Adv in Health Sci Educ
DOI 10.1007/s10459-008-9117-3

REFLECTIONS

Reflections on experimental research in medical education

David A. Cook · Thomas J. Beckman

Received: 18 March 2008 / Accepted: 8 April 2008

Abstract As medical education research advances, it is important that education researchers employ rigorous methods for conducting and reporting their investigations. In this article we discuss several important yet oft neglected issues in designing experimental research in education. First, randomization controls for only a subset of possible confounders. Second, the posttest-only design is inherently stronger than the pretest–posttest design, provided the study is randomized and the sample is sufficiently large. Third, demonstrating the superiority of an educational intervention in comparison to no intervention does little to advance the art and science of education. Fourth, comparisons involving multifactorial interventions are hopelessly confounded, have limited application to new settings, and do little to advance our understanding of education. Fifth, single-group pretest–posttest studies are susceptible to numerous validity threats. Finally, educational interventions (including the comparison group) must be described in detail sufficient to allow replication.

Keywords Medical education · Research methods · Research design · Experiment · Curriculum

Introduction

Research in medical education is gaining vitality and momentum. The Best Evidence in Medical Education movement (Harden et al. 1999) has seen the publication of several systematic reviews, the impact factors of medical education journals are rising, a call for papers for a medical education supplement to the *Journal of General Internal Medicine* generated over 130 submissions, and attendance at the annual meeting of the Association of Medical Education in Europe has grown from 375 in 1997 to 1700 in 2007 (personal communication, Pat Lilley, 28 February 2008). At the same time, editors and authors have

D. A. Cook (✉) · T. J. Beckman
Division of General Internal Medicine, Mayo Clinic College of Medicine, Baldwin 4-A, 200 First Street SW, Rochester, Minnesota 55905, USA
e-mail: cook.david33@mayo.edu

noted that much education research could benefit from greater attention to careful planning and conduct (Education Group for Guidelines on Evaluation 1999; Hutchinson 1999; Norman 2003; Cook et al. 2007; Dauphinee and Wood-Dauphinee 2004; Shea et al. 2004). It seems that some authors view education research as a soft science that does not require rigorous research methods, while others seek to conduct rigorous education research by applying clinical research methods without adaptation.[1] We agree that medical education research should be rigorous. However, certain aspects of study design require special consideration and emphasis.

Descriptive, correlational, causal-comparative, and qualitative study designs are appropriate for many education research questions (Fraenkel and Wallen 2003). However, we will focus on experimental research, which seems particularly problematic for medical education investigators. Campbell and Stanley (1963), in their classic taxonomy of experimental study designs, defined an experiment as, "research in which variables are manipulated and their effects upon other variables observed." Although nonrandomized studies are often referred to as "quasi-experimental," we will use the term "experiment" to refer to all randomized, nonrandomized, and uncontrolled designs in which variables are manipulated and observations made.

We will address six common misconceptions and errors in designing and reporting education experiments: undue confidence in randomization; over reliance on the pretest; widespread use of no-intervention comparisons, multifactorial interventions, and single-group pretest–posttest studies; and failure to explicitly define the interventions. These points may seem elementary to seasoned researchers, but our experience as editors, reviewers, and consumers of the literature suggests that these issues merit attention.

Randomization is not a panacea

The purpose of randomization is to control for differences, both known and unknown, between study groups. However, differences in participant characteristics constitute only one threat to the internal validity of a research study. Campbell and Stanley (1963) list additional threats, including history (simultaneous, unplanned events unrelated to the study), maturation (changes in participants over time), testing (the effect of taking a pretest on study outcomes), instrumentation (changes in scoring system, calibration, or rater fatigue), regression to the mean (Bland and Altman 1994), selection (bias in the recruitment and/or assignment of individuals to different groups), and mortality (loss to follow-up) (see Table 1). Other authors have added to this list additional factors such as differences in location (implying differences in the environment or available resources), participant attitude/motivation (for example, the Hawthorne effect), implementation (variation in expertise or preconceptions of the instructors and in the "quantity" of instruction), and learning outside the curriculum (which is often inequitably distributed among intervention groups) (Beckman and Cook 2004; Fraenkel and Wallen 2003; Norman 2003).

