Do we really need Confidence Intervals in the new statistics?

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

Abstract

This paper compares the use of confidence intervals (CIs) and a sensitivity analysis called the number needed to disturb (NNTD), in the analysis of research findings expressed as 'effect' sizes. Using 1,000 simulations of randomised trials with up to 1,000 cases in each, the paper shows that both approaches are very similar in outcomes, and each one is highly predictable from the other. CIs are supposed to be a measure of likelihood or uncertainty in the results, showing a range of possible effect sizes that could have been produced by random sampling variation alone. NNTD is supposed to be a measure of the robustness of the effect size to any variation, including that produced by missing data. Given that they are largely equivalent and interchangeable under the conditions tested here, the paper suggests that both are really measures of robustness. It concludes that NNTD is to be preferred because it requires many fewer assumptions, is more tolerant of missing data, is easier to explain, and directly addresses the key question of whether the underlying effect size is zero or not.

Introduction

This paper considers how to represent some of the key strengths or weakness of a research result, especially one expressed as an 'effect' size. Effect sizes can be of many kinds including odds ratios for cross-tabulated categorical variables, and correlation coefficients such as R^2 for real numbers related linearly, and all can be encompassed by the argument within this paper. However, for simplicity, the discussion here focuses on one type of effect size – the difference between two means divided by their overall standard deviation. Such an 'effect' size could be used in a cross-sectional study to illustrate the difference between the incomes of those based in urban and rural locations, or the examination results of two ethnic groups. The difference would not actually be an 'effect', because area of residence may not *cause* income level, for example. Effect sizes could be called standardised differences, and some kind of scaling is necessary (Gorard 1999). But they are generally called 'effect' sizes and so that is the term used in this paper. The term might be more appropriate when effect sizes are used to present the results of an experimental design.

In a typical experiment there will be two groups of roughly equal size to which cases will have been randomised – one for the intervention or 'treatment' and one for the control or 'business as normal' treatment. There will be a pre-specified measurable outcome, and the post-intervention mean score for that outcome will be calculated for each group separately, along with the overall standard deviation of the score. One way of presenting the results is to find the difference between the two means and divide the result by the overall standard deviation. If the resultant 'effect' size is at or near zero then the treatment is deemed to have had no effect relative to the control, or there is no difference between the groups in a comparative study. If there is a substantial effect size (whether positive or negative), then one plausible explanation is that the treatment has had a differential effect, or there is a noticeable difference between two groups in a comparison. The key judgement for analysts is then whether the effect size is large enough to assume that the true effect of the intervention was not zero.

This paper considers in turn the use of significance tests, judgement of research quality, confidence intervals (CI), and a measure of sensitivity called the number of counterfactuals needed to disturb the finding (NNTD). All of these have been used by analysts to help judge how much to trust a result and whether an effect size should really be treated as zero or not. The paper presents the results of simulated trials comparing the CI and NNTD for 1,000 large comparisons between two groups, and considers the implication of these results for the meaning of CIs and the use of NNTD.

Problems with p-values

In many fields, the results of numeric studies including experiments have long been assessed in terms of their supposed probabilistic uncertainty expressed as a p-value from a test of significance – such as a t-test, ANOVA, chi-squared or similar. Given a set of cases randomly allocated to the two group(s), no measurement error in the outcome, and no missing data or missing cases, it is possible to calculate the probability of finding an 'effect' size at least as large as that obtained in the study, as long as the underlying or true difference between the two groups is zero. This probability is the p-value computed by any form of significance test.

These p-values have two major problems for analysis. First, the conditions for computing them are never or hardly ever met in real-life. They only work with complete randomisation of cases, and so cannot be used with population data, convenience samples, or any samples with missing cases, or with cases missing data (Freedman 2004, Filho et al. 2013, Glass 2014). As Colquoun (2014, p.3) states about p-values from significance tests:

> 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.

Therefore, neither significance tests nor confidence intervals (see below) can be used in the situation of comparing income by area of residence or attainment by ethnic group. Such groups are naturally occurring, and have not been randomised. There is no probabilistic uncertainty about the differences between them.

Second, it is not clear what use can be made of the probability computed. Analysts want to know whether the underlying or true difference between the two groups in an experiment or other comparison is zero, or not, given the size of the difference they measured. The p-value is the probability of obtaining the difference they measured given that the true value is zero. These two are very different values, and one can be small and the other very large (Colquoun 2016). Without further information, that is never available in usual practice, it is not possible to convert one probability (of the effect size given that the real difference is zero) into the other (of the difference being zero given the effect size). The tests cannot provide the answer analysts really want (Falk and Greenbaum 1995), but the desire to include some quantitative measure of credibility has prompted an illegitimate use of statistical significance as an inadequate and misleading surrogate for such a measure (Matthews 2001, Pharoah et al. 2017, White and Gorard 2017).

