Inference from RCTs: did it work, will it work, 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 0.7 favouring the intervention arm.

The first question that may arise is, 'are these differences due to the intervention or are they due to

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

Arm	Maths	Bullied
0	38.4	10.7
1	40.1	7.4
	0.2	0.7

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 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%, typically be in the 60-80% range for a good test. Similarly, the bullying measure may have imperfect sensitivity and specificity, so there are false positives and false negatives in our samples. 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.

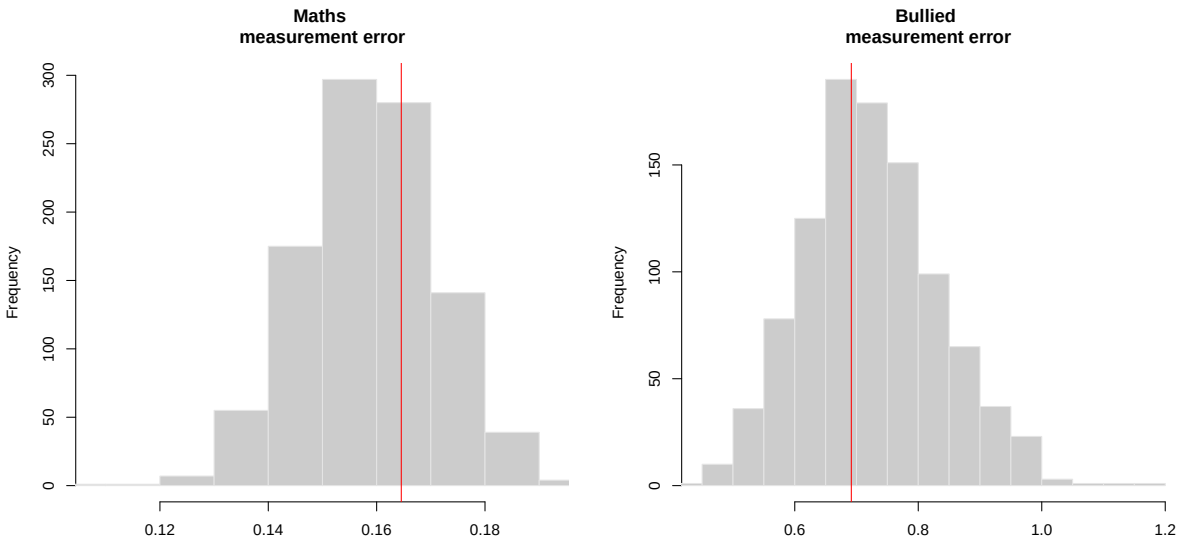

Figure 1: Histograms of the results after applying the measurment error, with the observed results shown as a red vertical line.

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.

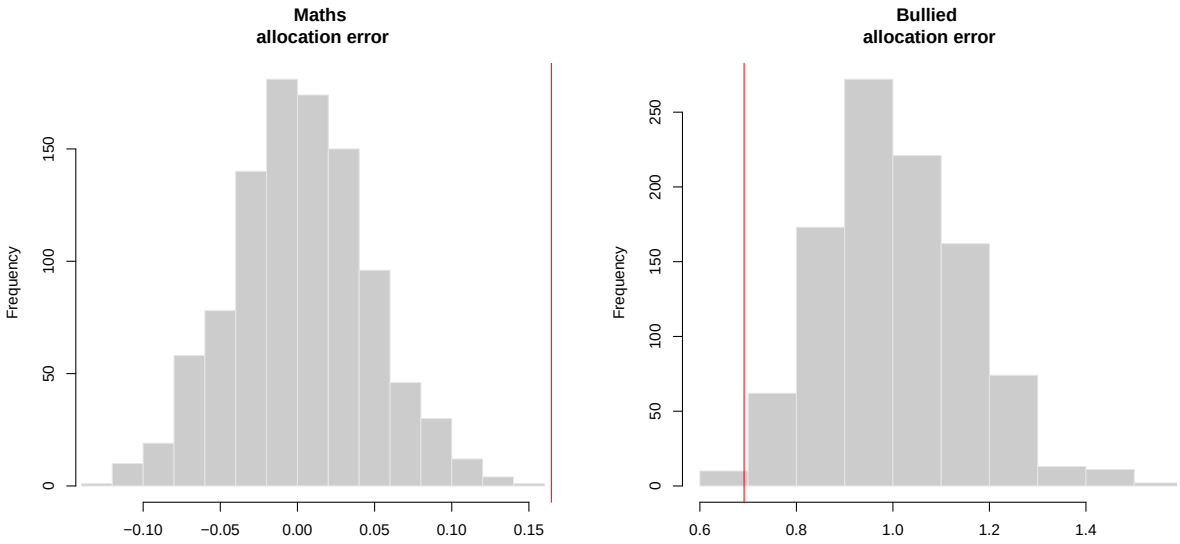

Figure 2: Histograms of the permuted-allocation results with the observed results shown as a red vertical line.

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 practice) 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.