A comparison group can help mitigate many of these threats, but randomization is only *required* to control threats from selection and maturation. On the other hand, randomization cannot control for mortality,[2] location, attitude, or implementation threats. Thus, the

[1] For a recent discussion of whether medical education is a hard or soft science, see Gruppen (2008).

[2] Randomization cannot control for mortality (loss to follow-up), but it can facilitate analyses seeking to explore the implications of high participant dropout.

Table 1 Threats to the internal validity of education research studies

Threat	Description	How to minimize the threat
Subject characteristics	Differences among participants at start of study	Randomization
Selection bias	Biased assignment to experimental groups	Randomization
Maturation	Changes in participants over time unrelated to particular events	Randomization
History	Unplanned events unrelated to the intervention that might impact outcome	Concurrent control group
Instrumentation	Changes in scoring rubric or instrument calibration, including rater fatigue	Control group
Regression to the mean	Participants selected or groups assigned based on high or low performance will be closer to average upon subsequent testing (Bland and Altman 1994)	Control group (assignments not based on baseline performance)
Testing	The effect of taking a pretest on study outcomes (familiarizes participants with questions on posttest, stimulates learning/study to the test, and heightens awareness of intent of study)	No pretest[a]
Mortality (loss to follow-up)	Participants leave study	Prevent loss; collect information on those lost[a]
Location	Differences between groups in the environment or available resources	Collect information on potential difference[a]
Participant attitude and motivation	Learners involved in something they consider novel, or who are being observed, tend to be more motivated (conversely, those in comparison group may be demotivated)	Blind participants to study hypothesis[a]
Implementation	Variation in the learning experience e.g. differences in the expertise of the instructors, the opinions of instructors regarding the efficacy of the intervention, or the actual amount of instruction received (e.g. did participants skip class?). Learning outside the curriculum (how much learners studied on the topic beyond that intended by the intervention) falls into this category as well	Careful planning of study interventions; collect information on actual experiences (both within and without the study)[a]

See Fraenkel and Wallen (2003), Campbell and Stanley (1963), and Norman (2003) for more information on these threats

[a] Control groups (with or without randomization) do not control for these threats

mere presence of randomization does not guarantee a study's validity. Cook and Campbell (1979), in their classic text, noted, "The case for random assignment cannot be made on the grounds that it is a general facilitator of high-quality research." Cronbach subsequently endorsed this statement and added, "Randomization may be achieved at the expense of relevance." (Cronbach 1982) More recently, Norman (2003) argued that in much education research the variance introduced by ambiguity and heterogeneity in treatments (implementation threat) outweighs the variance arising from differences among participants (i.e. the threats addressed by randomization).

All this should not be construed as an argument against randomization. On the contrary: randomization is required for the strongest study design, and should be performed whenever feasible. Yet valid inferences can certainly be derived from well-designed non-randomized studies. In fact, reviews of research in both education (Wilson and Lipsey 2001) and clinical medicine (Benson and Hartz 2000; Concato et al. 2000) suggest that bias from non-randomized study designs is likely small.

In judging inferences from research results, Cronbach (1982) proposed the following thought experiment exploring *reproducibility*. First, would similar conclusions be justified if the experiment were repeated using the same procedures, (a) on the same group of participants, and (b) on a different sample? Second, would similar conclusions be justified if the same research question were addressed using different procedures? Reflecting upon the answers to these questions will help researchers identify potential threats.

A non-randomized study that carefully controls for key threats is likely to support more reproducible inferences than a randomized study that fails to adequately attend to potential sources of invalidity. We encourage greater attention to the entire spectrum of validity threats.

Pretests often weaken the study design

Many educators hold the pretest–posttest, randomized design—in which participants are randomized to two or more conditions, take a pretest, undergo an intervention of some type, and then take a posttest—as the gold standard in education research. We wish to dispel this myth, and suggest that greater attention be given to the randomized posttest-only design. In 1963, Campbell and Stanley (1963) wrote, "While the pretest is a concept deeply embedded in the thinking of research workers in education and psychology, it is not actually essential to true experimental designs… [The randomized posttest-only design] is usually to be preferred to [the randomized pretest–posttest design] unless there is some question as to the genuine randomness of the assignment. [It] is greatly underused in educational and psychological research." More recently, Fraenkel and Wallen (2003) state, "[The randomized posttest-only design] is perhaps the best of all designs to use in an experimental study, provided there are at least 40 participants in each group."

