Published in final edited form as:

JAMA. 2007 September 5; 298(9): 1038–1045. doi:10.1001/jama.298.9.1038.

Trends in Study Methods Used in Undergraduate Medical Education Research, 1969–2007

Amy Baernstein, MD, Hillary K. Liss, MD, Patricia A. Carney, PhD, and Joann G. Elmore, MD, MPH

Division of General Internal Medicine (Drs Baernstein, Liss, and Elmore) and Department of Obstetrics and Gynecology (Dr Liss), University of Washington School of Medicine, and Department of Epidemiology, University of Washington School of Public Health and Community Medicine (Dr Elmore), Seattle; and Departments of Family Medicine and Public Health and Preventive Medicine, Oregon Health and Science University, Portland (Dr Carney)

Abstract

Context—Evidence-based medical education requires rigorous studies appraising educational efficacy.

Objectives—To assess trends over time in methods used to evaluate undergraduate medical education interventions and to identify whether participation of medical education departments or centers is associated with more rigorous methods.

Data Sources—The PubMed, Cochrane Controlled Trials Registry, Campbell Collaboration, and ERIC databases (January 1966–March 2007) were searched using terms equivalent to *students, medical* and *education, medical* crossed with all relevant study designs.

Study Selection—We selected publications in all languages from every fifth year, plus the most recent 12 months, that evaluated an educational intervention for undergraduate medical students. Four hundred seventy-two publications met criteria for review.

Data Extraction—Data were abstracted on number of participants; types of comparison groups; whether outcomes assessed were objective, subjective, and/or validated; timing of outcome

©2007 American Medical Association. All rights reserved.

Corresponding Author: Amy Baernstein, MD, Division of General Internal Medicine, Harborview Medical Center, 325 Ninth Ave, Box 359702, Seattle, WA 98104-2499 (abaer@u.washington.edu).

Author Contributions: Dr Baernstein had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Baernstein, Liss, Carney, Elmore.

Acquisition of data: Baernstein, Liss.

Analysis and interpretation of data: Baernstein, Liss, Carney, Elmore.

Drafting of the manuscript: Baernstein, Liss, Carney, Elmore.

Critical revision of the manuscript for important intellectual content: Baernstein, Liss, Elmore.

Statistical analysis: Baernstein, Carney.

Obtained funding: Elmore.

Administrative, technical, or material support: Elmore.

Financial Disclosures: None reported.

Additional Contributions: The Division of General Internal Medicine, University of Washington School of Medicine, provided administrative support in obtaining documents. The statistical analysis was reviewed by Jan Carline, PhD, Department of Medical Education and Biomedical Informatics, University of Washington. Sherry Dodson, MLS, Health Sciences Libraries, University of Washington, assisted in designing the search strategy. Lisa Johnson, MN (Medicos Sin Fronteras, Guatemala), Angelika Koch-Liebmann, MD (Department of Medicine, University of Washington), Wendell de Moura, Hiroo Takayama, MD (Department of Surgery, University of Washington), and Melanie Tratnik, JD (Washington State Office of the Attorney General) translated foreign-language publications and Morrie Bills, Raymond Harris, PhD, and R. J. Lambert (University of Washington) provided technical assistance. None of these individuals was compensated for their role in the study.

assessments; funding; and participation of medical education departments and centers. Ten percent of publications were independently abstracted by 2 authors to assess validity of the data abstraction.

Results—The annual number of publications increased over time from 1 (1969–1970) to 147 (2006–2007). In the most recent year, there was a mean of 145 medical student participants; 9 (6%) recruited participants from multiple institutions; 80 (54%) used comparison groups; 37 (25%) used randomized control groups; 91 (62%) had objective outcomes; 23 (16%) had validated outcomes; 35 (24%) assessed an outcome more than 1 month later; 21 (14%) estimated statistical power; and 66 (45%) reported funding. In 2006–2007, medical education department or center participation, reported in 46 (31%) of the recent publications, was associated only with enrolling more medical student participants ($P = .04$); for all studies from 1969 to 2007, it was associated only with measuring an objective outcome ($P = .048$). Between 1969 and 2007, the percentage of publications reporting statistical power and funding increased; percentages did not change for other study features.

Conclusions—The annual number of published studies of undergraduate medical education interventions demonstrating methodological rigor has been increasing. However, considerable opportunities for improvement remain.

Evidence-based medicine, an accepted construct that asserts that medical interventions should be guided by data from rigorous studies, was described at least 15 years ago.¹ The ideal that medical educators should choose teaching approaches based on evidence soon followed.^{2–4} However, educators cannot be guided by evidence if the studies providing that evidence have significant shortcomings.

Concerns about the rigor of medical education research have been identified, but previous reviews have been limited in scope. Limitations identified in medical education research include insufficient sample sizes, lack of generalizability when interventions are assessed at only 1 institution or are conducted only once, lack of appropriate control groups, use of subjective and unvalidated instruments to assess outcomes, and assessing only short-term outcomes.^{5–12}

Previous reviews of medical education that examine study methods have been limited to specific educational topics^{6,13–15}; have looked at only 1 aspect of studies, such as funding or outcomes^{12,16,17}; or have reviewed only studies published in specific journals.^{12,16–18} None have examined the timing of outcome assessments, whether interventions were conducted more than once, whether statistical power was estimated, or whether participation of medical education departments or centers had an impact on methods. Comprehensive reviews have not included non-English-language research or explored whether study methods are improving over time.