There are many other problems with significance tests, including data fiddling, publication bias, false positives, misuse of power analyses to deny non-significant results, *post hoc* dredging, and the multiple use of tests designed for one-off use (Halsey et al. 2015). The p-values do not provide an easy picture of the scale of any finding, or its substantive importance. None of these observations are new. The fact that significance tests do not work as used has been clear for 100 years in the social sciences and beyond (Boring 1919, Berkson 1938, Rozeboom 1960, Bakan 1966, Meehl 1967, Morrison and Henkel 1970, Cox 1977, Carver 1978, Berger and Sellke 1987, Loftus 1991, Daniel 1998, Tryon 1998, Nickerson 2000, Gorard 2006, Hubbard and Meyer 2013). Significance tests have no useful role in day-to-day research, are misleading, and should not be published any further (Walster and Cleary 1970, Guttman 1985, Hunter 1997, Nix and Barnette 1998, Selke et al. 2001, Fidler et al. 2004, Starbuck 2016).

Judging trustworthiness

Moving on from the flawed approach of significance testing should encourage researchers to consider and report a much wider range of issues – such as the possible importance and methodological soundness of their findings (Kline 2004). One possible generic approach is proposed in Gorard et al.

(2017). Assuming full and honest reporting, we can judge the quality of each research report and therefore the trustworthiness of its findings based on its design, scale, attrition, outcome measurements, appropriateness, fidelity and validity. The design should fit the research question(s). A cross-sectional design with two or more naturally occurring groups could be used to address a comparative question, and an experimental or equivalent design with randomisation of cases to treatment groups would be suitable for a causal question. Ideally, the number of cases in each treatment group should be large (at least in hundreds). No cases should refuse to take part, be missing or drop out, and no data should be missing from any of the existing cases. All missing data from a planned study is a source of bias in the results, because there will be a reason why the data is missing (it is not a random phenomenon). The outcomes should be clearly measured, pre-specified where possible to prevent dredging and, in a trial, standardised rather than tied to the intervention being experimented on. There should be no diffusion of the intervention between treatment groups, and no vested interests for the researchers. These and other ideal factors are arranged into a 'sieve' in Gorard (2014a), to aid the judgement process by comparing the achieved study with this ideal template.

The 'new' statistics

The 'new' statistics has emerged as an alternative to using significance tests. It uses the scale of the findings, often based on standardised 'effect' sizes (see above). Effect sizes (ES) can give a good idea of the size of any difference, pattern, trend, or correlation (Coe 2002). They do not rely on the same unlikely assumptions as significance tests – so can be used legitimately with non-random cases, where substantial data is missing, and so on. In themselves, ES offer no direct assessment of a finding arising by chance, they need care to be used with appropriate forms of data, give no indication of the scale or quality of the study that led to the finding, and are open to publication bias just like significance tests (Chowl and Ekholm 2018). However, used in conjunction with consideration of the trustworthiness of research (as above), they can be a useful way of presenting individual study results. In order to provide readers with an idea of the likelihood of any finding arising by chance, advocates of the new statistics have suggested also presenting confidence intervals (CIs) for effect sizes (Cumming 2013).

CIs

Perhaps the biggest single drawback of effect sizes is that they contain no indication of the scale of the study the led to them – a crucial factor in judging the trustworthiness of any research finding. The effect sizes discussed in this paper are computed from the difference between two means and their overall standard deviation. Of course they should always be presented with the number of cases on which they are based (N), but CIs combine the elements of effect sizes with N more directly. For an effect size from a large sample, its confidence interval is computed as the effect size plus or minus 1.96 times the overall standard deviation (of the mean of all cases) divided by the square root of N. This is for a 95% CI, since 95% of the area of the normal distributions lies within 1.96 standard deviations of its mean. Other approaches are possible with CIs specifically for effect sizes (see Appendix).

CIs are intended to be an estimate of probabilistic uncertainty, and so are closely related to significance tests, and mainly have the same underlying assumptions. The wider the CI the more uncertain the finding is interpreted as. Given full randomisation, no underlying difference between the two groups, no missing data, and no measurement error (as with significance tests), the 95% has an odd meaning (the same mathematical assumptions as with significance tests). It suggests that if many other samples of the same size from the same population were to be created and their means and CIs calculated, then around 95% of the CIs of these other samples would include the mean of the sampling distribution (which can be construed as an estimate of the population mean). This is a imaginary recursive definition because it uses a very large number of CIs to define any one CI. It is also reverse logic since it does not say that the sampling distribution mean is 95% likely to be within the achieved CI (Lindley and Phillips 1976). It merely says that 95% of very many similar CIs would contain the mean. This is hard to understand and explain to a wide audience, and often misportrayed even in what are intended to be training resources (Morey et al. 2016). See, for example, where the UK Data Service describes CIs as the probability of containing the true population figure

(https://dam.ukdataservice.ac.uk/dataskills/surveys/6/story_html5.html). There are no interpretations of this concept that are at once simple, intuitive, correct, and foolproof (Greenland 2016). In fact, CIs contain little or no information about the false discovery rate, and are simply a different way of expressing p-values from significance tests (Colquoun 2014).

