IOSR Journal of Research & Method in Education (IOSR-JRME)
e-ISSN: 2320–7388, p- ISSN: 2320-737x Volume 13, Issue 2 Ser. IV (Mar. – Apr. 2023), 50-56
www.iosrjournals.org

A Sensitivity Test Does Everything That a Significance Test Does, And Better

Stephen Gorard
Durham University Evidence Centre for Education

Abstract
This paper reminds readers of some of the problems in using significance testing, and of using "effect" sizes instead. It looks at a simple sensitivity test for effect sizes (the number of counterfactuals needed to disturb a finding or NNTD). Using 1,000 simulations of two sets of 100 random numbers each, the paper shows that the p-values from significance tests and the results from an NNTD analysis are equivalent and interchangeable. Both are really a scaled "effect" size, based on a difference between means, their variance, and the number of cases in the comparison. A similar point could be made for all effect sizes, including R^2 from correlation or regression, and odds ratios from tables of categorical variables. As a measure of sensitivity NNTD should be preferred to p-values for several key reasons. NNTD requires fewer, if any, assumptions about the data, permits missing data and measurement error, assesses the robustness of findings in the face of missing data, directly addresses the key question of whether the underlying effect size is zero or not, and is much easier to explain and understand. It has an everyday meaning. Perhaps more importantly as an implication for research methods, a significance test is meant to provide a measure of the probabilistic uncertainty in a research finding, that could have been produced by random sampling variation alone. As used in practice, and illustrated in this paper, it is really nothing of the sort.

Keywords: *Sensitivity test, significance test, scaled effect size, counterfactual score, uncertainty, NNTD*

Date of Submission: 12-04-2023 Date of Acceptance: 25-04-2023

I. Introduction

This paper illustrates several ways in which the key strengths or weakness of a research result can be expressed, focusing on the difference between two sets of numbers. The usual way to summarise the difference between two sets of numbers is to present their means and standard deviations. This can then be standardised as an "effect" size by dividing the difference between means by their overall standard deviation (in one version). Some analysts still also publish p-values or probabilities based on significance tests such as the t-test for independent samples. These p-values are meant to express the probabilistic uncertainty in the result. Both effect sizes and p-values, especially p-values, face problems of applicability and validity.

The paper then presents a simple way of judging a research finding, like a difference between means, in terms of how robust it is, using a simple sensitivity test called the number of counterfactuals needed to disturb the finding (or NNTD). This is a kind of scaled "effect" size. A repeated simulation is then used to show that p-values and NNTD are really doing the same job, and expressing the same thing. The paper ends by discussing the implications of this finding for research, researchers, and research users.

II. Problems with significance testing

Although still widely used and published, the use of significance tests (and related phenomena) in social science is clearly problematic. Significance tests, such as the t-test, ANOVA, and chi-squared, are meant to help to estimate the probability of finding an "effect" size (or other data) 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 only calculable on the further basis that the cases have been fully randomised to each group, that no cases or values are missing, and that there is no measurement error. Some tests have further requirements, such that the datasets in each group are normally distributed and have the same standard deviations. The mathematical computation underlying a significance test is premised on full randomisation, as a matter of mathematical necessity, and if this premise is not true then anything that follows from a significance test is automatically wrong or meaningless (Berk and Freedman 2001, Gorard 2019a).

As may be imagined, these prior conditions for conducting the analysis are rarely, if ever, met in practice, and therefore significance tests should rarely, if ever, be used in practice (Freedman 2004, Filho et al. 2013,

Colquoun 2014). In fact, I have never seen a piece of real-life research in which all of the required assumptions were met, even where significance tests had nevertheless been incorrectly conducted and reported by authors.

