Adv in Health Sci Educ
DOI 10.1007/s10459-012-9381-0

EDITORIAL

If you teach them, they will learn: why medical education needs comparative effectiveness research

David A. Cook

Received: 22 May 2012 / Accepted: 22 May 2012

If you teach a medical student, can they learn? The answer may seem self-evident. After all, undergraduates don't make it into medical school without demonstrating a remarkable capacity to learn and perform well on tests. So asking if medical students (or other health professionals and students) are capable of learning should be a superfluous question. Yet education researchers seem compelled to repeatedly ask this question. And surprisingly (or not), they repeatedly come up with the same answer.

Figure 1 shows the results of over 750 studies, summarized from 4 separate meta-analyses (Cook et al. 2010a, 2008b, 2011a, McGaghie et al. 2011), comparing various forms of training with no intervention. For example, a meta-analysis of Internet-based education found 126 studies comparing training with no intervention (either a single-group pretest–posttest study, or a no-intervention comparison group; Cook et al. 2008b). Only 2 studies failed to favor the training group for outcomes of knowledge, and the average effect size was 1.0—which according to Cohen (1988) would be considered a large effect. Results were similarly large for outcomes of skills and behaviors. Another meta-analysis found similar results for computer-based virtual patients (Cook et al. 2010a). Most recently, two meta-analyses of simulation-based education confirmed similarly strong benefits, with effect sizes ranging 0.8–2.0 (Cook et al. 2011a; McGaghie et al. 2011). Moreover, these results held true across various learner subgroups (medical students, postgraduate physician trainees, physicians, nurses and nursing students, and others), study designs (there were over 150 randomized trials), and multiple other subgroup analyses. Only when actual impact on patients was considered was a lower effect size noted, and then it was still a moderately-large 0.50. Even after adjusting for possible publication bias

D. A. Cook
Office of Education Research, Mayo Medical School, Rochester, MN, USA

D. A. Cook (✉)
Division of General Internal Medicine, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55905, USA
e-mail: cook.david33@mayo.edu

Published online: 14 June 2012

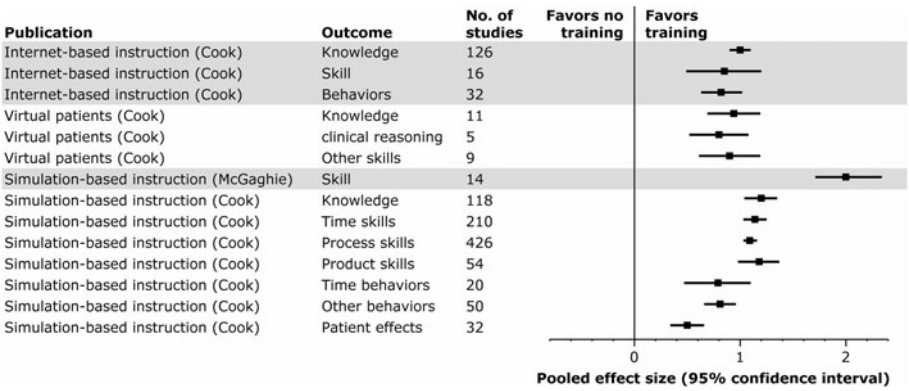

Fig. 1 Pooled effect sizes for studies comparing training with no training. Effect sizes represent Cohen's d or the nearly-equivalent Hedges' g from random-effects meta-analysis; > 0.80 is large, 0.50–0.79 is moderate. Data derive from meta-analyses of Internet-based instruction (Cook et al. 2008b) virtual patients (Cook et al. 2010a), simulation-based instruction (McGaghie et al. 2011), and simulation-based instruction (Cook et al. 2011a)

(the tendency of small studies showing null effects to remain unpublished), these meta-analyses showed large effects favoring instruction.

The answer seems clear: if you teach students, they will learn.

Why do we keep asking if students can learn?

Why, then, do educators persist in asking this question? I suspect three reasons contribute to this practice. First, they can. Nearly two-thirds of the studies in these meta-analyses were single-group pretest–posttest studies, using the learners' baseline as their own control. Such studies are relatively easy to conduct in the course of normal instruction, pose minimal ethical risk, and require little advance planning. Even for studies with a control group, the logistics of withholding an intervention from one group are simpler than planning and orchestrating an alterative active comparison. Given the lack of funding for education research (Reed et al. 2005), it comes as no surprise that educators gravitate to comparisons with no intervention.

