Is experimental research passé

Geoff Norman

Received: 11 August 2010 / Accepted: 11 August 2010 / Published online: 7 September 2010
© Springer Science+Business Media B.V. 2010

In perusing the contents of this issue, I discovered that, once again, the majority of the papers are not experimental. There are several qualitative studies—of teachers' attitudes to student feedback, reflective learning for continuing education, residents' perceptions of outpatient teaching, and assessment expert's view of the framework for assessments, cohort study looking at predictors of motivation for medicine, a survey of teachers' perceptions of student feedback, a correlational study of concurrent validity of a personality test, a review paper on the issue of whether Asian students are rote learners, and 3 randomized trials. One looked at the use of simulation in addition to a PBL session for management of respiratory and cardiac distress, another at web-based learning, and a third examined the effect of testing on retention of CPR skills. So of the 11 original studies, 3 were experimental.

I suppose I should be pleased. Todres et al. (2007), in a widely cited review of 2 years of Medical Education and Medical Teacher articles, found only 5% were randomized controlled trials. So at $3/11 = 27\%$, we're doing much better.

Better? Pardon a second. Should we be striving for 100% randomized trials? Is this any measure of quality of research? After all, a study of faculty perception of student feedback is unlikely to benefit from having a control group (randomized of course) who were not given feedback and then asked their perceptions of not getting feedback. For that matter, randomizing students to be Asian or not is a bit impractical. In short, the variety of research designs in the papers in this issue reflect the nature of the research questions asked, which really span the gamut of educational research, and only a few are amenable to experimentation.

But what of quality? One carryover from clinical research is a perverse hierarchy of research designs where randomized controlled trial is at the top in quality, case series is at the bottom, and cohort studies are somewhere in the middle. The first problem with this approach is that it is a very limited taxonomy, suitable for epidemiologic studies and little else. The common approach to a study of assessment, looking at reliability and validity,

---

G. Norman (✉)
Clinical Epidemiology and Biostatistics, McMaster University, 1200 Main St. W, Hamilton,
ON, Canada
e-mail: norman@mcmaster.ca

must be classified as a “cohort” study, which scarcely does justice to it. Moreover, an abstract hierarchy does not examine the appropriateness of a design for the question being asked, but instead infers quality from the label put on the design. Why should this be an index of quality? Indeed, why should we presume, as do Todres et al. (2007), and Torgerson and Torgerson (2001), that the published literature in medical education is inherently inferior because we don’t do enough RCTs? After all, the papers in this issue, represent the 20% or so that survived a rigorous screening beginning with one editor (Henk van Berkel) who screened out 50–60% before review, then 2 or 3 expert reviewers, then an associate editor, then yours truly. And we were not exactly scraping the bottom of the barrel. Even after the revisions were accepted (only about 1–2 papers a year are accepted without revision), these articles still sat around in the publication queue for a year or so. If we are to pass judgment that our research methods are inferior, either the collective judgment of the 5–6 people involved in every paper review is consistently flawed, since we are apparently unable to identify good studies from bad, or the criteria used by the critics are wrong.

Randomized controlled trials have a critically important role in clinical research. But this is a consequence of a number of preconditions, most of which are absent in education. To begin with, let me state the obvious—even in clinical research, they are useful for study of therapies or interventions. They are not useful for study of risk factors, natural history, prognosis, etc.

Some additional special characteristics of RCTs in clinical research:

- (1) The mechanism of action is well understood. For drug interventions, this was worked out by the molecular biologists, biochemists and pharmacologists long before the trial began. Otherwise, it is difficult to interpret any evidence of an effect (or non-effect).
- (2) The endpoint is easily objectified. In cardiovascular research it may be death, which is well quantified, although more often it may be cardiac death, or cardiac events, which are less so. Although these are relatively unambiguous, still trialists will go to great lengths in setting up criteria and committees of oversight to ensure that the judgments cannot be challenged. Other objective endpoints such as blood sugar, blood pressure or range of motion can be used, but are less favoured as the link to outcome is less well understood.
- (3) Those who will benefit most from the therapy are easily identified. The diagnosis is a beginning point, but rarely sufficient. Often the trial will include additional inclusion and exclusion criteria to isolate a subgroup who will show a maximal treatment response. This is often necessary, because:
- (4) The treatment effect is very small. Although relative risk reductions may appear large, when this is converted to effect sizes, the numbers are often small. For example, the physician study of effect of aspirin on heart attacks had an effect size of 0.07 (Rosenthal 1991). As a result, these small treatment effects can often be lost in possible biases, so it is at this point that randomization is critical to average the biases across groups.
- (5) The effect of the treatment is independent of who is administering it. It makes sense to talk of 300 mg. of a drug t.i.d. However, when the therapy does depend on personal factors, as in surgery, where surgeon expertise can be a major determinant of outcome (Devereaux et al. 2005), interpretation is much more difficult. Further, for therapies like exercise, smoking cessation, or diet, both patient and physician factors may be a major determinant of outcome.
- (6) Finally, the consequences of a successful trial can be enormous. The rapid adoption of new drug classes like H₂ antagonists for peptic ulcer disease or statins for

