*This document gives a fuller treatment of the mathematics behind Lehr’s approximation, the exact calculation of minimum sample size, and what the exact test should be in the examples of 50, 100 and 540 cases in each group along with how the outcomes of the simulations match the predicted outcome in the theory.*

**Lehr’s approximation**

In Gorard’s textbook, published only in 2013, he discusses at some length the minimum sample size needed to identify any particular effect size and also the reasons why it is important to use an appropriate size of sample. For some reason, his article contains no mention of the sample size needed to reliably return a significant test outcome when there is the specific effect size at work (80% is the power he uses for this informal use of ‘reliably’).

Gorard claims that in 10,000 replications his simulation produced 12% significant differences under conditions in which the logic of hypothesis testing says there should be (approximately) 24% significant differences. However he has used a difference of means of 10% in the simulation using U(0, 1) and U(0, 1.1) distributions

The calculation of Lehr’s approximation for 10% difference in means is:



So Gorard is using a sample size which is a fraction of what his own textbook says should be used to reliably (80% power) identify a 10% difference in means because the effect size for a 10% difference in means in his simulation is tiny – at 0.173.

**Exact sample size calculation for test to have 80% power:**

There is actually no need to use an approximation in this case as the exact value of minimum sample size for 80% power can be computed quite straightforwardly:
With sample size *n* in each group the critical values for a two-tailed 5% test will be .

If the second group is now uniform on interval (0, 1.1) the standard error of the difference in means will be .

In order for 80% of repetitions to reject the null hypothesis when there is a 10% increase in the mean (from 0.5 to 0.55), *ngroup* has to be such that 

Solving this gives *ngroup* = 539.64, and simulating the situation Gorard describes using 540 in group 1 and in group 2, does give 80% (on average) of repetitions generating a statistically significant result.

The test now is: ‘Accept H0 (no difference in means) if ’ and when the second group has a 10% increase in the mean you have:





Meaning that the theory predicts that 80.05% of the simulated tests will show a statistically significant difference in means between the two groups with a sample size of 540. The simulations we provide do return this proportion – around 80% (on average).

**Gorard’s example:**

Using the Central Limit Theorem the distribution of will be approximately Normal, giving 

The (two-tailed 5%) test is:
 ‘Accept H0 (no difference in means) if ’ since under H0 the two groups have the same distribution. [[1]](#footnote-1)

Investigating a 10% difference in means, using the Central Limit Theorem the distribution of is also approximately Normal, based on an exact distribution of *U*(0, 1.1):






So Gorard’s observation of 12% significant results in simulating this situation is nowhere near the ballpark expected (24%), even you when you take into account that he is using a *t* test with (we think) equal variances assumed which will systematically underestimate the proportion of significant outcomes.

The simulations we provide do return this proportion – around 24% (on average).

**Did Gorard actually use 50 cases but report using 100 cases in each group?**

Using the Central Limit Theorem the distribution of will be approximately Normal, giving 

The (two-tailed 5%) test is:
 ‘Accept H0 (no difference in means) if ’ since under H0 the two groups have the same distribution.

Investigating a 10% difference in means, using the Central Limit Theorem the distribution of is also approximately Normal, based on an exact distribution of *U*(0, 1.1):






So Gorard’s observation of 12% significant results in simulating this situation is close to the ballpark expected (15%), when you take into account that he is using a *t* test with (we think) equal variances assumed which will systematically underestimate the proportion of significant outcomes.

The simulations we provide do return this proportion – around 15% (on average).

**Summary:**

Gorard’s paper makes no mention of the importance of the sample size in being able to identify any given shift in mean, despite writing in detail about this in a recently published textbook.

Simulations of the situation described by Gorard, using 50, 100 and 540 as the sample sizes, give proportions of statistically significant differences which are consistent with what is predicted by the theory that Gorard claims is discredited by his experimental results.

Gorard impugns the competence and integrity of people who use significance tests, claiming they are acting unethically, but we believe he has grossly misrepresented what his simulations show by using a sample size he should know is far too small to reliably identify a 10% shift in mean in the really bad model he has chosen to work with.

We believe that our simulations show that the hypothesis test logic is not flawed at all – it will consistently return proportions of significant outcomes which are in line with the theory.

1. Note that the critical value for the test is greater than the 10% difference in mean that Gorard hopes to be able to detect with this test so it is clear that this sample size will give a significant test result less than half the time. [↑](#footnote-ref-1)