Medical Teacher, Vol. 28, No. 6, 2006, pp. 497–526

informa
healthcare

BEME GUIDE

A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME Guide No. 8

YVONNE STEINERT[1], KAREN MANN[2], ANGEL CENTENO[3], DIANA DOLMANS[4], JOHN SPENCER[5], MARK GELULA[6] & DAVID PRIDEAUX[7]

[1]McGill University, Montreal, Canada; [2]Dalhousie University, Halifax, Canada; [3]Austral University, Buenos Aires, Argentina; [4]University of Maastricht, Maastricht, The Netherlands; [5]University of Newcastle upon Tyne, Newcastle, UK; [6]University of Illinois at Chicago, Chicago, USA; [7]Flinders University, Adelaide, Australia

ABSTRACT Background*: Preparing healthcare professionals for teaching is regarded as essential to enhancing teaching effectiveness. Although many reports describe various faculty development interventions, there is a paucity of research demonstrating their effectiveness.*

Objective*: To synthesize the existing evidence that addresses the question: "What are the effects of faculty development interventions on the knowledge, attitudes and skills of teachers in medical education, and on the institutions in which they work?"*

Methods*: The search, covering the period 1980–2002, included three databases (Medline, ERIC and EMBASE) and used the keywords: staff development; in-service training; medical faculty; faculty training/development; continuing medical education. Manual searches were also conducted.*

Articles with a focus on faculty development to improve teaching effectiveness, targeting basic and clinical scientists, were reviewed. All study designs that included outcome data beyond participant satisfaction were accepted. From an initial 2777 abstracts, 53 papers met the review criteria.

Data were extracted by six coders, using the standardized BEME coding sheet, adapted for our use. Two reviewers coded each study and coding differences were resolved through discussion.

Data were synthesized using Kirkpatrick's four levels of educational outcomes. Findings were grouped by type of intervention and described according to levels of outcome. In addition, 8 high-quality studies were analysed in a 'focused picture'.

Results: *The majority of the interventions targeted practicing clinicians. All of the reports focused on teaching improvement and the interventions included workshops, seminar series, short courses, longitudinal programs and 'other interventions'. The study designs included 6 randomized controlled trials and 47 quasi-experimental studies, of which 31 used a pre-test–post-test design.*

Key points: *Despite methodological limitations, the faculty development literature tends to support the following outcomes:*

- *Overall satisfaction with faculty development programs was high. Participants consistently found programs acceptable, useful and relevant to their objectives.*
- *Participants reported positive changes in attitudes toward faculty development and teaching.*
- *Participants reported increased knowledge of educational principles and gains in teaching skills. Where formal tests of knowledge were used, significant gains were shown.*
- *Changes in teaching behavior were consistently reported by participants and were also detected by students.*
- *Changes in organizational practice and student learning were not frequently investigated. However, reported changes included greater educational involvement and establishment of collegiate networks.*
- *Key features of effective faculty development contributing to effectiveness included the use of experiential learning, provision of feedback, effective peer and colleague relationships, well-designed interventions following principles of teaching and learning, and the use of a diversity of educational methods within single interventions.*

Methodological issues: *More rigorous designs and a greater use of qualitative and mixed methods are needed to capture the complexity of the interventions. Newer methods of performance-based assessment, utilizing diverse data sources, should be explored, and reliable and valid outcome measures should be developed. The maintenance of change over time should also be considered, as should process-oriented studies comparing different faculty development strategies.*

Conclusions: *Faculty development activities appear highly valued by participants, who also report changes in learning and behavior. Notwithstanding the methodological limitations in the literature, certain program characteristics appear to be consistently associated with effectiveness. Further research to explore these associations and document outcomes, at the individual and organizational level, is required.*

Correspondence: Yvonne Steinert, PhD, Centre for Medical Education, Faculty of Medicine, McGill University, Lady Meredith House, 1110 Pine Avenue West, Montreal, Quebec, H3A 1A3, Canada. Tel: (514) 398-2698; fax: (514) 398-6649; email: yvonne.steinert@mcgill.ca

ISSN 0142–159X print/ISSN 1466–187X online/06/060497–30 © 2006 Informa UK Ltd.
DOI: 10.1080/01421590600902976

Introduction

> Academic vitality is dependent upon faculty members' interest and expertise; faculty development has a critical role to play in promoting academic excellence and innovation. (Wilkerson & Irby, 1998)

Faculty development, or staff development as it is often called, has become an increasingly important component of medical education. Whereas it was once assumed that a competent basic or clinical scientist would naturally be an effective teacher, it is now acknowledged that preparation for teaching is essential. Given the increasing complexity and pressures of healthcare delivery, new approaches to teaching and learning, and competing demands on teachers' time, faculty members require a broad range of teaching and learning strategies that can be used in diverse settings. To help faculty members fulfill their multiple roles, a variety of faculty development programs and activities have been designed and implemented. These activities include workshops and seminars, short courses and site visits, fellowships and other longitudinal programs. Many of these activities have been designed to improve teacher effectiveness across the medical education continuum (e.g. undergraduate and postgraduate education), and they have been offered to healthcare professionals at local, regional and national levels (Clark *et al.*, 2004; Skeff *et al.*, 1997). However, despite numerous descriptions of program development and implementation, there is a paucity of research demonstrating the effectiveness of faculty development interventions. The goal of this report is to present the results of a systematic review of the impact of faculty development initiatives on teaching effectiveness in medical education. It is hoped that such a review of existing research will help to synthesize our knowledge of the field and guide future program development and evaluation.

Faculty development

Faculty development has been defined as that broad range of activities that institutions use to *renew* or *assist* faculty in their roles (Centra, 1978), and includes initiatives designed to improve the performance of faculty members in teaching, research and administration (Sheets & Schwenk, 1990). In many ways, faculty development is a planned program to prepare institutions and faculty members for their academic roles, including teaching, research, administration, writing and career management (Bland *et al.*, 1990). Faculty development is also meant to improve practice and manage change (Bligh, 2005), by enhancing individual strengths and abilities as well as organizational capacities and culture.

Faculty development programs have been classified in different ways. Ullian & Stritter (1997) describe a typology that includes organizational strategies, fellowships, comprehensive local programs, workshops and seminars, and individual activities. Wilkerson & Irby (1998) offer a different classification, ranging from professional orientation for new faculty members to instructional development, leadership development and organizational development. These authors also suggest that all four elements comprise a comprehensive approach to faculty development that is fundamental to academic vitality. Bligh (2005) has made a similar suggestion, stating that faculty development programs are outward signs of the inner faith that institutions have in their workforce, and that successful faculty development is expected to result in improved teaching performance and better learning outcomes for students or doctors. Examples of such improvements include the development of new teaching skills or assessment techniques, better ways of planning or implementing curricula, new ways of thinking about the student–teacher relationship, and increased commitment to educational scholarship.

To date, a number of publications have reviewed the effectiveness of faculty development activities. In 1984, Sheets & Henry observed that despite the growth in faculty development programs, evaluation of these initiatives was a rare occurrence, usually consisting of short questionnaires tapping participants' satisfaction. In 1990, Sheets & Schwenk reviewed the literature on faculty development activities for family medicine educators and made a similar observation, calling for more rigorous evaluations based on observed changes in participant behavior. In 1992, Hitchcock *et al.* summarized earlier reviews of the faculty development literature (e.g. Stritter, 1983; Bland & Schmitz, 1986; Sheets & Schwenk, 1990) and concluded that the concept of faculty development was evolving and expanding. In particular, they observed that teaching skills were a prominent aspect of faculty development, that fellowships were being used effectively to recruit and train new faculty, and that the efficacy of faculty development needed better research documentation. In 1997, Reid *et al.* reviewed 24 papers (published between 1980 and 1996) and concluded that despite some positive outcomes for fellowships, workshops and seminars, methodological weaknesses precluded definitive conclusions regarding faculty development outcomes. In 2000, Steinert highlighted the need for faculty development to respond to changes in medical education and healthcare delivery, to continue to adapt to the evolving roles of faculty members, and to conduct more rigorous program evaluations. She also commented that faculty development programs need to broaden their focus, consider diverse training methods and formats, and foster new partnerships and collaborations.

Notably, none of the above authors conducted a systematic review of the literature, and none of the reviews followed a predetermined protocol. In addition, few reviews considered the impact of faculty development on the organizations/institutions in which individuals work.

Best Evidence Medical Education

The Best Evidence Medical Education (BEME) Collaboration involves an international group of individuals, universities and organizations committed to moving the education of physicians from ''opinion-based education to evidence-based education'' (Harden *et al.*, 1999). Its goal is to provide medical teachers and administrators with the latest findings from scientifically grounded educational research to provide a basis for informed decisions. The international BEME Collaboration has three main purposes: to produce systematic reviews of medical education research studies that capture the best evidence available; to disseminate information worldwide; and to create a culture of

best-evidence medical education among teachers, administrators, and educational institutions (http://www.bemecollaboration.org/).

In 2001, the BEME Collaboration established a Faculty Development Topic Review Group (TRG) to review the 'best evidence' in faculty development. The TRG was deliberately international in its composition, and consisted of individuals with an expertise in faculty development, medical education and research methodology. The following report describes the review process and synthesizes its results.

This report is structured in the following way:

- **Objectives**—which summarizes the overall objectives of this review.
- **Review question**—which describes the evolution of the review question.
- **Review methodology**—which includes group formation, the pilot process, the development of a conceptual framework for faculty development and assessing outcome, inclusion/exclusion criteria, the search strategy and sources of papers, and selection methods and judgment of methodological quality.
- **Data management techniques**—which includes data extraction, analysis and synthesis.
- **Review findings**—which includes an overview of the studies included in this review, narrative comments on both the review results and the methodological quality of the studies, and a summary of the results, by program type and for a select group of eight studies that received the highest scores for study quality and strength of findings.
- **Discussion**—which highlights the major results of this review by summarizing outcomes, describing 'key features' of faculty development activities, and discussing observations regarding faculty development interventions and methodological issues.
- **Conclusion**—which describes implications for practice and future research as well as the strengths and limitations of this review.

Objectives

The goal of this review is to determine the effect of faculty development activities on faculty members' teaching abilities and to assess the impact of these activities on the institutions in which these individuals work. We focused specifically on programs designed to improve faculty members' teaching abilities because the majority of faculty development programs have targeted this particular role (Hitchcock *et al.*, 1992; Irby 1996); instructional effectiveness is central to the mission of medical education; and we wanted to limit the scope of our search to a feasible task. We did not examine faculty development programs designed to improve research or writing skills, administrative or management skills, or professional academic skills (career development). We also chose to limit the review to faculty development programs designed for teachers in medicine, and did not examine those programs specifically designed for residents or other healthcare professionals (e.g. nurses; dentists). All types of faculty development interventions (e.g. workshops, short courses and seminars, and fellowships) were included in the review.

Review question

The selection of the topic review question required several iterations. The BEME Steering Committee originally gave the TRG the following question:

What are the features of faculty development that make it effective?

After initial discussion and a pilot review of five papers (which will be described in the following section), the TRG revised the review question as follows:

Does faculty development make a difference?

- What makes for effective faculty development?
- Does participation in faculty development improve faculty members' teaching, research and administrative skills?
- Does faculty development have an impact on the institutional climate and organization?

However, after a more extensive pilot review of 30 papers (also described in the next section) and the 'state of the art' of the literature in 2002, the TRG refined the question as follows:

What are the effects of faculty development interventions on the knowledge, attitudes and skills of teachers in medical education, and on the institutions in which they work?

In addition, we also explored the following questions:

- What characterizes the faculty development activities that have been described?
- What are the methodological strengths and weaknesses of the reported studies?
- What are the implications of this review for faculty development practices and ongoing research in this area?