cholesterol-lowering testify to the rapid dissemination of effective therapies. The link with financial gain clearly explains why drug companies are willing to invest the enormous sums in these studies. Of course, because of financial imperative, there are far more trials of therapies for myocardial infarction than myasthenia gravis.

In education, none of these special circumstances exist. In turn:

- (1) How people learn, how non-cognitive factors like motivation or stress (Harvey et al. 2010) interact with learning, how individual differences like gender or ethnicity (see this issue) affect learning, how individual differences and teaching interact (Garg et al. 2002; Levinson et al. 2007), are subject of ongoing research and are an appropriately large share of the studies we do in education.
- (2) Educational outcomes are very diverse, and interventions that improve some outcomes may reduce others (e.g. Needham and Begg 1991). Not surprisingly, one of the most active and effective areas of research is assessment, which explores the strengths and weaknesses of multiple approaches. But this complicates the choice of outcome.
- (3) It would seem that different kinds of knowledge or skill (taken as analogous to the diagnosis) might require different kinds of interventions. However, we are at a rudimentary stage in tailoring instructional intervention to specific content or skill.
- (4) The large effect sizes associated with successful educational and psychological interventions are well established (Lipsey and Wilson 1993). In their survey of 400 systematic reviews, the average effect size was about 0.5. In fact, they found no overall effect of randomization on effect size. Thus concern about baseline differences, the *raison d'être* for randomization, is less critical.
- (5) Who is delivering the intervention can be a major determinant of success. In one review, teacher effects accounted for much more variance in outcome than curriculum effects (Darling-Hammond and Youngs 2002). Moreover the content of a curriculum may be a larger determinant of outcome than the specific curriculum type. At McMaster, successive curricula, all small-group, PBL, with 5–6 students per group and by and large the same teachers, have shown effect sizes on the national examination ranging from $-.53$ to $+.26$ (Norman et al. 2010). Educational interventions tend to assume that all will benefit equally (for that matter, so do trialists). In education, some approaches like learning style attempt to individualize treatments, but by and large this has been unsuccessful. More promising is the specific exploration of aptitudes, such as spatial ability (Garg et al. 2002).
- (6) Rarely is a new approach adopted widely, to anyone's economic benefit. While curricula like PBL have achieved fairly wide penetration, this has been a slow process, and is clearly not a consequence of any demonstration of effectiveness. By contrast, standardized educational tools like high-fidelity simulators have seen slow and sparse acceptance (Issenberg et al. 2005; Friedman 1997), and the attempts to create widely available e-learning modules have met with little adoption (anatomical mysteries). Not surprisingly, therefore, there are no companies waiting in the wings eagerly awaiting their turn to invest in educational trials.

In short, randomized trials have seen a very limited adoption in education, for very good reason. They are not a terribly informative way to address many educational questions. Nevertheless, they do have a role to play, in examining such relatively standardized therapies as e-learning (Cook et al. 2008), and simulation (maybe) (Issenberg et al. 2005). Even here, however, just about every attempt to conduct a systematic review of an

educational area to determine overall if it “works”, has become mired in specificity. The reviews by Cook et al. (2008) and Ruiz et al. (2009) related to e-learning and animation are exceptional. But of the 9 published BEME reviews, only one was able to estimate an overall effect. Ironically, this was a review of predictive validity of assessment methods, based on correlational studies (Hamdy et al. 2006).