However, even if all of the key assumptions for a significance test were to be met, it is not clear what use the resulting p-value would be to a researcher. It is not the desirable and scientifically useful probability that the difference between the two sets of scores is zero, or p of zero|data (Falk and Greenbaum 1995). One of the assumptions of the test is that the difference is precisely zero, so the test only provides the probability of the data occurring given that the different is zero, or p of data|zero, instead. This second conditional probability is not the same as the first one. Nor can it be converted into the first one with the facts available when doing a significance test. One probability can be small and the other large, or *vice versa*, or some other combination (Colquoun 2016). There is no way of knowing p of zero|data from the p-value (p of data|zero) alone. Therefore, knowing the probability of the data given that difference is zero is of no help in estimating the probability of the difference actually being zero.

There are many other problems with significance tests as they are routinely used, including data dredging, 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). Importantly, the use of p-values does not provide an easy picture of the scale of any finding, or its substantive importance. They are also hard to understand (Pocock and Ware 2009). 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, Meehl 1967, Morrison and Henkel 1970, Carver 1978, Berger and Sellke 1987, Daniel 1998, Nickerson 2000, Gorard 2006, Hubbard and Meyer 2013).

All of these reasons and more are why methods' experts in medicine, psychology, sociology, and education, the American Psychological Association (APA), American Sociological Association, and other bodies advise against the use of significance tests (Fidler et al. 2004, Lipsey et al. 2012, Siegfried 2015). Journals such as the American Journal of Public Health, Epidemiology, Basic and Applied Psychology, and numerous others, including most US medical journals, have therefore banned the publication of significance tests (Walster and Cleary 1970, Guttman 1985, Hunter 1997, Nix and Barnette 1998, Starbuck 2016).

III. Problems with effect sizes

Instead of p-values, the so-called "new statistics" movement has called for the use of effect sizes (Cumming 2013). "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 when related linearly (Funder and Ozer 2019). This paper focuses on effect sizes based on the difference between two means divided by their overall standard deviation, but the key points made are applicable to effect sizes of any kind.

Effect sizes can give us a good idea of the size and direction of any difference, pattern, trend, or correlation in the data (Gorard 2021). They do not rely on the same unlikely assumptions about full randomisation as significance tests – so can be used safely with non-randomised cases, or where substantial data is missing, and so on. They can be a useful way of presenting the results of individual studies.

However, effect sizes have some drawbacks. Despite their name, "effect" sizes are really standardised differences, or patterns, based on the data. In themselves they do not represent a cause and effect model. For example, if an effect size is used to compute the difference in school exam results between the top (most talented) and bottom (less talented) class in a school, then it is not true to say that the difference is the "effect" of being in one of these two classes. It may be, but it may also be due to the difference in talent that led to the students being placed in those classes at the outset, or to their teachers, or a host of other explanations. The term effect size is most appropriate when it is used to summarise the impact of an intervention in a good experimental design (Gorard 2013).

Because an effect size is a "standard" score, some commentators have suggested that this makes effect sizes comparable between studies using different measures and approaches. This is not true, and assuming it to be true can be misleading (Morris 2019). The value of a standardised difference between scores relies on a number of factors such as the research design, scale of the study, missing data, and the nature and quality of the underlying measurements (Gorard 2021). This means that effect sizes cannot reasonably be aggregated or averaged, except where all of the studies involved were of the same design and quality. Effect sizes are also open to publication bias just like significance tests (Chowl and Ekholm 2018). They also give no indication of the likelihood of a research finding arising by chance, but then nor do p-values in fact, as explained above.

Effect sizes give no indication of the quality of the study from which they emerged, although there is evidence that smaller and weaker studies tend to yield larger-seeming effect sizes (Wolf et al. 2020). The biggest single problem of effect sizes is that they contain no indication in themselves of the scale of the study that led to them – which is a crucial factor in judging the trustworthiness of any research finding. Of course, effect sizes can and should always be presented with the number of cases on which they are based, for each comparison group.

IV. The NNTD as scaled effect size

One way of improving the information contained in an effect size is to incorporate the scale of the underlying dataset into it directly, and so convert it from a standardised difference to an estimate of the number of standardised counterfactual scores that would be needed to make the effect size disappear. An example is used to illustrate this idea.