Review methodology

Group formation

An international Topic Review Group (TRG) of individuals representing six countries was constituted. Three criteria were used to invite individuals for TRG participation: international diversity; practical experience in faculty development and medical education; and expertise in educational research methodology.

The pilot process

A two-step pilot process was undertaken to prepare for the formal, systematic review.

Pilot I: All TRG members reviewed five articles (chosen by the lead reviewer) to determine the scope of the review, to refine the review question, and to assess the applicability of the BEME Coding Sheet (http://www.bemecollaboration.org/). Following this initial step, we identified areas of the BEME Coding Sheet that required adaptation for our review (e.g. target population; stated intervention; expected learning outcomes; impact of the intervention; and study design); highlighted areas for reviewer training; and further refined

the review question. Modifications to the BEME Coding Sheet were required in most categories.

Pilot II: The second step consisted of a pilot review of 30 articles that addressed all aspects of faculty development (i.e. a focus on teaching, research and administration). Two TRG members reviewed each paper, which enabled us to 'test' our faculty development BEME Coding Sheet, determine a process for working together and further refine the review question. At this stage, we decided to focus specifically on faculty development designed to enhance teaching rather than other faculty roles. This step also helped us to finalize our coding sheet, identify additional needs for reviewer training to increase inter-rater reliability, and determine the full scope of the literature search.

Development of a conceptual framework

The pilot phase led to the development of a conceptual framework that guided this review (see Figure 1). This framework acknowledges the different roles of faculty members, of which teaching is one. It also highlights the fact that many mediating factors beyond specific faculty development activities can influence teacher effectiveness, and that outcome can be observed at a number of levels.

To classify and analyze outcomes, we used Kirkpatrick's model of educational outcomes (Kirkpatrick, 1994), which offers a useful evaluation framework for this purpose (see Figure 2). The model describes four levels of outcome: learners' reaction (to the educational experience); learning (which refers to changes in attitudes, knowledge and skills); behavior (which refers to changes in practice and the application of learning to practice); and results (which refers to change at the level of the learner and the organization). In his original work, Kirkpatrick (1967) asserted that these outcomes were not hierarchical and that the model is intended to provide a more holistic and comprehensive evaluation that can inform policy and program development. The model has also been used by other BEME groups (e.g. Issenberg *et al.*, 2005) as well as other review groups (e.g. Freeth *et al.*, 2003), and with some modifications, was well suited to our review.

Inclusion/exclusion criteria

Based on the pilot studies, the following criteria guided the selection of articles for review:

- *Faculty development focus*—Within our focus on faculty development interventions designed to improve teaching effectiveness, all types of activities, of whatever duration, were included. Faculty development activities that focused only on the teaching of specific content areas (e.g. addiction medicine; geriatric medicine) were excluded, unless they also addressed methods of teaching and learning.
- *Target population*—Faculty development activities for both basic science and clinical faculty in all areas of medicine were selected for this review. Interventions designed to improve teaching effectiveness of residents-in-training or

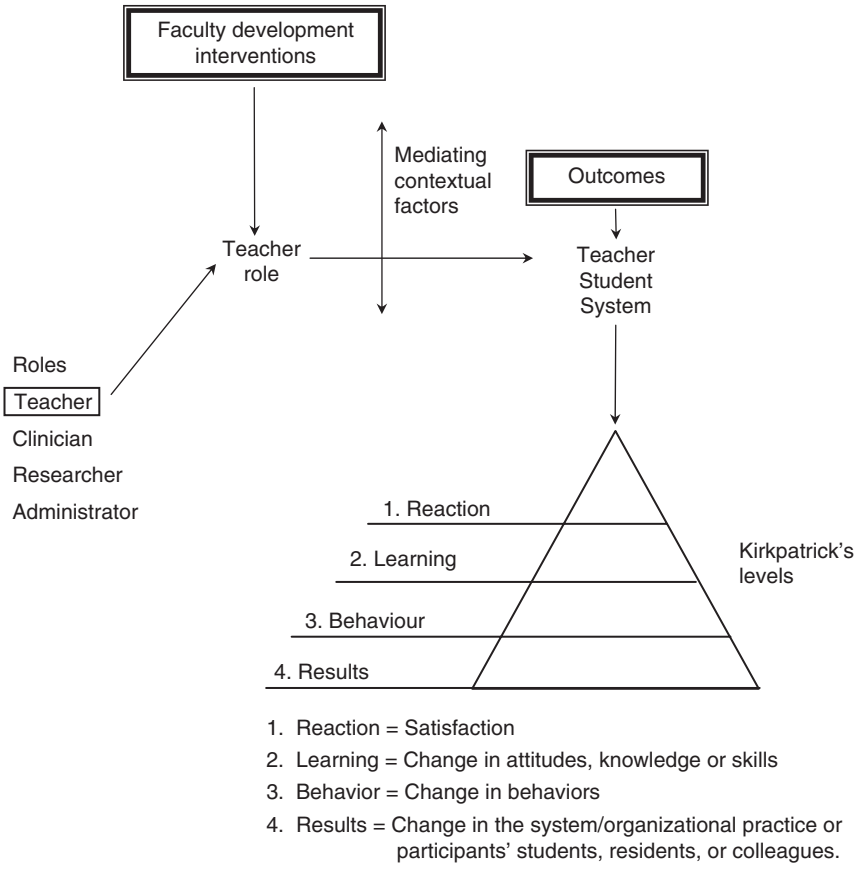

Figure 1. Conceptual framework for review.

Level 1	**REACTION**	Participants' views on the learning experience, its organization, presentation, content, teaching methods, and quality of instruction
Level 2A	**LEARNING**—Change in attitudes	Changes in the attitudes or perceptions among participant groups towards teaching and learning
Level 2B	**LEARNING**—Modification of knowledge or skills	For *knowledge*, this relates to the acquisition of concepts, procedures and principles; for *skills*, this relates to the acquisition of thinking/problem-solving, psychomotor and social skills
Level 3	**BEHAVIOR**—Change in behaviors	Documents the transfer of learning to the workplace or willingness of learners to apply new knowledge & skills
Level 4A	**RESULTS**—Change in the system/ organizational practice	Refers to wider changes in the organization, attributable to the educational program
Level 4B	**RESULTS**—Change among the participants' students, residents or colleagues	Refers to improvement in student or resident learning/performance as a direct result of the educational intervention

Figure 2. Kirkpatrick's model for evaluating educational outcomes.*
*Kirkpatrick's model (1994) was modified by Freeth et al. (2003) and was adopted by the BEME Collaboration. This model was further adapted for this review to include students, residents and colleagues (instead of patients) at level 4B.

other healthcare professionals (e.g. nursing) were excluded.

- *Study design*—We included all study designs across the positivist (empirical observation and measurement), interpretist (construction of understanding), and participatory (action research) paradigms (Creswell, 2003; Freeth *et al.*, 2005). However, only studies that included outcome data beyond participant satisfaction were examined. While participant satisfaction is important, we wished to explore evidence of learning and change.
- *Year of publication*—All articles assessing faculty development interventions from 1980–2002 were included in the search. 1980 was chosen based on the TRG's knowledge of the literature and the appearance of reports describing faculty development initiatives. The selection of papers for review was completed in 2002.
- *Language and geography*—The search was conducted to include all languages and sites of practice. The review, however, was limited to articles published in English, French, Spanish and German.

Search strategy and sources of papers

A literature search was conducted on Medline and ERIC using the following key words: staff development; in-service training; medical faculty; faculty training/development; and continuing medical education. (A copy of the search strategy is included in Appendix I, which is available on the BEME website: http://www.bemecollaboration.org) Only original research articles and reviews were retrieved. Editorials and essays were excluded. In addition, we conducted manual searches of the following journals: *Academic Medicine*, *Medical Education*, *Medical Teacher*, *Teaching and Learning in Medicine*, and *Advances in Health Sciences Education*. We also hand searched Proceedings of the Ottawa Conferences on Medical Education, reference lists of all review articles, and experts' recommendations of papers to be included. A search of EMBASE, using the same key words, did not yield any additional references.

Selection methods and judgment of methodological quality

The literature search resulted in a total of 2777 abstracts. A two-stage process was employed in the selection of studies eligible for review (Freeth *et al.*, 2003) and is outlined in Figure 3. Initially, each abstract was evaluated by the lead reviewer (YS) and another reviewer (AC), to ascertain whether the article related to faculty development and to teaching improvement. This resulted in 324 (12%) articles related to faculty development, of which 226 were related to teaching. Discrepancies in judgment between the two reviewers were resolved through discussion. A subsequent hand search (of all reference lists and the lead reviewer's own files) resulted in an additional 130 articles related to faculty development, of which 77 (60%) were related to teaching. Full texts of all 303 articles that related to teaching effectiveness and improvement (based on the literature search and the hand search) were retrieved for examination.

For the second step, the lead reviewer (YS), together with another reviewer (KM), reviewed all of the articles to apply the inclusion criteria. Fifty-three articles related to teaching improvement and included outcome data beyond satisfaction ratings. The remaining articles described faculty development programs with no evaluation data or consisted of

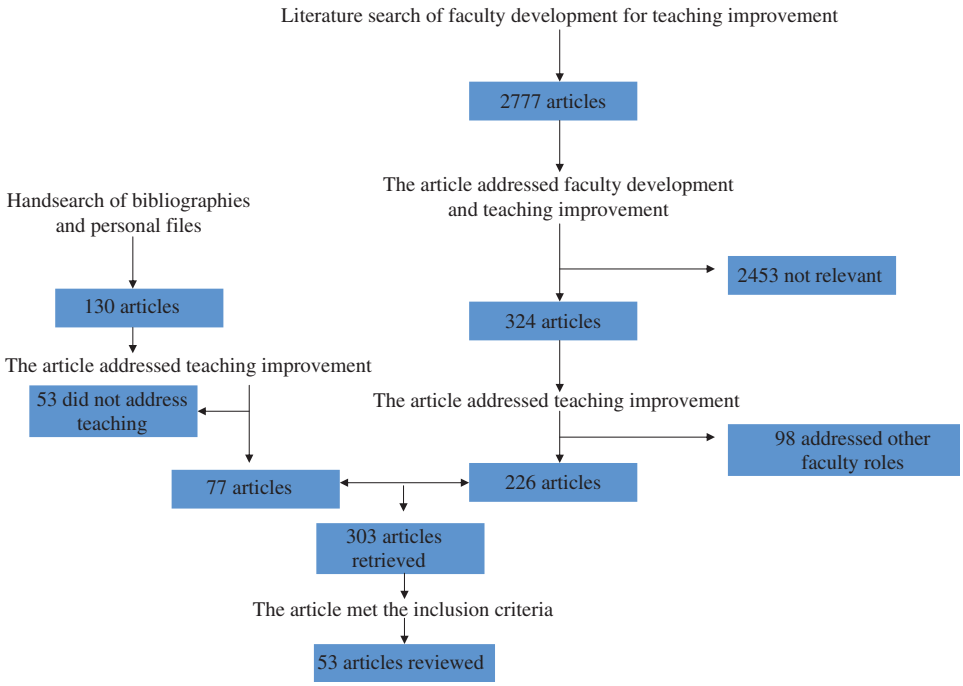

Figure 3. Literature review and selection of articles for review.

conceptual approaches to professional development; they were all eliminated. However, to contribute to a systematic cataloguing of the literature retrieved for this review, all articles were entered into *Reference Manager*. The use of *Reference Manager* as a bibliographic database has also been cited in other reviews (Reeves *et al.*, 2002).

