

## The end of educational science?

Geoff Norman

Published online: 21 October 2008  
© Springer Science+Business Media B.V. 2008

This has been an interesting week. As well as welcoming in a new (academic) year, I also came across three papers in recent publications that have, in my view, enormous implications for how we go about our trade. If we take the message of these papers to heart, it will no longer be “business as usual” in medical educational research.

In the September issue of *Medical Education*, Colliver et al. (2008) reanalyzed studies of effectiveness of PBL derived from a meta-analysis by Gijbels et al. (2005). He found that 10 of 11 studies in the review were “quasi-experimental”, and this led to measurable bias. In two studies, assignment to PBL or other was not randomized, resulting in demonstrable baseline differences in favour of PBL. In five others, there was a confounding between intervention and outcome, where PBL students had practice with the outcome format, and one study had both problems (and a few more). Colliver’s conclusions go beyond the old “PBL-other” debate, however; he challenges the whole idea that anything of value can come from systematic reviews directed to the calculation of an effect size, and instead argues that “the field may be better served, in most cases, by systematic narrative reviews that describe and individually evaluate individual studies and their results rather than reviews that obscure biases and confounds by averaging.”

Eva, editor of *Medical Education*, concurred. In his editorial (2008), he states, “A good educational research literature review, in my opinion, is one that presents a critical synthesis of a variety of literatures, identifies knowledge that is well-established, highlights gaps in understanding, and provides some guidance regarding what remains to be understood.” Well, surprisingly, Colliver and I, who have historically found ourselves in protracted debates about many things, are on exactly the same side on this one. In a recent ASME monograph, Eva and I (2008) argued that systematic reviews in education have serious problems. Moreover, we noted a paradox, that “If one cannot combine the findings in some systematic way as a result of heterogeneity of outcomes to the point of having to describe each study independently, then the only thing separating systematic reviews from critical narrative reviews is the amount of time and resources spent searching for

---

G. Norman (✉)

Program for Educational Research and Development, McMaster University, Room 3519, MDCL, 1200  
Main St. West, Hamilton, ON, Canada L8N 3Z5  
e-mail: norman@mcmaster.ca

information.” Indeed, that is what has happened with several of the systematic reviews commissioned by the BEME collaboration (Issenberg et al. 2005; Veloski et al. 2006). In fact, only one of the BEME reviews attempts to compute overall effect sizes (Hamdy et al. 2006), but this is addressing questions of the predictive validity of measures, not experimental interventions.

So systematic reviews in education are overrated. Well, maybe not. Also on my desk this week was the Sept. 8 copy of JAMA, containing the most extensive and careful systematic review in health sciences education I’ve ever seen (Cook et al. 2008), looking at effect of e-learning. They began with 2,200 studies and ended up with 206, which enabled them to look in detail at various aspects. They examined two kinds of study designs—e-learning versus nothing and e-learning versus some other learning, and four kinds of outcomes—knowledge, skills, “practice behaviour and patient outcomes”, and satisfaction, and, remarkably, had sufficient data to make estimates in all of these areas. E-learning versus nothing had effect sizes from 0.82 to 1.0, which is very large, but is frankly of little interest, as the authors appreciated “...there appears to be limited value on further research comparing Internet-based interventions against no-intervention comparison groups” (p. 1195). However, when Internet-based learning is compared to an alternative (usually face-to-face or paper), effect sizes were much less impressive: +.10 for satisfaction, +.12 for knowledge, +.09 for skills, and +.51 for the six studies looking at outcomes, with no significant effects except knowledge.

Interestingly, the effects of non-randomization were not as one-sided as Colliver showed; for satisfaction, randomized studies and high quality studies had significantly larger, not smaller, effects and for the remainder there was no significant difference, although the same trend was present in looking at patient outcomes (The opposite was true in the meta-analysis of studies comparing to a no intervention arm). This is also consistent with the huge meta-meta-analysis of Lipsey and Wilson (1993), that involved over 300 meta-analyses and over 14,000 studies and found no effect of study quality or randomization. Not to say that Colliver’s analysis is wrong, but simply that one cannot assume all biases act in favour of the intervention.

Despite the large number of studies in Cook’s review, at the end of the day, the null hypothesis remained supreme in comparing one medium to another. In this regard, it will undoubtedly join the depressingly long summary of 355 studies and reports, dating back to the 1920s on the “No significant difference phenomenon” (Russell 2008), showing again and again that one medium is as good or bad as another.

