. Rather than going through the arguments of Gorard one-by-one, we assess the

consequences of these arguments. We follow the same general line as above: (1) in response to the critique of the principles of significance testing based on p-values we look at whether p-values are really the cause of the damage, (2) in response to the critique of how researchers use significance testing we look at the relative merit of p-values compared to alternatives, and (3) in response to the critique of the wider culture of significance, we discuss the tone in which the discussion is held.

Are p-values the cause of damage?

Gorard's argument against the scientific merit of p-values rests in parts on the "inverse probability fallacy" inherent to *null-hypothesis* significance testing, but his key arguments relate to the misuses of p-values (for example, section 3.1). Whilst we acknowledge several of the pitfalls summarized in and exemplified by the article based on such misuses of p-values, we disagree both with the reasoning for and the corollary of the conclusion.

The reasons against the p-value mentioned in the article (i.e. flawed logic and assumption of significance testing, observable misuse, and culture of significance; described in more detail in the previous section Damaging Real Lives (DRL)- revisited) do not provide the necessary link between the method used and damage done. Rather, it gives examples of research that failed to stand up to scrutiny or the test of time. However, this is and – we would argue – should be a normal part of the whole scientific endeavour. If anything this should be strengthened through post-publication peer-review and easy access to replication information (also see Andrew Gelman, 2013). Even though post-publication peer-review may still not happen as often as we would like, journals like *SAGE Open* explicitly welcome this by stating in their Submission Guidelines that "Readers and the academic community at large will then have the power to continue the peer review process after online publication" (SAGE Open 2016).

Similarly as a corollary of his conclusions, Gorard seems to imply that damage will be done inevitably. Since significance testing does have appropriate interpretations and does provide information (see Wasserstein and Lazar 2016 below), this appears to us overstated. This holds even if significance testing should only be used in the context of other analyses and (like all research) on basis of well-designed data collection efforts. Nicholson and McCusker (2016) describe the rationale of significance testing in detail. By contrast the claim that does follow from Gorard's arguments and conclusions is that researchers using, reviewing, and teaching significance testing, should use it appropriately and be aware that there might be better alternatives.

The relative merit of significance testing

Following from the conclusions drawn, Gorard argues that there is the logically and ethically compelling need to abandon significance testing. As he acknowledges the recommendation that significance testing should be abandoned has been made repeatedly, albeit not always as strongly formulated as Gorard does (see e.g. Hunter 1997; Halsey et al. 2015). Whilst we agree that there is misuse of p-values and that there is much to be improved on in both

research and teaching, we do not – at the moment - agree with Gorard's demand that p-values should be abandoned altogether. Instead, it seems much more appropriate to follow Gerd Gigerenzer's (2004) recommendation that researchers should use and be taught the whole "toolbox of statistics". In this toolbox significance testing does have a place, but should be taught amongst other research strategies, such a stronger focus on thorough descriptive statistics, effect sizes, and Bayesian inferences. These different methods together will then inform us on the best evidence regarding a hypothesis. Our focus then would be on how to improve teaching and use of significance testing, rather than abandoning it. Gorard dismisses this approach outright. As applied researchers it seems immediately sensible to have more, rather than less analytical tools available.

Gorard's arguments seem to be even more problematic in that it lacks the *comparison with alternatives*. Even though Gorard rightfully states that a lack of a good alternative is "irrelevant to whether significance tests work or not" (Gorard, 2016: 6.6), it is important for applied researchers to know how they compare, even if only to the alternative of not using significance testing at all. Gorard rightly states there are many alternatives, some of which he himself proposed (Gorard and Gorard 2016), and this may be the reason he did not list these alternatives here. However, how should readers of his article be convinced that the alternatives to significance testing are indeed better when it is unclear what their *relative merit* is? Simply, we would need evidence that systematically better research is/would be produced when it is done in different ways. Then we would find it more convincing that abandoning e.g. p-values would lead to better research.