Data management techniques

Data extraction, analysis and synthesis

Data extraction involved the manual completion of an abstract sheet for each study; this also allowed for a summary of the content of each paper reviewed (Freeth *et al.*, 2003). The Coding Sheet, which was based on the original prototype provided by the BEME Steering Committee, was modified to facilitate content specificity and data extraction. These modifications were informed by the pilot study, the TRG members' research experience and knowledge of the field, and texts on research methods (e.g. Dawson & Trapp, 2001; Creswell, 2002). (See Appendix II on the BEME website: http://www.bemecollaboration.org for a copy of the Faculty Development Coding Sheet.) Data were collected on the following items:

- expected learning outcomes;
- context of the intervention;
- description and impact of the intervention;
- evaluation methods, including study design, data-collection methods and data sources;
- study quality and strength of findings;
- avenues for further research;
- new insights and implications for faculty development.

For each report, reviewers were also asked to make a judgment and answer the following question: "Based on this intervention, does faculty development make a difference?"

Members of the TRG reviewed and coded each article in pairs. Results were entered into a central EXCEL database and verified for completion and accuracy. The EXCEL summary was then returned to one reviewer per team who was asked to resolve coding differences. Where necessary, the lead reviewer (YS) assisted in resolving differences; she also read all of the articles and coding sheets to ensure uniformity in approach.

Review findings

Overview of studies included in review

This review is based on 53 articles, all of which focused on faculty development to improve teaching effectiveness. Table 1 summarizes all of the interventions that were reviewed for this report.

This section will be organized into two main components:

(a) *Description of the interventions and expected outcomes*—which will be further divided into: setting, professional discipline, focus of the intervention, program type, instructional methods, duration, and level of outcome assessed.

(b) *Methodological quality of the studies*—which will be further divided into: study goal and theoretical framework, study design, data-collection methods, data sources, and study quality and strength of findings.

(a) Description of the interventions and expected outcomes

Setting: Of the 53 papers reviewed, 38 studies (72%) took place in the US, the remainder being in Canada, Egypt, Israel, Malta, Nigeria, the UK, Switzerland and South Africa. Most activities were delivered in a university, hospital or

Table 1. Summary of faculty development interventions reviewed.

Ref #	Authors[1]	Intervention type[2]	Setting & population	Study design[3]		Outcomes	Outcome Level[4]					
							1	2A	2B	3	4A	4B
1	Andriole, D.A., Evans, S.R., Foy, H.M., Atnip, R.G. & Mancino, A.T. (1998)	Workshop	USA, Mixed surgical specialties	Single group	Post-test, delayed post-test	Self-reported changes in teaching behavior.	√		√	√		
2	Baroffio, A., Kayser, B., Vermeulen, B., Jacquet, J. & Vu, N.V. (1999)	Workshop (2 levels)	Switzerland, Health science PBL tutors	Single group [& non-equivalent control group]	Pre-test, post-test	Improved student ratings of teaching behavior. Ratings improved with experience; greatest improvement among tutors with lowest ratings.				√		
3	Baxley, E.G., Probst, J.C., Schell, B.J., Bogdewic, S.P. & Cleghorn, G.D. (1999)	Workshop	USA, Family medicine teachers	Single group	Post-test only	Self-reported gains in knowledge (e.g. clinical teaching; curriculum planning) and teaching skills.	√		√			
4	Bing-You, R.G., Renfrew, R.A. & Hampton, S.H. (1999)	*Other:* site visits	USA, Community-based preceptors	Single group	Delayed post-test only	Self-reported changes in attitudes towards teaching (e.g. promoting reflection) & collegiality.	√	√				
5	Bird, J., Hall, A., Maguire, P. & Heavy, A. (1993)	Workshop	UK, Mixed clinical specialties	Single group	Delayed post-test only	Increased motivation and self-awareness as teachers; self-reported gains in teaching skills; new educational initiatives planned.	√	√	√	√		
6	Bland, C.J. & Froberg, D.G. (1982)	Seminar series	USA, Family medicine teachers	Single group, repeated measures	Pre-test, post-test	Self-reported gains in knowledge & skills in several areas (e.g. teaching; research; administration).	√		√			
7	Coles, C.R. & Tomlinson, J.M. (1994)	Workshop	UK, GP teachers	Single group	Post-test, delayed post-test	Self-reported changes in learner-centered teaching behaviors (e.g. negotiating objectives).	√			√		
8	DaRosa, D.A., Folse, J.R., Reznick, R.K., Dunnington, G.L. & Sachdeva, A.K. (1996)	Short course	USA, Surgical teachers	Single group	Post-test, delayed post-test	Self-reported changes in educational activities (e.g. curricular development & planning; educational administration).	√			√		

(*continued*)

Table 1. Continued.

Ref #	Authors[1]	Intervention type[2]	Setting & population	Study design[3]		Outcomes	Outcome Level[4]					
							1	2A	2B	3	4A	4B
9	Dennick, R. (1998)	Workshop	UK, Basic scientists & clinical teachers	Single group	Post-test only	Self-reported increases in knowledge about teaching as well as changes in teaching behavior (e.g. questioning learners, formulating objectives).	√		√	√		
10	DeWitt, T.G., Goldberg, R.L. & Roberts, K.B. (1993)	Seminar series	UK, Pediatric community preceptors	Single group	Pre-test, post-test	Increased familiarity with educational terms & principles; increased ability to analyse videotaped teaching encounters; self-reported changes in teaching behavior.	√		√	√		
11	Elliot, D.L., Skeff, K.M. & Stratos, G.A. (1999)	Longitudinal program	USA, Clinical teachers (family medicine & internal medicine)	Single group	Post-test only	Self-reported changes in attitudes & use of educational terms; self-reported changes in educational roles/ responsibilities.	√	√	√	√		
12	Gordon, G.H. & Levinson, W. (1990)	Short course	USA, Internal medicine teachers	Single group	Pre-test, post-test, delayed post-test	Self-reported changes in attitudes to learner-centered learning, with a decrease over time.	√	√				
13	Hewson, M.G. (2000)	Seminar series	USA, Clinical teachers (medicine & pediatrics)	Single group [& non-equivalent control group]	Pre-test, post-test	Changes in self-assessment of teaching competencies (e.g. clarifying expectations; giving feedback); increases in student & resident ratings of teacher behavior; new educational initiatives sustained.	√		√	√	√	
14	Hewson, M.G. & Copeland, H.L. (1999)	Seminar series	USA, Mixed clinical specialties	Single group [& non-equivalent control group]	Pre-test, post-test [& post-test only]	Improved student & resident ratings of teaching effectiveness.				√		
15	Hewson, M.G., Copeland, H.L. & Fishleder, A.J. (2001)	Seminar series	USA, Mixed clinical specialties	Single group	Pre-test, post-test	Self-reported gains in teaching skills (e.g. feedback; stimulating independent learning); improved student & resident ratings of teacher behavior.			√	√		

16	Hitchcock, M.A., Lamkin, B.D., Mygdal, W.K., Clarke, C.M. & Clarke, S.O. (1986)	Fellowship	USA, Family medicine teachers	Single group [& non-equivalent control group]	Pre-test, post-test	Self-reported changes in self-image as a clinical teacher (e.g. capability & authority).		√				
17	Irby, D.M., Vontver, L.A. & Stenchever, M.A. (1982)	Workshop	USA, Clinical teachers (Obstetrics and Gynecology)	Single group	Post-test only	Self-reported gains in enthusiasm about teaching & cohesiveness among faculty; self-reported gains in teaching skills & behavior; improved student ratings of teacher behavior.	√	√	√	√		
18	Johnson, D.H. & Zammit-Montebello, A. (1990)	Fellowship	Malta, Family medicine teachers	Single group	Pre-test, post-test	Self-reported changes in attitudes towards primary care & health promotion and teaching behaviors; observed changes in problem solving & communication skills; new educational practices & creation of an academic professional body.	√	√	√	√	√	
19	Litzelman, D.K., Stratos, G.A., Marriott, D.J., Lazaridis, E.N. & Skeff, K.M. (1998)	*Other:* Augmented feedback	USA, Clinical teachers	Randomized control trial	Pre-test, post-test	Improved student ratings of teacher performance; complex interactions with baseline performance.				√		
20	Lye, P.S., Simpson, D.E., Wendelberger, K.J. & Bragg, D.S. (1998)	*Other:* One-hour conferences	USA, Clinical teachers (Pediatrics)	Non-equivalent control group	Pre-test, post-test	Improved student and resident ratings of clinical teaching effectiveness.	√			√		
21	Mahler, S. & Benor, D.E. (1984)	Workshop	Israel, Basic science & clinical teachers	Single group	Pre-test, post-test, delayed post-test	Observed changes in teaching behavior (e.g. greater encouragement of verbal activity & greater ability to stimulate higher level of cognitive activity), maintained over time.				√		
22	Mahler, S. & Neumann, L. (1987)	Workshop	Israel, Basic science & clinical teachers	Single group	Pre-test, post-test	Observed changes in teaching behavior (e.g. increases in higher level cognitive activities & increases in cognitive diversity).				√		

(*continued*)

Table 1. Continued.

Ref #	Authors[1]	Intervention type[2]	Setting & population	Study design[3]		Outcomes	Outcome Level[4]					
							1	2A	2B	3	4A	4B
23	Marvel, M.K. (1991)	*Other:* Individual feedback session	USA, Family medicine teachers	Single group	Pre-test, post-test, delayed post-test	Observed & resident-reported changes in teaching behavior (e.g. increased reinforcement of learner efforts & use of open-ended questions); patients' ratings of residents showed some increases in 5 of 7 interview behaviors.	√			√		√
24	McDermott, J.F. & Anderson, A.S. (1991)	Workshop	USA, Clinical teachers	Single group	Pre-test, post-test	Gains in knowledge & skills related to PBL.			√			
25	Nasmith, L. & Steinert, Y. (2001)	Workshop	Canada, Clinical & basic science teachers	Non-equivalent control group	Post-test, delayed post-test	Self-reported increase in number of interactive lectures & use of interactive techniques; observations indicated greater use of 2 interactive techniques (e.g. questionning; using audience responses) but no significant differences in overall presentation skills.	√		√	√		
26	Nasmith, L., Steinert, Y., Saroyan, A., Daigle, N. & Franco, E. (1997)	Workshop	Canada, Family medicine teachers	Randomized control trial	Pre-test, post-test, delayed post-test	Non-significant increases in knowledge about small group teaching; self-reported changes in use of small group teaching methods but no observed significant differences.	√		√	√		
27	Nathan, R.G. & Smith, M.F. (1992)	Workshop	USA, Family medicine teachers	Single group	Pre-test, post-test, delayed post-test	Improved student ratings of teacher interactivity; no significant changes in student exam scores.	√			√		√
28	Olmesdahl, P.J. & Manning, D.M. (1999)	Workshop	South Africa, Health science teachers	Single group	Post-test only	Increased awareness of educational challenges in PBL groups (e.g. dispensing information; sharing expertise).	√	√				

29	Olukoya, A. (1986)	Workshop	Nigeria, All health professionals	Single group	Pre-test, post-test	Increased knowledge of educational methods.	√		√			
30	Pinheiro, S.O., Rohrer, J.D., & Heimann, C.F.L. (1998)	Fellowship	USA, Community-based osteopaths	Single group	Pre-test, post-test	Increased skills in collaborative teaching methods; observed changes in learner-centered approach.			√	√		
31	Pololi, L., Clay, M.C., Lipkin, M., Jr., Hewson, M., Kaplan, C. & Frankel, R.M. (2001)	Short course	USA, Mixed clinical specialties	Single group	Pre-test, post-test, delayed post-test	Increased awareness of learner-centered approaches & motivation to participate in ongoing faculty development; increased collegiality.	√	√				
32	Quirk, M.E., DeWitt, T., Lasser, D., Huppert, M. & Hunniwell, E. (1998)	Workshop	USA, Community-based teachers	Single group	Pre-test, post-test, delayed post-test	Increased familiarity with educational concepts (e.g. teaching style); self-reported changes in use of specific teaching behaviors (retained at 3 months). Observed changes in ability to analyse educational encounters.	√		√	√		
33	Rayner, H., Morton, A., McCulloch, R., Heyes, L. & Ryan, J. (1997)	Seminar series	UK, Mixed clinical specialties	Single group	Post-test only	Self-reported increases in awareness of teaching issues and methods; committee to link network of teachers established.	√	√			√	
34	Sachdeva, A.K. & Kelliher, G.J. (1994)	Workshop	USA, Health science teachers	Single group	Delayed post-test only	Self-reported changes in teaching style (e.g. more student-centered, interactive); self-reported changes in leadership roles, scholarly activities & receipt of teaching awards; dissemination of skills in participants' institutions.			√	√	√	
35	Schmidt, H. G., Bouhuijs, P. A. J., Khattab, T. & Makladi, F. (1989)	Workshop	Egypt, Clinical & basic science teachers	Single group	Pre-test, post-test	Significant changes in attitudes towards teaching & learning (e.g. from teacher-oriented to student-oriented).		√				

(*continued*)

Table 1. Continued.