We conducted a structured longitudinal review of the undergraduate medical education literature from nearly 4 decades to characterize historical and current study methods and to evaluate whether participation of medical education departments or centers is associated with more rigorous methods. We hypothesized that we would observe a trend for increasingly rigorous studies over time. We further hypothesized that publications demonstrating participation of medical education departments or centers would exhibit more features of rigorous methods.

METHODS

Data Sources and Study Selection

We analyzed interventions involving medical students rather than medical education more generally because interventions for this group are most similar internationally and, thus, most comparable. Many curricular topics are relevant to all medical students, whereas much educational content for residents and practicing physicians is specialty-specific. We used PubMed, the Cochrane Controlled Trials Registry, the Campbell Collaboration, and the ERIC databases to identify publications indexed under the exploded Medical Subject Heading terms *students*, *medical* or *education*, *medical* or equivalent terms. We crossed these search terms with descriptors of relevant study designs (Box). We included publications in all languages. To assess trends over time, we selected publications from academic years (July 1–June 30) 1969–1970, 1974–1975, 1979–1980, 1984–1985, 1989–1990, 1994–1995, 1999–2000, and 2004–2005. We then added the most recent 12 months (April 1, 2006–March 31, 2007).

Publications were excluded if they (1) were not about undergraduate medical students; (2) were opinion pieces, editorials, or critiques of prior studies; (3) were reviews summarizing or interpreting preexisting data; or (4) did not evaluate an educational intervention. Examples of publications in this last group include studies that examined characteristics of medical students such as health status or demographics, analyses of predictors of student outcomes such as comparisons of Medical College Admission Test scores to specialty choice, and studies that assessed the performance of testing instruments.

Data Abstraction

We developed and piloted a standardized data abstraction form that documented the number of medical student participants, number and nationality of institutions involved, whether a control or comparison group was used, methods for assigning controls, outcome variables assessed, use of validated assessment instruments, timing of outcome assessments, number of times the intervention was assessed, whether statistical power was estimated, and cost of the intervention. Sources of funding were also recorded. Author affiliations and acknowledgments were examined to identify whether a medical education department or center had participated in the study.

Study Definitions

Features of rigorous study methods were defined based on published guidelines for appraising the effectiveness of medical education curricula.^{10,19,20} More rigorous study methods were defined as (1) greater number of student participants; (2) multi-institutional enrollment of participants; (3) having a control or appropriate comparison group; (4) measuring an objective outcome; (5) measuring a validated outcome; (6) measuring some outcome at least 1 month after the intervention; (7) conducting the intervention more than once; and (8) estimating statistical power.

Publications were categorized as multi-institutional if they recruited participants from more than 1 medical school (including studies that used one school as the intervention group and another as the control). Objective outcome measures were defined as any evaluation other than self-assessment by students, including (1) tests of knowledge; (2) course grades; (3) objective structured clinical examinations or observations of standardized patient interactions; (4) assessment of performance with a real patient, such as graded observation of patient interactions or patient feedback; (5) objective clinical outcomes such as blood pressure control in patients; (6) performance with mannequins, computer simulations, or laboratory animals; and (7) psychological inventories. We also noted assessment of

subjective outcomes, such as student satisfaction, self-assessment of competence, or attitude. Specialty choice or success in obtaining desired residency position was recorded as an outcome but was not categorized as either objective or subjective. The study outcomes were further classified into the levels of behavioral change of the Kirkpatrick²¹ hierarchy: (1) participation; (2a) modification of attitudes or perceptions; (2b) modification of knowledge and skills; (3) behavior change in the workplace; (4a) change in organizational practice; and (4b) benefit to patients.

Outcome measures were considered to be validated if the authors stated that their evaluation tool was validated beyond face validity or used tests that are generally known to be validated, such as United States Medical Licensing Examinations. Funding was noted if the publication's text, acknowledgments, or separate disclosure section reported any financial support for the study. Sources of funding were also recorded. Participation of a medical education department or center was defined as having any author or acknowledged person listed with an affiliation with a medical education department or center or if a medical education department or center was acknowledged.

Data Quality

All publications were abstracted by 1 author (A.B. or H.K.L.). To assess the validity of data abstraction, 10% of publications were selected using a random number generator and abstracted by 2 authors (A.B. and H.K.L.). Variables collected were compared for consistency. Discrepancies in coding of dual-reviewed publications were resolved by consensus. Each abstractor was blinded to the other's abstraction but not to our study hypotheses.

Data Analysis

Descriptive analysis compared categorical variables over time using the χ^2 test for trend. Continuous variables were evaluated with the t test (if compared with dichotomous variables) or 1-way analysis of variance (if comparing over time). Data from the earliest years were grouped according to an a priori scheme when sample sizes were too small (<5 expected counts per cell) to allow analyses; this was necessary because there were few publications in the earliest periods. Percentage agreement and the Cohen κ statistic were used to assess interrater agreement between abstractors. Statistical significance was set at 2-sided $P < .05$. The sample size of 472 articles was large enough to provide 80% power to detect small differences (effect size, 0.12) for the variable "participation of medical education departments and centers" with an α level of .05 for a 2-tailed χ^2 test.²² The analysis was performed using SPSS software, version 14.0 (SPSS Inc, Chicago, Illinois).