Suppose we were comparing the average test scores for two schools (Table 1). We collect 100 scores from each school. School A has a mean of 70 and School B a mean of 75. There appears to be a difference in scores between the two schools, and we want to conclude that scores in school B are substantially larger than those in school A, on average. The effect size is around -0.5 (or 5/10) which seems substantial, and worthy of further consideration. Just before we go ahead and make this claim though, we can work out the number of counterfactual test scores that would be needed to make this effect size disappear (NNTD).

Table 1 – Difference between mean test scores in two schools

	N	Mean test score	Standard deviation
School A	100	70.00	10
School B	100	75.00	10
Overall	200	72.25	10

We pick the smaller group, and estimate its counterfactual. Here both groups are the same size so for illustration we will use School A. The standardised counterfactual for School A could be one overall standard deviation (10) above the mean for School A (70). This would be 80. If we add one such imaginary case to School A, with the counterfactual score of 80, we get the figures in Table 2. Adding a score of 80 has increased the mean for School A by a small amount to 70.1, and the effect size would go down to 0.49 (the difference between 75 and 70.1, divided by 10). NNTD could be defined as the number of times this operation would have to be repeated for the mean score in School A to reach 75, and so for the effect size to become zero. This would a measure of the stability of the result.

Table 2 – Difference between mean test scores in two schools, with one counterfactual case

	N	Mean test score	Standard deviation
School A	101	70.10	10
School B	100	75.00	10
Overall	200	72.25	10

Of course the counterfactual does not have to be one standard deviation from the mean, but it is preferrable to pick a standard unit to enhance the ability of readers to comprehend what NNTD means, and to allow easier comparison between studies. Other alternatives include picking cases from each group at random, to swap over. But this is more complex and harder to follow, and repeated testing has shown it to be equivalent (although on a different scale) to NNTD based on one standard deviation.

In practice, and using the one standard deviation counterfactual, it is easiest to compute NNTD as the original effect size multiplied by the number of cases in the smallest group. For the figures in Table 1, the estimate would be 50 (effect size of 0.5 times N of 100). This suggests that it would take 50 quite extreme scores inconvenient for the finding for the effect size to disappear. Based on a large number of studies (Gorard et al. 2017), a NNTD of 50 can be considered a reasonably secure finding, given how tough this definition is.

As illustrated here, NNTD is a combination of the effect size and the scale of the study, and the effect size itself is a combination of the difference between means and the variability of the test scores. NNTD shows both the scale of the *finding*s and the scale of the *study*.

This combination is easy to compute, and has a simple meaning – the number of inconvenient scores (running against the finding) that it would take to make the finding disappear. Because NNTD is measured on a scale based on a number of cases it can be compared directly to the number of missing cases, or cases missing values, in any study. NNTD tells analysts whether the missing scores, if seriously inconvenient for the main finding, would be a feasible explanation for that surface finding. For example, in the schools' comparison if five scores were missing from each school, the total of 10 missing scores can be compared to an NNTD of 50. It is clear that even if all 10 missing scores were strongly counterfactual there would still be a sizeable effect size, in favour of School B. The number of missing cases is less than (only 20% of) the NNTD.

This kind of analysis is called sensitivity testing. Sensitivity analyses are sometimes used in economics, agriculture clinical trials, and natural sciences to help decide whether a finding is robust, and so worth taking notice of substantively (Thabane et al. 2013, Pannell (1997). This includes estimating the proportion of cases that would have to be replaced with counterfactual data to invalidate the inference being made (Frank et al. 2013). Taking this idea further, NNTD converts sensitivity into a real number representing how different any missing cases or data would have to be in order for the effect size to become zero, and allows an easy comparison to the

number of cases missing data (Gorard and Gorard 2016). NNTD also permits a researcher planning a new study to estimate the number of cases they would need to in order to detect an effect size of a given magnitude, with an NNTD of at least 50 (or any figure required). See Gorard (2019b).