CIs also share with the flawed approach of significance testing the problem that they do not and cannot address the most important analytical questions about any one sample result. Systematic biases in design, measurement or sampling are generally more substantively important than random variation when deciding on the trustworthiness of any research results. For example, a confidence interval cannot distinguish between a sample of 100 cases with a 50% response rate and a random sample of 100 cases with a 100% response rate. Both will be assessed by a CI as providing equal 'confidence' despite the latter being far superior in quality, and therefore in the level of trust that can be placed in it (a true concept of "confidence"). Similarly, all other things being equal, the CI for a sample of 200 cases with a response rate of 20% (i.e. when 1,000 cases were approached) will be reported as markedly superior to the CI for a random sample of 100 cases with a 100% response rate. In both examples, CIs should not be used since at least one of the samples is no longer random. The non-response cannot be assumed to be random, and careful research has shown than non-responding cases are not a random sub-set of all others. Their mere existence and occurrence creates a ready potential for bias (Brunton-Smith et al 2014).

The impact of missing data can be illustrated through simulations (see Gorard 2014b). Using 100 samples of 10,000 random numbers, around 95% of their computed CIs did contain the true population mean. However, when 10% of the highest or lowest scores were ignored then the CI success rate dropped to 59%. When these scores were instead replaced by the sample mean, the CI success rate dropped to 43%. This means that, even if the logic of CIs were accepted, an analyst reporting a 95% CI for a sample with a 90% response rate could actually be citing a 43% CI or worse. Just as with any significance tests, confidence intervals take no account of bias or missing data. And this bias or potential bias is a far bigger threat to the security of findings than random sampling variation. Anyway, just as with any significance tests, the accuracy of a CI depends on the prior probability. Even in their own terms and with perfect data, CIs are less than 50% accurate (contain the appropriate mean) when the prior is as low as 0.07 (Pharoah et al. 2017).

NNTD

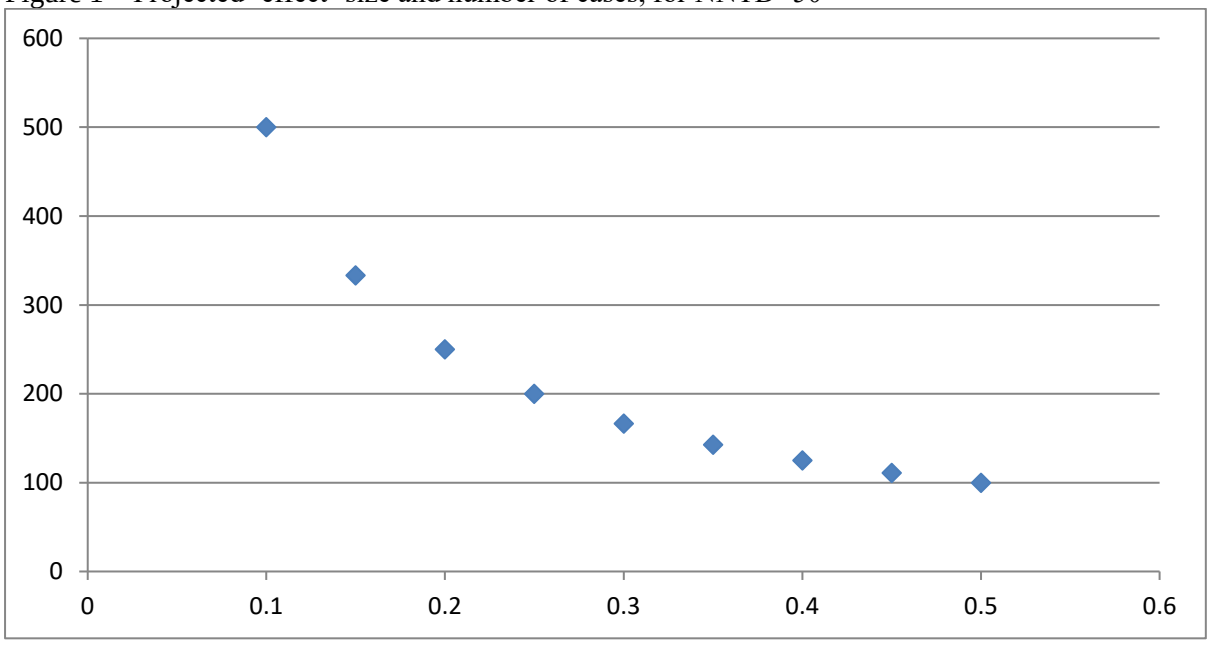
One way to encapsulate how trustworthy any effect size is, according to several of the factors above, involves sensitivity analysis. Sensitivity analyses are used in fields including economics, agriculture clinical trials, and natural sciences to support decision making, such as whether a finding is robust and so worth taking notice of substantively (Thabane et al. 2013). This usually involves looking at the impact on the substantive research findings of varying assumptions about errors and related factors, and according to Pannell (1997) even the simplest of such approaches are useful. One approach is to assess the proportion of cases that would have to be replaced with counterfactual data to invalidate the inference being made (Frank et al. 2013). This is the approach discussed in this paper, and compared to confidence intervals. Taking the idea further than those above, converting it into a number representing how different any missing cases or data would have to be in order for the effect size to become zero, the approach also compares the number of such counterfactual cases with the amount of missing data (Gorard and Gorard 2016).

