Did it work and will it work? (But not: does it work?)

Calum Davey, LSHTM

16 December, 2019

There is a common misunderstanding about the uncertainty regarding the results of randomised trials. Most trials are reported with a confidence interval to indicate the degree of uncertainty around the point estimate. However, there are multiple sources of uncertainty with different implications for the interpretation of the result. These are: uncertainty about the measures, uncertainty arising from random allocation, and uncertainty about sampling. The first two of these relate to uncertainty about *did the intervention work*, also known as the internal validity. Only the last of these relates the uncertainty about whether the intervention *will work* elsewhere in the future, also known as external validity. Confusion arises when commonly used methods to describe uncertainty in trials mixes these together. With the help of an illustration, I will show how these different sources of uncertainly relate to the kinds of inferences we hope to make from trials, and propose alternatives to the common standard-error-based method.

Imagine that you have have completed a trial of an educational intervention, with 1000 children in each arm. The effect on the primary outcome, maths ability, was 0.2 standard deviations of difference on average between arms (see table 1). There was also a secondary outcome, experience of bullying, which was binary, and there was a risk ratio of 0.8 favouring the intervention arm.

The first question that may arise is, 'are these differences due to the intervention or are they due to something else?'. Assume we are able to rule out any systematic biases, such as missing data or lack of blinding. We are still left with the possibility that some or all of the differences observed were due

Table 1: Mean maths score and percentage reporting bullied in each arm.

Arm	Maths	Bullied
0	37.7	9.3
1	39.9	7.7
	0.2	0.8

to chance. Note that the uncertainty is not about what the measured differences were — those are fixed by the results of the study — but the uncertainty is whether those were entirely caused by the intervention alone.

At this point, the evaluator may now estimate the confidence intervals around these point estimates by calculating the standard error, which is proportional to the standard deviation for the continuous primary outcome and the proportions for the binary secondary outcome, and inversely proportional to the sample size. This, however, is too hasty, as risks lumping too many sources of uncertainty into one measure. The standard error is based on sampling theory and captures the uncertainty regarding the internal validity and the external validity in one metric.

Let us instead break the uncertainty down into its different sources. The first concern is that the measures themselves have uncertainty. The test-re-test reliability of the maths exam that we used will not be 100%, and would typically be in the 60-80% range for a good test. Similarly, the bullying measure that we used may have imperfect sensitivity and specificity, so that there are some false positives and some false negatives in our samples. Therefore, there is a risk that some of the difference between the arms of the trial is due to the random measurement error. Most evaluators do not assess the reliability of the measures in their sample and so the analysis will need to use measures external to the trial, ideally published by other independent researchers.

