What to do instead of significance testing? Calculating the 'number of counterfactual cases needed to disturb a finding'

Stephen Gorard and Jonathan Gorard
School of Education
Durham University
s.a.c.gorard@durham.ac.uk

Abstract

This brief paper introduces a new approach to assessing the trustworthiness of research comparisons when expressed numerically. The 'number needed to disturb' a research finding would be the number of counterfactual values that can be added to the smallest arm of any comparison before the difference or 'effect' size disappears, minus the number of cases missing key values. This way of presenting the security of findings has several advantages over the use of significance tests, effect sizes and confidence intervals. It is not predicated on random sampling, full response or any specific distribution of data. It bundles together the sample size, magnitude of the finding and the level of attrition in a way that is standardised and therefore comparable between studies.

Introduction: the problems with current approaches to analysis

The use of sampling theory statistics as aids to analysis and for presenting the security of research findings has been around for one hundred years. But from their inception, their use has been questioned and their logic pulled apart. More recently, there has been an emphasis instead on presenting results as effect sizes. This is certainly an improvement but the approach has almost as many problems as statistical testing does.

This paper outlines the limitations of both before presenting a new approach – the number of counterfactual cases needed to disturb a finding. The latter forms the substance of this brief paper, which is not intended to cover again the full argument against the use of significance tests (both in themselves and in disguised form such as with modelling, power calculations and confidence intervals). This argument has been fully covered elsewhere (such as in Gorard 2015). The paper is therefore primarily concerned with the scale, strength or trustworthiness of any research findings (an activity internal to the study) rather whether the findings might be more generally true for cases not involved in the study.

Significance tests

The use of significance tests, such as t-tests, is still widespread in the social sciences. The approach was developed to try and assess how unlikely it is that a research finding arose solely by chance, due to the vagaries of random sampling. The algorithm for calculating significance assumes and requires complete random samples (Siegel 1956). However, the approach is widely misused and misunderstood (Watts 1991). It is routine to come across significance tests reported in the social science literature where the cases involved have not been randomly selected – representing instead population data, convenience samples, and incomplete attempted random samples. Much of the literature therefore cites 'significance' of a kind that is corrupt and misleading.

Even when used as intended with complete random sampling, the results of significance tests are still widely misinterpreted (Carver 1978). In fact, significance tests are really only another way of presenting the scale of a piece of research – saying little or nothing about the magnitude or importance of a finding (Kuhberger et al. 2014). For example, a t-test may be used to examine the difference between the means of two sub-groups of the overall sample. It is extremely unlikely that any two sub-groups in the population will have exactly the same mean, and so the sample means will tend to differ, even if only by a tiny amount (Meehl 1967). If a significance test gives a 'significant' result it means only that the sample size was large enough to enable the study to pick up the difference that must exist in the population.

The tests generate a probability of obtaining by chance a finding (difference, pattern, or trend) at least as large as the one found in new research, assuming that this finding does *not* occur in the population from which the cases were drawn at random. Yet such test results are widely mis-reported, or utilised, as being the probability of the finding occurring in the population, given the size of the finding in the sample cases. These two probabilities are entirely different, and one may be large and the other small or *vice versa*. Because social scientists generally want to know the probability of their finding being true of the population, this fact alone should make significance testing a waste of time for them. No one wants to know the probabilistic answer the tests actually provide (Falk and Greenbaum 1995). This is presumably why so many researchers envisage that the significance test results that they *can* calculate (the probability of data given the null hypothesis) are the results they actually want (the probability of the null hypothesis given the data).

Significance tests are anyway based on the standard error for the sampling distribution of the population, which will never be known in practice, and the inaccuracy in estimating it from one achieved sample can make test results misleading even in their own terms (Gorard 2015).

Finally, it is clear from empirical studies that the threshold nature of 'significance' leads to serious publication bias. A very large number of reported p-values are just inside the traditional 5% significance threshold (Kuhberger et al. 2014). And very few non-significant findings are ever reported formally. There must be a better approach.