RESULTS

The PubMed search strategy identified 10 213 publications spanning a 38-year period (Figure). Of these, 2974 publications were within the defined periods of our sampling strategy and were selected for review. The PubMed search yielded 459 publications appropriate for full abstraction. The ERIC, Cochrane, and Campbell databases were searched using equivalent strategies, resulting in 13 nonduplicate publications appropriate for full abstraction. Thus, 472 publications were included in the analysis. Reasons for exclusion are shown in the Figure. Between 9% and 19% of publications retrieved by PubMed from a given year met inclusion criteria, and there was no change over time in the percentage included.

Approximately half of included publications appeared in journals dedicated to medical education, such as *Academic Medicine* and *Medical Education* (46%). The rest were

published in specialty journals, such as *Annals of Internal Medicine* (36%); other types of journals, including basic science journals (11%); and general medical journals, such as *JAMA* and *BMJ* (7%).

Eighteen publications suitable for inclusion were published in non-English languages, including German (9), Spanish (5), Arabic (1), Japanese (1), Portuguese (1), and Swedish (1). These publications were translated prior to inclusion in the analysis. Studies were conducted in many regions of the world, including the United States (263), Europe (101), Asia (36), Canada (27), Australia (23), Latin America (11), Africa (10), and multiple regions (1).

Interrater agreement on the decision to include or exclude publications was 98% ($\kappa = 0.91$). Agreement on data from included publications ranged from 77% to 100% (mean [SD], 92% [5.8%]), with κ values ranging from 0.57 to 1.00 (mean [SD], 0.81 [0.12]), indicating excellent agreement.²³

Methods Used in Recent Publications (2006–2007)

Of the publications, 147 were published between April 2006 and March 2007 (Table 1). The mean number of medical student participants was 145 (median, 81; range, 4–1663). Recruitment of participants from more than 1 institution was present in 6% of publications. A little more than half of the publications ($n = 80$ [54%]) used a comparison group with the following designs: randomized controls, 37 (25%); nonrandomized concurrent controls, 28 (19%); historical controls, 11 (7%); comparison with national norm, 2 (1%); and more than 1 kind of comparison group, 2 (1%). Publications with no comparison group were divided between 24 (16%) with pretest/posttest designs and 43 (29%) with posttest data only.

The outcomes assessed in recent publications were subjective in 56 (38%), objective in 38 (26%), and both in 53 (36%), with types of objective outcomes shown in Table 2. Validated outcomes were uncommon ($n = 23$ [16%]). The interval between the intervention and the last assessment was less than 1 month in the majority of publications ($n = 102$ [69%]). More than 1 presentation of the intervention was evaluated in 71 (51%) of the publications (Table 3). It was rare to have reporting of the statistical power of the study ($n = 21$ [14%]) or the cost of the intervention ($n = 5$ [3%]).

Among the recent publications, 109 (74%) measured an outcome at level 1 or 2a of the Kirkpatrick hierarchy²¹ (participation; modification of attitudes or perceptions), 75 (51%) measured an outcome at level 2b (modification of knowledge and skills), 5 (3%) measured an outcome at level 3 (behavior change in the workplace), and none measured an outcome at level 4b (benefit to patients). Five (3%) examined subsequent practice among an underserved population, which may fit Kirkpatrick level 4a (change in organizational practice).

Sixty-six (45%) of the recent publications reported receiving funding (Table 3). Sources of funding included government agencies ($n = 31$ [47% of funded studies]), private foundations ($n = 28$ [42%]), internal grants and awards ($n = 25$ [38%]), and pharmaceutical or device companies ($n = 10$ [15%]). Of the publications, 22 reported multiple funding sources.

Changes Over Time

The annual number of publications increased over time from 1 (1969–1970) to 147 (2006–2007). Study size was stable over time, with the mean number of medical student participants ranging from 105 (1969–1970) to 387 (1984–1985). The number of publications demonstrating the other features of rigorous methods, reporting funding, and reporting

participation of medical education departments and centers increased over time (Table 1, Table 2, and Table 3).

An increasing percentage of publications per year estimated statistical power (range, 0% in 1969–1985 to 14% in 2006–2007; $P = .003$) and reported funding (range, 0% in 1969–1970 to 45% in 1999–2000 and 2006–2007; $P = .002$) (Table 3). However, the percentage of published studies incorporating most of the features of rigorous study methods has not changed over time (Table 1, Table 2, and Table 3).

Participation of Medical Education Departments and Centers

Percentage of publications with participation of medical education departments or centers did not change over time (Table 3). Participation, documented in 139 (29%) of the publications, was associated with measuring an objective outcome (76% of publications with participation vs 66% of publications without participation; $P = .048$) (Table 4). This association was not significant when examining only 2006–2007 publications ($P = .14$). However, 2006–2007 publications enrolled more medical student participants if a medical education department or center participated (mean number of medical students, 218 vs 111; $P = .04$). Publications in medical education journals compared with other types of journals enrolled more medical student participants (mean number of medical students, 217 vs 137; $P = .007$) and were more likely to measure a validated outcome (22% vs 13%; $P = .02$).