Ref #	Authors[1]	Intervention type[2]	Setting & population	Study design[3]		Outcomes	Outcome Level[4]					
							1	2A	2B	3	4A	4B
36	Sheets, K.J. (1985)	Fellowship	USA, Allopathic & osteopathic physicians	Single group	Pre-test, post-test, delayed post-test	Gains in cognitive knowledge; observed changes in teaching behavior.	√		√	√		
37	Sheets, K.J. & Henry, R.C. (1984)	Short course	USA, Family medicine teachers	Single group	Pre-test, post-test, delayed post-test	Gains in knowledge, maintained over time (with slight drops at delayed post-tests); self-reported improvement in presentation skills, clinical teaching skills & AV techniques; slight improvement in observed teaching behaviors; correlation between cognitive tests, self-ratings & observations.	√		√	√		
38	Sheets, K.J. & Henry, R.C. (1988)	Short course	USA, Allopathic & osteopathic physicians in family medicine	Single group	Pre-test, post-test, delayed post-test	Gains in knowledge; self-reported changes in presentation skills, clinical teaching skills & AV techniques; slight improvement in observed teaching behaviors.	√		√	√		
39	Skeff, K.M. (1983)	*Other:* Intensive feedback method	USA, Clinical teachers (Internal medicine & neurology)	Randomized control trial	Pre-test, post-test	Significant changes in awareness of teaching abilities and desire to evaluate & improve teaching in feedback groups; significant observed improvement in targeted teaching behaviors and overall teaching performance in intensive feedback group; no significant differences in student & resident ratings of teacher behavior.	√	√		√		

40	Skeff, K.M., Campbell, M., Stratos, G.A., Jones, H.W. & Cooke, M. (1984)	*Other:* Seminar method	USA, Clinical teachers (Internal medicine)	Randomized control trial	Pre-test, post-test, delayed post-test	Significant changes in attitudes towards teaching & awareness of strengths & limitations; self-reported changes in knowledge of teaching & learning as well as teaching behavior.	✓	✓	✓	✓		
41	Skeff, K.M. & Stratos, G.A. (1985)	*Other:* Seminar or intensive feedback method	USA, Clinical teachers	Randomized control trial	Pre-test, post-test	Self-reported gains in instructional skills & use of teaching methods; observed changes in teaching behaviors; no significant differences in student & resident ratings.	✓		✓	✓		
42	Skeff, K.M., Stratos, G. A. & Bergen, M.R. (1992a)	Seminar series	USA, Clinical teachers	Single group	Pre-test, post-test	Retrospective ratings revealed more significant changes in teaching performance & attitudes than traditional pre-post comparisons; student ratings were more consistent with faculty's retrospective pre-post than traditional pre-post comparisons.		✓		✓		
43	Skeff, K.M., Stratos, G.A., Bergen, M.R., Albright, C.L., Berman, J., Farquhar, J.W. & Sox, H.C. (1992b)	Short course	USA, Clinical teachers	Single group	Post-test only	Majority of participants conducted seminar series in own settings & reported changes in educational practices; 67 facilitators trained >500 faculty and 200 residents in groups of 6–10.	✓			✓	✓	

(*continued*)

Table 1. Continued.

Ref #	Authors[1]	Intervention type[2]	Setting & population	Study design[3]		Outcomes	Outcome Level[4]					
							1	2A	2B	3	4A	4B
44	Skeff, K.M., Stratos, G.A., Bergen, M.R. & Regula, D.P. (1998)	Seminar series	USA, Basic science teachers (pathologists)	Single group	Pre-test, post-test, delayed post-test	Self-reported changes in philosophy & enthusiasm about teaching, knowledge of teaching principles, ability to analyse teaching & teaching methods (maintained over time); self-reported & observed changes in teaching behavior; improved student ratings of teacher behavior.	✓	✓	✓	✓		
45	Skeff, K.M., Stratos, G.A., Bergen, M.R., Sampson, K. & Deutsch, S.L. (1999)	Workshop	USA, Clinical teachers (Internal medicine, pediatrics, family medicine & sub specialties)	Single group	Pre-test, post-test	Self-reported changes in attitudes & enthusiasm towards teaching as well as knowledge of teaching principles and ability; significant self-reported changes in teaching behaviors (e.g. provision of positive learning climate & feedback).	✓	✓	✓	✓		
46	Skeff, K.M., Stratos, G.A., Berman, J. & Bergen, M.R. (1992c)	Seminar series	USA, Clinical teachers	Single group	Pre-test, post-test	Self-reported changes in attitudes towards teaching; self-reported gains in awareness of personal strengths & weaknesses; student & resident ratings indicated some changes in teaching behavior.	✓	✓	✓	✓		
47	Skeff, K.M., Stratos, G.A., Campbell, M., Cooke, M. & Jones, H.W., III (1986)	*Other:* Seminar method	USA, Mixed clinical specialties - from 4 sites	Randomized control trial	Pre-test, post-test, delayed post-test	Significant changes in attitudes towards teaching; self-reported changes in teaching behavior; observed changes in teaching behavior with significant improvements in learning climate & control of session; student & resident ratings of own behaviors indicated some changes (e.g. doctor-patient communication skills).	✓	✓		✓		✓
48	Snyder, S. (2001)	Workshop	USA, Family medicine	Single group	Post-test only	Self-reported changes in skills related to curriculum development; majority of participants developed curriculum projects; changes to residency curricula made.	✓		✓	✓	✓	

49	Steinert, Y., Nasmith, L., Daigle, N. & Franco, E.D. (2001)	Workshop	Canada, Mixed health professions	Single group	Pre-test, post-test, delayed post-test	Self-reported changes in attitudes towards working with "problem" residents; self-reported changes in knowledge & skills of working with ''problem'' residents (maintained over time).	✓	✓	✓			
50	Stratos, G.A., Bergen, M.R., Albright, C.L., Skeff, K.M. & Owens, D.K. (1997)	Seminar series	USA, Ambulatory care physicians	Single group	Pre-test, post-test	Self-reported changes in teaching skills (e.g. feedback; stimulating independent learning); self-reported increases in confidence & ability to teach medical decision-making; some changes made in own setting.	✓		✓		✓	
51	Stuart, M.R., Orzano, A.J. & Eidus, R. (1980)	*Other:* One-on-one faculty coaching	USA, Community preceptors	Single group	Pre-test, post-test	Observed changes in teachers' preceptor skills.	✓			✓		
52	Valdiserri, R.O., Koziol, S.M., Korth, W.W. & Haynes, L.T. (1986)	Workshop	USA, Pathology	Single group	Pre-test, post-test	Self-reported gains in knowledge (e.g. learning objectives; use of questions; evaluation).			✓			
53	Wilkerson, L. & Sarkin, R.T. (1998)	Workshop	USA, Ambulatory care teachers	Single group	Pre-test, post-test	Significant gains in knowledge about ambulatory teaching.			✓			

[1] This table is organized alphabetically by author.

[2] In most instances, intervention type reflects the authors' exact wording.

[3] In two-part studies, both study designs are reported, with the second one in brackets.

[4] Outcome levels are based on Kirkpatrick's model for evaluating educational outcomes, as described in Figure 2.

community setting, with several initiatives offered by professional associations.

Professional discipline: The majority of faculty development interventions targeted practicing clinicians, with a preponderance of activities in family medicine and internal medicine. Interestingly, 21 of the faculty development initiatives (40%) welcomed more than one clinical discipline. Five interventions (10%) were designed for both clinicians and basic scientists; an additional two (4%) targeted basic scientists only. The number of participants in the interventions (which does not equal respondents for the evaluative component) ranged from six to 399, with a mean attendance of 60. In programs that extended over time, some participants attended only one session; a few attended all. The majority of teachers participated on a voluntary basis.

Focus of the intervention: As a result of the selection criteria, all of the reports focused on teaching improvement. The majority aimed to improve clinical teaching, with a secondary emphasis on feedback and evaluation, small-group teaching and lecturing skills. Several studies highlighted 'learner centeredness' as an outcome, and several others focused on the teaching of specific content areas in addition to general teaching improvement (e.g. communication skills and medical interviewing; principles of family medicine and preventive medicine). Although the primary focus of these reports was instructional improvement, many also addressed personal/career development, organizational change, administration and educational leadership, and research skills.

Program type: The majority of activities were workshops ($n = 23$; 43%), of varying duration. Ten (19%) of the interventions were described as a seminar series and six (11%) as a short course. Five (10%) were described as a longitudinal program (e.g. fellowship) and nine (17%) fell under 'other', which included a seminar method, individual or augmented feedback, or site visits. An inconsistent and variable use of terms (e.g. workshops and seminars; seminars and short courses), complicated this classification; however, whenever possible, the authors' terminology was used.

Instructional methods: All reports described a wide range of instructional methods that included lectures, small-group discussions, interactive exercises, role plays and simulations, films and videotape reviews of performance. No programs were completely lecture-based, and the majority included an experiential component with opportunities for guided practice with feedback (i.e. micro-teaching). Some programs offered on-site training opportunities where teachers could readily apply what they learned. Few described a direct link to teachers' ongoing educational activities, although educational projects and *in vivo* practice were part of several interventions (most notably seminars and short courses). Needs assessments were used sparingly.

Duration: The faculty development interventions ranged in duration from one hour to one year. Workshops, which were generally one-time interventions, ranged in duration from three hours to one week, with a median duration of two days. The seminar series, which occurred over time, ranged in duration from 12 hours to one month (with a median duration of 14 hours), and the short courses ranged from one week to one month. Fellowships were both full time and part time in nature, and one intervention, entitled a 'longitudinal program', was 50 hours in length over 18 months.

Table 2. Summary of faculty development outcomes by Kirkpatrick level.*

Reaction	74%
Learning	77%
19/53 assessed self-reported changes in attitudes	
31/53 assessed self-reported changes in knowledge/skills	
Behavior	72%
13/53 assessed self-reported changes in behavior	
25/53 assessed observed changes in behavior	
Results	19%
7/53 assessed change in organizational practice	
3/53 assessed change in students/residents	

*Numbers may not equal 100% as some studies assessed outcomes in more than one way.

Level of outcome assessed: Table 2 shows that 39 studies (74%) assessed *reaction*, which included participant satisfaction, perception of program usefulness and acceptability, and value of the activity. Forty-one studies (77%) assessed *learning*, which included changes in attitudes, knowledge or skills. Thirty-eight (72%) assessed change in *behavior*. At the *results* level, seven studies (13%) reported change in organizational practice and three (6%) assessed change in student or resident learning.

