Getting granted

Geoff Norman¹

Published online: 31 March 2018

© Springer Science+Business Media B.V., part of Springer Nature 2018

The academic game is not easy. As I've said elsewhere, it is a star system that makes the NFL, NHL or Hollywood look easy. Even after you've cleared the many hurdles that accompany undergraduate and graduate work on your way to a B.Sc. or Ph.D., many more challenges remain. I am not thinking about the major obstacles of promotion and tenure, but the ongoing day to day hassles associated with the research game and the soul-destroying rejections that frequently accompany the submission of grant proposals and publications. The ecology of these two activities is sufficient to induce chronic depression. I haven't been able to get much statistics about funding rates at national agencies; Canadian Institute of Health Research (CIHR) indicated a funding rate of 15% in 2016. Acceptance rates at the mainstream medical journals—JAMA, BMJ and NEJM are about 5%. In medical education, acceptance rates of the top journals are all less than 20%, which means that the chance that bright idea will end up in print might be about $.15 \times .20 = 3\%$.

I have published a couple of editorials to help authors get their papers published. One highlighted errors authors make (Norman 2014), another described strategies to deal with amateur statisticians who raise all sorts of predictable objections to analytical methods, most of which are specious and irrelevant (Norman 2010). However, both commentaries are really directed at the publication stage, and aren't much use if you have nothing to publish. So this issue's editorial is directed at the first hurdle—getting permission to do the study in the first instance.

Permission to do studies is not simply a matter of getting funding from a grant agency. Even before that happens, most agencies require evidence that the study has been approved by the appropriate research ethics board. I'm not going to pursue that avenue; lots has been written about ethics of educational research. And while we all accept that ethics review is a "good thing", it does sometimes go off the rails. Recently a researcher was going to ask an opinion of her co-investigators about some issue related to the research they were doing and the issue arose whether she had to get ethics approval to talk to her co-investigators. It could be more extreme I guess. In the tradition

✉ Geoff Norman
norman@mcmaster.ca

¹ McMaster University, Hamilton, ON, Canada

of “autoethnography” (Varpio et al. 2012) does one have to sign a consent form to ask oneself a question?

A second internal hurdle that occasionally arises is to receive permission from the programs involved. In particular, if we are studying students, it seems right and just for the program in which the students are registered to review the proposal and ensure that it does not cause undue hardship or waste of time.

Now it would seem that issues of methodology are the purview of the funding agency, ethics is handled by the local REB and logistics by the relevant program. The three Venn circles don’t overlap. But in the real world of academe (is that an oxymoron?), the circles DO overlap—sometimes a lot. Again, some of this is reasonable. After all, a study with inadequate methodology is, by definition, unethical. And the program people, who are closer to the students than the REB, may be sensitive to some ethical issues that the REB missed. But what this means is that there are at least three critical places where you may encounter the kinds of questions and criticisms I’ll talk about below. So this editorial is designed to help you get prepared and have some reasoned responses.

This is not hypothetical. Some of the questions arose from an internal committee that recently reviewed a study proposal we had put together. The study was actually a reasonable learning task; we wanted to see if getting medical students to learn about sensitivity, predictive value and likelihood ratios of diagnostic tests, then work through ten clinical cases with feedback, would give them an understanding of the role of test characteristics in patient care. But the committee’s questions were not at all at this level. Below is the list of questions and the answers I didn’t send (in part because they didn’t ask some of these questions). Following each I explain why it was the wrong question to ask.

1. How will participants be allocated to groups?

We will use a die and cup approved by the Ontario Lottery and Gaming Commission. To ensure blinding the die will be rolled by one research associate wearing a blindfold and read by a second RA. If it lands on 1 or 2, they will be assigned to Group A; 3 or 4 to Group B, and 5 or 6 to the control group To ensure no accumulated bias, the die and cup will be replaced by a new one every 72 hours

Explanation Well, we could use a die and cup, but none of this is really necessary.

From time to time someone will write a paper lambasting educational researchers for not doing enough randomized controlled trials, under the assumption that the RCT is the best way to achieve scientific nirvana. A rational response would point out that, while an RCT may be a very good design to examine the effectiveness of an intervention when it can be standardized, and when there is uniform agreement about a reasonable outcome measure, many questions in education (and for that matter, in clinical research) do not lend themselves to experimental design.

On the other hand, we really are doing an intervention here. So what’s the problem? None, really. We could use the throw of the dice to get folks into different groups. But that would require having someone there to throw the dice. Further, if we were doing an intervention where there is some face to face instruction, it would be very difficult to

deal with possible contamination and unblinding, as students could talk to each other. It would be a lot easier to have all students at each site do one arm of the study. But of course this is not true randomization.

This brings us to the real point, that random allocation is a means to an end, not an end unto itself. The goal is to ensure that whatever strategy got students into different groups, it is unlikely to be related to the study outcome. So in my case, if students picked a clerkship site based on proximity to their home, I'm pretty sure that the neighbourhood you live in is not related to how well you can learn about diagnostic tests.