COMMENT

This review, which quantifies the methods used to evaluate the efficacy of undergraduate medical education, found that educational researchers use experimental and epidemiologic study designs, measure objective and validated outcomes, and assess long-term outcomes. We found that the number of publications evaluating undergraduate medical education interventions has increased over time, and every feature of stronger methods we examined was used in more studies each year. Even though the percentage of studies demonstrating many of the features of strong methods has not changed over time, educators have an ever larger pool of rigorous published studies to guide their curricular decision making.

Participation of medical education departments and centers was associated with use of objective outcomes and number of participants in our analysis. Detecting associations with other features of study rigor may have been limited by our inability to detect either affiliations with medical education departments and centers that were not documented in the publications or informal consultations. Physicians who teach face conflicting pressures and expectations among their clinical duties, teaching responsibilities, and scholarly requirements; they may be able to produce more rigorous evaluations of curricular innovations when supported by medical education departments and centers.^{6,11,24,25} Creating national centers for educational research excellence that can facilitate collaboration, review emerging evidence, disseminate key findings, and advise credentialing agencies may improve the quality of educational research.^{7,26}

The majority of published undergraduate medical education research examines local processes and relies on student satisfaction and short-term acquisition of knowledge to measure educational efficacy. There was a paucity of research linking educational interventions to improved patient outcomes. While it may be difficult to demonstrate improvements in patient outcomes because of confounding factors, including individual student characteristics, the multitude of intervening experiences between medical school and practice, and the influences of varying practice environments, the analytic techniques that help overcome such obstacles in clinical research can also be applied in educational settings.²⁷

An essential step toward making the medical education research endeavor more relevant to society is to identify patient-level outcome measures that can be linked to educational interventions.^{11,26,28,29} In biomedical research, outcomes such as mortality, quality-adjusted life-years, and glycemic control are recognized as meaningful. Equivalent measures should be used in medical education. Educators currently use validated knowledge-based outcome measures (such as board examinations) as intermediary end points. However, whether they are appropriate surrogate markers for the larger goal of improving the health of a population is still unclear.

A research agenda for medical education that is based on patient-level clinical outcomes has been described.⁷ Collaboration with experts in health services has been identified as essential to establish outcome-driven education.^{11,29} It is not practical to assess all undergraduate medical education interventions using patient-level outcomes, but identifying deficiencies and opportunities for improvement in current health care delivery should play a much larger role in guiding curriculum development.²⁸

Of the publications we reviewed, 2 used an objective clinical outcome to demonstrate the effectiveness of their intervention. These studies, which demonstrated that it is possible to link education to patient outcomes, used patients' weight loss in Japan³⁰ and adherence with tuberculosis therapy in South Africa³¹ to judge the effectiveness of a practicum in public health. Most clinical outcomes will not be evident until students have finished training and moved on to practice, but only 5% of recent publications assessed an outcome more than 1 year after the intervention. The effect of a curriculum on health care delivery or patient outcomes cannot be ascertained in that time frame; the length of follow-up must increase markedly to measure such educational effects. Medical education research can be made more meaningful by (1) planning prospective studies with mechanisms for following participants over many years, (2) making better use of rigorous retrospective studies, and (3) clarifying linkages between educational processes and patient health outcomes.

There is significant unrealized potential for medical education research to improve the quality of future physicians. For example, medical schools periodically undertake complete overhauls of their undergraduate curricula,^{32–36} providing opportunities to evaluate the effect of curricula on participant learners. While major curricular reforms such as those at Harvard,³⁷ University of Missouri–Columbia,³⁸ McMaster,^{39–41} Maastricht,⁴² and the University of New Mexico^{43,44} have been evaluated on long-term outcomes, they have not included objective measures of graduates' practice behavior and performance.

Educational efficacy can be evaluated with study designs other than randomized controlled trials.^{9–11,45–48} Different study designs are appropriate for different types of research questions. For example, qualitative research^{9,19,49} can explore why an intervention does or does not work and is an effective way to generate hypotheses. Investigators can reduce bias and confounding while testing hypotheses with nonexperimental quantitative designs, such as cohort and case-control studies.²⁷ Crossover designs eliminate the ethical concern of depriving one group of a potentially better educational experience.

We found that a greater percentage of publications over time reported receiving funding. This is encouraging, as high-quality research in any field is challenging without funding, and increasing funding has been proposed as a specific strategy for improving medical education research.^{6,9,10,16,17,26} Research funding buys faculty time, statistical expertise, and administrative support for long-term studies that have sufficient power to measure meaningful outcomes. Reed et al¹⁷ investigated medical education studies conducted from 2002–2003 and found that 30% were funded, whereas 45% of recent publications in our sample reported funding. The difference may be due to inclusion criteria: our sample was

international and limited to undergraduate medical education, whereas their sample was US researchers and spanned the continuum of undergraduate, graduate, and continuing medical education. The most common sources of funding in both samples were government agencies and private foundations. We were not able to determine the amount of funding for each publication or measure association of funding with study methods.

Educators considering a curricular change need to know the cost as well as the potential benefits of the program they are considering.^{10,12,14,20} Although calculating the cost of an educational program requires considerable effort,^{17,50} some costs, such as equipment or standardized patient salaries, can be estimated. However, we found that only 3% of recent publications reported the cost of any part of their program, reducing their usefulness to educators making cost-benefit analyses before adopting new curricula.