Why is this so? Recall that proper randomization should equalize participant differences—including baseline knowledge and skills—among the study groups. Thus, the pretest is not necessary to ensure that randomized groups are equivalent at baseline (except for chance). Furthermore, using a pretest creates several disadvantages (see Table 2). Exposure to the pretest will affect performance on an identical posttest through familiarity with the questions, and may also influence learning during the intervention (e.g. "studying for the test"). These issues constitute the "testing threats" described above. Use of an alternate version of the test is plagued by likely differences in difficulty, and yet does not entirely avoid testing threats.

Furthermore, contrary to common misconceptions, pretests cannot adequately correct for baseline differences between study participants (Cronbach and Furby 1970), and the notion that a pretest increases sensitivity or statistical power is not necessarily true (Norman and Streiner 2007). While subtracting pretest from posttest does remove stable individual differences, it does so at the cost of introducing measurement error twice—from pretest and posttest (although if the reliability coefficient is >0.5, appropriate statistical analysis may offset this liability).

Table 2 When to use a pretest in education research

Disadvantages to using a pretest	Consider using a pretest when
• The pretest imposes extra burden on learners	• The pretest constitutes an integral part (or was already a part) of the intervention
• The pretest influences learning during the intervention	• Using a nonrandomized design
• Previous exposure affects performance on the posttest; alternate test form may have different level of difficulty	• Sample size is small (less than 40 per arm)
• The pretest compounds the unreliability of scores	• Anticipating a high dropout rate
	• Hoping to generalize to a different population

This does not mean that pretests are necessarily bad. Cronbach (1982) suggested, "To investigate pretest-plus-instruction is obviously suitable when the pretest is a natural part of the operating program. Pretests added for the sake of the evaluation are the suspect ones." Likewise, Campbell and Stanley observed that "for educational research frequent testing is characteristic of the universe to which one wants to generalize," and noted that *if pretests are already part of the educational program* they can increase statistical power and explore interactions between the intervention and pretest performance.

Pretests are useful under certain circumstances.[3] When randomization is not possible, pretests can help to judge the similarity of groups (but *not* to correct for differences found in other variables). In randomized trials, pretests may be useful when sample size is small (because randomized groups are likely to be rather different [due to chance] for small samples, and also to increase statistical power [caveats above in mind]), when researchers anticipate a high dropout rate (to compare dropouts to those who remain), or when trying to generalize results to learners who may not be similar to the study population. Except in these situations, researchers may wish to avoid pretests in randomized, controlled trials in education.[4]

No-intervention and placebo-controlled studies have limited application

The vast majority of experimental medical education research makes comparison with a no-intervention control. Unfortunately, such studies are similar to trials comparing a drug to no intervention when other effective drugs exist. Comparing amlodipine to no intervention may show that the drug lowers blood pressure, but comparison with another drug such as hydrochlorothiazide is required to establish amlodipine's role in clinical practice.

[3] When pretests are used, researchers should not calculate the difference between pretest and posttest scores and statistically analyze the difference or change scores. Although this method is commonly used (indeed, we are guilty of having used it), it is inferior to the more appropriate use of the pretest as a covariate (along with treatment group and other relevant variables) in multivariate statistical models. See Cronbach and Furby (1970) and Norman and Streiner (2007) for detailed discussions.

[4] Pretests may also be useful in randomized trials comparing active interventions if no treatment effect is found, by providing evidence that the lack of effect is not due to similarly *ineffective* interventions or an insensitive measurement tool (an exploration of the absolute effects of the treatments rather than the relative effects between groups). However, this analysis parallels the single-group pretest-posttest study with all attendant limitations.

Similar logic applies to placebo (inactive intervention) controls (we hereafter use placebo and no-intervention interchangeably because the differences are not relevant to this discussion). Likewise, while comparison with a no-intervention control demonstrates proof of concept of a novel educational intervention, it does little to inform educational practice.