Such comparison is particularly important, since the damaging outcomes mentioned by Gorard may also exist when p-values are not used. Let us look at just one of the issues with p-values: publication biases (see e.g. Egger et al. 1997 on the impact of publication bias on accumulating information). Head et al. (2015) claim in their study on misuses of p-values that "many of the problems with publication bias reoccur with other approaches, such as reporting effect sizes and their confidence intervals or Bayesian credible intervals. Publication biases are not a problem with p-values per se. They simply reflect the incentives to report strong (i.e., significant) effects." (p.2). This suggests to us that we need to be wary and critical of any research, and that a simple solution as abandoning p-values may actually be no solution at all or as Savalei et al.'s (2015) asks: "Is the call to abandon p-values the red herring of the replication crisis?"

More general, any method will have its problems, potential biases or give wrong results due to oversight or inappropriate usage of the method. Thus, in order to demonstrate the *relative merit* it should be demonstrated that alternative methods to significance testing do in fact produce superior results. An example for what we have in mind would be Howard et al. (2000) who juxtapose and compare conclusions of analyses using significance testing, a (Frequentist) meta-analysis of studies and a number of Bayesian specifications. Such studies are necessary to assert more convincingly the relative merit of a method.

To summarise our arguments thus far; for us it neither follows that significance testing is the cause of the damages mentioned by Gorard, nor can we assess the relative merit of alternatives. We would be better able to judge and, we believe, more likely to change our practices (if necessary), if we were convinced that significance testing is the source of the problems and per definition better research would be done using different approaches or significance testing.

Counterproductive language

We mostly see the article itself as an *intervention to improve scientific practice*. The main goal of this paper does not seem to be to provide arguments in a debate on the usefulness of significance testing, hoping to persuade researchers to Gorard's side of the argument, but rather to intervene in and change current research practices. Similarly, Trafimov et al. (2015) explicitly justified their ban of significance testing in the *Journal of Basic and Applied Psychology* as an attempt to improve research: "We hope and anticipate that banning the NHSTP will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking"(ibid. p.2). And indeed, to change research practices strong interventions may be required. However, to us, it seems unlikely that this paper or the demand to ban significance testing are convincing given the extreme and uncompromising terms in which they are presented.

To start with, Gorard alleges throughout the article an *"obstinacy"* on part of researchers, described as the unwillingness to change. This is so central to his argument, that it is part of the title "Damaging Real Lives Trough Obstinacy". For example, he describes "the problem of the obstinacy of significance users *cannot* be logical or mathematical either. Significance testing derives from a psychological flaw. 'It does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!' (Cohen 1994). Schmidt (1996) considers it an addiction to false belief." (Gorard 1996: 6.4). We wonder whether this is really the case and what would be a good indicator for obstinacy – other than disagreement with and uncertainty about the objections. An argument against obstinacy is that the research practice *is* changing, for example by an increasing focus on effect sizes and dealing with limitations of the data rather than 'just' reporting p-values or mere statistical significance (e.g. American Psychological Association. 2010).

Stating further that individuals are "*deliberately*" causing damage is a very strong judgement, even if it is only targeted at the reader supposedly enlightened by the present article. This comes in addition to claiming a causal link between "significance" methods and poor research, a link we claim to seem problematic (see above). These judgments make us wary about the demands put on us as readers. We do not appreciate the way the argument does not want to convince but demands us to be convinced. This demand demotivates the interest in learning and understanding, and takes away from the energy needed to be better and more critical users of the statistical tool kit. As social researchers who use statistical methods to find answers to our substantive questions, we want to

learn from developments in statistics and increase our statistical understanding. However, this article seems to question our integrity in trying to do this. To us this counters the intention of improving research and teaching.

Questioning the integrity of applied researchers based on their method use is especially problematic given that for each paper that says that there is no use of significance testing (as is the argument of Gorard's paper), there appears to exist another paper defending significance testing (e.g. Murtaugh 2014), with some researchers agreeing (e.g. de Valpine 2014; Cumming 2014) and others not (e.g. Burnham and Anderson 2014; Morey et al. 2014). Thus, for us applied researchers it is not clear-cut how we should stand on this. We can see value in arguments from both sides. Therefore, we end up with a somewhat more practical question, namely the extent to which it substantially matters. To what degree do the different philosophies behind the approaches lead to different bodies of literature? As applied researchers, that is what we are most interested in and this is the evidence we are missing most at the moment. Gorard's (2016) does not seem to provide us with clarity on the substantive importance, but rather presents us with the conclusion that the debate is settled and that any uncertainty is due to obstinacy grounded in some psychological need for specific answers (see e.g. Sect 6.4).