But no one would suggest that we abandon educational experiments although a few people (Regehr 2010; Mennin 2010), have suggested, or at least implied, just that. I don’t agree. Educational experiments can provide useful information that cannot be learned any other way. Cook et al. (2008) review is one example of the utility of experiments; the active research program of Mayer (1997) is another. But it is critical to recognize that is is not randomization per se that is critical to the quality of educational experiments, or that the methods of clinical experimental research can and should be adopted wholesale into the educational setting.

David Cook, in this issue’s Reflections article, reflects carefully on the specific characteristics of the educational environment, and examines the question of what aspects of design are critical in the conduct of experimental studies in education. In doing so, he has created a masterpiece of scholarship, drawing from sources spanning several decades, from the classic tomes of experimental design (which, by the way, originated in education, with Campbell and Stanley 1983), to contemporary sources. He points out the fallacy in the assumption that randomization is the magic bullet, and carefully elucidates various special aspects of the educational environment that require different approaches to design and analysis.

This is an informative, state of the art, paper. It should be required reading for anyone venturing into the dangerous waters of educational interventions. I m delighted to see it in the pages of AHSE.

References

- Campbell, D., & Stanley, J. (1983). *Experimental and quasi-experimental designs for research on teaching*. London: Rand-McNally.
- Cook, D. A., Levinson, A. J., Garside, S., Dupras, D. M., Erwin, P. J., & Montori, V. M. (2008). Internet-based learning in the health professions: A meta-analysis. *Journal of the American Medical Association*, 300, 181–196.
- Darling-Hammond, D., & Youngs, P. (2002). Defining “highly qualified” teachers: What does “scientifically-based” research actually tell u? *Educational Researcher*, 31, 13–25.
- Devereaux, P. J., Bhandari, M., Clarke, M., Montori, V. M., Cook, D. J., Yusuf, S., et al. (2005). Need for expertise based randomised controlled trials. *British Medical Journal*, 330, 88–93.
- Friedman, C. P. (1997). Perspectives on computing and medical education Introduction. *Academic Medicine*, 72, 511.
- Garg, A. X., Norman, G. R., Eva, K. W., Spero, L., & Sharan, S. (2002). Is there any real virtue of virtual reality? The minor role of multiple orientations in learning anatomy from computers. *Academic Medicine*, 77, S97–S99.
- Hamdy, H., Prasad, K., Anderson, M. B., et al. (2006). Predictive values of measurements obtained in medical schools and future performance in medical practice. *Medical Teacher*, 28, 103–116.
- Harvey, A., Nathens, A. B., Bandiera, G., & Leblanc, V. R. (2010). Threat and challenge: Cognitive appraisal and stress responses in simulated trauma resuscitations. *Medical Education*, 44, 587–594.
- Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Lee, G. D., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher*, 27, 10–28.
- Levinson, A. J., Weaver, B., Garside, S., McGinn, H., & Norman, G. R. (2007). Virtual reality and brain anatomy: A randomised trial of e-learning instructional designs. *Medical Education*, 41, 495–501.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *American Psychologist*, 48, 1181–1209.

- Mayer, R. C. (1997). Multimedia learning: Are we asking the right questions. *Educational Psychologist*, 32, 1–19.
- Mennin, S. (2010). Self-organization, integration and curriculum in the complex world of education. *Medical Education*, 44, 20–30.
- Needham, D. R., & Begg, I. M. (1991). Problem-oriented training promotes spontaneous analogical transfer: Memory-oriented training promotes memory for training. *Memory & Cognition*, 19, 543–557.
- Norman, G., Neville, A., Blake, J. M., & Mueller, B. (2010). Assessment steers learning down the right road: Impact of progress testing on licensing examination performance. *Medical Teacher*, 32, 496–499.
- Regehr, G. (2010). It's NOT rocket science: Rethinking our metaphors for research in health professions education. *Medical Education*, 44, 31–39.
- Rosenthal, R. (1991). Meta-analysis: A review. *Psychosomatic Medicine*, 53, 247–271.
- Ruiz, J. G., Cook, D. A., & Levinson, A. J. (2009). Computer animations in medical education: A critical literature review. *Medical Education*, 43, 838–846.
- Todres, M., Stephenson, A., & Jones, R. (2007). Medical education research remains the poor relation. *British Medical Journal*, 335, 333–335.
- Torgerson, C. J., & Torgerson, D. J. (2001). The need for randomised controlled trials in educational research. *British Journal of Educational Studies*, 49, 316–328.