Effect sizes

One issue for significance tests is that even if they worked and were used as intended they convey no easily interpreted information about the size of the finding (difference or pattern). Most commentators now advocate not using significance testing for the reasons given above, and instead presenting findings in terms of standardised 'effect' sizes (Lipsey et al. 2012). This is certainly a preferable approach. It is not so open to abuse since the figures presented are easier to understand (not *modus tollendo tollens* as with significance test results – see Gorard 2010a). Effect sizes make sense even with non-random samples, and they overtly portray the strength of the pattern or difference in the data.

However, effect sizes are also open to publication bias and abuse. Effect sizes are theoretically independent of sample sizes. Yet in a study of 1,000 articles, a correlation of -0.45 was found between the sample size and reported effect size (Kuhberger et al. 2014).

Small samples are more likely to produce aberrant or extreme 'effect' sizes which, like significant results, are then more likely to be published (Slavin and Smith 2009).

Effect sizes offer no assessment of the likelihood of a finding arising by chance and give no indication of the quality of the study that led to the finding. They take no account of scale (sample size), and are often predicated on assumptions such as a normal distribution of the data that are just as unrealistic in practice as having a complete random sample. Some commentators, in scrapping significance tests, have suggested using confidence intervals (CIs) with effect sizes, on the basis that CIs offer at least a likelihood estimate (Cumming 2013). But the problems with CIs are largely the same as for significance tests. CIs are, in fact, significance tests. They apply only to complete random samples and, like p-values, are the reverse of what most commentators want and believe them to be, and again are largely the size of the sample converted into a less comprehensible format. This approach, of CIs with effect sizes, requires two sets of figures to be presented together, but still does not allow easy interpretation of the findings by the intended user.

Missing data

Perhaps most importantly of all, neither significance testing or effect sizes take any account of missing data in the findings (Berk and Freedman 2001). In real-life research there will always be missing data, including inaccurate recording of values, missing values and missing cases. Cases will refuse to participate, have lost records, and drop out during research. There is no reason to assume that these hard-to-reach cases and values will be a random sub-set of the total (Peress 2010). Therefore, missing data must be assumed to lead to bias in research findings, such as in the reported differences between groups, or the apparent relationships between variables when modelling with datasets.

Some commentators attempt to defend the use of traditional statistics with incomplete samples by drawing a purported distinction between data missing randomly (an unlikely scenario), which they term missing completely at random, and data missing at random, which they claim is data whose 'missingness' does not depend upon its value (Brunton-Smith et al. 2014). This then 'permits' them to use their favoured statistical approaches with over 60% of the relevant data missing (for example, see Pampaka et al. 2014). The distinction is a false one. Randomness is a very simple concept, meaning that events happen by chance. Since we are sure from empirical studies that missing data does not occur by chance, it cannot and must not be assumed to be random in nature (Hansen and Hurwitz 1946, Sheikh and Mattingly 1981). This has been repeatedly demonstrated when population or administrative data are compared to surveys of the same cases, or where different studies using the same selected 'sample' have different response rates (Dolton et al. 2000). These differences can have a 'dramatic impact' on the data, and lead to 'puzzingly different results' (Behaghel et al. 2009, p.1). Any bias in the substantive results caused by missing data generally cannot be corrected by technical means (Cuddeback et al. 2004). In fact fiddling with the data *post hoc*, rather than improving the situation, usually creates further bias by emphasising the cases that are not missing (Gorard 2014a).

Introducing the 'number needed to disturb'

A better alternative to all of the above is to consider how different the data obtained would have to have been in order for the finding of interest to disappear. If little would have to change then the finding is clearly not a strong one, unless the finding is that there is no difference (pattern or trend). There are a number of ways of implementing this approach. The method proposed here is simple, standardised, and takes into account in one summary figure the sample size, the magnitude of the finding and the level of missing data. It involves calculating the number of counterfactual cases needed to disturb the finding.

Imagine a piece of research comparing the mean scores of two groups. Perhaps these are the outcomes scores in a simple trial comparing an intervention and a control group. Assuming that the study design was good, the measurements secure, and that there are no obvious threats to validity, then the difference between the means is an estimate of the impact of the intervention given to one group but not the other (Gorard 2014b).