The number of counterfactual cases needed to disturb the finding (NNTD) involves creating a counterfactual score, such as one standard deviation away from the mean of the smaller group, in the opposite direction to the 'effect' size. The number of these counterfactual scores (running against the finding) that can be added to the smallest group in the comparison before the effect size disappears is then a standard measure of the strength of the 'effect'. This number can be calculated more easily as the effect size multiplied by the number of cases in the smallest group. NNTD is a useful measure of the sensitivity of the scale of the *finding*s (and their variability as represented by the standard deviation used to compute the 'effect' size), taking into account the scale of the *study* (N). All three elements are

combined in one summary figure measured in a number of cases that can be directly compared to the number of missing cases.

Based on a large number of studies (Gorard et al. 2016), a NNTD of 50 can be considered a very strong and secure finding, given how tough this definition is. Using this as a working assumption, Figure 1 shows the number of cases needed in each comparison group (assuming equal size) for 'effect' sizes from 0.1 to 0.5, in order to obtain NNTD=50 with no attrition. As expected, the number of cases needed decreases exponentially as the projected 'effect' size increases. For a typical education intervention or comparison, with an expected effect size of 0.25, the number of cases needed per comparison group is 200. Larger effect sizes in other social science studies would permit correspondingly smaller samples sizes, still with NNTD of 50.

Figure 1 – Projected 'effect' size and number of cases, for NNTD=50



NNTD is measured as a number of cases. This means that NNTD can be used to make a direct comparison with the number of initial cases missing data in any study (from refusal, non-response and attrition). The number of missing cases, or cases missing data, can be subtracted from NNTD. If the result is still greater than zero then it means that even in the unlikely situation that all missing data were in the opposite direction to the main finding (counterfactual) the effect size would still be non-zero. This would suggest a strong finding. The larger the NNTD is, after attrition has been subtracted, the stronger the finding. NNTD is much newer and less widespread than CIs, but has been used successfully in several national projects. NNTD is not meant to be an alternative to CIs, because they clearly have theoretically different purposes, but it is interesting to compare them here for the same datasets.

Methods used for this paper

The rest of the paper is based on a comparison between the CI and NNTD for a large number of simulated comparative studies with two groups. The simulations here involved 1,000 datasets of uniform random numbers between generated in Excel, each portraying a study or trial with between 1 and 1,000 cases (N selected at random). For each dataset an 'effect' size (ES) outcome was generated as the difference in the means between the two groups divided by their overall standard deviation. For simplicity both groups had the same number of cases. The number of counterfactual cases needed to disturb this ES was calculated as the ES multiplied by N/2. The CI for the same datasets was calculated as 1.96 times the overall standard deviation, divided by the square root of N (i.e. this was one 'arm' of