More recently, in a different field, Frank et al. (2021) published a similar if somewhat more complex approach termed "Robustness of Inference to Replacement" or RIR. Both RIR and NNTD are improvements on significance testing, or to the use of effect sizes alone. And both approaches, and others like them, have a close relationship to apparent significance test results, as is illustrated in the rest of this paper.

V. Methods used for this paper

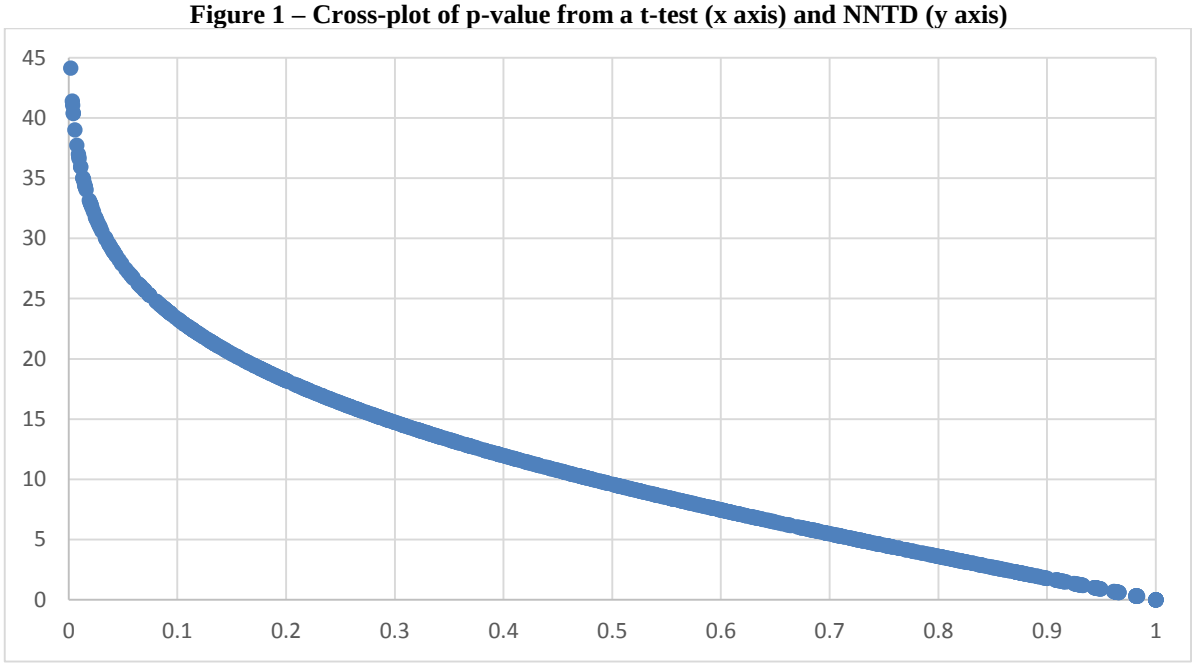
The simple new demonstration analysis in this paper involves a comparison between the p-values generated by t-tests and the results for a NNTD sensitivity test, for 1,000 simulations of a comparison between two sets of scores. Each set of scores consists of 100 uniform random numbers between 0 and 10. The means of each set of 100 scores were computed, along with their overall standard deviations. Then the effect sizes for each of the two sets in a pair was calculated as the differences between the means for each set divided by their overall standard deviations. The NNTD was computed for each effect size by multiplying it by 100 (the size of the smallest group in this simulation). An independent samples t-test was also completed for each of the two pairs of 100 scores (see Appendix for more detail).

The 1,000 pairs of p-values and NNTDs were then cross-plotted, and their Pearson's R correlation coefficient computed. For illustration, the p-values less than 0.05 and their associated NNTDs were also cross-plotted, and the correlation coefficient computed, and the same was done with p-values clustered around 0.5, from 0.475 to 0.525.