A no-intervention-controlled educational experiment showing significant effect tells us only that learning *can* occur. However, we already know that if learners spend time learning they will learn (see Fig. 1). So confirming a benefit from the latest workshop on the topic *du jour* is rarely exciting because it tells us nothing of why the intervention worked, how it integrates into current practice, how it compares with existing interventions, or how it can be improved for the next go-round. In short, such "justification" studies (Cook et al. 2008) do not advance our understanding of how to teach. In fact, if a no-intervention-controlled experiment shows no benefit from a learning intervention this is likely due as much to an insensitive outcome measure or inadequate sample size as it is to an ineffective intervention. Even if we believe the intervention was ineffective, such research tells us that this particular intervention failed to work in this particular setting, but does not tell us why it failed, how to improve future interventions, or whether it might have worked in another environment.

To meaningfully inform practice, head-to-head comparisons of carefully designed active interventions will be necessary. The differences between the two interventions will need to be focused, explicitly defined, and replicable. Interventions should be based on theory or evidence suggesting that the instructional design is appropriate for the learning objectives. For example, a recent study comparing two methods for teaching neurological physical diagnosis found that a method focusing on pathophysiological explanations for exam findings was slightly superior to a method that emphasized disease probabilities for each finding (Woods et al. 2005). These investigators grounded their hypotheses in theories of cognition, and concluded that, "the basic science information, because of its conceptual coherence, was itself more memorable." This finding *clarifies* our understanding of learning

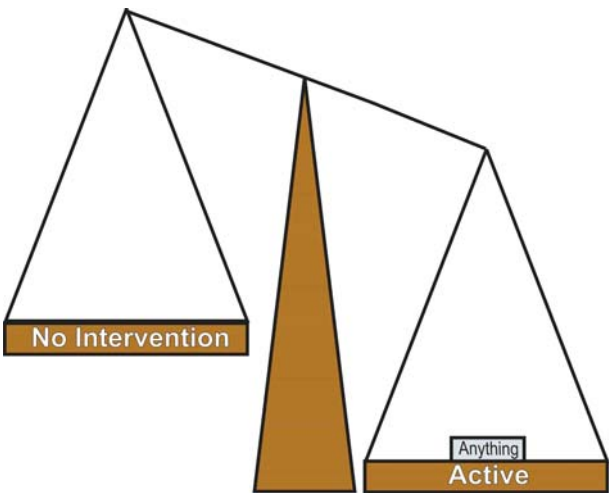

Fig. 1 The futility of no-intervention-controlled (and placebo-controlled) experiments. This figure illustrates a critical limitation of experiments with no-intervention or placebo controls, namely that any virtually educational intervention will make a difference of some sort. Failure to detect this difference is more likely due to an insensitive outcome measure or underpowered study than an ineffective intervention

and recall, and can be generalized to other settings. More "clarification" studies of this sort are needed to advance the science of medical education (Cook et al. 2008).

There may be exceptions to this rule. For example, many interventions in continuing medical education have failed to demonstrate expected effects on physician behavior and patient outcomes. Such studies have highlighted the ineffectiveness of weak instructional designs and encouraged the development of stronger interventions. However, we maintain that our understanding of how to effectively change physician behavior (and other "higher-order" outcomes (Kirkpatrick 1996)) will be better advanced through comparisons of carefully-selected active interventions.

Multifactorial interventions are hopelessly confounded

Experimental interventions should be focused. Many educational interventions (here we refer to both study and comparison interventions) involve multiple simultaneous or sequential learning methods and experiences, such as a workshop that employs lecture, small group sessions, and practice with a standardized patient. When complex interventions show significant benefit, they demonstrate that a specific outcome (e.g. knowledge or behavior) *can* be modified but tell us little about which components of the intervention (e.g. instructional methods and experiences) determined this change (see Fig. 2). Such investigations have only limited generalizability (Norman 2003) because the multifactorial intervention cannot be replicated precisely, and implementing only a portion of the intervention may or may not be effective. A recent article discussed this issue in greater detail in regard to computer-based learning (Cook 2005).