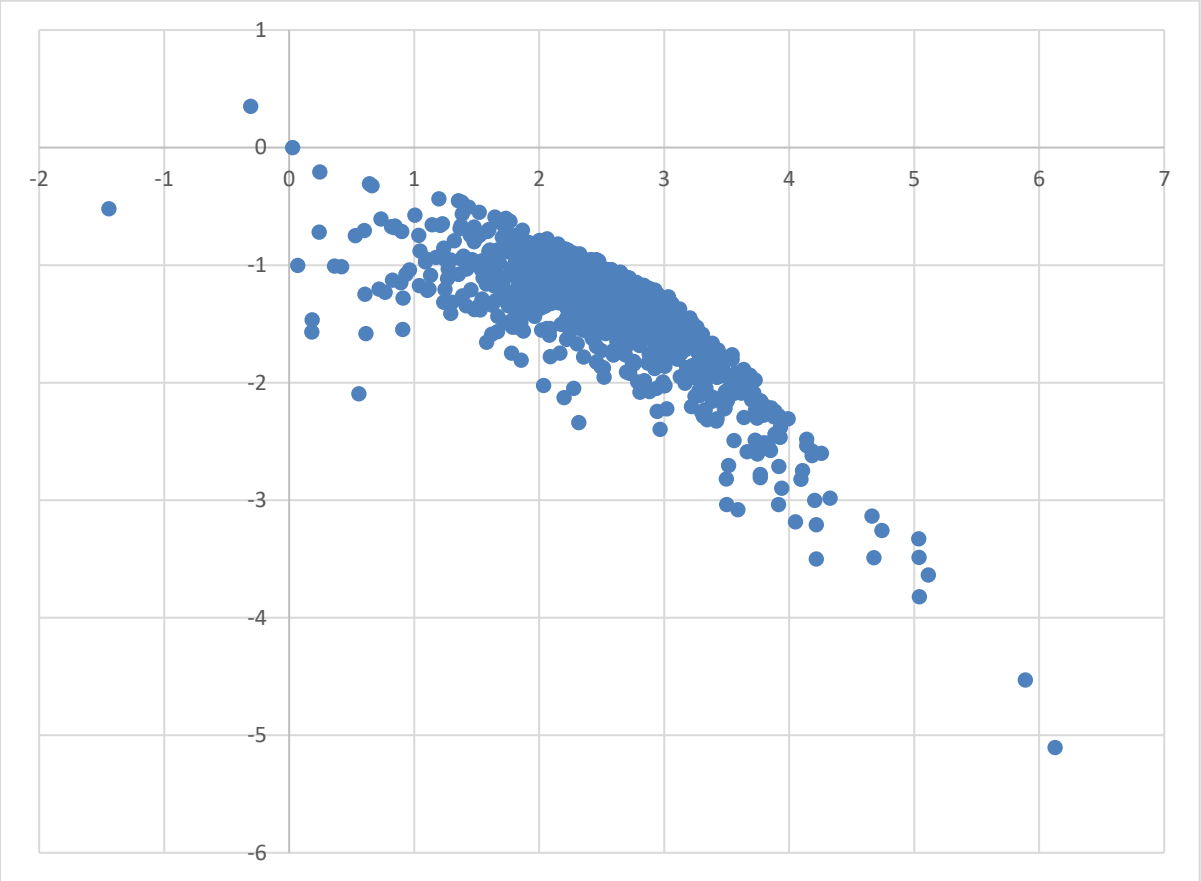
the interval). Both measures are therefore based on N and the standard deviation in some form. For more details see the Appendix.

The 1,000 pairs of NNTD and CI measures were cross-plotted and their Pearson's R correlation coefficients computed. The cross-plotted graph is curvilinear with the high NNTD figures linked to near zero CIs, and low NNTDs linked to increasingly large CIs. Because the two figures are of different scales, and the range of NNTD is so large, the cross-plot was redone with the logs of the two sets of measures, creating a clearer linear pattern, and Pearson's R was recalculated. The latter are the findings presented here.

Comparing NNTD and CIs

Under these conditions, as shown in Figure 2, the measures for NNTD and for CIs are each highly related to each other, and each could be predicted very accurately from the other. When NNTD is larger (positive log) the corresponding CI is always smaller (negative log). When NNTD is smaller, even notionally fractional (negative log), the CI is always larger. For the 1,000 trials illustrated the correlation between the two measures was 0.96 (so that they appear to have over 92% of variation in common). This is typical over many simulations.

Figure 2 – Cross-plot of log of NNTD (x axis) and log of CI (y axis)