VI. Comparing p-values and NNTD

As shown in Figure 1, the p-values and NNTD values NNTD from these 1,000 trials are very closely related. The two measures correlate with each other at R= 0.96, with R^2 over 0.92, meaning that over 92% of the variation in each measure is common to the other measure. This relationship holds over many simulations, and variations (see Appendix). The two values are in effect telling the same story about a scaled effect size.

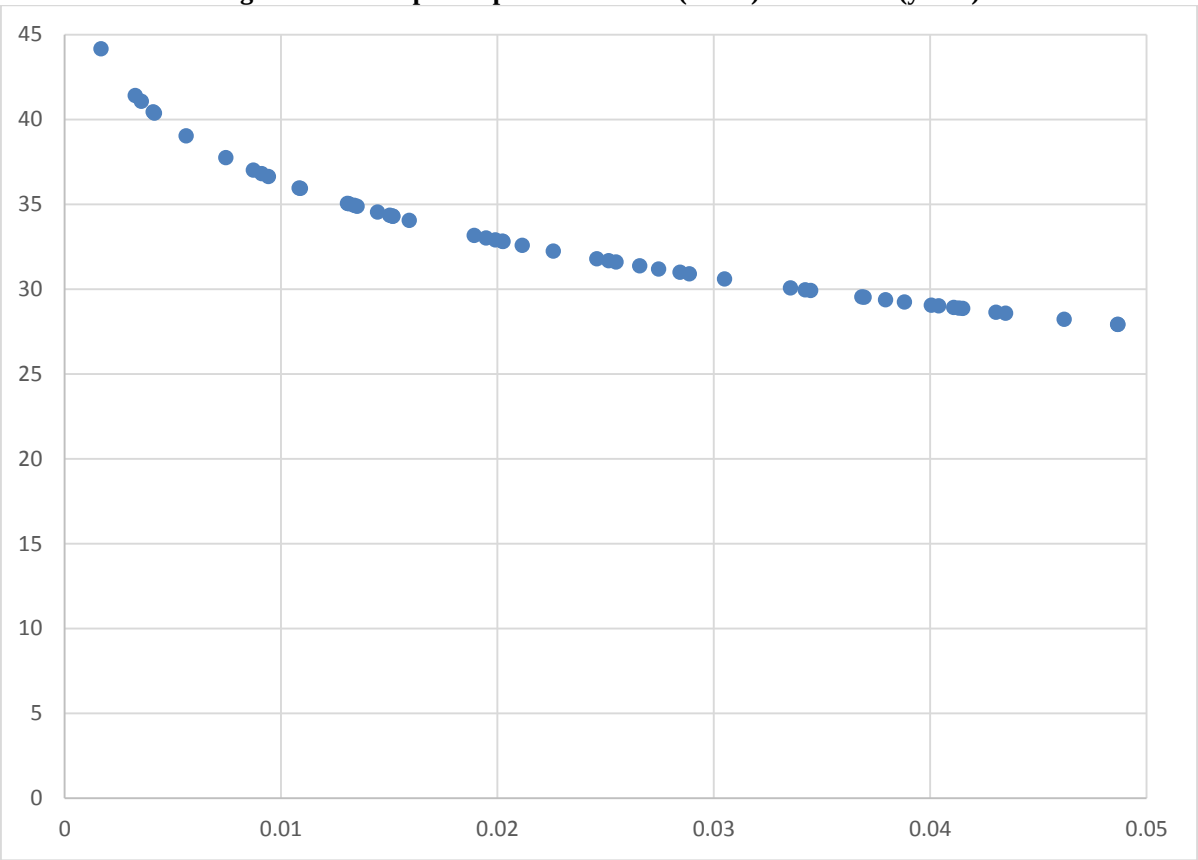
Figure 1 – Cross-plot of p-value from a t-test (x axis) and NNTD (y axis)



One or both scores could be transformed, perhaps through use of logarithms, to create an even flatter line, and an even higher R^2.