Limitations of our study need to be considered. We conducted a current review, updating our list of studies 3 months after the latest publication date, although some articles may not yet have been indexed. The criteria we used to define rigorous methods are important but are not the only ones relevant for generating meaningful data. We did not examine whether a study asked an important question, chose apt study methods to address the question, or used an appropriate conceptual framework.^{10,19,51} However, because such criteria cannot be objectively determined, they were not appropriate for this analysis.

Strengths of our study include reviewing a large sample of publications, examining all topics in the undergraduate curriculum, and reviewing publications from 4 databases and with no language restrictions. Other authors have systematically reviewed specific educational topics and have reached similar conclusions about insufficient rigor of research in medical education.^{5,6,13–15} We have added to these findings by demonstrating that these problems are present throughout the undergraduate medical education literature and by showing ways in which important research features have changed over time. While the percentage of published research with strong methods has not changed appreciably over the years, more education studies are being conducted every year; therefore, the absolute number of rigorous studies available is increasing. Appropriate application of study design and development of clinically meaningful outcomes have the potential to make medical education research more relevant to the health needs of society.

Search Terms Used
Students, medical or

Education, medical

crossed with

Evaluation studies^{a,b} (includes but is not limited to the MeSH terms found below this term in the MeSH tree: clinical trials,^{a,b} controlled clinical trials, ^{a,b} randomized controlled trials,^{a,b} multicenter studies,^{a,b} program evaluation,^a reproducibility of results^a) or

Epidemiologic studies^a (includes but is not limited to the MeSH terms found below this term in the MeSH tree: case-control studies^a, retrospective studies^a, cohort studies^a, longitudinal studies^a, follow-up studies ^a, prospective studies ^a, cross-sectional studies^a) or

Intervention studies^a or

Validation studies^b

^aMedical Subject Headings (MeSH).

^bPublication types.

REFERENCES

1. Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA*. 1992; 268(17):2420–2425. [PubMed: 1404801]
2. van der Vleuten CP, Dolmans DH, Scherpbier A. The need for evidence in education. *Med Teach*. 2000; 22(3):246–250.
3. Hutchinson L. Evaluating and researching the effectiveness of educational interventions. *BMJ*. 1999; 318(7193):1267–1269. [PubMed: 10231262]
4. Harden RM, Grant J, Buckley G, Hart IR. Best evidence medical education—BEME guide No. 1. *Med Teach*. 1999; 21(6):553–562. [PubMed: 21281174]
5. Issenberg SB, McGaghie W, Petrusa E, Lee Gordon D, Scalese R. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach*. 2005; 27(1):10–28. [PubMed: 16147767]
6. Price EG, Beach MC, Gary TL, et al. A systematic review of the methodological rigor of studies evaluating cultural competence training of health professionals. *Acad Med*. 2005; 80(6):578–586. [PubMed: 15917363]
7. Chen FM, Bauchner H, Burstin H. A call for outcomes research in medical education. *Acad Med*. 2004; 79(10):955–960. [PubMed: 15383351]
8. Lurie SJ. Raising the passing grade for studies of medical education. *JAMA*. 2003; 290(9):1210–1212. [PubMed: 12953006]
9. Murray E. Challenges in educational research. *Med Educ*. 2002; 36(2):110–112. [PubMed: 11869436]
10. Reed D, Price EG, Windish DM, et al. Challenges in systematic reviews of educational intervention studies. *Ann Intern Med*. 2005; 142(12 pt 2):1080–1089. [PubMed: 15968033]
11. Shea JA, Arnold L, Mann KVA. RIME perspective on the quality and relevance of current and future medical education research. *Acad Med*. 2004; 79(10):931–938. [PubMed: 15383348]
12. Prystowsky JB, Bordage G. An outcomes research perspective on medical education: the predominance of trainee assessment and satisfaction. *Med Educ*. 2001; 35(4):331–336. [PubMed: 11318995]
13. Dornan T, Littlewood S, Margolis SA, Scherpbier A, Spencer J, Ypinazar V. How can experience in clinical and community settings contribute to early medical education? a BEME systematic review. *Med Teach*. 2006; 28(1):3–18. [PubMed: 16627313]
14. Letterie GS. Medical education as a science: the quality of evidence for computer-assisted instruction. *Am J Obstet Gynecol*. 2003; 188(3):849–853. [PubMed: 12634669]
15. Ogrinc G, Mutha S, Irby DM. Evidence for longitudinal ambulatory care rotations: a review of the literature. *Acad Med*. 2002; 77(7):688–693. [PubMed: 12114141]
16. Carline JD. Funding medical education research: opportunities and issues. *Acad Med*. 2004; 79(10):918–924. [PubMed: 15383346]
17. Reed DA, Kern DE, Levine RB, Wright SM. Costs and funding for published medical education research. *JAMA*. 2005; 294(9):1052–1057. [PubMed: 16145025]
18. Wolf FM. Methodological quality, evidence, and research in medical education (RIME). *Acad Med*. 2004; 79 suppl(10):S68–S69. [PubMed: 15383394]
19. Bordage G, Caelleigh AS, Steinecke A. Joint Task Force of Academic Medicine and GEA-RIME Committee. Review criteria for research manuscripts. *Acad Med*. 2001; 76(9):897–978. [PubMed: 11565558]
20. Green ML. Identifying, appraising, and implementing medical education curricula: a guide for medical educators. *Ann Intern Med*. 2001; 135(10):889–896. [PubMed: 11712879]
21. Kirkpatrick, DL. Evaluation of training. In: Craig, RL.; Bittel, LR., editors. *Training and Development Handbook*. New York, NY: McGraw-Hill; 1967. p. 87–112.