(b) Methodological quality of the studies

Study goal and theoretical framework: All 53 reports stated their objective, sometimes quite broadly (e.g. to describe, implement and evaluate a faculty development initiative). Some reports described more specific objectives, outlining a particular study question such as assessing the effectiveness of a faculty development program on teaching behaviors (Hewson, 2000) or attitudes (Schmidt *et al.*, 1989). One study examined the effect of experience on workshop gains (Baroffio *et al.*, 1999), and several others assessed different methods of assessment (Nasmith *et al.*, 1997; Hewson *et al.*, 2001) and program evaluation (Sheets, 1985). All but seven cited the relevant literature, though often in a very limited fashion. Thirty reports (57%) placed their work within a conceptual or theoretical framework, primarily drawing upon principles of adult learning, instructional design, experiential learning and reflective practice.

Study design: Of the 53 papers reviewed, there were six (11%) randomized controlled trials. The majority of studies ($n = 47$; 89%) were quasi-experimental in design, with two including a comparison group in the main part of the study. Of the 45 single-group designs, 31 (69%) employed a pre-test–post-test design. Fourteen studies (26%) used a post-test only. None of the reports used a qualitative approach only,

though 11 (21%) incorporated a qualitative method (or analysis) in their design.

Data collection methods: Methods to evaluate faculty development programs included end-of-workshop questionnaires, pre- and post-test measures to assess attitudinal or cognitive change, student, resident and self-assessment of post-training performance, and direct observations of teaching behavior. Questionnaires were the most popular method of data collection. All but four of the interventions used a survey or questionnaire. Twenty-nine (55%) of the interventions used a questionnaire only; 20 (38%) used a questionnaire and another method (e.g. observation; expert opinion). Most questionnaires were designed for a particular study, and few reports described psychometric properties. Sixteen studies (30%) included direct observation (of live or videotaped teaching sessions) as part of their assessment methodology.

Data sources: The majority of programs relied on self-reported ratings of teaching, with a limited use of performance-based measures of change. Fifteen studies (28%) employed student or resident ratings to assess changes in teaching behaviors. An additional two used expert opinions to assess outcomes. One study assessed student exam scores; another included patient ratings of resident behaviors. In many studies, the response rates for outcome measures were low or unspecified; statistical methods or differences were often not described.

Study quality and strength of findings: Study quality was rated on a five-point scale (1 = low; 5 = high), and reviewers were asked to indicate study strengths and weaknesses. We had originally included subscales to rate the evaluation methods (e.g. appropriateness of and implementation of study design; appropriateness of data analysis), but this did not yield reliable results. We therefore chose to use an overall rating for this variable. Strength of findings was rated on a five-point scale with specific anchors (1 = no clear conclusions can be drawn; 3 = conclusions can probably be based on results; 5 = results are unequivocal). The mean rating for study quality was 3.14, with a range from 1 to 5. The mean rating for strength of findings was 2.88 (with a range of 1–4).

Summary of findings by intervention type

We present the study findings according to the type of intervention. Within each classification, of workshop, short course, seminar series, longitudinal program and fellowship, we describe the measures generally used and the results obtained at each level of Kirkpatrick's model of educational outcomes. We did not perform this analysis for programs in the 'other' category, as the diversity of interventions in this group precluded such an analysis. Following this summary, we will present a 'focused picture' in which we describe the findings of eight studies that received the highest overall reviewer ratings for both study quality and strength of findings.

(a) Workshops

Twenty-three of the interventions reported were described as workshops, most commonly a single intervention of varying duration (Irby *et al.*, 1982; Mahler & Benor, 1984; Olukoya, 1986; Valdiserri *et al.*, 1986; Mahler & Neumann, 1987; Schmidt *et al.*, 1989; McDermott & Anderson, 1991; Nathan & Smith, 1992; Bird *et al.*, 1993; Coles & Tomlinson, 1994; Sachdeva & Kelliher, 1994; Nasmith *et al.*, 1997; Andriole *et al.*, 1998; Dennick, 1998; Quirk *et al.*, 1998; Wilkerson & Sarkin, 1998; Baroffio *et al.*, 1999; Baxley *et al.*, 1999; Olmesdahl & Manning, 1999; Skeff *et al.*, 1999; Nasmith & Steinert, 2001; Snyder, 2001; Steinert *et al.*, 2001). Only seven of the 23 stated a theoretical or conceptual framework.

Level 1—Reaction: At level one, satisfaction was usually measured on a Likert scale, of 4–5 points, or a comparable categorical scale, from poor to excellent. The majority of participants consistently rated the workshops as helpful, relevant and useful in providing an opportunity for sharing with other teachers. While many aspects of the workshops were found to be of value, micro-teaching and working on specific skills (i.e. the opportunity to practice) were very well received.

Level 2a—Learning: Outcomes at this level addressed attitudes. Participants reported increased motivation, self-awareness and enthusiasm. They also reported increased understanding of, and intent to try, learner-centered techniques.

Level 2b—Learning: Changes in knowledge and skill, from pre-test to post-test measures, were frequently reported for this outcome. More specifically, a greater understanding and use of specific teaching skills and behaviors (e.g. questioning skills; increasing student participation) were noted, primarily through self-report. Very few interventions used a control group. In those which did (Nasmith *et al.*, 1997; Nasmith & Steinert, 2001), no statistically significant differences were reported, although the experimental groups tended to report greater familiarity with concepts.

Level 3—Behavior: 15 reports evaluated outcomes at level 3 (Irby *et al.*, 1982; Mahler & Benor, 1984; Mahler & Neumann, 1987; Nathan & Smith, 1992; Bird *et al.*, 1993; Coles & Tomlinson, 1994; Sachdeva & Kelliher, 1994; Nasmith *et al.*, 1997; Andriole *et al.*, 1998; Dennick, 1998; Quirk *et al.*, 1998; Baroffio *et al.*, 1999; Skeff *et al.*, 1999; Nasmith & Steinert, 2001; Snyder, 2001), primarily through self-reports. Teachers reported improvements in their teaching abilities and use of specific approaches to teaching. In one case, they reported that they had undertaken new curriculum projects (Snyder, 2001). Student ratings were reported by three authors (Irby *et al.*, 1982; Nathan & Smith, 1992; Baroffio *et al.*, 1999). In Baroffio *et al.*'s study (1999), student ratings of teacher behavior improved significantly. Another study (Irby *et al.*, 1982) found that self-reports and observable behaviors matched; a third (Nasmith *et al.*, 1997) did not. In those studies where post-tests and delayed post-test comparisons were made, changes appear to have been maintained (Mahler & Benor, 1984; Steinert *et al.*, 2001).

Level 4a—Results: Outcomes at this level assessed change in organizational practice. Only three reports (Nathan & Smith, 1992; Sachdeva & Kelliher, 1994; Snyder, 2001) examined

outcomes at this level. In one study, faculty members reported curriculum development and enhancement (Snyder, 2001). Another study described the dissemination of skills at the participants' home institutions (Sachdeva & Kelliher, 1994).

Level 4b—Results: Only one study assessed change among the participants' students (Nathan & Smith, 1992). This study found no difference in student examination performance.

(b) Short courses

Six of the 54 interventions (Sheets & Henry, 1984, 1988; Gordon & Levinson, 1990; Skeff *et al.*, 1992b; DaRosa *et al.*, 1996; Pololi *et al.*, 2001) were in the form of a short course, ranging in duration from one week to one month. All had a stated objective and all but one provided a theoretical framework.

Level 1—Reaction: As in workshops, participants' reactions were generally measured on a 5- or 10-point Likert scale, with most respondents indicating a high level of satisfaction and strong recommendations for peers to attend similar events.

Level 2a—Learning: Both participants and faculty developers reported increased positive attitudes to learner-centered learning (Gordon & Levinson, 1990; Pololi *et al.*, 2001), although this decreased in one study's delayed post-test (Gordon & Levinson, 1990). A sense of increased self-awareness and collegiality was also reported in one study, with over 90% of participants deciding to apply for a year-long follow-up program (Pololi *et al.*, 2001).

Level 2b—Learning: Two studies measured outcome at this level (Sheets & Henry, 1984, 1988). In both, knowledge improved from pre-test to post-test, and this change was maintained over time despite a small decrease in scores. Both also indicated self-rated improvement in presentation skills, clinical teaching and the use of audiovisual techniques.

Level 3—Behavior: Four of the six short-course studies collected data relating to level 3 outcomes (Sheets & Henry, 1984, 1988; Skeff *et al.*, 1992b; DaRosa *et al.*, 1996). In one study, more than half of the participants reported taking on additional educational activities related to the initial course (DaRosa *et al.*, 1996). Two other studies (Sheets & Henry, 1984, 1988), which relied on analysis of videotaped and simulated teaching sessions to document change, showed slightly continued improvements at four and nine months, although there is no statistical support for this claim. The fourth study (Skeff *et al.*, 1992b) described self-reported changes in educational practices.

Level 4a—Results: One study (Skeff *et al.*, 1992b), which aimed to foster change in the participants' institutions, tracked dissemination activities following a month-long course and found that 67 participants trained more than 500 faculty and 200 students (in groups of six to 10) in their own settings.

(c) Seminar series

Ten studies described a seminar series characterized by the fact that the sessions were spaced over time (Bland & Froberg, 1982; Skeff *et al.*, 1992a, 1992c; DeWitt *et al.*, 1993; Rayner *et al.*, 1997; Stratos *et al.*, 1997; Skeff *et al.*, 1998; Hewson & Copeland, 1999; Hewson, 2000; Hewson *et al.*, 2001). Eight of these reported a theoretical framework (Bland & Froberg, 1982; Skeff *et al.*, 1992a, 1992c; DeWitt *et al.*, 1993; Stratos *et al.*, 1997; Skeff *et al.*, 1998; Hewson & Copeland, 1999; Hewson, 2000).

Level 1—Reaction: All but three reports presented data on the participants' reactions. As with the other interventions, ratings of satisfaction and perceived usefulness of the intervention were high. In particular, participants valued skill-building activities such as micro-teaching and group discussions, exchanges regarding teaching, and the development of an educational framework. When asked, participants reported that they would recommend the seminar to colleagues.

Level 2a—Learning: Impact on attitudes and perceptions was reported by four studies (Skeff *et al.*, 1992a, 1992c, 1998; Rayner *et al.*, 1997). This included raised awareness of teaching issues, teaching methods and theory (Rayner *et al.*, 1997) as well as positive effects on enthusiasm and attitudes towards teaching. Where specific attitudes were measured (e.g. desire to evaluate and improve teaching; satisfaction with teaching), they increased significantly following the intervention. Of interest is that changes were greater when both pre- and post-intervention ratings were gathered following the intervention (Skeff *et al.*, 1992a).

Level 2b—Learning: In the studies that assessed knowledge and skills (Bland & Froberg, 1982; DeWitt *et al.*, 1993; Hewson, 2000; Hewson *et al.*, 2001), positive results were shown in both. Hewson *et al.* (2001) observed improvement in instructional skills such as clarifying expectations, checking learners' understanding, providing feedback and tailoring instruction to learners' needs. Teachers also reported an increased ability to assess their strengths and weaknesses (Skeff *et al.*, 1992c), enhanced knowledge regarding teaching principles and skills in analyzing teaching (DeWitt *et al.*, 1993; Skeff *et al.*, 1998), and an improvement in their ability to influence the learning environment (Stratos *et al.*, 1997).

Level 3—Behavior: Level 3 results reporting changed behavior were presented in seven studies (Skeff *et al.*, 1992a, 1992c; DeWitt *et al.*, 1993; Skeff *et al.*, 1998; Hewson & Copeland, 1999; Hewson, 2000; Hewson *et al.*, 2001). Some improvements were self-reported. However, student ratings of teachers' behaviors also changed. In one study, ratings by participants' students and residents improved significantly from pre-test to post-test (Hewson, 2000). Moreover, median ratings for participants significantly exceeded those of the control group in two reports (Hewson & Copeland, 1999; Hewson, 2000). In two other studies, student ratings of teacher behavior were significantly improved in certain areas (e.g. specifying expectations and establishing effective teacher–student relationships) (Skeff & Stratos, 1985; Hewson *et al.*, 2001). Self-reported increases

were not consistently reflected in student and resident ratings.