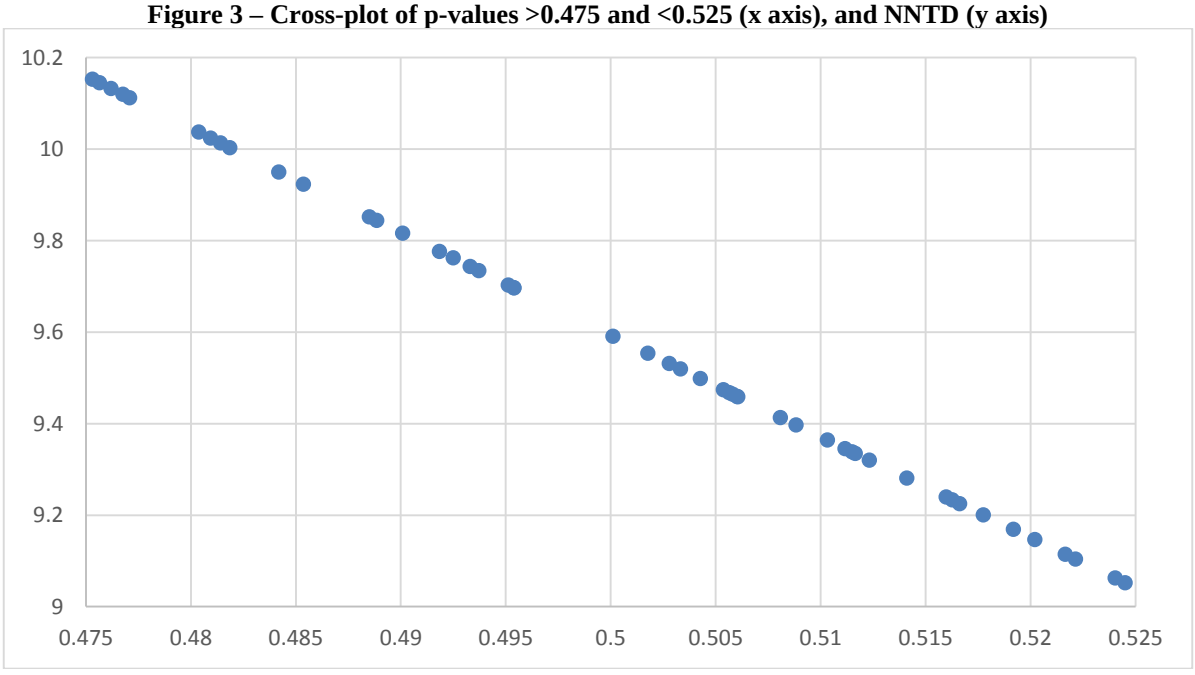
If we magnify the section of the graph at the bend, the head of the "hockey stick", and examine the relationship in the region where p-values are traditionally thought to be "significant", the relationship is as shown in Figure 2. In the 1,000 simulations there were 53 trials where p was less than 0.05. These correlated with their paired NNTD scores at R=0.95, R^2>0.90. The relationship is still curvilinear, but the expansion suggests that each section of the curve in Figure 2, even this part, is still relatively flat.

Figure 2 – Cross-plot of p-values <=0.05 (x axis) and NNTD (y axis)



If instead we magnify the area around the value of p=0.5, there were 49 relevant trials in the 1,000 simulations, and the pattern looks like Figure 3. Here the relationship is a straight line as far as it is possible to tell, with a correlation of 1.0. This means that NNTD is directly calculable from the p-value and *vice versa*. Why does this happen?