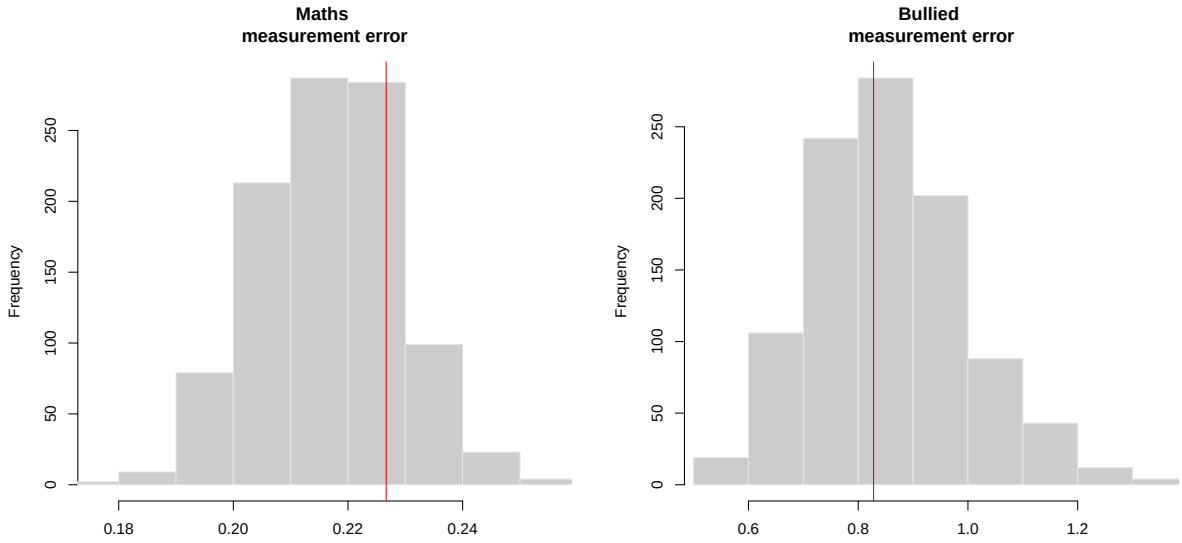

Figure 1: Histograms of the results after applying the measurment error, with the observed results shown as a red vertical line.

Looking again at our trial, we can model the implications of the reliability of the measures on our certainty that the results were due to the intervention alone. We can estimate the distribution of effects that we might have seen under different realised measures. Figure 1 shows the distribution of effects that would be observed if the tests were repeated on the same children. Here we have assumed that the test-re-test distribution for the maths test has a standard deviation of two, and the specificity and the sensitivity of the bullying test are 95% and 90%, respectively. This figure gives us an impression of the range of results that could have arisen from this one group of chilren because of measurement error alone.

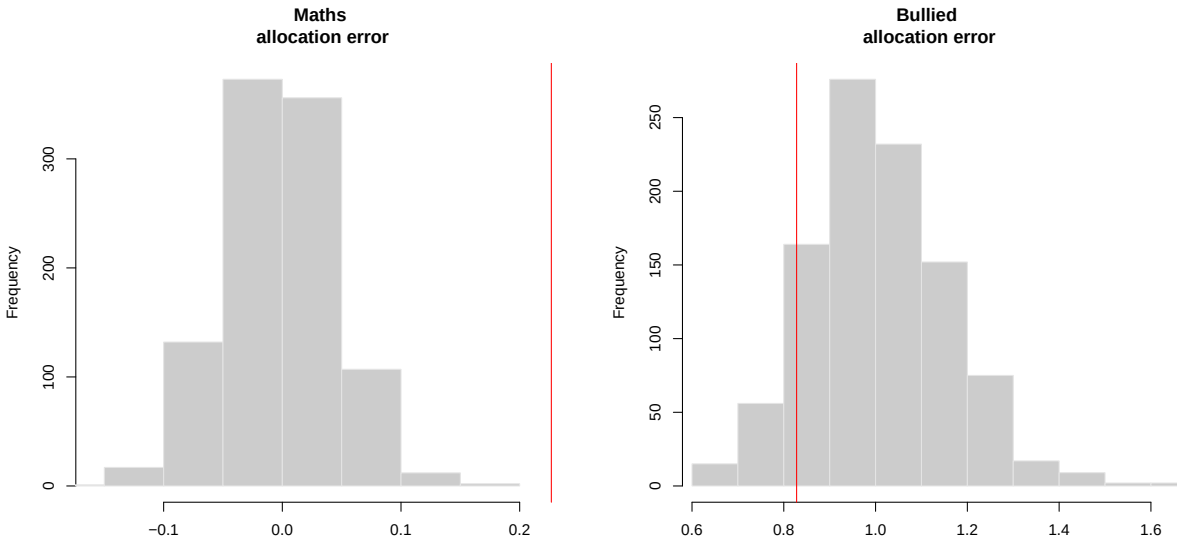

Figure 2: Histograms of the permuted-allocation results with the observed results shown as a red vertical line.

The second source of random error that is relevant to the question 'did it work here' is the uncertainty arising from the random allocation. This source of random error is the closest to the ideal random-sampling error found in textbooks (but rarely in practive) since the true randomness is actually observed (assuming that there is no missing data or drop-out). This error raises the possibility that the difference observed between the arms could be in part or entirely due to the random variation in the mean outcome that would arise whatever selection of participants were allocated to the treatment or to the control. Fortunately, there is a very simple way to assess how the size and direction of the effects observed compares to the variation that would arise naturally without any causal effect. All we need to do is to repeat the random allocation procedure (i.e. including any stratification or restrictions) many times, and see how the observed effect compares to the distribution of possible alternatives. If there is a high proportion of random allocations that would have resulted

in as large or larger effects in the same direction as the observed result then this might suggest that some or all of the observed result was just due to random variation and not because of a causal effect. However, if the observed result was larger or smaller than the vast majority of the other possible allocation results then this provides evidence that the effect was not due to this source of variation. The results of permuting the allocatoin of the 2000 children to the two arms are shown in Figure 2.