The proposed procedure is explained first using a small imaginary dataset, and then illustrated with real data in the next section. In this small imaginary dataset there are two groups each of 10 cases with no missing data and no design problems (Table 1). The mean average of Group 2 (5) is larger than the mean average of Group 1 (3.9). How ‘robust’ is this difference between the means?

Table 1 - A comparison between two groups, each of 10 cases

Case number	Group 1	Group 2
1	3	6
2	8	9
3	0	7
4	3	6
5	6	1
6	3	1
7	5	9
8	1	2
9	4	6
10	6	3
Mean	3.9	5.0
Standard deviation	2.3	2.9

In terms of Hedge’s ‘effect’ size, the difference between the two groups would be expressed as the difference between their means (-1.1) divided by their pooled standard deviation (2.74). This would be -0.4, which is a fairly substantial effect size for the social sciences. To calculate a standardised number needed to disturb that finding, the first step is to create a counterfactual score from the group with the most cases to apply to the group with the fewest cases. Since both groups have N=10 here, it does not matter which group is used to create the counterfactual score. We will use Group 1 data to create the counterfactual, which is defined as the mean score for that group (3.9) minus the overall SD (2.74) or 1.16. The SD is subtracted to make the counterfactual even smaller than the mean for Group 2. If we had used Group 2 to create the counterfactual then we would have *added* the mean for that group (5) to the SD (2.74) to create a counterfactual of 7.74 (see summary below).

If we now increase Group 2 to 11 cases by adding a counterfactual case of 1.16, then the mean of Group 2 drops from 5.0 to 4.65. The difference between the means of Groups 1 and 2 would now be -0.75. The counterfactual case can be added repeatedly to the existing cases in Group 2 until the difference in means between the groups becomes zero or changes sign. At that stage the Hedge’s effect size also becomes zero or changes sign. The number of counterfactual cases needed to disturb the original finding in this way becomes a standardised measure of the robustness of the original ‘effect’ size. The larger this number is, the safer and more trustworthy is the finding. The generic steps are summarised below.

- Take the mean and standard deviation (SD) of the group with the most cases.
- If this mean is larger than that of the group with the fewer cases, then add the overall SD to that larger mean. If the mean of the group with the most cases is smaller than that of the group with the fewer cases, then subtract the overall SD from the smaller mean. This forms the standard ‘counterfactual’ score.
- Calculate, by iteration or estimation, how many of these counterfactual cases can be added as cases to the smaller group before the difference between the group means disappears or reverses. This is the ‘number (of counterfactual cases) needed to disturb’ the result (assuming that there is no dropout).

Using the imaginary data from Table 1, it would take 5 counterfactual scores of 1.16 added to the original cases in Group 2 to make the mean of Group 2 less than or equal to the mean of Group 1. Similarly, and obviously, it would also take 5 counterfactual scores of 7.74 added to the original cases in Group 1 the make the mean of Group 1 greater than or equal to the mean of Group 2.

In general, how much attention is paid to the difference between means in situations like this should depend on the number of cases in the smallest group, the size of the difference in the outcomes between groups, and the level of missing data. Large studies with little missing data are preferable to small studies or studies with high rates of attrition. For an effect to be meaningful the difference between the means must be considered substantial, and robust in face of missing data, chance and design issues such as bias. As explained further below, the number needed to disturb the finding (NNTD) allows all of these factors to be encompassed in one summary figure.

A heuristic approach

The basic NNTD can be assessed exactly by repeatedly adding a counterfactual case to the smaller group and recalculating the effect size until the difference between the means disappears (as above). This is relatively simple using syntax in SPSS or similar. Even easier is to approximate it using a heuristic, which our repeated tests have found to be reasonably accurate with group sizes of 50 or more cases. This means that the calculation can be done in one step with no iteration, and it can even be done when the original data are not published but the means and standard deviations are known.