Norman suggests that medical education research will provide meaningful results when the "various factors that contribute to a result are systematically varied … based on a

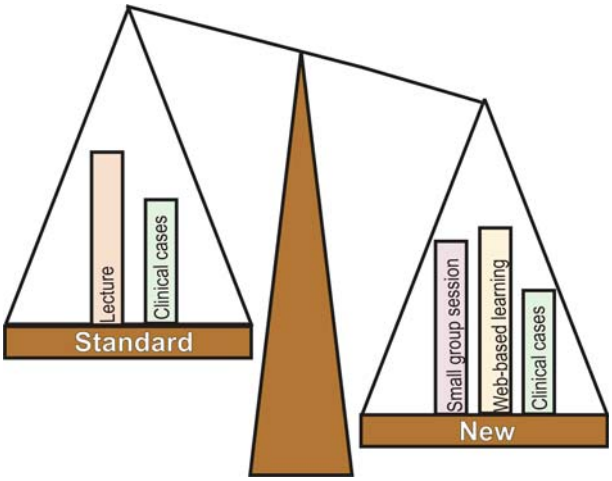

Fig. 2 The challenge of interpreting multifactorial interventions. This figure illustrates a hypothetical experiment in which a new intervention incorporating small group sessions, Web-based learning, and clinical cases was compared to the standard intervention comprised of lecture and the same clinical cases. Although the new intervention proved superior ("heavier"), it is impossible to know the relative contribution ("weight") of each individual component and thus it will be difficult to apply these findings to future educational endeavors

theory of causation." (Norman 2003) We agree, and suggest that studies of multifactorial educational interventions be replaced, or at least complemented, by more theory- and evidence-based comparisons of specific instructional methods. Cook (2005) presented a model for systematic variation of computer-based instructional designs that we believe readily translates to educational research in general. Using this model, a recent study compared a series of Web-based learning modules to a second series of modules, identical to the first save for case-based questions interspersed throughout the text. The results confirmed the study hypotheses by showing that the case-based questions improved test scores and were preferred by the majority of learners (Cook et al. 2006). More studies of this type will provide evidence to assist educators when selecting from among multiple possible instructional methods to teach a specific objective.

Single-group pretest–posttest designs suffer from many validity threats

Single-group pretest–posttest studies are experiments in which participants act as their own control (Campbell and Stanley 1963). Such studies are ubiquitous in medical education (Baernstein et al. 2007; Cook et al. 2007). Yet without a concurrent control group, this design is susceptible to numerous validity threats including history, maturation, testing, instrumentation, regression, location, and attitude. Collectively, these threats seriously constrain the inferences that can be drawn from research using this design. We recognize that many educational settings preclude the use of stronger designs, but remind researchers to use concurrent controls whenever possible.

Interventions must be thoroughly described

Educational interventions are often inadequately described (Cook et al. 2007; Price et al. 2005). In contrast to clinical trials on drugs and procedural interventions, which typically use standardized preparations or codified protocols, education research involves workshops, lectures, small group sessions, and computer-based applications that can each be implemented in various ways. While it would be unreasonable to expect a report to contain the entire curriculum for each topic addressed, authors must outline the key instructional methods and the theory and evidence upon which they are based in sufficient detail that someone at another institution could replicate the intervention (Des Jarlais et al. 2004). Highlighting this problem, a recent study (Cook et al. 2007) found that 8% of intervention descriptions were incompletely described. It is also important to carefully describe what happens to the comparison group. Even when "no intervention" is intended, learners typically engage in some kind of alternate education such as the standard curriculum or self-directed learning. Yet this information is frequently omitted in reports of education experiments—20% incomplete or absent in the study above (Cook et al. 2007). Researchers should carefully anticipate and document the learning opportunities likely to be experienced by all study participants.

Final thoughts

We recognize the increasing use of qualitative research methods (Bordage 2007; Harris 2003; Shea et al. 2004), and calls for the use of case–control and cohort study designs

(Carney et al. 2004). These methods can help answer questions that experiments cannot (Callahan et al. 2007; Kennedy and Lingard 2007; Papadakis et al. 2005; Tamblyn et al. 2007). Rigorous research using such designs will add greatly to the evidence base.

In conclusion, education research requires rigorous methods. While education researchers can learn much from clinical research approaches, certain aspects of study design and reporting require special consideration. Increased attention to these fundamental issues will improve the quality of educational experiments and advance the art and science of medical education.