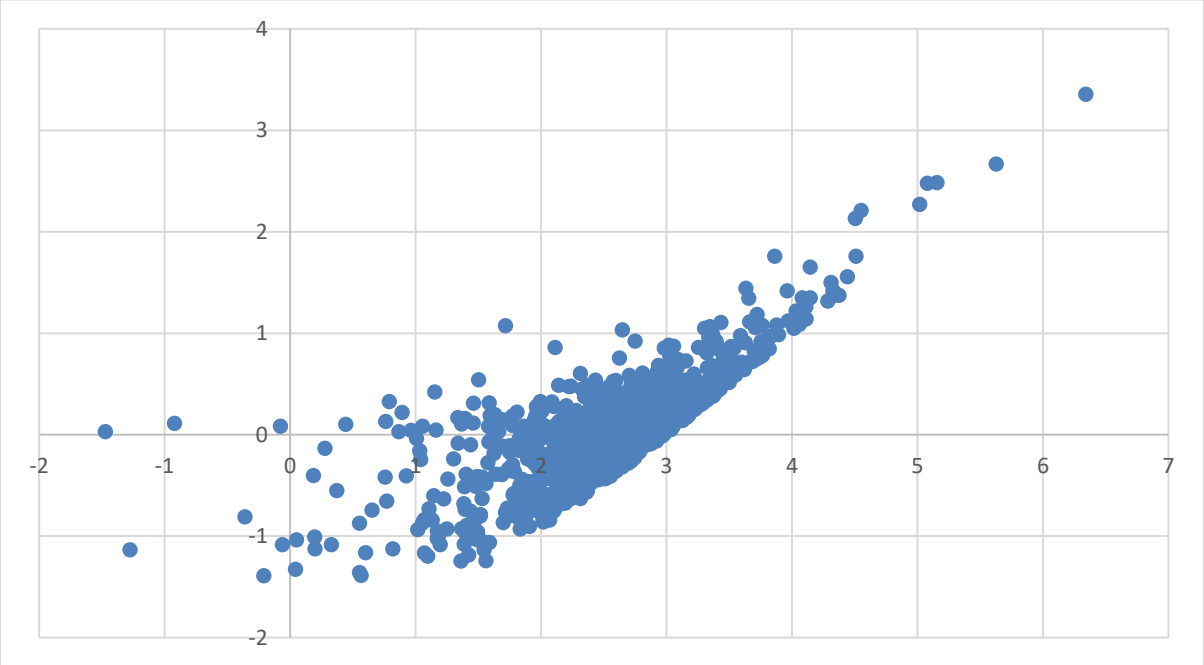


The point representing a dataset with the smallest NNTD (log of less than -1) has a relatively low CI compared to the others. This is an example of a large trial with an ES outcome at or near zero (a perfectly valid result), leading to a small NNTD. There are a few other scores like this but less extreme. These few are perfectly proper results and would probably happen more often in real-life trials than with these random datasets representing trials. With experimental data for example, the US IES has found only about 12% of the interventions it has evaluated to have an ES substantially different to zero (Gorard et

al. 2017). However, these examples do mean that the correlation between CIs and NNTD is not perfect, as well as still being slightly curved rather than a perfectly straight line.

In real datasets, the mean and standard deviation are often of the same order of magnitude and this creates an association between their size. Other than that though, the CI in Figure 2 does not involve the effect size (only N and the standard deviation). The same kind of simulation is therefore redrawn in Figure 3, again plotting NNTD on the x axis, and the effect size plus the CI (both as logs, as above). If anything this creates an even clearer line, and a slightly larger correlation.

Figure 3 – Cross-plot of log of NNTD (x axis) and log of ES-CI (y axis)



Both CI (however computed) and NNTD also include all three elements – the mean, standard deviation and the number of cases. It is not surprising therefore that to a great extent, NNTD and CIs portray the same thing, and one can be predicted reasonably accurately from the other. But one is presented by its advocates as a measure of probabilistic uncertainty and the other as the measure of robustness in the finding. So, which is more accurately portrayed, and which is preferable in practice?

NNTD directly addresses the key question for analysts of whether the ES is substantially distinguishable from zero, where the achieved ES or N or both are so large that it is hard to envisage such a result if the real ES were zero. As illustrated here, the CI is a close proxy for this, but as a range it does not directly address the issue of whether the ES is zero. Instead it gives a rather wide range of possible effect sizes, within which the true ES may or may not be included. In practical terms, this is almost useless. NNTD could be used to create a similar range, such as the largest and smallest ES achieved if 10% counterfactual cases were added to the dataset, for example. This would show the same information as NNTD itself, but be harder to interpret. If, on the other hand, a CI is used to decide whether the ES is likely to be zero in reality, by seeing whether the interval contains zero or not, which is standard practice, then the CI process simply becomes a significance test (see above). Significance tests have already been abandoned in the new statistics, and so CIs should also be adandoned.

Implications

Insofar as NNTD and CI are measures of the same thing, they might be considered interchangeable to a great extent. They used the same three values combined to present a context for an effect size finding. However, they are not theoretically interchangeable. CIs purport to present an estimate of probabilistic

uncertainty in the effect size. To do so they require that the data has no values or cases missing, no measurement error, and is based purely on randomisation. These requirements are built into their computation and interpretation. But such requirements are unrealistic in social science, and are never met in practice. NNTD, on the other hand, has no such requirements and can be computed freely for population, incomplete and convenience datasets. It does not claim, or attempt, to measure probabilistic uncertainty at all. It is a simple measure of robustness, or the sensitivity of the ES to changes such as the addition of cases inconvenient for the effect size. Since NNTD gives such a similar result to CIs, and is based on the same values, perhaps this means that CIs are also measures of robustness and not of uncertainty. This is much more likely than that NNTD is somehow an undiscovered measure of probabilistic uncertainty.

As shown above, as soon as any data or cases are missing and the potential for bias is present, CIs do not work even in their own terms and with randomised perfectly-measured data. NNTD, on the other hand, is measured in a number of cases, and so can be compared easily and directly with the number of cases missing data or missing entirely. It addresses the question of whether the apparent ES could be explained by the missing data or not in a way that CIs simply cannot. It is also free of the assumptions needed for CIs. It can be used for any comparison between groups, even where these are drawn from convenience or other non-random samples. And it does depend on the precise type of effect size, being usable with effect sizes based on the mean absolute deviation rather than the standard deviation, and those based on real numbers and categorical data (Gorard 2015).