If μ1 and μ2 denote the means of the smaller and larger groups respectively, and the overall standard deviation is σ, then the standard counterfactual case is given by:

$$c = \mu_2 \pm \sigma$$

It is established that with a sufficiently large group of N cases with mean μ1, adding each of the counterfactual cases c to the smaller group will change the mean by approximately:

$$\frac{|\mu_1 - c|}{N}$$

in either direction, where N is the number of cases in the smaller group.

To achieve a 'disturbance', the total change in µ1 must equal or exceed the difference between the two means |µ1 - µ2|. Or put another way:

$$\text{NNTD}\ \frac{|\mu_1 - c|}{N} \geq |\mu_1 - \mu_2|$$

Hence, with large N, we can approximate NNTD with the heuristic:

$$NNTD \geq \frac{N|\mu_1 - \mu_2|}{|\mu_1 - c|}$$

Where $$c = \mu_2 + \sigma \text{ if } \mu_1 < \mu_2$$
$$c = \mu_2 - \sigma \text{ if } \mu_1 \geq \mu_2$$

The result will clearly be an under-estimate for the true value since the actual N in the smaller group will increase by one on each step of the iterative true algorithm, slightly reducing the impact of each added counterfactual, even with large N. Using the data from Table 1 (above), the minimum estimated NNTD using this heuristic is 3 (rather than the actual figure of 5). This would be calculated as 10*1.1 divided by either |3.9-7.74| or |5-1.16|.

Including scale of missing data in the NNTD

The number needed to disturb (NNTD), however computed, is expressed as a number of cases. Therefore, it can be compared directly with the number of cases missing key data, or missing entirely through dropout or non-response. The initial result from the steps above should be reduced by the number of cases missing data. The final result would then take attrition into account as well, on the assumption that any missing data will tend to be counterfactual. This last assumption is a somewhat pessimistic one, but it does mean that any finding that still looks robust after this process deserves to be taken seriously. It means that the ‘effect’, difference or trend cannot have been caused simply by missing data, even if *all* of the missing data had been counterfactual.

Real examples of NNTD in use

As an example, a recent evaluation of a new literacy intervention (Switch-on Reading) initially had 314 cases in two equal groups, and an effect size in terms of gains scores of +0.24 (Gorard et al. 2014a). The effect size used was Hedge’s g – the difference between the mean scores for the two groups divided by their pooled standard deviation. The control group was slightly smaller (153 cases) than the intervention group (155 cases) so the intervention group is used to compute the ‘counterfactual’ score. This is 4.40 (the mean) plus 7.45 (the standard deviation), or 11.85 (Table 2). If 35 of these scores are added to the 153 existing scores for the control group then the ‘effect’ size is still just above zero. However, if 36 such scores are added then the 'effect' size becomes slightly negative. Therefore, the simple number needed to disturb the finding is +36. The heuristic approximation underestimates this as +30, calculated as (153*1.81)/|2.59-11.85|.

Table 2 - Estimated impact of Switch-on Reading Programme

Treatment group	N	Gain	Standard deviation	'Effect' size
Intervention	155	4.40	8.18	-
Control	153	2.59	6.53	-
Overall	308	3.50	7.45	+0.24

The number of cases randomised to one of the two groups but for whom there is no post-test scores was 5. Subtracting this from +36 yields a complete NNTD-minus-attrition of +31. On the basis of working with random values in simulations, +31 seems very unlikely to arise by chance or to be caused by the very low level of attrition, even if *all* of the missing scores would have been counterfactual to the overall finding.

Another recent evaluation of a literacy intervention (Response to Intervention) involved 373 cases in two groups, with fewer cases in the intervention group than in the control (Table 3). It produced an effect size of in terms of gain scores of +0.12 (Gorard et al. 2014b). The overall standard deviation (9.73) is subtracted from the mean for the control group (2.80) to create the counterfactual (-6.93). It would take the adding of 21 of these counterfactual scores to those of the treatment before the effect size disappeared and became slightly negative. The initial NNTD is therefore +21. The heuristic approximation underestimates this as +19 calculated as (178*1.14)/|3.94+6.93|.