References

Baernstein, A., Liss, H. K., Carney, P. A., & Elmore, J. G. (2007). Trends in study methods used in Undergraduate Medical Education Research, 1969–2007. *Journal of the American Medical Association, 298*, 1038–1045.

Beckman, T. J., & Cook, D. A. (2004). Educational epidemiology. *Journal of the American Medical Association, 292*, 2969.

Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine, 342*, 1878–1886.

Bland, J. M., & Altman, D. G. (1994). Statistic notes: Regression towards the mean. *British Medical Journal, 308*, 1499.

Bordage, G. (2007). Moving the field forward: Going beyond quantitative–qualitative. *Academic Medicine, 82*(10 suppl), S126–S128.

Callahan, C. A., Hojat, M., & Gonnella, J. S. (2007). Volunteer bias in medical education research: An empirical study of over three decades of longitudinal data. *Medical Education, 41*, 746–753.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carney, P. A., Nierenberg, D. W., Pipas, C. F., Brooks, W. B., Stukel, T. A., & Keller, A. M. (2004). Educational epidemiology: Applying population-based design and analytic approaches to study medical education. *Journal of the American Medical Association, 292*, 1044–1050.

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine, 342*, 1887–1892.

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., 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. (2008). Description, justification, and clarification: A framework for classifying the purposes of research in medical education. *Medical Education, 42*, 128–133.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cook, D. A., Thompson, W. G., Thomas, K. G., Thomas, M. R., & Pankratz, V. S. (2006). Impact of self-assessment questions and learning styles in web-based learning: A randomized, controlled, crossover trial. *Academic Medicine, 81*, 231–238.

Cronbach, L. J. (1982). *Designing evaluations of educational and social problems*. San Francisco: Jossey-Bass.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—or should we? *Psychological Bulletin, 74*, 68–80.

Dauphinee, W. D., & Wood-Dauphinee, S. (2004). The need for evidence in medical education: The development of best evidence medical education as an opportunity to inform, guide, and sustain medical education research. *Academic Medicine, 79*, 925–930.

Des Jarlais, D. C., Lyles, C., & Crepaz, N. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health, 94*, 361–366.

Education Group for Guidelines on Evaluation. (1999). Guidelines for evaluating papers on educational interventions. *British Medical Journal, 318*, 1265–1267.

Fraenkel, J. R., & Wallen, N. E. (2003). *How to design and evaluate research in education*. New York, NY: McGraw-Hill.

Gruppen, L. D. (2008). Is medical education research 'hard' or 'soft' research? *Advances in Health Sciences Education: Theory and Practice, 13,* 1–2.

Harden, R. M., Grant, J., Buckley, G., & Hart, I. R. (1999). BEME Guide No. 1: Best evidence medical education. *Medical Teacher, 21*, 553–562.

Harris, I. (2003). What does "the discovery of grounded theory" have to say to medical education? *Advances in Health Sciences Education, 8*, 49–61.

Hutchinson, L. (1999). Evaluating and researching the effectiveness of educational interventions. *British Medical Journal, 318*, 1267–1269.

Kennedy, T. J., & Lingard, L. A. (2007). Questioning competence: A discourse analysis of attending physicians' use of questions to assess trainee competence. *Academic Medicine, 82*(10 suppl), S12–S15.

Kirkpatrick, D. (1996). Revisiting Kirkpatrick's four-level model. *Training and Development, 50*(1), 54–59.

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

Norman, G. R., & Streiner, D. L. (2007). *Biostatistics: The bare essentials* (Vol. 3). Hamilton: BC Decker.

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.

Price, E. G., Beach, M. C., Gary, T. L., Robinson, K. A., Gozu, A., Palacio, A., et al. (2005). A systematic review of the methodological rigor of studies evaluating cultural competence training of health professionals. *Academic Medicine, 80*, 578–586.

Shea, J. A., Arnold, L., & Mann, K. V. (2004). A RIME perspective on the quality and relevance of current and future medical education research. *Academic Medicine, 79*, 931–938.

Tamblyn, R., 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.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6*, 413–429.

Woods, N. N., Brooks, L. R., & Norman, G. R. (2005). The value of basic science in clinical diagnosis: Creating coherence among signs and symptoms. *Medical Education, 39*, 107–112.