Figure 3 – Cross-plot of p-values >0.475 and <0.525 (x axis), and NNTD (y axis)



Both NNTD and p-values include three common elements – the mean, the standard deviation and the number of cases. It is therefore not surprising, in one way, that they portray substantively the same thing - and that knowing one can lead to a very accurate estimate of the other. However, NNTD was devised as a sensitivity test for an effect size based on any kind of numeric dataset, and its calculation involves no formal assumptions about the dataset used. Significance tests, on the other hand, were devised as a method of assessing probabilistic uncertainty in results based only on fully randomised cases with no measurement error. And, as outlined above, they involve many other underlying assumptions. The two things should not be equivalent, but they are. What does this mean?

VII. Implications

NNTD directly addresses the key question for analysts of whether an effect size is substantially distinguishable from zero. It is not a probability. A big number (at least 50 perhaps) means that the effect size is so large, and/or the study so large-scale, and the number of missing cases/values are so small, that it is hard to envisage such a result if the real underlying effect size were truly zero (or approximately zero).

The computation of NNTD makes no assumptions about the distribution of the data, the randomisation, of cases, the lack of measurement error, or that there are no missing cases or values. NNTD can be computed freely for population, incomplete and convenience datasets. NNTD is much easier than a p-value to compute, and can be easily worked out *post hoc* for all relevant studies when carrying out a structured review of existing literature (needing only that the N per comparison group, and an effect size are reported). NNTD is also much easier than p-values for a wider audience to understand, and does not rely on the inverse logic of *modus tollendo tollens* as required in significance tests. NNTD can be compared directly to the scale of missing values, and used to assess the scale needed for a new study. It is clearly preferable to the use of p-values in many ways.

In a way, NNTD provides an answer to the question that arises from advocates of p-values whenever these are criticised – "so, what shall we do instead?" (Gorard 2021). As the simulation in this paper shows, using NNTD instead of p-values means that no valuable information is lost. There is only gain and greater flexibility in analysis.

Perhaps more fundamentally, even though they are thought not to be theoretically interchangeable, the equivalence shows that p-values are not really probabilities at all. Since NNTD gives such a similar result to p-values this means that p-values are also (overly hard to comprehend) measures of robustness, and not of uncertainty. They are disguised effect sizes, scaled by the number of cases. This is much more likely than that NNTD is somehow an undiscovered measure of probabilistic uncertainty. NNTD does, however, provide an indication of the empirical uncertainty in any given finding.

NNTD does not 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, 2021). There may be better alternatives to NNTD in the future, but for the present NNTD can be considered preferable to the use of p-values and related paraphernalia for all of the reasons given. An increasing number of studies are using it successfully. It provides at least as much information, is more practical, and is easier to use and understand for a wide readership.

References

[1]. 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

[2]. Berk, R. and Freedman, D. (2001) Statistical assumptions as empirical commitments, http://www.stat.berkeley.edu/~census/berk2.pdf

[3]. 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

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

[5]. Carver, R. (1978) The case against statistical significance testing, Harvard Educational Review, 48, 378-399

[6]. 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

[7]. 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

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

[9]. Cumming, G. (2014) The new statistics: why and how, Psychological Science, 25, 1, 7-29

[10]. 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

[11]. Falk, R. and Greenbaum, C. (1995) Significance tests die hard: the amazing persistence of a probabilistic misconception, Theory and Psychology, 5, 75-98

[12]. 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

[13]. 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

[14]. Frank, K., Lin, Q., Maroulis, S., Mueller, A., Xu, R., Rosenberg, J., Hayter, C., Marynia, M. and Zhang, L. (2021) Hypothetical case replacement can be used to quantify the robustness of trial results, Journal of Clinical Epidemiology, 134, 150-159