Table 3 – Estimated impact of Response to Intervention

Treatment group	N	Mean	Standard Deviation	Effect Size
Intervention	178	3.94	10.62	-
Control	195	2.80	8.83	-
Overall	373	3.35	9.73	+0.12

However, a total of 89 cases did not complete the outcome tests, and so are missing key data. This means that the NNTD-minus-attrition is actually -68. This suggests that the difference between the means is very insecure, and could easily have arisen even if only some of the missing cases would have been counterfactual to the overall finding. The level of attrition indicates a failed trial due to dropout, and so the difference between means must be treated as very far from robust.

Expanding the applicability of NNTD

The illustrations so far have been deliberately simple, and based on real numbers as measurements. However, the calculation of NNTD can be extended to almost any analytical situation where there is a concern to assess the robustness of a difference, trend, or other pattern in the data. As already shown, the process does not depend on equal sized groups. Nor does the data have to represent that from a random sample, or mimic a specific distribution (such as the normal curve). It can be used where there are more than two groups to compare, simply by pairwise consideration. It can even be used in modelling of the kind based on correlation/regression by asking how much the data would have to have differed to disturb a coefficient, or eliminate an increase in R linked to a specific variable.

It is also reasonably simple to apply NNTD to data based on frequencies, since frequencies are real numbers in the same way as measurements are. In analytical terms the difference, between a comparison of groups of measurements classified in terms of which group they are in (as above) and a comparison of groups of frequencies classified in terms of which group they are in, is often over-emphasised (Gorard 2010b). Here the illustration again involves two groups, although the process could easily be extended to more complex situations. The example is imaginary, but a fuller explanation of these calculations appears in Chapter 3 of Gorard (2003). Suppose that a sample of 100 people were asked a number of questions including their sex, and whether they had visited their GP (doctor) in the past two years. The results show that 59% of males (24/41) and 49% of females (29/59) had visited their GP (Table 4).

Table 4 - Cross-tabulation of GP visits by sex

	Visit GP	Not visit GP	Total
Male	24	17	41
Female	29	30	59
Total	53	47	100

How robust is this apparent difference between males and females, given that there are more females in the sample? For the present, we will ignore the issues of sample quality and dropout and consider only the number of counterfactuals needed to disturb this apparent difference. One way to approximate the NNTD examples above would be to add new counterfactual cases (not visiting GP) to the smallest group (males). If 8 such cases are added, the percentage of males reporting visiting their GP would become 24/49 which is just under 49% (the same percentage as for females).

However, it would probably make more sense to continue to use the difference between the observed and expected values in each cell, as is traditional, and which can be computed using the marginal totals. Here, each cell differs from expectation by 2 cases, so the total would still be 8 cases needing to change (or 4 cases needing to *ex*change) in order to disturb the finding. Either way, the NNTD can then be compared directly to levels of non-response and missing data, also expressed as a number of cases.

Discussion

Of course, the approach presented in this paper for the first time still requires development, and consideration of issues such as how to estimate the sample size needed in order to detect a pre-specified ‘effect’ size with a pre-specified NNTD. And it does not, in itself, address other crucial factors for analysts such as the quality of the measurements, the accuracy of the classifications and the design quality and biases in any study (but see Gorard 2014b). Perhaps most importantly, readers and users of research would need practice in judging what the numbers in NNTD mean. Four examples are used in this paper yielding figures of -68,5, 8, and 31. The first of these should more accurately be zero, showing quite clearly that the 'effect' size or difference can be quite easily explained by missing data or chance. The fact that it is negative also implies that the research has failed as a piece of evaluation. The second and third figures represent situations where the pattern or difference is very weak. The findings should be treated as indicative at most. Only the fourth figure (+31 after the missing cases are accounted for) suggests a finding

This paper has introduced the computation of a 'number of counterfactual cases needed to disturb the finding' (NNTD) to help assess the robustness of a pattern, trend or difference in numeric data. The idea of the number of counterfactual cases needed to disturb the safety of an empirical finding is a simple one. It is presented here as a superior alternative to the failed approach of significance testing and confidence intervals, and the limitations of 'effect' sizes alone. Unlike the permutation test it is not predicated on complete random sampling. In fact, creating NNTD requires no particular underlying assumptions, creates no threshold of acceptance/rejection, is comparable between studies of different sizes using different outcome measures, and uniquely it takes into account in one figure the sample size, magnitude of the finding and the level of missing data. It can be used with any form of numeric data, and any number of groups or classifications. It is easy to understand, and will be simple to compute once operationalised in software. The heuristic approach allows a minimum value to be estimated even when there is no access to the actual data (such as when conducting a review or meta-analysis). Regardless of how the details are worked out, the principle of NNTD has a lot to recommend it.