22. Cohen, J. *Statistical Power Analyses for the Behavioral Sciences*. 2nd ed. New York, NY: Academic Press Inc; 1988. p. 235
23. Jekel, J.; Elmore, J.; Katz, D. *Epidemiology, Biostatistics and Preventative Medicine*. 2nd ed.. Philadelphia, PA: Harcourt Health Sciences; 2001. p. 114
24. Davis MH, Karunathilake I, Harden RM. AMEE education guide No. 28: the development and role of departments of medical education. *Med Teach*. 2005; 27(8):665–675. [PubMed: 16451885]
25. van der Vleuten CP, Dolmans DH, de Grave WS, et al. Education research at the Faculty of Medicine, University of Maastricht: fostering the interrelationship between professional and education practice. *Acad Med*. 2004; 79(10):990–996. [PubMed: 15383361]
26. Wartman SA. Revisiting the idea of a national center for health professions education research. *Acad Med*. 2004; 79(10):910–917. [PubMed: 15383345]
27. Carney PA, Nierenberg DW, Pipas CF, Brooks WB, Stukel TA, Keller AM. Educational epidemiology: applying population-based design and analytic approaches to study medical education. *JAMA*. 2004; 292(9):1044–1050. [PubMed: 15339895]
28. Glick TH. Evidence-guided education: patients' outcome data should influence our teaching priorities. *Acad Med*. 2005; 80(2):147–151. [PubMed: 15671318]
29. Whitcomb ME. Using clinical outcomes data to reform medical education. *Acad Med*. 2005; 80(2): 117. [PubMed: 15671312]
30. Okuda N, Okamura T, Kadowaki T, Tanaka T, Ueshima H. Weight-control intervention in overweight subjects at high risk of cardiovascular disease: a trial of a public health practical training program in a medical school. *Nippon Koshu Eisei Zasshi*. 2004; 51(7):552–560. [PubMed: 15446673]
31. Williams RL, Reid S, Myeni C, Pitt L, Solarsh G. Practical skills and valued community outcomes: the next step in community-based education. *Med Educ*. 1999; 33(10):730–737. [PubMed: 10583763]
32. Litzelman DK, Cottingham AH. The new formal competency-based curriculum and informal curriculum at Indiana University School of Medicine: overview and five-year analysis. *Acad Med*. 2007; 82(4):410–421. [PubMed: 17414200]
33. O'Connor Grochowski C, Halperin EC, Buckley EG. A curricular model for the training of physician scientists: the evolution of the Duke University School of Medicine curriculum. *Acad Med*. 2007; 82(4):375–382. [PubMed: 17414195]
34. Loeser H, O'Sullivan P, Irby DM. Leadership lessons from curricular change at the University of California, San Francisco, School of Medicine. *Acad Med*. 2007; 82(4):324–330. [PubMed: 17414186]
35. Goldstein EA, Maclaren CF, Smith S, et al. Promoting fundamental clinical skills: a competency-based college approach at the University of Washington. *Acad Med*. 2005; 80(5):423–433. [PubMed: 15851451]
36. Neville AJ, Norman GR. PBL in the undergraduate MD program at McMaster University: three iterations in three decades. *Acad Med*. 2007; 82(4):370–374. [PubMed: 17414193]
37. Peters AS, Greenberger-Rosovsky R, Crowder C, Block SD, Moore GT. Long-term outcomes of the New Pathway Program at Harvard Medical School: a randomized controlled trial. *Acad Med*. 2000; 75(5):470–479. [PubMed: 10824772]
38. Hoffman K, Hosokawa M, Blake R Jr, Headrick L, Johnson G. Problem-based learning outcomes: ten years of experience at the University of Missouri–Columbia School of Medicine. *Acad Med*. 2006; 81(7):617–625. [PubMed: 16799282]
39. Shin JH, Haynes RB, Johnston ME. Effect of problem-based, self-directed undergraduate education on life-long learning. *CMAJ*. 1993; 148(6):969–976. [PubMed: 8257470]
40. Tolnai S. Continuing medical education and career choice among graduates of problem-based and traditional curricula. *Med Educ*. 1991; 25(5):414–420. [PubMed: 1758318]
41. Tolnai S. Lifelong learning habits of physicians trained at an innovative medical school and a more traditional one. *Acad Med*. 1991; 66(7):425–426. [PubMed: 2059274]
42. Schmidt HG, Vermeulen L, van der Molen HT. Longterm effects of problem-based learning: a comparison of competencies acquired by graduates of a problem-based and a conventional medical school. *Med Educ*. 2006; 40(6):562–567. [PubMed: 16700772]