If we are not convinced by a paper, that demands change in our research practices, is there an alternative? An example for the way the critique could have been more fruitfully expressed and been a motivation to adapt, innovate, or adopt, is Ioannidis' by now classic article "Why Most Published Research Findings Are False" (2005). The overall claim is equally explicit, however it follows from the arguments presented and provides substantive amounts of suggestions for innovation, rather than stating that these are sufficiently documented elsewhere. Similarly, Greenland et al. (2016) provide a clear guide of common misinterpretations and how to avoid them, rather than dismissing p-values all together and questioning the ethics of applied researchers. It provides, in our view, a nice overview on what you can and cannot say with p-value based tests and seems to be written with the intention of helping the field rather than condemning it. This is not a statement about whether the arguments are correct, but rather that it matters how arguments are presented, and to give examples of papers that have the potential to change and (hopefully) improve our work.

In short, we are committed to improving how we work, but the issues do not appear to us as clear-cut as Gorard (2016) would have us believe. There are alternative forms of critical interventions with helpful suggestions, as for example made by Greenland et al. (2016) or Simmons et al. (2011). These do not attack our scientific integrity, and they are as a result more likely to be convincing to us.

Towards improving social science research

Fundamentally, we agree with the author on the need to improve research practices to avoid the misuses and poor research outcomes such as ones he

mentions. So far we disagreed with the identification of significance testing as the cause of the problem and the likely effectiveness of Gorard's article in achieving more robust and innovative research as well as teaching practices. This begs the question wherein we would see the source of poor research and what would be effective interventions?

The answer for use lies with a holistic view on the research process rather than 'just' focussing on one element of it (whether someone reports a p-value or uses other aspects of significance testing). It is well known from what is now called "meta-research" literature that bias in estimated associations or effects is closely linked to study designs, their implementation and their reporting (Moheret al. 1995, Schulz et al. 1995)[REFs]). Related, Nicolson and McCosker's (2016) earlier response highlights the importance of explicitly formulating and stating alternative hypotheses for significance testing.

We will briefly describe a way of conceptualizing the common quantitative research process (similar to those in widely used social science research methods textbooks e.g. Bryman (2012) or Frequentist introductions to statistics e.g. Field (2013)). Readers should note that these deductive steps are also frequently followed in qualitative studies.

A research process for quantitative data-analysis

The quantitative research process can be simplified into six steps most readers will be familiar with:

1) Reviewing existing literature
2) Formulating or stating a hypothesis and planning research
3) Collection or selection of data
4) Description and analysis of the data
5) Robustness checks
6) Reporting and documenting

In our eyes, poor research can originate at any of these six stages and thus in our view, the focus should not just be on the analysis of data, but rather on the full research process. For example, taking the critique against misrepresenting the p-value seriously also means that we should be more critical when reviewing previous literature, inspecting the entire context and research (as far as possible) and critically interpreting the presented statistics ourselves and the conclusions as offered in the paper (cf. Greenland et al. 2016). The step of formulating and stating hypotheses is not only generally crucial, but also central for the appropriate use of statistical methods and their unbiased reporting, together with pre-planning of data-collection (or –selection) and analyses (Wagenmakers et al. 2012). Moreover, we believe in the importance of robustness checks and also reporting these robustness checks. This enables the reader to assess how much findings are dependent on specific method choices. Examples for robustness checks relate to model-assumptions (such as unobserved heterogeneity (Karlson 2011)) and increasingly part of new statistical procedures (e.g. sensitivity analyses for mediation (Imai et al 2010)) It is of course important to report all robustness checks that the researcher has done; not only the ones in favour of the result presented in the paper. This would

give us information for the best evidence regarding a hypothesis. Finally, whilst the ordering of the research steps is no panacea, it is nonetheless a potential source of bias, e.g. if researchers return to their data-collection *after* an initial analysis. Correspondingly, trial registrations or analysis plan pre-publication are increasingly becoming important (Monogan 2013, also see for a critical appraisal, Scott et al 2015).

By contrast, Gorard's critique does not address the role significance testing plays in all of the steps of the research process. He focuses on one step only; the analysis of the data. Nevertheless, he uses problems from other steps, such as data-collection, to outright dismiss the validity of the analytical logic (Gorard 2016: 3.1). His critique seems strangely detached from the other dimensions of the research process.