References

Behaghel, L., Crepon, B., Gurgand, M. and Le Barbanchon, T. (2009) *Sample attrition bias in randomized surveys: a tale of two surveys*, IZA Discussion Paper 4162, http://ftp.iza.org/dp4162.pdf, accessed 060714

Berk, R. and Freedman, D. (2001) *Statistical assumptions as empirical commitments*, http://www.stat.berkeley.edu/~census/berk2.pdf, accessed 030714

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

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

Cuddeback, G.. Wilson, E., Orme, J. and Combs-Orme, T. (2004) Detecting and statistically correcting sample selection bias, *Journal of Social Service Research*, 30, 3, 19-30

Cumming, G. (2013) The new statistics: why and how, *Psychological Science*, doi:10.1177/0956797613504966

Dolton, Lindeboom, M. and Van den Berg, G. (2000) *Survey Attrition: A taxonomy and the search for valid instruments to correct for biases*, http://www.fcsm.gov/99papers/berlin.html

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

Gorard, S. (2003) *Quantitative methods in social science: the role of numbers made easy*, London: Continuum

Gorard, S. (2010a) All evidence is equal: the flaw in statistical reasoning, *Oxford Review of Education*, 36, 1, 63-77

Gorard, S. (2010b) Measuring is more than assigning numbers, pp.389-408 in Walford, G., Tucker, E. and Viswanathan, M. (Eds.) *Sage Handbook of Measurement*, Los Angeles: Sage

Gorard, S. (2014a) Confidence intervals, missing data and imputation, *International Journal of Research in Educational Methodology*, 5, 3, 693-698

Gorard, S. (2014b) A proposal for judging the trustworthiness of research findings, *Radical Statistics*, 110, 47-60

Gorard, S. (2015) Rethinking "quantitative" methods and the development of new researchers, *Review of Education*, 3, 1, (forthcoming)

Gorard, S., See, BH and Siddiqui, N. (2014a) *Switch-on Reading evaluation report*, http://educationendowmentfoundation.org.uk/uploads/pdf/FINAL_EEF_Evaluation_Report_-_Switch-on_-_February_2014.pdf

Gorard, S., Siddiqui, N. and See, BH (2014b) *Response to Intervention evaluation report*, http://educationendowmentfoundation.org.uk/uploads/pdf/FINAL_EEF_Evaluation_Report_-_Response_to_Intervention_-_February_2014.pdf

Hansen, M. and Hurwitz, W. (1946) The problem of non-response in sample surveys, *Journal of the American Statistical Association*, *41*, 517–529

Kuhberger, A., Fritz, A. and Schemdl, T. (2014) Publication bias in psychology, *PLOSone*, doi: 10.137/journal.pome.0105825

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, p.13

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

Pampaka, M., Hutcheson, G. and Williams, J. (2014) Handling missing data, *International Journal of Research and Method in Education*, doi: 10.1080/1743727X.2014.979146

Peress, M. (2010) *Correcting for Survey Nonresponse Using Variable Response Propensity*, Journal of the American Statistical Association, http://www.rochester.edu/College/faculty/mperess/Nonresponse.pdf

Sheikh, K. and Mattingly, S. (1981) Investigating nonresponse bias in mail surveys, *Journal of Epidemiology and Community Health*, *35*, 293–296

Siegel, S. (1956) *Nonparametric statistics for the behavioural sciences*, Tokyo: McGraw Hill

Slavin, R. and Smith, D. (2009) The relationship between sample sizes and effect sizes in systematic reviews in education, *Educational Evaluation and Policy Analysis*, 31, 4, 500-506

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