Perhaps most importantly the use of NNTD is much easier to explain to a wider audience than a CI is (Watts 1991). CIs are recursive (defined in terms of CIs) and use the same perplexing inverse logic as significance tests. It is very rare to find a methods resource, let alone a piece of research, that describes CIs accurately. There may be better alternatives to NNTD in the future, but for the present NNTD is preferable to CIs for all of these reasons. It provides at least as much information, is more practical and easier to use and understand.

Acknowledgements

With thanks to Jonathan Gorard for helpful discussions.

References

Bakan, D. (1966) The Test of Significance in Psychological Research, *Psychological Bulletin*, 77, 423-437

Berger, J. and Sellke, T. (1987) Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence (with comments), *Journal of the American Statistical Association,* 82, 1, 112–39

Berkson, J. (1938) Some difficulties of interpretation encountered in the application of the chi-square test, *Journal of the American Statistical Association*, 33, 526–536

Boring, E. (1919) Mathematical vs. scientific importance, *Psychological Bulletin*, 16, 335–338

Brunton-Smith, I., Carpenter, J,, Kenward, M. and Tarling, R. (2014) Multiple imputation for handling missing data in social research, *Social Research Update*, 65, Autumn 2014

Chowl, J. and Ekholm, E. (2018) Do published studies yield larger effect sizes than unpublished studies in education and special education?, *Educational Psychology Review*, 30:727–744, https://doi.org/10.1007/s10648-018-9437-7

Coe, R. (2002) *It's the Effect Size, Stupid: What effect size is and why it is important*, Paper presented at the British Educational Research Association annual conference, Exeter, 12September, http://www.cem.org/attachments/ebe/ESguide.pdf

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

Cumming, G. (2014) The new statistics: why and how, *Psychological Science*, 25, 1, 7-**29**
Daniel, L. (1998) Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals, *Research in the Schools*, 5, 2, 23-32
Falk, R. and Greenbaum, C. (1995) Significance tests die hard: the amazing persistence of a probabilistic misconception, *Theory and Psychology*, 5, 75-98
Fidler, F., Thomson, N., Cumming, G., Finch, S. and Leeman, J. (2004) Editors Can Lead Researchers to Confidence Intervals, but Can't Make Them Think: Statistical Reform Lessons From Medicine, *Psychological Science***,** 15, 2, **119-126**
Filho, D., Paranhos, R., da Rocha, E., Batista, M., da Silva, J., Santos, M. and Marino, J. (2013) *When is statistical significance not significant?,* http://www.scielo.br/pdf/bpsr/v7n1/02.pdf
Frank, K., Maroulis, S., Doung, M. and Kelcey, B. (2013) What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences, *Educational Evaluation and Policy Analysis*, 35, 4, 437-460
Freedman, D. (2004) Sampling, in M. Lewis-Beck, A. Bryman and T. Liao (Eds) *Sage Encyclopaedia of Social Science Research Methods* (Thousand Oaks, CA: Sage), 987–991
Glass, G. (2014) Random selection, random assignment and Sir Ronald Fisher, *Psychology of Education Review*, 38, 1, 12-13
Gorard, S. (1999) Keeping a sense of proportion: the "politician's error" in analysing school outcomes, *British Journal of Educational Studies*, 47, 3, 235-246
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. (2014a) A proposal for judging the trustworthiness of research findings, *Radical Statistics*, 110, 47-60, http://www.radstats.org.uk/no110/Gorard110.pdf
Gorard, S. (2014b) Confidence intervals, missing data and imputation, *International Journal of Research in Educational Methodology*, 5, 3, 693-698, http://cirworld.org/journals/index.php/ijrem/article/view/2105/pdf_51
Gorard, S. (2015) Introducing the mean absolute deviation 'effect' size, *International Journal Research and Methods in Education*, 38, 2, 105-114, http://www.tandfonline.com/eprint/NMYudhtEmTaDUnwspE9P/full
Gorard, S. and Gorard, J. (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*, 19, 4, 481-489
Gorard, S. and See, BH and Morris, R. (2016) *The most effective approaches to teaching in primary schools*, Saarbrucken: Lambert Academic Publishing
Gorard, S., See, BH and Siddiqui, N. (2017) *The trials of evidence-based education*, London: Routledge
Greenland, S, Senn, S., Rothman, K., Carlin, J., Poole, C., Goodman, S. and Altman, D. (2016) Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations, *European Journal of Epidemiology*, 31, 4, 337-350
Guttman, L. (1985) The Illogic of Statistical Inference for Cumulative Science, *Applied Stochastic Models and Data Analysis*, 1, 3-10
Halsey, L., Curran-Everett, D., Vowler, S. and Drummond, G. (2015) The fickle p value generates irreproducible results, *Nature Methods*, 12, 3, 179-185
Hedges, L. and Olkin I. (2014) Statistical methods for meta-analysis, Orlando: Academic Press Inc
Hubbard, R. and Meyer, C. (2013) The rise of statistical significance testing in public administration research and why this is a mistake, *Journal of Business and Behavioral Sciences*, 25, 1
Hunter, J. (1997) Needed: A Ban on the Significance Test, *Psychological Science*, 8, 1, **3-7**
Jaynes, E. (1976) Confidence intervals vs Bayesian intervals, in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, The University of Western Ontario Series in Philosophy of Science Volume 6b, pp. 175-257
Kline, R. (2004) *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*, Washington, DC: American Psychological Association
Lindley, D. and Phillips, L. (1976) Inference for a Bernoulli process, *The American Statistician*, 30, 3, 112-119