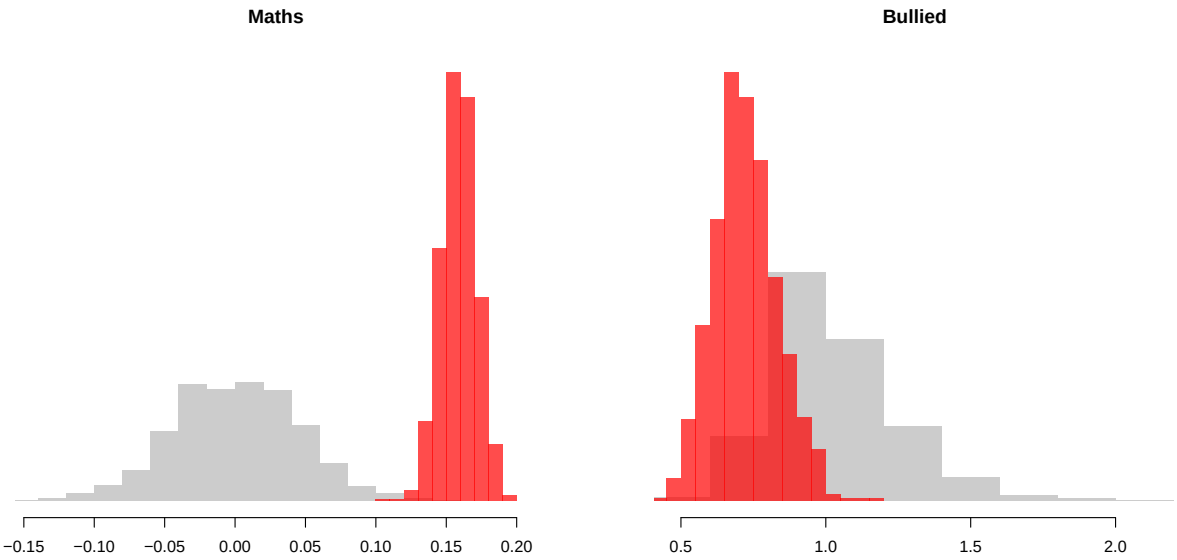

Figure 3: Combined plot of the permuted-allocation results with measurement error superimposed around the observed effect.

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 given the error from measurement and allocation and thus derive a range of effects that should include the true causal effect of the intervention in this evaluation. Figure 3 shows the distribution of effects that could be observed in this sample of children given the measurement and permulation error effects (in grey), with the observed effect, with measurement error, supermimposed (red). It may be possible to use the combined distribution shown in grey to generate summary statistics of likely effect sizes[1]. If a range of probably effects was presented, e.g. 95%, the point estimate could be ommitted since the chance that the true effect very close to the observed effect

[1] *Not sure degree to which would need Bayesianism to formalise this step* – CD

is small and presenting it could over-emphasise its importance for inference.

Having established whether the intervention most likely did or did not have an effect in this sample, the evaluator can then ask, 'will this result be seen elsewhere?' This is the issue that is often called the 'external validity' of the result. Evaluations are almost never strictly 'representitive' of a wider population because the participants (or clusters) are rarely randomly sampled.[2] Evaluators should resist the temptation, and requests of funders, to provide an answer to the question '*does* this work?', and focus on *did* it work and offer advice on whether it *will* work for other specific populations (including the same population at a different time) based on the evaluation. 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, and not only the work of statistics. It is essentially a prediction problem, and even with the best science, prediction is difficult.

Despite nuanced discussion of external validity, evaluators will often simply 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 variation discussed above — of which allocation is truly random — and sampling variation in the selection of the participants — which is almost 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 highlighted the problem that the observed sample was not drawn at random and the assumptions behind the standard error are not met. The evaluator asking 'will this work elsewhere?' should not use confidence intervals unless the sampling was at random. However, it is also problematic that the standard error muddles internal validity and external validity. 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 is a different 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.

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

[2] And they are never randomly sampled in time.

1. inference regarding 'did it work?' requires careful assessment of systematic biases, assessmet of 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.

1 Code Appendix

```r
# Generate the simulation data
#============================

# Load packages
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
#==================

# Assume the reliability of the measures

mt_sd <- 2 # SD of the test around true value in expectatoin
bmss <- c(.95, .9) # Bullying measure sensitivity and specificity

# Re-sample the data based on the reliability of the measures

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']])
})

# Plot the results

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') # Add the observed trial result

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')

# Allocation error
#=================

# Create function to run permutations of the analysis

permute_analysis <- function(arm=data[[2]], maths, bullied){
  analysis(sample(arm), maths, bullied)
}

# Run permutations

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

# Plot the results

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') # Add observed trial result

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')
# Measurement and permuation error
#==================================

# Permute the re-samples of the data based on the measurement reliability
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']])
})

# Plot the results
par(mfrow=c(1,2), cex=.9)
b <- c(max(hist(rs1[1,], plot=F)$density),
       max(hist(rs1[1,], plot=F)$breaks))

# PLot the permutations
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])))

# Plot the observed values with measurement error
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))

# PLot the permutations
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])))

# Plot the observed values with measurement error
hist(rs1[2,], add=T, freq=F, col=alpha('red',.7), border=NA)
```