Explaining the number of counterfactual cases needed to disturb a finding: a reply to Kuha and Sturgis

Stephen Gorard (Durham University) and Jonathan Gorard (King's College, London)

We want to make statistics both more logical and easier to comprehend, and in doing so reduce the real damage caused by vanishing breakthroughs, publications bias and other consequences of using standard errors incorrectly. Our paper on the number of counterfactual cases needed to disturb a finding (NNTD) was a part of that. We are grateful to Kuha and Sturgis (K&S) for suggesting an 'exact' formula, for calculating the NNTD (used when the concern is about a difference between means). It is not, in fact, a formula for calculating the iterative NNTD as we described it. Rather, it concerns the number of cases in the smaller group in the comparison that would have to change to being more counterfactual (by one standard deviation) before the apparent difference disappeared. But it is simple, and seemingly more accurate than our heuristic - at least for small to middle N. We actually prefer it, and it strengthens our argument for the use of NNTD.

K&S complain that we did not define terms like the 'trustworthiness' of a finding. But we did - in the paper referenced in our original, and less than opaquely entitled 'A proposal for judging the trustworthiness of research findings' (Gorard 2014). How secure a finding is depends on having an appropriate research design for the research question, the quality of the measurements, and a number of other factors such as diffusion, initial balance, and blinding. None of these are addressed by NNTD, or by t-tests or similar. They require judgement. In addition, the security of a finding depends on its scale and variability, the number of cases, and the threat of bias caused by missing data. All of the latter are covered by NNTD in one statistic – that is its beauty. The results from different studies can be compared directly, and the logic encompasses frequencies as well as real numbers (as illustrated in the original paper), and all forms of 'effect' sizes including R^2.

NNTD is about internal validity only (the security of a finding), and not about random sampling variation or generalisation to a population of cases not involved in the study. Therefore, all of the rest of the K&S commentary misses the point and should be ignored. It is, perhaps, not surprising that they try to defend the use of significance tests, because they both use them routinely and incorrectly. In recent years they have jointly and singly published papers using significance tests and quoting standard errors for data from populations (such as the early British birth cohort studies) or from samples with 40% attrition which they simply ignore (e.g. Sturgis et al. 2013). NNTD is nothing to do with standard errors – there are no standard errors in the cases used as examples in our paper, just as there can be none for population data or incomplete random samples. Assessing security and generalising to cases not involved in a study are very different processes, even though p-value advocates routinely conflate them, just as they either ignore missing data or assume that their significance tests have somehow dealt with it.

K&S have a long section in their commentary introducing and defining symbols, in which the needless notational confusion may lose some readers (we used Greek symbols precisely because we were not intending to generalise to a bigger population). They could have summarised all of this in one sentence. Their formula for N* (close to our NNTD) is the simply the absolute value of the 'effect' size multiplied by the number of cases in the smaller group in the comparison. In the example from Table 2 in our original paper the smallest group had 153 cases, and the effect size was +0.24, therefore N* is 37 (24% of 153). The 'true' NNTD calculating by adding rather than changing cases is slightly different. But we are happy to work with the N* version.

This in turn means that NNTD provides a clear and useful interpretation of this kind of effect size. An effect size (here Hedges g) represents the proportion of cases in the smaller group that need to be made more counterfactual (or that need to be added as counterfactual cases) before the 'effect' disappears. Multiplying the effect size by the smaller N yields a single statistic that incorporates the absolute size of the difference, the variability in measurements, and the scale. And as explained in the first paper, this figure can be directly compared to the number of missing relevant values to judge the security of the finding in face of attrition. An 'effect' size can be defined *in terms* of NNTD. For example, a study with NNTD of 100 might have an effect size of +0.5 and a cell size of around 200, or an effect size of around -0.25 and a cell size of 400. Both can be considered equivalently secure, before attrition and other factors are considered.

References

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

Sturgis, P., Brunton-Smith, I., Kuha, J. and Jackson, J. (2014) Ethnic diversity, segregation and the social cohesion of neighbourhoods in London, *Ethnic and Racial Studies*, 37, 8, 1286-1309