Second, it appears important. Studies making comparison with no intervention seem like a good idea—after, we want to know whether or not the new course/curriculum/tool "works," right? A pretest–posttest comparison of change, or—better—comparison with a group that received no additional training, would clearly demonstrate such effectiveness, and justify continued use.

Third, they see others do it. Not only is the education literature rife with no-intervention comparison studies, but the clinician-teachers who often conduct such studies see ubiquitous examples in their reviews of the clinical literature. If the *New England Journal of Medicine* or *JAMA* publishes studies comparing tamoxifen or lisinopril with placebo, shouldn't the top medical education journals publish studies comparing a novel instructional tool with no intervention?

The problem, of course, is that appearances can be deceiving. Feasibility is not a good reason, in itself, to warrant a study. Studies justifying the use of a new course/curriculum/tool do little to advance the science of education because they generally fail to inform the

development of the *next* course/curriculum/tool (Cook et al. 2008a). And while the clinical paradigm does apply readily to education research, it requires a minor shift in perspective to identify the important questions.

Reframing the question

In clinical medicine, once a therapy has been established as efficacious, subsequent research focuses on the comparative effectiveness of alternative approaches, i.e., comparison of two active interventions. Indeed, in the absence of equipoise it would be unethical to conduct a no-intervention-comparison study. Comparative effectiveness research (Ellis et al. 2007) seeks to understand how to apply efficacious therapies in practice. Such studies directly compare two active interventions head-to-head (Hochman and McCormick 2010), such as comparing aspirin with clopidogrel for stroke prevention, comparing different doses of atorvastatin for cardiovascular protection, or comparing two breast cancer screening strategies. Such comparisons are most meaningful—and most likely to yield useful results—when guided by theoretical and empirical evidence of what should work (Giacomini 2009). Comparative effectiveness research evaluates not only the bottom-line outcomes, but also the processes (both effective and ineffective) that led to those outcomes. Thus, they consider the costs, barriers, unforeseen consequences, and effective strategies associated with implementing therapies in practice.

The same principles apply in medical education. The key difference is that rather than administering known quantities of a specific drug, in education research we assemble our interventions from a menu of assorted ingredients (lecture, bedside teaching, feedback, repetition, practice problems, etc.) with varying levels of specificity and activity. It's relatively safe to say that any single ingredient or combination thereof will have *some* degree of activity. But to assemble an *optimally* effective intervention for a given learning objective we need to understand the strengths, weaknesses, and cost of each of the potential components, how their effectiveness varies for different applications, and how they interact with one another. For this, we need theory-guided and theory-building studies comparing active interventions (Albert et al. 2007). and robust evaluations that provide practical guidance for subsequent implementations (Cook 2010). These studies must account not only the benefits of an instructional or assessment intervention, but also the financial expense (Zendejas et al. in press), adverse effects, and other costs of therapy.

Of course, not all clinical interventions involve drugs; many involve care pathways, procedures, and psychosocial interventions that bear a striking resemblance to education research (Boutron et al. 2008), and educators can learn much from such research programs (Hawe et al. 2004). In both clinical medicine and medical education, comparative effectiveness "of alternative methods to prevent, diagnose, treat, and monitor a clinical [or educational] condition or to improve the delivery of care" (Sox and Greenfield 2009) will be of substantial value in advancing the science and the art.

Caution in research with active comparisons

However, in transitioning to a comparative effectiveness paradigm researchers must be wary of at least two potential pitfalls. First, studies comparing two active interventions will require much larger sample sizes than studies making comparison with no intervention, because the expected effect size will be smaller. The difference in effectiveness between