In our opinion, taking the critique seriously for the whole research process will improve scientific practice. By increasingly reporting the whole research process, we think we would get a better understanding of the results presented in the limited space of a published article and we perceive this more fruitful than banning significance testing. Our readers might wonder whether this is the case, but we would maintain that fuller reporting would provide an alternative against a ban, and would allow assessment of any effects as well as a sense of the relative merit of various methods, something we claimed earlier to be necessary.

Readers may say that it is not possible to report more fully as there is a limit to what can be placed in a journal article. However, it is increasingly possible to provide web appendices for published articles, and this would allow for an increase in the information available to interested readers, not only to assess the best evidence there is for (or against) a certain hypothesis, but also on the process by which it was generated. For example, the journal *Work, Employment, and Society,* explicitly asks for this in their Author's Guide: "authors are strongly encouraged to include additional results and analysis in an appendix submitted alongside the article" (British Sociological Association 2014). Even if a journal does not accept a web appendix, researchers may think about publishing a web appendix on their own website or stable repositories (see e.g. Open Science Framework: <www.osf.io>).

Using the debate on significance testing to innovate the research process

It is easy to argue that broadening the debate to include the whole research process might be too ambitious and just another version of the "lazy" argument that "everything is more complicated". Therefore we will now briefly look at two alternative interventions to improve research that have been proposed previously. Moreover we also point to the need to look at the whole research process again. The first is specifically concerned with significance testing but embeds it in the wider research process, the second acknowledges explicitly the need to change the whole research process.

Alternative 1 – Improving on p-value use.

The first alternative is the recent statement by the American Statistical Association in response to the most recent wave of discussion about significance testing and specifically the use of p-values (Wasserstein and Lazar 2016). The statement aims to clarify the correct use of p-values, which as far as we understand challenges the outright dismissal of significance testing argue that "Significance tests just do not work - even when used as their advocates intended" (Gorard 2016: 8.3). As such it seems to us a better intervention to improve scientific practice. The five central points of the statement are listed below.

> "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." (Wasserstein and Lazar 2016)

The third and fourth point apply essentially to any quantitative research and thus provide the broader focus on the overall research process. Conclusions and decisions should correspond to all aspects of the research process, its complete and open documentation, and ultimately should be the result of replications, meta-analysis (but see Egger et al (1997) on bias in meta-analyses) or a more general summary of evidence.

By contrast the other points highlight the analytical need for several measures; hark back to the need of any statistical analysis to employ a variety of 'tools' to arrive at a conclusion.

Alternative 2 - Improving replicability and synthesis through better reporting

We criticized Gorard's article due to the lack of assessing *relative merit* and *being a poor intervention to improve research practice.* The second particularly powerful alternative way of achieving improved research practice is the requirement for transparent reporting, judged against the standard of replicability and the ability of knowledge accumulation to better understand the best evidence for or against a hypothesis, *irrespective of analysis method used.*

Key examples for this are journals that require the deposit of data and syntax (e.g., American Economic Association 2016), as well as requiring the use of reporting guidelines (such as CONSORT (Moher, Schulz, and Altman 2001) or

STROBE (von Elm et al. 2007)), that cover all aspects of a study's research design.

These requirements for standardized reporting provide a more fruitful social process of changing current research practice. Moreover they can be judged against the standard of replicability. Genuine transparency rather than a different analysis method enables knowledge accumulation, the lack of which Gorard laments and blames on significance testing. This is a point Gorard himself is well-aware given his work on research synthesis methods such as systematic reviews and meta-analyses (e.g. Gorard and Taylor 2004).

We have no direct way of comparing Gorard's proposed ban of significance testing and the two alternative interventions mentioned here in their effectiveness. However, if we assess to which degree "damage" can be reduced through the degree by which transparency and replicability are achieved, the relative merit of these two interventions should be clear. Transparency also would ensure that mistakes and wrong interpretations of significance testing could be spotted, corrected, and results be compared to alternative not significance-based analyses.