43. Mennin SP, Kalishman S, Friedman M, Pathak D, Snyder J. A survey of graduates in practice from the University of New Mexico's conventional and community-oriented, problem-based tracks. *Acad Med.* 1996; 71(10):1079–1089. [PubMed: 9177642]
44. Santos-Gomez L, Kalishman S, Rezler A, Skipper B, Mennin SP. Residency performance of graduates from a problem-based and a conventional curriculum. *Med Educ.* 1990; 24(4):366–375. [PubMed: 2395429]
45. Prideaux D. Researching the outcomes of educational interventions: a matter of design: RTCs have important limitations in evaluating educational interventions. *BMJ.* 2002; 324(7330):126–127. [PubMed: 11799017]
46. Drescher U, Warren F, Norton K. Towards evidence-based practice in medical training: making evaluations more meaningful. *Med Educ.* 2004; 38(12):1288–1294. [PubMed: 15566540]
47. Torgerson CJ. Educational research and randomised trials. *Med Educ.* 2002; 36(11):1002–1003. [PubMed: 12406259]
48. Norman GR, Schmidt HG. Effectiveness of problem-based learning curricula: theory, practice and paper darts. *Med Educ.* 2000; 34(9):721–728. [PubMed: 10972750]
49. Britten N. Making sense of qualitative research: a new series. *Med Educ.* 2005; 39(1):5–6. [PubMed: 15612893]
50. Kelly M, Murphy A. An evaluation of the cost of designing, delivering and assessing an undergraduate communication skills module. *Med Teach.* 2004; 26(7):610–614. [PubMed: 15763850]
51. Education Group for Guidelines on Evaluation. Guidelines for evaluating papers on educational interventions. *BMJ.* 1999; 318(7193):1265–1267. [PubMed: 10231261]

Figure.
Study Selection

Table 1

Study Designs Used to Evaluate Undergraduate Medical Education Interventions, 1969–2007 (N = 472 Publications)^a

	1969– 1970	1974– 1975	1979– 1980	1984– 1985	1989– 1990	1994– 1995	1999– 2000	2004– 2005	2006– 2007	P Value for Percentages ^b
No. of studies	1	19	15	19	23	46	67	135	147	
Mean (SD) No. of participants	105	178 (227)	228 (263)	387 (667)	147 (83)	162 (193)	246 (501)	137 (164)	145 (223)	.90 ^c
Multi-institutional ^d	0	2 (11)	1 (7)	0	3 (13)	4 (9)	5 (8)	9 (7)	9 (6)	.52
Used a comparison group	1 (100)	7 (37)	9 (60)	11 (58)	9 (39)	25 (54)	46 (69)	76 (56)	80 (54)	.42
Study designs										
Randomized controls	1 (100)	1 (5)	1 (7)	4 (21)	1 (4)	10 (22)	16 (24)	27 (20)	37 (25)	
Nonrandomized, concurrent controls	0	2 (11)	5 (33)	3 (16)	2 (9)	10 (22)	16 (24)	26 (19)	28 (19)	
Historical controls	0	3 (16)	3 (20)	2 (11)	1 (4)	3 (7)	8 (12)	15 (11)	11 (7)	
National norm	0	1 (5)	0	1 (5)	2 (9)	0	1 (1)	2 (1)	2 (1)	
>1 Comparison group	0	0 (0)	0	1 (5)	3 (13)	2 (4)	5 (7)	6 (4)	2 (1)	
Pretest/posttest (no comparison group)	0	4 (21)	0	5 (26)	9 (39)	12 (26)	11 (16)	27 (20)	24 (16)	
Posttest only (no comparison group)	0	8 (42)	6 (40)	3 (16)	5 (22)	9 (20)	10 (15)	32 (24)	43 (29)	

^aData are presented as No. (%) except as noted. Percentages may not total 100% because of rounding. Studies examined July through June every 5 years from 1969–1970 through 2004–2005 as well as April 2006 through March 2007.

^bBy χ^2 test for trend, except as noted. *P* values were calculated only for categories designated a priori as hypotheses.

^cBy 1-way analysis of variance.

^dNumber of institutions contributing participants. Number of institutions was unclear in 8 studies.

Table 2

Outcomes Measured in Evaluating Undergraduate Medical Education Interventions, 1969–2007 (N = 472 Publications)^a