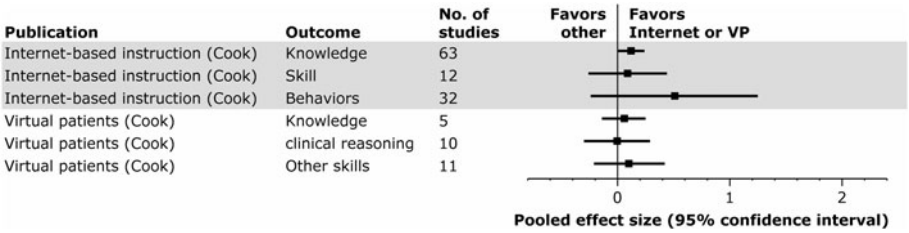

Fig. 2 Pooled effect sizes for studies comparing training with other training. Effect sizes represent Cohen's d or the nearly-equivalent Hedges' g from random-effects meta-analysis; 0.50–0.79 is moderate, 0.2–0.49 is small, <0.2 is negligible. Data derive from meta-analyses of Internet-based instruction (Cook et al. 2008b) and virtual patients (VP) (Cook et al. 2010a)

high-dose and low-dose atorvastatin will be less than the difference between atorvastatin and a sugar pill. Likewise, the difference in effectiveness between two approaches to delivering feedback during clinical teaching will be less than the difference between clinical teaching and no teaching. Based on data in Fig. 1, the latter effect size could be estimated at approximately 0.8–1.0. The effect size for the former will depend on the specifics of the approaches under study, but might reasonably range 0.1–0.4 (see Cook et al (2010b) for some examples). To show a statistically significant difference for an effect size of 0.8 a researcher needs 52 participants (26 per group); for an effect size of 0.3 the required sample is 351 (176 per group). Education researchers rarely report sample size calculations (Cook et al. 2011b) but even when they do, they often use unrealistically high anticipated effects.

The second pitfall is confounding. Confounding occurs when there is more than one explanation for the observed results. Much has been written on this topic (Norman 2003; Cook 2009; Cook and Beckman 2008; Regehr 2010) suffice to note that when multiple ingredients vary simultaneously it is impossible to know which ingredient or combination of ingredients accounted for the difference. For example, if we compare lecture with multiple practice problems and simple yes/no feedback versus bedside teaching with a single case and hands-on feedback, we won't know whether it was the format (lecture vs. bedside), the number of cases (multiple vs. one) or the feedback (simple vs. hands-on) that led to the observed difference—regardless of the magnitude or statistical significance. Randomization cannot compensate for confounding.[1] A large sample size, blinded outcome assessment, or multi-institutional enrollment cannot compensate for confounding (in fact, multi-institutional studies are particularly susceptible to confounding; see Regehr 2010). The only way to address confounding is to identify and control (eliminate or adjust for) the potential confounding variables. One particularly common example of confounded research is the comparison of a new educational technology (e.g., Internet-based instruction or virtual patient) against an older approach. That such "media-comparative" studies are hopelessly confounded was first articulated nearly 30 years ago (Clark 1983), and has been re-articulated multiple times (Friedman 1994; Cook 2005). Avoiding confounding requires researchers to consider not only the "study" intervention but also the comparison group (something they seem to often neglect; see Cook et al. 2011b, 2007) and most importantly the differences between the two.

[1] Articles such as "RCT = results confounded and Trivial"19. Norman (2003) do not condemn randomization per se, but rather the use of randomization to conduct confounded studies that yield uninterpretable result.

Figure 2 illustrates both of these problems using the pooled effect sizes of approximately 90 media-comparative studies. The individual effect sizes for these 90 studies varied widely, but the average effects were very small ($\leq$0.12 in all but one instance) and not statistically significant. But not only are the results trivial, they are also confounded. In nearly all instances there were multiple educationally-important changes in the instructional design between the two study arms, and subgroup analyses (Cook et al. 2008b) suggested that these changes could account for much of the observed differences.

Where do we go from here?

The future is bright for medical education research. Much remains to be learned about teaching and assessing health professionals. But education researchers need to move beyond the no-intervention comparison, they need to adequately power their studies, and they need to consider the possibility of confounding. Not all studies require a large sample size, or a tightly-controlled comparison, or any comparison group at all. But decisions regarding these issues should arise after careful deliberation and thoughtful consideration of how the decision will impact study interpretations.

If we teach them, they *will* learn. The question we face is how to make learning as painless, relevant, and efficient as possible.

References

Albert, M., Hodges, B., & Regehr, G. (2007). Research in medical education: Balancing service and science. *Advances in Health Sciences Education Theory and Practice, 12*, 103–115.