Another consideration is that trialists have good reason to worry a lot about whether their groups are equivalent, for the simple reason that the effects they are looking for are so small they can easily vanish in a sea of noise. A typical effect size for an intervention like antihypertensive therapy to avoid strokes is 0.03 (Meyer et al. 2001). Conversely, a meta-meta-analysis of 302 meta-analyses done a number of years ago (Lipsey and Wilson 1993) showed that the average effect size of educational interventions was .45. The study also showed that randomized and non-randomized interventions has exactly the same effect size. So the bottom line is that as long as the allocation to different groups is unlikely to be related to the outcome of interest, press on.

2. How will you address baseline confounding?

Ideally, we would like to have baseline variables that are related to performance on the diagnostic task. However, these do not exist. We could, perhaps get information on prior academic achievement, but this would create significant ethical problems, in terms of disclosure, as well as methodological problems in terms of determining equivalence in different programs. We might consider measuring IQ using a brief IQ test, but such tests are not available. So after some discussion, we have decided to use head circumference as a surrogate for intelligence. Of course, since head circumference is also related to physical size, we will measure other variables related to size, including weight, height and shoe size. And as males are on average larger than females, we will code for gender, recognizing that there are potentially 15 or 20 categories. And as weight increases with age, we will collect age data. All these will be used as covariates (6), as well as interactions (15).

Explanation This question could easily generate a whole chapter in response. First of all, a definition. I think what they mean by “baseline confounding” is that some unspecified variables may be (a) different between the groups and (b) potentially related to the outcome. At one level, this is a non-issue. If we did randomize to groups, then whatever baseline variables are kicking around, they will be equivalent among the groups except for the operation of chance, and that's precisely why we do statistics (if we don't randomize, see question 1 above). So baseline confounding arises purely by chance and is dealt with in the analysis.

This fundamental axiom reveals the fallacy in randomizing to groups then doing a statistical test on a bunch of baseline variables to see whether the “randomization failed”. Three problems arise. First, the whole logic is illogical. You do a statistical test to determine the likelihood that a difference in some baseline variable could have arisen by chance (That's what $p < .01$ means!). But it did arise by chance; the probability is 1, by definition. Second,

if you don't see a significant difference on any baseline variable, grab a bunch more. Sooner or later, one will be significant at $p < .05$ (Worrall 2010). And at that point a third problem arises—there is really no defensible way to correct for it anyway. You can ANCOVA away to your heart's delight but cannot ever know if you're over or under-correcting.

The second issue is that none of this matters if the baseline variable is not related to the outcome. But if you're doing a study like ours, where the instructional materials and outcomes are designed for the study itself, there is no way to know a priori whether any variable is related to the outcome anyway. You can only know that after the fact. And as above, even if the ones you look at are not related you can never be sure that there is some other variable lurking in the woods that is related. However it is probably a safe bet that variables like age, gender, marital status, etc. are not related to understanding of diagnostic tests, so don't bother measuring them.

There is one variable that usually is related to the outcome, better than any other—the outcome itself. In short, why not do a pretest–posttest design? Actually, there are a couple of very good reasons. First, in the situation we're in, it is unlikely that the students will have much prior knowledge of test characteristics, so the pretest may well not be related to the outcome. And if so, then subtracting a pretest score simply introduces noise and reduces degrees of freedom. Second, the pretest becomes part of the intervention and tells students precisely what the outcome will look like. So this really can be a source of contamination. Paradoxically, an attempt to correct for contamination by introducing a pretest may introduce contamination!

3. A sample size estimation is reasonable to justify enrolments and resources used. Do you have a sense of the magnitude of difference you expect to see between groups?

We have no information on how people will perform on these tests, since we developed them for the study. For a sample size calculation, we have arbitrarily decided to use a mean score of 70% with a SD of 15% in Group C. As no one has attempted an intervention of this sort, we have no information on treatment effect. So we have decided to use an estimated treatment effect of 9.49. This results in a sample size 40/group as we suggested in the original proposal. Our original calculation used a treatment effect of 5.0, but this resulted in a sample size of 144, which is too large. So instead we decided to go with 9.49.

Explanation Sample size calculations are yet another holdover from the RCT ethos. When you're doing a multi-zillion dollar study, every patient you enroll might cost thousands of dollars, so it makes sense to keep sample size as low as possible. On the other hand, the risk of this is a serious problem of non-replication (Ioannidis 2005). If you do a study and the p value is exactly .05, the chance of replicating the study and rejecting the null hypothesis is 50% (not 5%) since the best guess of the mean for the alternative hypothesis is right at the critical value where $p = .05$, giving a beta error of .50. Further, if it's a drug trial there is a very good chance that there have been other trials of similar drugs using similar outcomes (like death). So you have good data on which to base a sample size, and a good reason to do one.

But education isn't like that. As our response indicates, no study has ever looked at these interventions and none has ever used this outcome. To do a reasonable sample size

we would need to know (a) the difference between groups on the outcome, and (b) the standard deviation within groups. We have neither.