Level 4a—Results: Three reports of change at the level of impact on the organization were found (Rayner *et al.*, 1997; Stratos *et al.*, 1997; Hewson, 2000). This included the implementation of new educational activities, although these were not quantified (Hewson, 2000). Other changes at the organizational level included the formation of a network to support teachers in the clinical setting and increased cooperative interactions with colleagues (Rayner *et al.*, 1997). In one case (Stratos *et al.*, 1997), participants returned to their own settings to facilitate other faculty members' learning and implemented changes generated during the seminar.

(d) Longitudinal programs and fellowships

One report described a longitudinal program (Elliot *et al.*, 1999) and four described fellowships (Sheets, 1985; Hitchcock *et al.*, 1986; Johnson & Zammit-Montebello, 1990; Pinheiro *et al.*, 1998). All had stated objectives and all but one incorporated a theoretical framework.

Level 1—Reaction: Where reaction was assessed (Sheets, 1985; Johnson & Zammit-Montebello, 1990; Elliot *et al.*, 1999), high levels of satisfaction with the intervention were found. Participants in the longitudinal program commented positively on the value of meetings over time and their role in fostering involvement in the institution's teaching activities (Elliot *et al.*, 1999). Fellowship participants felt they had benefited through teacher training opportunities (Johnson & Zammit-Montebello, 1990). As well, practical learning sessions were rated more highly than the theoretically based ones (Sheets, 1985).

Level 2a—Learning: Participants in the longitudinal program reported positive changes in attitudes toward teaching (Elliot *et al.*, 1999). Two fellowships (Hitchcock *et al.*, 1986; Johnson & Zammit-Montebello, 1990) measured attitudinal change. In the first (Hitchcock *et al.*, 1986), participants rated themselves on their sensitivity, capability and authority as teachers, in comparison with their rating of the ideal faculty member. Significant pre–post-test differences were found on two levels: both perception of capability and authority moved closer to the ideal, while sensitivity remained the same. Participants in the second study (Johnson & Zammit-Montebello, 1990) reported positive attitudinal shifts in professional self-worth and beliefs about primary care and health promotion.

Level 2b—Learning: The longitudinal program participants reported change in use and understanding of educational terms, such as feedback (Elliot *et al.*, 1999). In two fellowships, participants demonstrated a gain in knowledge regarding problem-solving, teaching and communication skills, all of which improved post-course (Johnson & Zammit-Montebello, 1990; Sheets, 1985). In one study (Sheets, 1985), measures of knowledge included a 40-item short-answer test, and knowledge increases were sustained over six months. In another program, improved skills in collaborative teaching were noted (Pinheiro *et al.*, 1998).

Level 3—Behavior: The longitudinal program reported changes in educational roles as seven participants took on 23 new educational responsibilities following the intervention (Elliot *et al.*, 1999). In this study, participants also reported sustained comfort with the use of educational language and structure. The fellowship programs also demonstrated behavioral change. In one study, changes were noted in videotaped encounter performances where participants used many of the skills and techniques learned in the initial intervention (Johnson & Zammit-Montebello, 1990). They also increased their use of certain educational practices (e.g. teaching and organizing CME events). In another study, a positive relationship was noted between performance on knowledge tests and performance on ratings of videotaped teaching performance, and between self-ratings and trained rater scores (Sheets, 1985). In another fellowship, videotapes of participants (before and after the intervention) showed a change from a teacher-centered to a learner-centered approach to teaching (Pinheiro *et al.*, 1998).

Level 4a—Results: In the one study that reported outcomes at the level of the system, participants created an academic professional body following their fellowship experience (Johnson & Zammit-Montebello, 1990).

The focused picture

Eight articles scored 4 (or higher) for both study quality and strength of findings, and we chose to examine these separately in order to provide a more focused picture of faculty development. The following section summarizes these reports, which received an additional review by two TRG members (KM and YS) on the following dimensions: description of the intervention and expected outcomes, study design and outcomes assessed. A summary of these studies can be found in Table 3.

Four of the eight studies included in our focused review provided data that allowed for the calculation of effect size (Baroffio *et al.*, 1999; Skeff, 1983; Skeff *et al.*, 1986; Mahler & Benor, 1984). Mean scores and standard deviations were drawn from the data and were converted into effect sizes (d) using Cohen d's calculation (Cohen, 1988). These effects are shown in Table 3, where these studies are summarized. While effect sizes varied, moderate to high effect sizes were found in all four studies, highlighting the effects of the interventions, particular aspects of teaching that were affected, and groups of teachers who might benefit from the intervention.

(a) Description of the interventions and expected outcomes

The interventions described in these eight reports ranged from a 45-minute feedback session for clinical teachers (Marvel, 1991) to a month-long seminar series designed to facilitate dissemination of workshop concepts (Stratos *et al.*, 1997). One study described two workshops aimed at improving tutor behavior, each consisting of several phases (Baroffio *et al.*, 1999). Another study provided augmented feedback, consisting of norm-referenced graphic summaries of teachers' clinical teaching performance ratings, together with individually written clinical teaching effectiveness guidelines, to attending staff and residents (Litzelman *et al.*, 1998). Two studies assessed the benefits of a four-day workshop

Table 3. Summary of the eight most highly rated studies.

Baroffio *et al*. (1999) examined the effects of experience and faculty development workshops on tutorial skills. Students provided ratings of 88 tutors (all of whom had more than one year of tutor experience) using a 16-item questionnaire adapted from a previously validated instrument. Of the 88 tutors, all had attended a Level I workshop and 44 attended a more advanced Level II workshop. The Level I workshop was a three-phase preparation for tutoring that involved experiential and interactive learning; the Level II workshop was optional and addressed difficult tutorial experiences, which were analyzed jointly by the group. Student ratings of tutor performance after the Level I workshop generally increased with experience. The group ratings become more homogenous with experience, apparently due to greater improvement in those with lower scores. Despite the overall improvement, tutors did not improve either in provision of feedback or in assisting the tutorial group with problem synthesis. Tutors who attended the voluntary Level II workshop had higher baseline scores than the group attending Level I, suggesting that these higher baseline scores provided a motivation to improve. Among these higher scoring groups however, more improvement occurred in those tutors with lower skills. Comparing the post-test scores of those who attended the Level II workshop with those who did not, the authors concluded that the Level II workshop produced an effect greater than that of experience alone, especially for low-rated tutors, in terms of overall performance ($d = 0.94$) and achievement on problem synthesis ($d = 0.85$). The high magnitude of effect values calculated for low-rated tutors suggests that faculty development interventions may be particularly beneficial for this group of individuals.

Litzelman *et al*. (1998) evaluated the effect of augmented feedback on clinical teaching of attending staff and residents in an internal medicine teaching ward. Using an RCT design, the experimental group members received teaching evaluation summaries with individualized clinical teaching effectiveness guidelines to improve their teaching, both immediately prior to and midway through their four-week teaching assignment. Outcomes were measured at pre-and post-test using a clinical teaching performance instrument developed and validated in the Stanford Faculty Development Program (Skeff *et al*., 1992b). The control groups received the summaries only. Significant interactions were seen between the intervention and baseline teaching performance. Experimental group teachers with higher baseline teaching scores had teaching improvement scores that were significantly higher than the control group teachers with higher baseline scores. However, teachers in the experimental group with lower baseline scores had lower scores at the mid- and end-of-month scores than the control group teachers with the same baseline scores. The interaction of teacher experience and the intervention was also seen. Experienced teachers with higher baseline scores had higher scores than inexperienced teachers with similar baseline scores. However, experienced teachers with lower baseline scores had lower post-test scores than inexperienced teachers with similar baseline scores. This study highlighted the complex interactions that may occur between the intervention, teachers' experience and perceptions of their teaching. The study also suggests that baseline performance may be important in the planning of faculty development activities, and that these activities may need to be tailored to accommodate different needs.

Mahler & Benor (1984) studied the effect of teaching training workshops. A four-day multidisciplinary (basic and clinical science teachers), experiential and interactive teaching workshop aimed to improve teacher behavior in two dimensions: the activity dimension (increasing student verbalization vs. lecturing) and the cognitive dimension (increasing the cognitive level of verbal exchanges in the lesson). Baseline performance was measured. Following the workshop, 161 lessons of 60 teachers (approximately three per teacher) were observed and rated on: lesson time used by students vs. that used by teachers; who initiated the activity; and the kind of activity. Raters were trained and used validated methods and criteria. Post-workshop measures revealed a significant improvement in teacher performance on both the activity and the cognitive dimensions. The magnitude for effect of workshop was moderate to high ($d = 0.50$ to $d = 0.82$) for the activity-level domain and low to moderate ($d = 0.10$ to 0.54) for the cognitive-level domain. The observations occurred over 500 days, allowing an examination of whether the effect was sustained. No significant regression occurred in the activity dimension over time; moderate decreases occurred in the cognitive dimension, although not until after 270 days, probably beginning about 180–270 days post-intervention. This study is important in identifying when supplementary intervention might be needed.

Mahler & Neumann (1987) examined the effects of the above workshop (Mahler & Benor, 1984) on the cognitive dimension of instruction, noting increased cognitive versatility and activities at Bloom's higher taxonomy levels of comprehension, application and evaluation. There was a concomitant decrease in activities at the lower levels of Bloom's taxonomy. Sixty faculty members were observed. Trained, blinded sixth-year medical students assessed three videotaped lessons of each participant, taken before and after the intervention. Changes in teaching behavior and cognitive versatility were noted in all classroom settings.

Marvel (1991) conducted an evaluation of an intervention to improve teaching skills in a family practice residency program; 16 family physicians participated. The intervention consisted of individuals viewing videotapes of their teaching, using a checklist for self-assessment. An individualized feedback session was held, based on a 45-minute videotape. Videotapes of five consultations per faculty member and resident trainee ratings of faculty teaching skills were used as outcome measures. Patient ratings of residents formed a third data source, intended to examine whether improved teaching skills were seen in resident performance. Five of seven interview behaviors improved following the intervention. Individualized feedback was provided to each faculty member following baseline data collection. Patient ratings of residents increased, but not significantly.

(*continued*)

Table 3. Continued.

Skeff (1983) evaluated the effect of intensive feedback. A total of 64 ward attending physicians were randomly assigned to one of four groups: intensive feedback; videotape control; questionnaire feedback; and questionnaire control. The effects of two feedback methods to improve teaching experience were explored: intensive feedback (videotape review, trainee questionnaire feedback, and teacher self-assessment), and trainee questionnaire feedback alone. The experimental group received individualized feedback (Group 1) at mid-rotation accompanied by a one-hour discussion with an expert faculty developer. Group 2 had videotaped sessions and trainee ratings, but no self-assessment or conference. Group 3 received trainee evaluations at the middle and end of the rotation. Group 4 was rated by trainees at the middle and end of the rotation, but did not receive the feedback. Results showed that 75% of teachers in the intensive feedback group rated their experience as definitely beneficial (vs. 12%, 6%, 6% for other groups). The intensive feedback groups had higher post-treatment videotape ratings, and greater proportions of teachers improved. In fact, the magnitude of effect of post-treatment ratings for overall teaching performance for the intensive feedback group was larger ($d = 0.85$) than any other group. Unexpectedly, average videotape category ratings decreased post-treatment in the videotape only group, but remained stable in the intensive feedback group. Trainee ratings were not significantly different across all groups. This study showed that individual teachers can increase their performance, and that, without effective assistance, teaching problems are likely to persist even with feedback.