Boutron, I., Moher, D., Altman, D. G., Schulz, K. F., & Ravaud, P. (2008). for the CONSORT group. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: Explanation and elaboration. *Annals of Internal Medicine, 148*, 295–309.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*, 445–459.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cook, D. A. (2005). The research we still are not doing: An agenda for the study of computer-based learning. *Academic Medicine, 80*, 541–548.

Cook, D. A. (2009). Avoiding confounded comparisons in education research. *Medical Education, 43*, 102–104.

Cook, D. A. (2010). Twelve tips for evaluating educational programs. *Medical Teacher, 32*, 296–301.

Cook, D. A., & Beckman, T. J. (2008). Reflections on experimental research in medical education. *Advances in Health Sciences Education Theory and Practice*. doi:10.1007/s10459-008-9117-3 (Epub ahead of print 22 April 2008).

Cook, D. A., Beckman, T. J., & Bordage, G. (2007). Quality of reporting of experimental studies in medical education: A systematic review. *Medical Education, 41*, 737–745.

Cook, D. A., Bordage, G., & Schmidt, H. G. (2008a). Description, justification, and clarification: A framework for classifying the purposes of research in medical education. *Medical Education, 42*, 128–133.

Cook, D. A., Erwin, P. J., & Triola, M. M. (2010a). Computerized virtual patients in health professions education: A systematic review and meta-analysis. *Academic Medicine, 85*, 1589–1602.

Cook, D. A., Hatala, R., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., et al. (2011a). Technology-enhanced simulation for health professions education: A systematic review and meta-analysis. *JAMA, 306*, 978–988.

Cook, D. A., Levinson, A. J., & Garside, S. (2011b). Method and reporting quality in health professions education research: A systematic review. *Medical Education, 45*, 227–238.

Cook, D. A., Levinson, A. J., Garside, S., Dupras, D. M., Erwin, P. J., & Montori, V. M. (2008b). Internet-based learning in the health professions: A meta-analysis. *JAMA, 300*, 1181–1196.

Cook, D. A., Levinson, A. J., Garside, S., Dupras, D. M., Erwin, P. J., & Montori, V. M. (2010b). Instructional design variations in internet-based learning for health professions education: A systematic review and meta-analysis. *Academic Medicine, 85*, 909–922.

Ellis, P., Baker, C., & Hanger, M. (2007). *Research on the comparative effectiveness of medical treatments: Issues and options for an expanded federal role*. Washington, D.C.: Congressional Budget Office, Congress of the United States.

Friedman, C. (1994). The research we should be doing. *Academic Medicine, 69*, 455–457.

Giacomini, M. (2009). Theory-based medicine and the role of evidence: Why the emperor needs new clothes, again. *Perspectives in Biology and Medicine, 52*, 234–251.

Hawe, P., Shiell, A., & Riley, T. (2004). Complex interventions: How "out of control" can a randomised controlled trial be? *BMJ, 328*, 1561–1563.

Hochman, M., & McCormick, D. (2010). Characteristics of published comparative effectiveness studies of medications. *JAMA, 303*, 951–958.

McGaghie, W. C., Issenberg, S. B., Cohen, E. R., Barsuk, J. H., & Wayne, D. B. (2011). Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Academic Medicine, 86*, 706–711.

Norman, G. (2003). RCT = results confounded and trivial: The perils of grand educational experiments. *Medical Education, 37*, 582–584.

Reed, D. A., Kern, D. E., Levine, R. B., & Wright, S. M. (2005). Costs and funding for published medical education research. *JAMA, 294*, 1052–1057.

Regehr, G. (2010). It's NOT rocket science: Rethinking our metaphors for research in health professions education. *Medical Education, 44*, 31–39.

Sox, H. C., & Greenfield, S. (2009). *For the institute of medicine committee on comparative effectiveness research prioritization. Initial national priorities for comparative effectiveness research: Report brief.* Washington, DC: National Academies Press.

Zendejas, B., Wang, A. T., Brydges, R., Hamstra, S. J., & Cook, DA. Cost: The missing outcome in simulation-based medical education research: A systematic review. Accepted (in press), Surgery.