However, the story does not end there. One real strength of Cook’s paper was that they were able to be more fine-grained and could estimate effect size for various factors in addition to study design. So for example, we learn that when the internet instruction includes practice, online discussion and tutorials, and lasts longer, larger effects are observed on some measures.

The problem with these conclusions is that these secondary comparisons may be helpful in addressing the concerns of Colliver, Eva and me in terms of trying to understand what makes a particular intervention effective or ineffective. But the authors have to sift through each study to try to determine whether a particular intervention had, or didn’t have practice exercises, and here, massive lumping must take place. The psychology literature on transfer (Eva et al. 1998) can almost use the differences in ways to do practice exercises as a defining characteristic, yet in the systematic review, all practice problems are created equal. And they are not equal. Cook notes in many of his sub-analyses that there was treatment heterogeneity, and quotes an  $I^2$  of .95, which amounts to massive heterogeneity. Furthermore, inevitably, the kinds of secondary variables that are common enough and

well enough described to get into the review are likely present simply because they are already known to be effective. It really is not of much value to find out that feedback enhances learning (Issenberg et al. 2005; Veloski et al. 2006) or practice problems enhance learning (Cook et al. 2008)—we already know that.

Hence, the final irony. As I stated before, any curriculum level experiment, whether of high or low quality, randomized or not, is unlikely to yield useful any generalizable information because the nature of the beast is that any curriculum contains a *potpourri* of pedagogical nostrums, some ineffective, some not, interacting or not in some unspecified way (Norman 2003). By extension, any attempt to combine these and retrospectively tease out the impact of various dimensions will inevitably result in much injudicious lumping of dissimilar elements. So even though the individual experiment may be, in some way, protected from internal threats to validity, the systematic review looks more like a retrospective observational study than an experiment. The only difference is that its “subject” is a study, not a person, but it still requires heroic faith in homogeneity within categories in order to lump things into treatment variables like “feedback” or “practice”. As Cronbach (1975) said, “When a researcher says that such and such an effect is true, all other things being equal, he speaks from the experience of setting a great many other things equal.” And when a systematic reviewer says that feedback, or discussion, or tutorials work”, he speaks from the experience of assuming all tutorials are made equal.

So does the educational research mission find itself hoisted on the petard of systematicity as Eva (2008) claims? What is the point of trying to standardize things that are, by their nature unstandardizable? Were qualitative researchers right all along that you can’t generalize?

I don’t think so. But I do think it requires a different way of viewing the world. While studies at the curriculum level will always contain so much noise as to be almost uninterpretable, it’s not all that difficult to pursue a research program directed at identifying and analyzing the critical elements in learning. Cognitive psychologists do it all the time; they typically report anywhere from 2 to 10 experiments in each paper. Each study is small and quick, but by systematically varying the conditions, by the time they’re done, they have a pretty good idea of what’s going on. This can work in applied settings like e-learning. The work of Clark et al. (2006) and Mayer (2004) exemplifies how this approach can be used to really pursue understanding of the characteristics of e-learning modules that make them more or less effective. But to achieve this, you change your viewpoint. The goal is a research program, not a study. Each study is viewed as one data point in a much larger array. Successive studies devolve from a growing and dynamic theory, and each, in turn, informs the theory.

Many authors have commented on how health sciences education is sadly lacking theoretical underpinning, but it is not always clear why this matters. It matters because the theory provides a strategy for identifying the variables that do and don’t matter. Sweller’s (2005) ‘cognitive load theory’ is a good example. The simple theoretical notion is that the “rate-limiting step” in learning is short term working memory, with severe capacity limitations. A consequence of this is that any aspect of e-learning that increases cognitive load without adding to learning is likely deleterious. And from that has evolved a full research program exploring various aspects of media from background music to cartoons in the corner.

On the other hand, such an enterprise has some structural risks. It may well be that there are interactions so that factors cannot be considered singly. On the other hand, (a) a small carefully designed experiment is a good basis for designing for interactions, and (b) Cronbach (1975) showed that interactions, specifically aptitude-treatment interactions are

much rarer than we might think. The other risk is that these studies are inevitably a long way removed from the “real world” of education, and clearly do not satisfy the requirement of a link to outcome.