Which is, actually very good news. REBs typically demand sample size calculations. And given the uncertainties we have indicated, we can go ahead and do it with complete assurance that the sample size will come out exactly as we want it. If it doesn't, we can diddle things until it does.

And now for a very well kept secret. You can save a ton of money in statistical consultations by remembering the formula:

$N \text{ equals sixteen } S \text{ squared over } d^2 \text{ squared } (N = 16 s^2/d^2)$. (Lehr 1992)

As it turns out, for a simple two group comparison, with $\alpha = .05$ and $\beta = .20$, the formula comes out to almost exactly a multiplier of 16. So you can easily do sample size calculations for 2 groups in your head, dazzle your colleagues and save money on stats consults.

Finally, a simple pragmatic strategy. Go with history (Norman et al. 2012). Things in education of any value have effect sizes about .5. That works out to 64/group by the above formula. So go for a sample size of 30–90. If you can't find it with that, it's likely not worth finding.

4. How will you ensure that participants are blinded to the intervention?

The instruction will be read aloud by a confederate behind a tall screen that is blocking the participant's view. As a second level of blinding, there will be a 300 W spotlight shining in the participants' eyes.

Explanation Yet another RCT holdover. In clinical trials, blinding is very important because of (a) placebo effect, which is typically about a 30% improvement (hence the effectiveness of chiropractic, homeopathy, aromatherapy, massage therapy, therapeutic touch, etc., etc.) and (b) as above, even real effects are small. Further it's practical for medications—we can create pills that look like the real thing but aren't. A bit more complicated for surgery, but it has been done. And we can blind everyone who is involved (I once did a triple blind study—even the researchers didn't know who was who. But that's for another day).

In education, it is a very sad day if students in an educational group didn't notice what instruction they got. Enough said.

5. How will you ensure that any learning is not just a consequence of the novelty of being studied (Hawthorne effect)?

See 4 above. We will tell them that their performance counts toward the final grade, so they'll think it's just part of the course, so can be crammed at the last minute. To pass ethics, we cannot ask for any identifiers, but we'll hope they don't notice that we have no way of crediting their record with the score.

Actually, the Hawthorne effect is one of those many myths in education. A brief history is in order. The Hawthorne studies took place in the relay room at the Western Electric plant

in Hawthorne IL in the 1920s. To increase productivity, the psychologists decided to introduce some changes to ease the work environment. They turned up the light level—and productivity went up; they made more rest breaks—and productivity went up, they shortened the day—and productivity went up. Everything they did, including reverting to previous conditions, made productivity go up. The conclusion was that the actual intervention was that the researchers showed they cared about the workers. And the Hawthorne effect was born. However, years later it was discovered that other effects could explain the increase (Levitt and List 2011). But by then the legend was born. And in 1984, a systematic review (Adair 1984) revealed that 9 of 13 studies showed no evidence of a Hawthorne effect in education.

6. There is a good argument to the extent that any immediate outcome is not real learning. We suggest that you don't bother with the immediate test and instead bring participants back in 30 days.

While we would like to, they will be long gone and on a different rotation at that point. So compliance would be so bad we couldn't do the study.

Explanation The comment is worthwhile at some level. Indeed it would be nice if more studies built in some kind of follow up assessment to ensure that meaningful enduring learning had occurred. However the rejoinder is also worthwhile. Unless the participants are truly a captive audience and/or inducements are outrageous, follow up assessments are notorious for poor response rates. But a more central point is that if there is no effect on immediate test, there is very likely to be none later. Learning does not grow with the passage of time like sourdough bread—it decays. So there is little point in engineering an expensive followup unless a difference is demonstrated from the outset.

Conclusions

Hopefully these questions and answers may make it just a bit easier for researchers in education to get studies funded.

References

- Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, 69(2), 334–345.
- Ioannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, 294(2), 218–228.
- Lehr, R. (1992). Sixteen S-squared over D-squared: A relation for crude sample size estimates. *Statistics in Medicine*, 11(8), 1099–1102.
- Levitt, S. D., & List, J. A. (2011). Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments. *American Economic Journal: Applied Economics*, 3(1), 224–238.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48(12), 1181.

- Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*(2), 128.
- Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632.
- Norman, G. (2014). Data dredging, salami-slicing, and other successful strategies to ensure rejection: twelve tips on how to not get your paper published. *Advances in Health Sciences Education*, *19*(1), 1–5.
- Norman, G., Monteiro, S., & Salama, S. (2012). Sample size calculations: should the emperor’s clothes be off the peg or made to measure? *BMJ*, *345*, e5278.
- Varpio, L., Bell, R., Hollingworth, G., Jalali, A., Haidet, P., Levine, R., et al. (2012). Is transferring an educational innovation actually a process of transformation? *Advances in Health Sciences Education*, *17*(3), 357–367.
- Worrall, J. (2010). Evidence: Philosophy of science meets medicine. *Journal of Evaluation in Clinical Practice*, *16*(2), 356–362.