Skeff *et al.* (1986) examined the effect of a seminar method to improve ward teaching. Teachers were randomly allocated to experimental and control groups; the outcome measures were videotapes of ward rounds, teachers' subjective assessments of their experience, and trainee ratings. Experimental group performance significantly exceeded the control group on all ratings. Measures were taken early and late in the rotation with a six-month follow-up questionnaire. On videotape analysis, the experimental group performed significantly better in two categories compared with the control group (i.e. learning climate and control of session). Specifically, the magnitude of effect for experimental–control group differences on average videotape scaled scores (post-tests only) for learning climate, control of session and evaluation/feedback was $d = 0.60$, $d = 0.37$, and $d = 0.66$, respectively. This suggests that the seminar intervention had a moderate to high impact on aspects of faculty ward teaching. Further, student and house staff ratings were statistically significantly higher for the experimental group in control of the session and techniques to increase understanding. However, no overall difference in student ratings was seen between the two groups. Experimental group teachers (92%) reported changes in their teaching, compared with 24% of the control group. Six months later, 67% of respondents reported permanent changes in their teaching behavior. Changes in teacher attitudes and ratings of teacher impact significantly favored the experimental group; specifically, changes in the teachers' behavior had the most impact on residents' patient communication skills and collegial relationships.

Stratos *et al.* (1997) evaluated the effects of a disseminated faculty development program on 64 ambulatory care faculty members. Eight two-hour seminars were delivered at their home institution by 64 participants trained in the Stanford one-month faculty development program. There were three streams, of clinical teaching, medical decision-making and preventive medicine. Measures included self-reports of knowledge, skills and attitudes measured pre- and post-intervention, and teachers' evaluations of the seminars. In the clinical teaching stream, statistically significant pre-to-post improvements were found for several categories of teaching skills, using retrospective pre-test–post-test ratings. At the system level, 20 of 45 (44%) clinical teaching recommendations for improvement were judged by facilitators six months later as having made significant progress toward implementation.

designed to improve teachers' cognitive styles (Mahler & Benor, 1984; Mahler & Neumann, 1987), and two studies assessed the impact of an intensive feedback and seminar method on clinicians' teaching behaviors (Skeff, 1983; Skeff *et al.*, 1986).

All of the studies assessed behavioral change, targeting level 3 or 4 of Kirkpatrick's model. Four studies included participant satisfaction. Three studies examined changes in learning (i.e. knowledge, attitudes or skills); seven studies assessed change in teacher behavior and three assessed change at the level of the student or system. One study assessed outcome at all four levels (Skeff *et al.*, 1986).

(b) Methodological quality of the studies

Three of the eight studies (38%) were randomized controlled trials; the remaining five (62%) were single-group designs, with one study including a non-equivalent control group for one part of the intervention. All eight studies employed a pre-test–post-test design, with the addition of a delayed post-test in three.

Six of the eight studies (75%) used questionnaires (the majority of which were tested for reliability and based on a theoretical construct). Three of these same six studies also incorporated objective measures of performance. The two remaining studies used observed measures of performance only.

All of the eight studies used data sources other than participants' self-report. Five of the studies incorporated student and resident ratings of teacher behavior; five utilized trained observer ratings.

(c) Summary of findings

Level 1—Reaction: Four of the eight studies measured participant reaction to the interventions (Skeff, 1983; Skeff *et al.*, 1986; Marvel, 1991; Stratos *et al.*, 1997). In all cases, the intervention ratings were very positive. Facilitators' skills were also rated highly.

Level 2—Learning: Three studies evaluated learning, which included changes in attitudes, knowledge and skills. Skeff (1983) and Skeff *et al.* (1986) found significant positive

changes in awareness of teaching strengths and problems, in the desire to evaluate and improve teaching, and in satisfaction with and enthusiasm for teaching. Stratos *et al.* (1997) found self-reported changes in specific skills, as well as increased confidence and ability to teach medical decision-making. These authors also found that significant increases occurred in retrospective pre–post course ratings of several specific teaching behaviors, in pre–post tests of knowledge, and in participants' confidence in their ability to perform new skills.

Level 3—Behaviour: All eight studies evaluated the effects of their intervention on teaching performance, with most studies using more than one measure to assess behavioral change. Only one study (Skeff *et al.*, 1986) included self-reports of change. In five studies, behavior was measured using student and resident ratings (Skeff, 1983; Skeff *et al.*, 1986; Marvel, 1991; Litzelman *et al.*, 1998; Baroffio *et al.*, 1999). While these ratings revealed some positive changes in specific teaching behaviors (Mahler & Neumann, 1987; Litzelman *et al.*, 1998), in two studies the student ratings did not confirm differences observed on videotaped performance or on ratings of teacher impact on learning (Skeff, 1983; Skeff *et al.*, 1986). This result raised the question of whether the instruments used were sufficiently sensitive to change. There also appeared to be an interaction between some interventions (e.g. feedback) with baseline teaching ratings (Litzelman *et al.*, 1998) and with experience (Litzelman *et al.*, 1998; Baroffio *et al.*, 1999). These findings suggested that, in some circumstances, interventions can have negative as well as positive effects. Videotaped performance was also used to assess teaching performance in ward rounds and clinical settings (Skeff *et al.*, 1986; Marvel, 1991); and Mahler & Benor (1984) and Mahler & Neumann (1987) used trained raters to make classroom observations. Three studies employed delayed post-tests (Mahler & Benor, 1984; Skeff *et al.*, 1986; Stratos *et al.*, 1997). These results suggest that change may be sustained following some faculty development interventions, and that deterioration may not occur until at least six months post-intervention. One study utilized patient ratings of resident behavior to assess impact (Marvel, 1991). The majority of indicators increased (although not significantly), and one area of behavior decreased (significantly). In another study (Skeff, 1983), it was evident that teaching performance decreased in the absence of any intervention.

Level 4—Results: Three studies evaluated outcomes at the level of the system. Marvel (1991) found that patient ratings of residents generally improved, although there were no statistically significant improvements. Stratos *et al.* (1997) found that participants reported making changes in their own institution, six months following the faculty development intervention. In fact, 18 (of 25) participants were on their way to implementing changes in their own settings. Lastly, Skeff *et al.* (1986) found that students' and house staff ratings of teacher impact improved significantly in 'patient communication skills' and 'quality of inter-collegial relationships'.

Discussion

This review has focused on faculty development interventions designed to improve teaching effectiveness in medicine. Although many of the studies employed weak designs, making definitive statements about outcome difficult, the literature suggests positive changes in teachers' knowledge, attitudes and skills following participation in a faculty development activity. The impact on the organization (i.e. the learners and the systems in which our teachers work) is yet to be fully determined. Moreover, although many of the reported interventions were complex in nature (i.e. comprising different educational strategies and methods), few studies focused on teasing apart 'key features' of effective programs; however, some preliminary observations can be made. We can also make some general observations about the nature of faculty development programs reported to date and the 'state of the art' of research in this area.

Summary of outcomes

Despite the methodological limitations alluded to in earlier sections, the faculty development literature tends to support the following outcomes.

High satisfaction with faculty development programs: Overall satisfaction with faculty development programs was high. Notwithstanding the fact that the participants were volunteers, they consistently found the programs acceptable, useful and relevant to their personal objectives. The methods used, especially those with a practical and skills-based focus, were also valued by program participants.

Changes in attitudes towards teaching and faculty development: Participants reported a positive change in attitudes towards faculty development and towards teaching as a result of their involvement in a faculty development activity. They cited a greater awareness of personal strengths and limitations, increased motivation and enthusiasm for teaching, and a notable appreciation of the benefits of professional development. This impact was observed both in answers to open-ended questions and in pre–post measures of attitudinal change.

Gains in knowledge and skills: Participants often reported increased knowledge of educational concepts and principles as well as various aspects of teaching (e.g. specific teaching strategies; a more learner-centered approach). They also described gains in skills (e.g. assessing learners' needs, promoting reflection and providing feedback). Formal tests of knowledge, though infrequently used, also demonstrated positive changes.

Changes in teaching behavior: Self-perceived changes in teaching behavior were consistently reported. While student evaluations did not always reflect the changes that participants perceived, there was evidence that change in teaching performance was detectable. For example, changes in teaching behavior were reported for 15 (of 23) workshops and seven (of 10) seminar series. New educational initiatives, designed and implemented during the intervention, were also described.

Changes in organizational practice and student learning: Changes in student (or resident) behavior as well as organizational practice were not frequently investigated. However, in those few studies that examined organizational practice, participants reported a greater involvement in new educational activities and the establishment of new and improved networks of colleagues. The latter outcome was most frequently noted for the seminar series and longitudinal programs.

Summary of 'key features'

Although few reports (Skeff *et al.*, 1998; Pololi *et al.*, 2001) teased apart features of faculty development that make it effective, some preliminary conclusions can be made based on the literature reviewed. These features include the following.

The role of experiential learning: The importance of applying what has been learned (during the intervention and afterwards), practicing skills, and receiving feedback on skills learned was highlighted by several authors (Irby *et al.*, 1982; Coles & Tomlinson, 1994; Hewson, 2000), all of whom suggest that faculty members need to practice what they learn, and that immediate relevance and practicality is key (e.g. Sheets & Henry, 1984, 1988).

The value of feedback: The role of feedback in promoting change was evident in many of the reported interventions. In addition, several studies (Skeff, 1983; Litzelman *et al.*, 1998) specifically examined the use of feedback as an intervention strategy and found that systematic and constructive feedback can result in improved teaching performance. However, in one study (Litzelman *et al.*, 1998), augmented feedback was shown to have some negative effects; this potential effect should be considered and investigated further.

The importance of peers: A number of reports (DeWitt *et al.*, 1993; Elliot *et al.*, 1999) commented on the benefits of peer and collegial relationships. In particular, they highlighted the value of using peers as role models, the mutual exchange of information and ideas, and the importance of collegial support to promote and maintain change.

Adherence to principles of teaching and learning: Although many of the programs were not grounded in a theoretical or conceptual framework, many cited principles of adult learning (e.g. Knowles, 1988) and experiential learning (e.g. Kolb, 1984) as an organizing structure. In fact, there appears to be a developing consensus that adherence to these principles promotes more effective learning and teaching. Principles of instructional design were also frequently cited.

The use of multiple instructional methods to achieve objectives: As mentioned earlier, all of the interventions included a wide range of instructional methods (e.g. small-group discussions; interactive exercises; role plays and simulations) and none relied on lectures alone. Apparently, each program was aware of the need to accommodate different learning styles as well as the fact that different methods are required to meet diverse objectives.

At the same time, it is interesting to note that a number of important aspects of program development highlighted in the continuing medical education (CME) literature (e.g. Davis *et al.*, 1995; Oxman *et al.*, 1995) were not identified in this review. This included: the need for systematic needs assessments at the outset of any program; the value of reflection 'in action' and 'on action' (Schön, 1987); the value of application to practice; and the need for follow-up or 'booster' sessions. Although we believe that these features guided the design and delivery of many reported interventions (Bland & Froberg, 1982; Coles & Tomlinson, 1994; Bing-You *et al.*, 1999; Elliot *et al.*, 1999), they were not highlighted in the individual program descriptions.

Observations re faculty development interventions

In addition to the above 'key features', this review has also highlighted a number of issues that are worthy of further exploration—both for program design and for evaluation and research purposes.

The role of context: The majority of reports describe programs that were developed to meet the needs of a particular group of faculty members, in a particular context. To the extent that this development and 'match' were often successful, it is not surprising that there were many reports of changes in the desired direction. One lesson to be learned from this observation is that context is key, and that although the results of these studies may not be generalizable, the principles of faculty development might be.