Our objection to an outright ban of significance testing is thus, we believe, not based on obstinacy. Rather, since such ban only aims at one aspect of the research process, we argue it is too simplistic and does not contribute to being more critical researchers and increasing replicability and transparency. We therefore do not believe it ultimately will bring about the positive changes Gorard appears to want to achieve.[ii,iii] By contrast, we believe that by clarifying how to use significance testing and by demanding detailed documentation of its application (or a more general detailed and standardized reporting of the whole research process) the research community are likely to reduce "damage" through explicitly increasing transparency and replicability and nudging researchers towards being more critical on their own work.

Reviewing and Teaching

Gorard's critique goes beyond significance testing in research practices, but also aims at reviewing and teaching practices in relation to significance testing – "*allowing them to pass peer-review for publication in their journals, teaching them to new researchers*" (Gorard 2016: 8.4). Two brief responses to this.

We already expressed how a ban of significance testing from journals is not likely to result in the wished for improvement since it ignores the wider research process and how journal's requirements for transparency and documentation would be more effective. Consequently, reviewers, even those not well-versed in quantitative methods, should judge any manuscript based on reporting guidelines with the aim to improve transparency and replicability. They should require it in their responses, and journals should focus on this, rather than on a ban of significance testing. Moreover, they should pay special attention on whether p-values are correctly interpreted and that only claims are made that are warranted by the analyses done.

Teaching should continue to introduce and provide training in significance testing techniques, as part of a broad statistical training to equip students with a large analytical toolbox. However, we also want our students to be critical of what they are doing and rather than separating these techniques from both substantive courses and training focus on research design, there is a need for integration.[iv] Learning to use a variety of analytical tools entails learning about the whole research process they are embedded in and crucially how to report them. The challenge for us (and others who teach quantitative methods) is teaching the methods in a way that is understandable and that students (feel they) can apply as well as at the same time not losing important nuances in what you can and cannot say with the methods.

Teaching quantitative methods also provides opportunities to directly improve replicability. And a small but powerful change in our eyes could be a widespread use of replication assignments, for example as master thesis (e.g. Janz 2016). That way replication gets a more important role in academia and is embedded in a new generation of academics, and ultimately raises the bar for the quality of reported research.

Concluding remarks

Gorard (2016) provides a very vocal critique of current statistical practices, with pointed statements about for example research programs in epidemiology. Throughout our comment, we highlighted the need to see the discussion as an impetus for research innovation and improvement. Gorard makes it clear that this is his aim, too, albeit through removing a tool from the researcher's toolbox and without providing evidence as to the relative merit of alternatives of abandoning significance testing:

> *"The intention behind the proposed ban (above) would be to force researchers to consider and report a much wider range of issues – such as the possible importance and methodological soundness of any findings (…). What is needed is some idea of the scale of any difference (or pattern or trend), the methodological limitations underlying it, and a judgement about its substantive importance (and perhaps also the cost-effectiveness of accepting the finding's implications or not). What is needed is an 'effect' size evaluation"* (Gorard 2016: 3.4).

In our opinion, this is the more important message of this article and these points, rather than banning significance testing, should generally be the focus of how to improve statistical research practices. And indeed, we believe that the field is already moving in that direction. The alternatives mentioned above address the whole research process and can be explicitly assessed for their relative merit in improving transparency and replicability. Increasingly, journals provide reporting guidelines, systematic reviews are done with explicit tests for publication biases (which consists of more than just an issue of statistical significance!), sometimes study and analysis preregistration is required, and we can see a start of a replication culture (Bohannon 2015; Camerer et al. 2016). These are important developments and seem more powerful than a ban of a single analytical approach.

The again invigorated debate of significance testing, particularly in the social sciences, means that analyses will have to be more robust and better reported. We welcome this development. It will provide challenges for researchers, reviewers, and teachers of quantitative methods alike and it is unlikely that the discussion on how to improve scientific practice will be solved soon.

To conclude, we believe in expanding our statistical toolbox as well as learning how to use each tool appropriately. We believe that it is through a broader understanding of statistical methods that we can improve our research, reviewing, and teaching; not by completely disregarding one tool. Moreover, the improvements in the research process should not be limited to the analytical strategy chosen, but should encompass the full research process, including reviewing existing literature, hypothesis formulation, data-collection or -selection, robustness checks, and reporting and replication.

Acknowledgments

We like to thank Steven Roberts for the invitation to contribute to Sociological Research Online and Heejung Chung for the encouragement to write this discussion piece.