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K. and Busick, M. (2012) *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*, Washington DC: Institute of Education Sciences

Loftus, G. (1991) On the tyranny of hypothesis testing in the social sciences, *Contemporary Psychology*, 36, 102-105

Matthews, R. (2001) Methods for assessing the credibility of clinical trial outcomes, *Drug Information Journal*, 35, 1469–1478

Meehl, P. (1967) Theory - testing in psychology and physics: A methodological paradox, *Philosophy of Science*, 34, 103 – 115

Morey, R., Hoekstra, R., Rouder, J., Lee, M. and Wagenmakers, E. (2016) The fallacy of placing confidence in confidence intervals, *Psychonomic Bulletin & Review*, 23, 1, 103–123

Morrison, D. and Henkel, R. (1969) Significance tests reconsidered, *American Sociologist*, 4, 131-140

Nickerson, R. (2000) Null hypothesis significance testing: a review of an old and continuing controversy, *Psychological Methods*, 5, 2, 241-301

Nix, T. and Barnette, J. (1998) The data analysis dilemma: Ban or abandon, A Review of null hypothesis significance testing, Research in the Schools, 5, 2, 3-14

Pannell, D. (1997) Sensitivity analysis of normative economic models: Theoretical framework and practical strategies, *Agricultural Economics*, 16, 139-152

Pharoah, P., Jones, M. and Kar, S. (2017) P-values and confidence intervals: not fit for purpose?. *bioRxiv*, http://dx.doi.org/10.1101/180117

Rozeboom, W. (1960) The fallacy of the null hypothesis significance test, *Psychological Bulletin*, 57,

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

Thabane , L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., Thabane, M. et al. (2013) A tutorial on sensitivity analyses in clinical trials, *BMC Medical Research Methodology*, 13, 92, http://www.biomedcentral.com/1471-2288/13/92

Tryon, W. (1998) The Inscrutable Null Hypothesis, *American Psychologist*, 53, 796

Walster, G. and Cleary T. (1970) A Proposal for a New Editorial Policy in the Social Sciences, *The American Statistician*, 241, 16-19

Watts, D. (1991) Why is introductory statistics difficult to learn?, *The American Statistician*, 45, 4, 290-291

White, P. and Gorard, S. (2017) Against Inferential Statistics: How and why current statistics teaching gets it wrong, *Statistics Education Research Journal*, 16, 1, 55-65

Appendix

A simple way to recreate the simulation described in this paper is to use Excel, and create two sets of 1,000 random numbers between 0 and 1 to represent the mean difference between groups and standard deviations, of 1,000 trials or comparisons. A third set of 1,000 random integers between 1 and 1,000 represents the number of cases (N) in each trial. The effect sizes for each trial are represented by a fourth column computed by dividing the mean difference by the standard deviation. Then NNTD can be computed for each trial as half of N (assume equal size groups) multiplied by the effect size. The confidence interval, traditionally and as used in the paper, is:

$$CI=ES\pm1.96*SD/\sqrt{N}$$

However, Hedges and Olkin (2014) have proposed instead using a different approach for Cohen's d effect sizes:

$$CI=ES\pm1.96*\sqrt{((n+m)/(n*m)+ES^2/2*(n+m))}$$

Where n and m are the number of cases in each comparison group, and so n+m=N. The simulation has also been run 1,000 times with this version of the CI (again assuming for this purpose that the groups

are equal size, and so n=m). The results are the same. The traditional CI and the bespoke CI correlate at greater than R=0.99. Therefore, which version of the CI is used makes no difference to the findings of this paper. If the two groups had been unequal, and as n and m diverge from each other, so the CI increases, just as the smallest group decreases and so reduces the NNTD.

Once all columns have been created, the NNTD and CI can be converted to logs, and correlated or cross-plotted.