	1969– 1970	1974– 1975	1979– 1980	1984– 1985	1989– 1990	1994– 1995	1999– 2000	2004– 2005	2006– 2007	P Value for Percentages ^b
No. of studies	1	19	15	19	23	46	67	135	147	
Used an objective outcome	1 (100)	16 (84)	8 (53)	13 (68)	17 (74)	33 (72)	53 (79)	92 (68)	91 (62)	.10
Types of outcomes ^c										
Objective outcomes only	1 (100)	6 (32)	4 (27)	5 (26)	4 (17)	11 (24)	16 (24)	36 (27)	38 (26)	
Subjective and objective outcomes	0	10 (53)	4 (27)	8 (42)	13 (57)	22 (48)	37 (55)	56 (41)	53 (36)	
Subjective outcomes only	0	3 (16)	7 (47)	6 (32)	6 (26)	13 (28)	14 (21)	43 (32)	56 (38)	
Types of objective outcomes ^d										
Test of knowledge	1 (100)	9 (47)	3 (20)	10 (53)	11 (48)	23 (50)	29 (43)	59 (44)	42 (29)	
OSCE or performance with standardized patient	0	1 (5)	3 (20)	1 (5)	0	5 (11)	17 (25)	11 (8)	19 (13)	
Performance with mannequin, computer model, or laboratory animal	1 (100)	0	0	0	1 (4)	3 (7)	7 (10)	12 (9)	19 (13)	
Course grade	0	1 (5)	1 (7)	1 (5)	1 (4)	2 (4)	5 (7)	9 (7)	5 (3)	
Performance with actual patient	0	1 (5)	1 (7)	0	0	1 (2)	2 (3)	1 (1)	5 (3)	
Psychological inventory	0	2 (11)	0	1 (5)	2 (9)	4 (9)	2 (3)	3 (2)	2 (1)	
Objective clinical outcome	0	0	0	0	0	0	1 (1)	1 (1)	0	
Other objective outcome	0	4 (21)	2 (13)	5 (26)	8 (35)	8 (17)	9 (13)	26 (19)	20 (14)	
Validated outcome	0	4 (21)	3 (20)	4 (21)	3 (13)	8 (17)	17 (25)	21 (16)	23 (16)	.43
Outcome assessed > 1 month later ^e	1 (100)	5 (36)	4 (36)	5 (38)	4 (21)	15 (35)	16 (27)	23 (21)	35 (26)	.12
Interval between intervention and last assessment										
<1 Month	0	9 (47)	7 (47)	8 (42)	15 (65)	28 (61)	44 (66)	89 (66)	102 (69)	
1 Month–1 year	1 (100)	2 (11)	0	0	2 (9)	9 (20)	9 (13)	12 (9)	27 (18)	
>1 Year	0	3 (16)	4 (27)	5 (26)	2 (9)	6 (13)	7 (10)	11 (8)	8 (5)	
Uncertain	0	5 (26)	4 (27)	6 (32)	4 (17)	3 (7)	7 (10)	23 (17)	10 (7)	

Abbreviation: OSCE, objective structured clinical examination.

^aData are presented as No. (%). Percentages may not total 100% because of rounding. Studies examined July through June every 5 years from 1969–1970 through 2004–2005 as well as April 2006 through March 2007.

^b By χ^2 test for trend. *P* values were calculated only for categories designated a priori as hypotheses.

^c Subjective outcomes include self-assessment, satisfaction, and attitudes.

^d Multiple categories are possible for each study.

^e The denominator for these calculations is the number of studies minus studies in which data were uncertain.

Table 3

Other Features of Studies Evaluating Undergraduate Medical Education Interventions, 1969–2007 (N = 472 Publications)^a

	1969– 1970	1974– 1975	1979– 1980	1984– 1985	1989– 1990	1994– 1995	1999– 2000	2004– 2005	2006– 2007	P Value for Percentages ^b
No. of studies	1	19	15	19	23	46	67	135	147	
Intervention presented more than once ^c	0	10 (59)	8 (57)	10 (53)	11 (50)	22 (56)	37 (56)	78 (62)	71 (51)	.93
Statistical power estimated	0	0	0	0	2 (9)	4 (9)	3 (4)	17 (13)	21 (14)	.003
Cost estimated	0	0	1 (7)	1 (5)	2 (9)	6 (13)	5 (7)	3 (2)	5 (3)	
Funding reported	0	6 (32)	4 (27)	1 (5)	5 (22)	17 (37)	30 (45)	53 (39)	66 (45)	.002
Participation of a medical education department or center ^d	0	7 (37)	2 (13)	6 (32)	3 (13)	14 (30)	19 (28)	42 (31)	46 (31)	.29

^aData are presented as No. (%). Studies examined July through June every 5 years from 1969–1970 through 2004–2005 as well as April 2006 through March 2007.^bBy χ^2 test for trend. *P* values were calculated only for categories designated a priori as hypotheses.^cPercentages reflect that whether a program was presented more than once was uncertain in 29 studies.^dNoted if any author or acknowledged person listed an affiliation with a medical education department or center or if authors acknowledged a medical education department or center.

Table 4

Medical Education Departments/Centers and Methods Used to Evaluate Undergraduate Medical Education, 1969–2007 (N = 472 Publications)^a

Feature of Rigorous Methods	Participation of a Medical Education Department or Center ^b		<i>P</i> Value ^c
	No (n = 333)	Yes (n = 139)	
Study participants and designs			
Mean No. of participants	152	223	.08 ^d
Multi-institutional ^e	25 (8)	8 (6)	.62
Comparison group	177 (53)	87 (63)	.08
Outcomes			
Objective outcome	219 (66)	105 (76)	.048
Validated outcome	53 (16)	30 (22)	.18
Outcome assessed >1 month later ^f	73 (25)	35 (29)	.61
Other features			
Intervention presented more than once ^g	164 (53)	83 (62)	.13
Statistical power estimated	31 (9)	16 (12)	.58

^aData are presented as No. (%) except as noted. Studies examined July through June every 5 years from 1969–1970 through 2004–2005 as well as April 2006 through March 2007.

^bNoted if any author or acknowledged person listed an affiliation with a medical education department or center or if authors acknowledged a medical education department or center.

^cBy χ^2 test with continuity correction except as noted.

^dIndependent-sample *t* test.

^eRefers to number of institutions contributing participants. Number of institutions was uncertain in 8 studies.

^fTiming of last outcome assessment was uncertain in 62 studies.

^gWhether a program was presented more than once was uncertain in 29 studies.