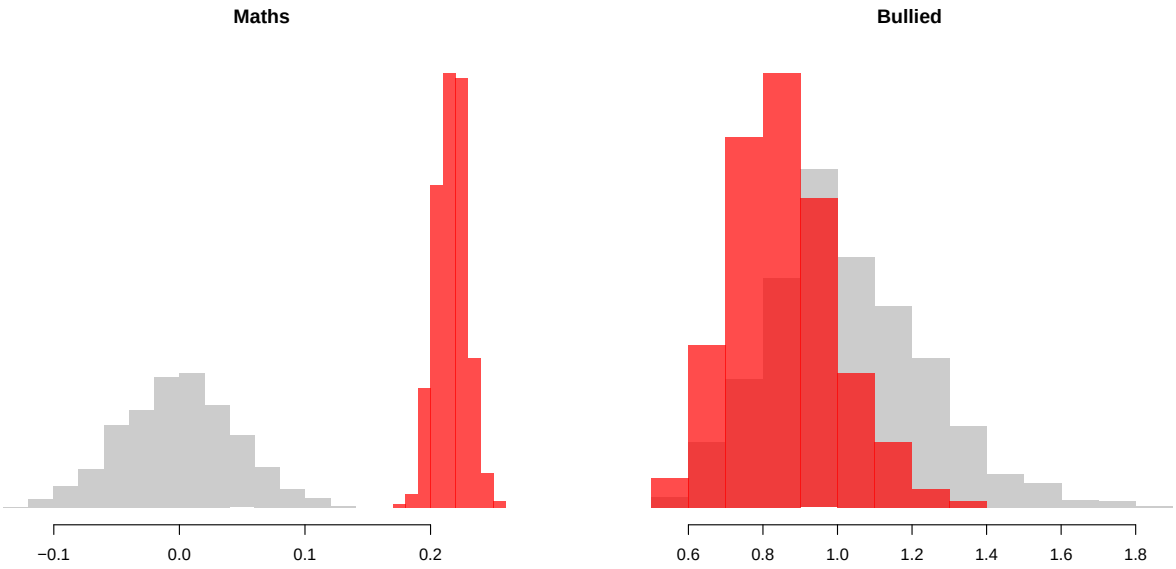

Figure 3: Combined plot of the permuted-allocation results with measurement error superimposed around the observed effect.

The two sources of variation discussed thus far relate to the question of whether the intervention had an effect on the outcomes in this case. This is the internal validity of the result, and before discussing the other evaluation question — will it work — we may want to develop a single summary of the internal validity regarding random error. Combining the effects of measurement and allocation error is most easily done with simulations and permutations to generate a single summary distribution of the possible effects and thus derive a range of effects that should include the true causal effect of the intervention in this evaluation. Figure 3 superimposes the error from the measurement on the error from the random allocation. The observed effect is the sum of the true causal effect, the result of measurement error, and the result of random variation in the mean outcome in each arm resulting from the allocation procedure. Since the measurement and allocation errors are random, by definition, the range will centre on the observed effect. Ideally, a reasonable range (e.g. 95%) should be presented without the observed effect itself because although the observed effect is the most likely causal effect, the likelihood that the true effect is or is very close to the observed effect is so small that presenting it

will vastly over-estimate its importance for inference.