Conflicts of Interest

The authors declare that the research was undertaken in absence of any commercial or financial relationships that could result in conflicts of interest. TFS is currently employed as postdoctoral research fellow funded by the Department of Social Policy and Intervention, University of Oxford. MvdH is currently employed as a post-doctoral research associate at the School of Social Policy, Sociology and Social Research, University of Kent. She currently works on various projects; none of which the authors believe would result in a conflict of interest.

Both authors have been using and teaching significance testing and worked or are working on projects that involve significance testing.

References

AMERICAN ECONOMIC ASSOCIATION. 2016. "The American Economic Review: Data Availability Policy." The American Economic Review. www.aeaweb.org/aer/data.php.

AMERICAN PSYCHOLOGICAL ASSOCIATION. 2010. Publication Manual of the American Psychological Association. American Psychological Association. 6th Edition.

BOHANNON, John 2015. "REPRODUCIBILITY. Many Psychology Papers Fail Replication Test." Science Vol. 349, No. 6251, p. 910–11.

BRITISH SOCIOLOGICAL ASSOCIATION. 2014. "Work, Employment and Society - Guidance on Main Research Articles." http://wes.sagepub.com/site/includefiles/WES_Research_Article_Guidance_2014.pdf.

BRYMAN, Alan 2012. Social Research Methods. 4th ed. Oxford, New York: Oxford University Press.

BURNHAM Ken P, and Anderson David R 2014. "P Values Are Only an Index to Evidence: 20th‐ vs. 21st‐century Statistical Science." Ecology Vol. 95, No. 3, p. 627‐30.

CAMERER, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." Science (New York, N.Y.) Vol. 351, No. 6280, p. 1433–36.

CUMMING, Geoff. 2014. "The New Statistics: Why and How." Psychological Science 25, No. 1, p. 7–29.

DE VALPINE, Perry. 2014. "The Common Sense of P Values." Ecology Vol. 95, No. 3, p. 617–21.

EGGER, Matthias, George Davey Smith, Martin Schneider, Christoph Minder, 1997 Bias in meta-analysis detected by a simple, graphical test, British Medical Journal,Vol 315, p.629.

FIELD, Andy. 2013. Discovering Statistics Using IBM SPSS Statistics. 4th ed. London: Sage.

GELMAN, Andrew. 2013. "Ethics and Statistics: It's Too Hard to Publish Criticisms and Obtain Data for Republication." Chance Vol. 26, No. 3, p. 49–52.

GORARD, Stephen. 2016. "Damaging Real Lives Through Obstinacy." Sociological Research Online Vol. 21, No. 1, p. 2. http://www.socresonline.org.uk/21/1/2.html.

GORARD, Stephen, and Jonathan Gorard. 2016. "What to Do instead of Significance Testing? Calculating the 'number of Counterfactual Cases Needed to Disturb a Finding.'" International Journal of Social Research Methodology Vol. 19, No. 4, p. 481–90.

GORARD, Stephen, and Chris Taylor. 2004. Combining Methods in Educational

and Social Research. Maidenhead, New York: Open University Press.

GREENLAND, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. "Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations." European Journal of Epidemiology Vol. 31, No. 4, p. 337–50.

HALSEY, Lewis G, Douglas Curran-Everett, Sarah L Vowler, and Gordon B Drummond. 2015. "The Fickle P Value Generates Irreproducible Results." Nature Methods Vol. 12, No. 3, p. 179–85.

HEAD, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, Michael D. Jennions, DM Barch, T Yarkoni, et al. 2015. "The Extent and Consequences of P-Hacking in Science." PLOS Biology Vol. 13, No. 3, e1002106.

HOWARD, George S., Scott E. Maxwell, and Kevin J. Fleming. 2000. "The Proof of the Pudding: An Illustration of the Relative Strengths of Null Hypothesis, Meta-Analysis, and Bayesian Analysis." Psychological Methods 5, No. 3, p. 315–32.

HUNTER, John E. 1997. "Needed: A Ban on the Significance Test." Psychological Science Vol. 8, No. 1, p. 3–7.

IMAI, Kosuke, Luke Keele, Dustin Tingley 2010 A General Approach to Causal Mediation Analysis, Psychological Methods, Vol. 15 , No. 4, p. 309–334.