The “outcomes” movement in medical education is strong (Harden 2007), although my personal view is that it is not particularly informative (Norman 2006). But if the goal is to show that some variables do or do not influence outcomes, the RCT—meta-analysis approach is only one means to this end. We have recently seen a number of examples of studies using large longitudinal data bases that can provide good evidence. Papadakis et al. (2005) showed that students who are identified with “problems” as undergraduates are more likely to be have malpractice suits as clinicians. Tamblyn et al. (2007) and Norcini et al. (2002) have shown a relation between exam performance and outcomes: malpractice claims and cardiac mortality. In a paper available in *Online First*, Collin et al. (2008) looked at the relation between MCAT scores, undergraduate GPA, and licensing examination (USMLE) scores in an international sample of 840,000 medical school applicants and about 100,000 students. Another paper from this group (Heckler and Violato 2008) showed, perhaps more convincingly than the several systematic reviews, that differences between schools explained less than 5% of variance in licensing exam performance, and curriculum effects (specifically PBL) contributed less than that.

Educational science is alive and well, and the papers cited in this editorial are outstanding examples of the diversity and quality of the endeavour. But the holy grail of randomized experiments followed by systematic reviews has been tried and has failed. We have more powerful and specific methods to find the answers we seek. The time is long overdue to abandon the worship of the false God of the RCT.

## References

- Clark, R., Nguyen, F., & Sweller, J. (2006). *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. San Francisco: Pfeiffer.
- Collin, V. T., Violato, C. & Heckler, K. (2008). Aptitude, achievement and competence in medicine: A latent variable model. *Advances in Health Sciences Education*, e-pub ahead of publication.
- Colliver, J. A., Kucera, K. & Verhulst, S. J. (2008). Meta-analysis of quasi-experimental research: Are systematic narrative reviews indicated? *Medical Education*, 42, 858–865.
- 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. *Journal of the American Medical Association*, 300, 1181–1196.
- Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, 30, 116–127.
- Eva, K. W. (2008). On the limits of systematicity. *Medical Education*, 42, 852–853.
- Eva, K. W., Neville, A. J., & Norman, G. R. (1998). Exploring the etiology of content specificity: Factors influencing analogical transfer and problem solving. *Academic Medicine*, 73, S1–S6.
- Gijbels, D., Dochy, F., Van den Bosche, P., Segers, M. (2005). Effects of problem based learning: A meta-analysis from the angle of assessment. *Reviews of Educational Research*, 75, 27–61.
- Hamdy, H., Prasad, K., Anderson, M. B., Scherpbier, A., Williams, R., Zwierstra, R., et al. (2006). BEME systematic review: Predictive values of measurements obtained in medical schools and future performance in medical practice. *Medical Teacher*, 28, 103–116.
- Harden, R. M. (2007). Outcome-based education: The future is today. *Medical Teacher*, 29, 625–629.
- Heckler, K., & Violato, C. (2008). How much do differences in medical schools influence student performance. A longitudinal study employing hierarchical linear modeling. *Teaching and Learning in Medicine*, 20, 104–113.
- Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Gordon, D. L., & 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.

- 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. E. (2004). Teaching of subject matter. *Annual Reviews of Psychology*, *55*, 715–744.
- Norcini, J. J., Lipner, R. S., & Kimball, H. R. (2002). Certifying examination performance and patient outcomes following acute myocardial infarction. *Medical Education*, *36*, 853–859.
- Norman, G. R. (2003). Commentary: RCT = results confused and trivial: The perils of grand educational experiments. *Medical Education*, *37*, 582–584.
- Norman, G. R. (2006). Outcomes, objective and the seductive appeal of simple solutions. *Advances in Health Sciences Education*, *11*, 217–220.
- Norman, G. R., & Eva, K. W. (2008). *Quantitative methods*. ASME Monograph, Association for Studies in Medical Education.
- Papadakis, M. A., Teherani, A., Banach, M. A., Knettler, T. R., Rattner, S. L., Stern, D. T., et al. (2005). Disciplinary action by medical boards and prior behavior in medical school. *New England Journal of Medicine*, *353*, 2673–2682.
- Russell, T. L. (2008). The no significant difference phenomenon. <http://www.nosignificantdifference.org/>.
- Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 19–30). New York: Cambridge University Press.
- Tamblyn, R. J., Abrahamowicz, M., Dauphinee, D., Wenghofer, E., Jacques, A., Klass, D., et al. (2007). Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *Journal of the American Medical Association*, *298*, 993–1001.
- Veloski, J., Boex, J. R., Grasberger, M. J., Evans, A., & Wolfson, D. B. (2006). Systematic review of the literature on assessment, feedback and physicians' clinical performance: BEME Guide no 7. *Medical Teacher*, *28*, 117–128.