Having established the range that most likely includes the true causal effect of the intervention in this case, the evaluator will ask, 'will this result be seen elsewhere?' This is the issue that is often called the 'external validity' of the result. And often evaluators will discuss the representativeness of the people and places included in the study of a wider population. It is, however, nearly always a discussion of various factors, since random selection is highly uncommon, or and usually impossible (not least because people are ethically obliged to be given the option to opt-out). This nuance is absent from what usually happens next, which is that evaluators calculate a confidence interval using a standard error. The standard error, and associated confidence interval (usually approximately two times the standard error above and below the observed effect estimate) captures the allocation variation discussed above — which is truly random — and sampling variation in the selection of the participants — which is never random — in an attempt to describe the distribution of effects that would be observed if random samples were drawn from a theoretical population. Many people have pointed out that this is problematic because the observed sample was not drawn at random and therefore is not a case in the theoretical random sample that could be taken from the theoretical population. However, what is also problematic is that the variation that is described muddles issues about the internal validity, the allocation variation, and external validity, the distribution of effects drawn at random from a population (were that possible). While the first issue has been rightly highlighted because interpreting the confidence interval as a distribution that should include a 'true' effect is unwise, the latter has has the reverse problem, where confidence intervals are used to infer whether the intervention had an effect *in the observed evaluation*. The liability of the estimate to random variation will be overestimated, since the confidence interval is designed to make inferences about a wider population, not particular cases.

The evaluator asking 'will this work elsewhere' should not use confidence intervals unless the sample was very nearly at random. Evaluators should resist the temptation, and requests of funders, to provide an answer to the question 'does this work' in the imperative sense, and focus on *did* it work and offering advice on wether it *will* work in other specific contexts. This should be discussed in terms of the similarities between places, the availability of necessary contextual factors, and the likelihood of factors supporting the key mechanisms that explain the effect being active in another context. This is hard work, and not the work of statistics. Answering 'will it work elsewhere?' is essentially a prediction problem, and even with the best science, prediction is difficult.

In summary, the following recommendations are offered to reduce confusion about random error in trials:

1. inference regarding 'did it work' requires careful assessment of systematic biases, data on measurement error, and the subjective interpretation of the results of permutation tests.
2. inference regarding 'will it work' requires data on the original and target contexts and the contextual factors that matter for making the prediction.
3. confidence intervals based on standard errors should only be used when the participants are randomly sampled from a population and the objective is to estimate an overall population-level effect of the intervention.

The implication of (1) is that off-the-shelf so-called statistical 'tests' will be insufficient. Statisticians can use simple simulation and permutation procedures instead, and example code is provided in an Appendix in Stata and R. Recommendation (2) implies that making inferences for elsewhere requires deep, grounded, theoretical scientific work, and cannot be achieved by statistics alone. The implication of (3) is that standard errors and confidence intervals should no longer appear alongside the result of 99% of trials, and thus hopefully the confusion about what it is they mean will end, and the confusion about what is being inferred will start to diminish. Together this should lead to better science.