[15]. 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
[16]. 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
[17]. Funder, D. and Ozer, D. (2019) Evaluating effect size in psychological research: Sense and nonsense, Advances in Methods and Practices in Psychological Science, 2, 156–168
[18]. Gorard, S. (2006) Towards a judgement-based statistical analysis, British Journal of Sociology of Education, 27, 1, 67-80
[19]. Gorard, S. (2013) Research Design: Robust approaches for the social sciences, London: SAGE
[20]. 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
[21]. Gorard, S. (2019a) Significance testing with incompletely randomised cases cannot possibly work, International Journal of Science and Research Methodology, 11, 2
[22]. Gorard, S. (2019b) Do we really need Confidence Intervals in the new statistics?, International Journal of Social Research Methodology, 22, 3, 281-291, https://www.tandfonline.com/doi/full/10.1080/13645579.2018.1525064
[23]. Gorard, S. (2021) How to make sense of statistics, London: SAGE
[24]. 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
[25]. Gorard, S., See, BH and Siddiqui, N. (2017) The trials of evidence-based education, London: Routledge
[26]. Guttman, L. (1985) The illogic of statistical inference for cumulative science, Applied Stochastic Models and Data Analysis, 1, 3-10
[27]. Halsey, L., Curran-Everett, D., Vowler, S. and Drummond, G. (2015) The fickle p value generates irreproducible results, Nature Methods, 12, 3, 179-185
[28]. 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
[29]. Hunter, J. (1997) Needed: A ban on the significance test, Psychological Science, 8, 1, 3-7
[30]. 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
[31]. Meehl, P. (1967) Theory - testing in psychology and physics: A methodological paradox, Philosophy of Science, 34, 103 – 115
[32]. Morris, P. (2019) Misunderstandings and omissions in textbook accounts of effect sizes, British Journal of Psychology, https://doi.org/10.1111/bjop.12401
[33]. Morrison, D. and Henkel, R. (1969) Significance tests reconsidered, American Sociologist, 4, 131-140
[34]. Nickerson, R. (2000) Null hypothesis significance testing: a review of an old and continuing controversy, Psychological Methods, 5, 2, 241-301
[35]. 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
[36]. Pannell, D. (1997) Sensitivity analysis of normative economic models: Theoretical framework and practical strategies, Agricultural Economics, 16, 139-152
[37]. Pocock, S. and Ware, J. (2009) Translating statistical findings into plain English, The Lancet, 373, 9679, 1926-1928
[38]. Rozeboom, W. (1960) The fallacy of the null hypothesis significance test, Psychological Bulletin, 57,
[39]. Siegfried, T. (2015) P value ban: small step for a journal, giant leap for science, Science News, https://www.sciencenews.org/blog/context/p-value-ban-small-step-journal-giant-leap-science
[40]. Starbuck, W. (2016) 60th Anniversary Essay: How journals could improve research practices in social science, Administrative Science Quarterly, 61, 2, 165–183
[41]. 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
[42]. Walster, G. and Cleary T. (1970) A proposal for a new editorial policy in the Social Sciences, The American Statistician, 241, 16-19
[43]. Wolf, R., Morrison, J., Inns, A., Slavin, R. and Risman,K. (2020) Average effect sizes in developer-commissioned and independent evaluations, Journal of Research on Educational Effectiveness, 10.1080/19345747.2020.1726537

Appendix

A simple way to recreate the simulation described in this paper is to use Excel, and create two sets of 100 random numbers between 0 and 10. Find the mean of each set, and their overall standard deviation. Convert to an "effect" size – the difference between the means divided by the overall standard deviation. Then multiply the effect size by the size of the smallest set of numbers (here both are 100). Also conduct an independent t-test comparing the same two sets, with a p-value as the outcome. Copy all of this (best done in one column on the spreadsheet) as many times as needed – here a further 999 times. Once all columns have been created, the corresponding NNTD and p-values can be correlated or cross-plotted.

Different size groups, and different scales of random number were used in varying simulations. All gave the same substantive findings as in this paper. If the total of 200 cases is held constant, but the two groups are made unequal, then as the smallest group decreases the p-value increases, while the NNTD reduces correspondingly. If the total is increased then the p-values drop and the NNTD scores increase.

The simulation can be tried with different counterfactuals, and different parametric tests of significance. All give the same substantive results as in this paper.