IOANNIDIS, John P. A. 2005. "Why Most Published Research Findings Are False." PLoS Medicine Vol. 2, No. 8, e124.

KARLSON, Kristian Bernt 2011 Multiple paths in educational transitions: A multinomial transition model with unobserved heterogeneity, Research in Social Stratification and Mobility, Vol. 29, p. 323-341.

JANZ, Nicole. 2016. "Bringing the Gold Standard into the Classroom: Replication in University Teaching." International Studies Perspectives. Online: doi: http://dx.doi.org/10.1111/insp.12104.

MOHER, David, Kenneth F Schulz, and Douglas G Altman. 2001. "The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomised Trials." The Lancet Vol. 357, No. 9263, p. 1191–94.

MOHER David, Pham Ba, Jones Alison, Cook Deborah J, Jadad Alejandro R, Moher Michael, Tugwell Peter, Klassen Terry P. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? The Lancet, Vol. 352, No. 9128, p. 609-13.

Monogan, James E. 2013 A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections, Political Analysis, Vol. 21, No. 1, p. 21-37.

MOREY, Richard D, Jeffrey N Rouder, Josine Verhagen, and Eric-Jan Wagenmakers. 2014. "Why Hypothesis Tests Are Essential for Psychological Science: A Comment on Cumming (2014)." Psychological Science Vol. 25, No. 6, p. 1289–90.

MURTAUGH, Paul A. 2014. “In Defense of P Values.” Ecology Vol. 95, No. 3, p. 611–17.

NICHOLSON, James, and Sean McCusker. 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 Vol. 21, No. 2, p. 11. doi:10.5153/sro.3985.

SAGE open. 2016. “Submission Guidelines.” https://uk.sagepub.com/en-gb/eur/journal/sage-open#submission-guidelines.

SAVALEI, Victoria, and Elizabeth Dunn. 2015. “Is the Call to Abandon P-Values the Red Herring of the Replicability Crisis?” Frontiers in Psychology Vol. 6, p. 245.

SCHMIDT, Frank L. 1996. “Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers.” Psychological Methods Vol. 1, No. 2, p. 15–29.

SCOTT Amelia, Julia J Rucklidge, Roger T Mulder , 2015 Is Mandatory Prospective Trial Registration Working to Prevent Publication of Unregistered Trials and Selective Outcome Reporting? An Observational Study of Five Psychiatry Journals That Mandate Prospective Clinical Trial Registration, PLoS one, Vol. 10, No. 8, e133718, doi:10.1371/journal.pone.0133718.

SIMMONS, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” Psychological Science Vol. 22, No. 11, p. 1359–66.

Schulz Kenneth F, Chalmers Iain, Hayes Richard J, Altman Douglas G. 1995 Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials, JAMA Vol. 273, No.5, p. 408-12.

TRAFIMOW, David, and Michael Marks. 2015. “Editorial.” Basic and Applied Social Psychology Vol. 37, No.1, p. 1–2.

VON ELM, Erik, Douglas G Altman, Matthias Egger, Stuart J Pocock, Peter C Gøtzsche, and Jan P Vandenbroucke. 2007. “The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies.” Preventive Medicine Vol. 45, No. 4, p. 247–51.

WAGENMAKERS, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L J van der Maas, and Rogier A Kievit. 2012. “An Agenda for Purely Confirmatory Research.” Perspectives on Psychological Science : A Journal of the Association for Psychological Science Vol. 7, No. 6, p. 632–38.

WASSERSTEIN, Ronald L, and Nicole A Lazar. 2016. “The ASA ’ S Statement on P-Values : Context , Process , and Purpose.” The American Statistician 1305 (April). doi:10.1080/00031305.2016.1154108.

Endnotes

[i] References to the original article are made with regard to sections in the online document, rather than pages.

[ii] It would be in our eyes an interesting test to see whether in general the quality of research has improved in journals banning significance testing. We would be skeptic that it has as it is only one and probably a less important aspect of the overall research process.

[iii] In that we also mirror Savalei and Dunns' (2015) critique of the Cumming's (2014) proposition to ban significance testing.

[iv] The step-change in quantitative methods teaching (Q-Step) initiative at fifteen UK universities, for example does this through embedding methods teaching in substantive course, e.g. www.kent.ac.uk/qstep/integration.html