1 Code Appendix

```
# Load data.table package
library(data.table)
library(scales)
# Generate example data
data <- data.table(
  ID=seq(1,2000,1), # Unique ID for each child
  Arm=rep(x=c(0,1),each=1000), # Control (0) and intervention (1) arms
  Maths=c(
    rnorm(1000, mean=38, sd=10), # Control-arm maths
    rnorm(1000, mean=40, sd=10)  # Intervention-arm maths
  ),
  Bullied=c(
    rbinom(n=1000, size=1, prob=0.10)*100, # Control-arm bullying
    rbinom(n=1000, size=1, prob=0.07)*100  # Intervention-arm bullying
  )
)

# Trial analysis
analysis <- function(arm=data[[2]], maths=data[[3]], bullied=data[[4]]){
  c((mean(maths[arm==1]) - mean(maths[arm==0]))/sd(maths),
     mean(bullied[arm==1])/mean(bullied[arm==0])
  )
}

# Trial results (Table 1)
m <- data[,c(mean=lapply(.SD, mean)), by=Arm, .SDcols=c('Maths', 'Bullied')]
m <- rbind(as.matrix(m[,1:3]), c(NA,analysis()))
# Measurement error
# mtrt <- .8 # Maths test-retest reliability
mt_sd <- 2 # SD of the test around true value in expectatoin
bmss <- c(.95, .9) # Bullying measure sensitivity and specificity

rs1 <- replicate(1000, {
  data[, prev_Bullied := mean(Bullied)/100
       ][, `:=` (Maths_error = rnorm(1, rnorm(1, Maths, mt_sd), mt_sd),
                 Bullied_ppv = (bmss[1]*prev_Bullied)/
                   ((bmss[1]*prev_Bullied+(1-bmss[2])*(1-prev_Bullied))),
                 Bullied_npv = (bmss[2]*(1-prev_Bullied))/
                   (bmss[2]*(1-prev_Bullied)+(1-bmss[1])*prev_Bullied)),
                 by=ID][,
                   Bullied_error := rbinom(n=2000, size=1,
                                           prob=ifelse(Bullied!=0,Bullied_ppv,(1-Bullied_npv)))*100
                 ][]
  analysis(maths = data[['Maths_error']], bullied = data[['Bullied_error']])
})

par(mfrow=c(1,2), cex=.9)
hist(rs1[1,], main='Maths \nmeasurement error', col='gray80', border='gray90', xlab='',
     xlim=c(min(min(rs1[1,]), m[3,2]),max(max(rs1[1,]), m[3,2])))
```

```
abline(v = m[3,2], col='red')
hist(rs1[2,], main='Bullied \nmeasurement error', col='gray80', border='gray90', xlab='',
     xlim=c(min(min(rs1[2,]), m[3,3]), max(max(rs1[2,]), m[3,3])))
abline(v = m[3,3], col='red')
permute_analysis <- function(arm=data[[2]], maths, bullied){
  analysis(sample(arm), maths, bullied)
}

rs2 <- replicate(1000,{
  permute_analysis(data[[2]], data[[3]], data[[4]])
})

par(mfrow=c(1,2), cex=.9)
hist(rs2[1,], main='Maths \nallocation error', col='gray80', border='gray90', xlab='',
     xlim=c(min(min(rs2[1,]), m[3,2]),max(max(rs2[1,]), m[3,2])))
abline(v = m[3,2], col='red')
hist(rs2[2,], main='Bullied \nallocation error', col='gray80', border='gray90', xlab='',
     xlim=c(min(min(rs2[2,]), m[3,3]), max(max(rs2[2,]), m[3,3])))
abline(v = m[3,3], col='red')

rs3 <- replicate(1000, {
  data[, prev_Bullied := mean(Bullied)/100
       ][, `:=` (Maths_error = rnorm(1, rnorm(1, Maths, mt_sd), mt_sd),
                 Bullied_ppv = (bmss[1]*prev_Bullied)/
                   ((bmss[1]*prev_Bullied+(1-bmss[2])*(1-prev_Bullied))),
                 Bullied_npv = (bmss[2]*(1-prev_Bullied))/
                   (bmss[2]*(1-prev_Bullied)+(1-bmss[1])*prev_Bullied)),
                 by=ID][,
                   Bullied_error := rbinom(n=2000, size=1,
                                           prob=ifelse(Bullied!=0,Bullied_ppv,(1-Bullied_npv)))*100
                 ]
  permute_analysis(maths = data[['Maths_error']], bullied = data[['Bullied_error']])
})

par(mfrow=c(1,2), cex=.9)
b <- c(max(hist(rs1[1,], plot=F)$density),
       max(hist(rs1[1,], plot=F)$breaks))
hist(rs3[1,], main='Maths', freq=F,
          col='gray80', border=NA, xlab='', yaxt='n', ylab='',
          ylim=c(0,b[1]),
          xlim=c(min(min(rs3[1,]), m[3,2]),
                 max(max(rs3[1,]), m[3,2], b[2])))
hist(rs1[1,], add=T, freq=F, col=alpha('red',.7), border=NA)

b <- c(max(hist(rs1[2,], plot=F)$density),
       max(hist(rs1[2,], plot=F)$breaks))
hist(rs3[2,], main='Bullied', freq=F,
          col='gray80', border=NA, xlab='', yaxt='n', ylab='',
          ylim=c(0,b[1]),
          xlim=c(min(min(rs3[2,]), m[3,3]),
                 max(max(rs3[2,]), m[3,3], b[2])))
hist(rs1[2,], add=T, freq=F, col=alpha('red',.7), border=NA)
```