Context is important in another way as well. According to Kirkpatrick (1994), four conditions are necessary for change to occur: the person must have the desire to change, knowledge of what to do and how to do it, a supportive work environment, and rewards for changing. Interestingly, the first two elements of change can potentially be achieved through faculty development activities; the last two cannot, and yet it is at this level that we expect change to occur. Consequently, the need to examine organizational characteristics, as well as the impact of faculty development on the organization, is critical. In looking ahead, it would be valuable to assess whether faculty development activities have an impact on the system at large and whether involvement in faculty development activities has an impact on career path. To date, we have only limited knowledge of this outcome level in two areas: in fellowship training, where we cannot draw any conclusions because of the lack of comparison groups, and in Skeff *et al.*'s work on the dissemination of faculty development activities (Skeff *et al.*, 1992b).

The nature of participation: Motivation to attend faculty development activities remains an unanswered question. What motivates participation? What determines whether someone will take advantage of specific offerings at a particular time? To date, the majority of participants are volunteers. Perhaps it is time for us to move beyond 'volunteerism' as we strive to enhance teaching and learning. It would also be worth exploring factors beyond the individual that encourage or impede attendance. As teaching is a 'social activity' (D'Eon *et al.*, 2000), the social determinants of participation merit further inquiry. It would

also be worthwhile to conduct further studies to determine what is learned through workshops vs. experience. For example, Litzelman *et al.* (1998) found an interaction between teaching experience and the faculty development intervention which suggested that teachers with more experience benefited more from the activity. On the other hand, Baroffio *et al.* (1999) discovered that the greatest improvement following their intervention occurred among inexperienced teachers. Further work in this area would shed light on this important, and complex, interaction.

The impact of participation on faculty development facilitators would also be worthy of investigation. It has been said that ''to teach is to learn twice''. Interestingly, no studies to date have examined the impact of participation on faculty development facilitators. It is our impression that facilitating a faculty development intervention requires a unique blend of skills and aptitudes that should be examined in greater depth.

The value of extended programs: Our review of findings by intervention type suggests that longer programs, extended over time (e.g. the seminar series), tend to produce outcomes not apparent in one-time interventions (e.g. short courses or workshops). For example, in several instances the seminar series resulted in the creation of networks and cooperative interactions among colleagues that are possible when a group meets over time (e.g. Rayner *et al.*, 1997). These interventions, as well as fellowships, also reported more involvement in educational activities following the faculty development activity, implying sustainability over time. A more rigorous comparison of 'short' and 'long' interventions would be beneficial to test out the hypothesis that extended programs yield more long-term changes.

The use of 'alternative' practices: The current literature demonstrates an over-reliance on traditional face-to-face methods such as workshops and seminars. Whereas these interventions seem to have the stated advantage of ease of scheduling, building a community of interested educators and increasing motivation, we should consider other methods that include online and self-directed learning, peer coaching (Flynn *et al.*, 1994) and mentorship (Morzinski *et al.*, 1996). It is interesting to note that some of the studies that scored highly on 'strength of findings' used alternative methods (e.g. individual feedback session).

Observations re methodological issues

The need for more rigorous designs: In 1992, Hitchcock *et al.* commented on the need to better evaluate faculty development programs and use sound qualitative and quantitative designs to document outcomes. The situation does not seem to have changed significantly since then. The results of this review suggest the need to conduct more rigorous research studies and overcome commonly encountered design problems. If possible, we should consider the use of randomized controlled trials, or at least comparison groups, so that we can make more generalizable statements about whether faculty development does, indeed, make a difference. We should also consider the systematic use of qualitative methods, or mixed designs, to capture the complexity of what occurs *during*, and *following*, faculty development interventions.

In reviewing the literature, we perceived an underutilization of rigorous qualitative methodologies. At the same time, many authors described an intuitive impression of enthusiasm, renewal and change following a particular faculty development activity. Current methods do not adequately capture these intuitions or anecdotal observations. Moreover, although there is general agreement that faculty development activities kindle interest in educational activities, how this is achieved, and what this inspires, needs to be examined more carefully. In many ways, a greater use of qualitative methods (e.g. Freeman *et al.*, 1992) would yield considerable benefits.

Faculty development activities represent complex interventions in complex settings (Drescher *et al.*, 2004). As noted in our conceptual framework, many intervening, mediating variables (e.g. personal attributes; teacher's status and responsibilities) interact with uncontrollable, extraneous factors. This is one of the many reasons that evaluation of effectiveness is difficult (for even if changes are noted, they may not definitively be attributed to the program) and that new research methodologies are required (e.g. Campbell *et al.*, 2000). Blumberg & Deveau (1995) have developed a model by which to evaluate an educational innovation/intervention that looks at academic dissemination, product development and implementation. This is something that we should consider in faculty development. We should also consider the value of examining anticipated and 'unanticipated' outcomes (e.g. Blumberg & Deveau, 1995), including impact on the organization.

Attention to participant satisfaction: It is time to re-affirm the value of participant satisfaction data. Although reaction to the program is an elementary level of evaluation, it is fundamental for change to occur. Participant satisfaction is important if faculty members are to be motivated to learn and to attend professional development activities. It also gives valuable feedback to program planners. As Belfield *et al.* (2001) have said, participant satisfaction is a crude proxy for the substantive effects of education. However, information on the reactions of participants to a specific program provides valuable information, as long as the purpose and use of such information is made explicit. In our opinion, we must build on the value of participant satisfaction rather than discredit it completely. Applying qualitative methodologies to participants' experiences and stories (e.g. analysis of narratives; critical incident technique) is another approach worth pursuing as we try to understand participants' reactions to faculty development offerings.

Outcome measures: The literature to date suggests an over-reliance on self-assessments and survey questionnaires to assess change. To move forward, we should consider the use of novel assessment methods. For example, Simpson *et al.* (1992) have developed standardized teaching situations to develop faculty teaching skills; Zabar *et al.* (2004) have utilized objective structured teaching examinations to evaluate impact. Given the increased fidelity of these teaching simulations, we should consider their potential use as an educational strategy and outcome measure, before and after a faculty development intervention.

Accurately measuring change requires reliable and valid measures. The majority of studies in this review used questionnaires for which psychometric properties were not reported. Faculty developers and researchers interested in assessing change should consider using questionnaires that have already been tested for validity and reliability, or work to establish these measures. For example, a number of scales and measures of teacher effectiveness have been developed in education (e.g. Gibbs & Coffey, 2004). Whenever possible, we should try to make use of these assessment tools and collaborate in order to share resources more consistently.

We should also try to correlate different measures of performance (e.g. self-assessment questionnaires and videotape recordings; student assessments and faculty self-ratings) so that we do not need to include all measures of change in every study. For example, several studies (e.g. (Mahler & Benor, 1984; Sheets & Henry, 1984) found a strong correlation between videotape ratings (albeit sometimes based on single observations) and knowledge tests. These findings, if corroborated, suggest the possibility of conducting reliable evaluations without always using direct observation (which can be costly and time-consuming). Based on similar results, we might be able to use student or resident evaluations of teachers' performance (together with knowledge tests) instead of videotaped observations. However, the value of triangulation to validate results cannot be understated. Some of the most highly rated studies (Skeff, 1983; Skeff *et al*., 1986) used multiple measures to assess outcome (e.g. self-ratings, videotaped observations and student ratings).

An important outcome of faculty development is improved student performance. We must therefore work to seek evidence of a relationship between changes in faculty members' teaching behaviors and learner outcomes. That is, we need to collect student and resident data (including indices of learner behaviour) more rigorously. Student evaluations of teaching competencies are invaluable; they need to be augmented, however, by a careful assessment of changes in students' and residents' own knowledge, attitudes and skills.

Attention to response shift bias: The notion of 'response shift bias' warrants more careful attention. As noted by Skeff *et al*. (1992a), post-course self-ratings are often lower than expected, and occasionally decrease, when increases are expected. This may occur because individuals overrate themselves at the beginning of a course, and then after the course (when they have a better idea of what is meant by different aspects of teaching and learning), they rate themselves more accurately (Nayer, 1995). As Skeff *et al*. have argued, we should more systematically consider the value of retrospective pre–post testing to overcome this possible response shift bias. In an interesting study (Skeff *et al*., 1992a), retrospective pre-tests correlated better with students' pre-workshop evaluations of their teachers' performance than did the regular pre-test. In addition, the retrospective pre- and post-tests showed significant differences in attitudes towards teaching that were not apparent in more traditional pre- and post-tests.

Assessment of change over time: A few studies assessed the maintenance of change over time. Most of them (Mahler & Benor, 1984; Skeff *et al*., 1986; Steinert *et al*., 2001) demonstrated that a number of changes were maintained, for as long as one year. Two studies (Mahler & Benor, 1984; Skeff *et al*., 1986) also indicated at what point reinforcing interventions might be helpful. It would be important to explore further the durability of change, those factors which help to sustain it, and the value of specific activities such as 'booster' sessions or other follow-up activities (Bland & Froberg, 1982).

Comparison of faculty development strategies: Although we have attempted to tease apart key 'features' of effective faculty development, there is little comparative research on which components of faculty development interventions are most useful (e.g. micro-teaching; role plays) and whether one method (e.g. seminar series) is more effective than another (e.g. short courses). For example, although workshops are one of the most common methods, many have suggested that they are too short to bring about lasting change. At the same time, they persist as a method of choice. Our findings suggest that longer interventions may have more durable outcomes. This, too, requires further investigation.

Grounding faculty development in theory and practice: Based on the findings of our review, we should caution ourselves against searching for the single 'perfect intervention'. In fact, an array of approaches exists and their appropriate use may differ from activity to activity and across settings. However, the work of faculty development should be grounded in both theory and empirical evidence. While educational theory has not yet provided us with a unified understanding of how learning occurs, there are well-supported models and principles of learning that can inform us in planning interventions, measuring outcomes and analysing effects (Mann, 2002). These include principles that draw on the science of cognition (e.g. how individuals make meaning of information and store it in memory) (Regehr & Norman, 1996); on understandings of social learning (e.g. how learning occurs from and with others; the influence of the learning environment) (Bandura, 1986); learning through experience (Kolb, 1984); and making meaning of learning and experience through reflection (Schön, 1987; Moon, 1999). More recently, the idea of learning through participation in communities of practice has also been explored (Lave & Wenger, 1991; Boud & Middleton, 2003), and this notion will have important implications for faculty development.

In a recent discussion of knowledge translation and continuing medical education, Davis *et al*. (2003) stated that ''a large gulf remains between what we know and what we practice''. The same may be said of some of the studies reviewed, as educational principles were not always applied in a systematic fashion. However, where positive and/or lasting effects on teacher performance were found, they were often associated with interventions that involved active and experiential learning over time. This could be explained by the fact that repeated interventions over time allow for cumulative learning and practice; they also enable the development of a trusted network of colleagues and a community of teachers. These considerations are critical to faculty development program design.

Collaborating across programs and disciplines: The value of sharing resources and collaborating across programs has been

highlighted earlier in this review. There is also much for us to learn from colleagues in the field of education. For example, many of our findings resemble what has been found in reviews of research on training of university teachers (Gibbs & Coffey, 2004); in many ways, it would be wise to learn from these studies and incorporate their methodologies (and findings) into our work. We should also build on lessons learned in the CME literature (e.g. Davis *et al.*, 1995). To accomplish our objectives for scholarly work in faculty development, we should develop local research teams and regional networks, to develop—and implement—a collaborative research agenda that is grounded in practice.

Conclusion

Using the review findings

Based on the review findings, we offer the following suggestions for practice and research.

Implications for practice:

We need to:

- Build on our successes. The literature describes successful programs, with recognizable, replicable elements. It is now important to tease apart the elements that work.
- Make more deliberate use of theory (particularly theories of learning) and educational principles in the design and development of our faculty development programs. Further, we need to link theory with practice, in an iterative cycle of asking questions in practice, studying these questions and testing our answers. We also need to better understand teachers' educational practices and the real problems that teachers encounter so that we can use this knowledge to inform theory, which can help us in developing improved interventions and evaluating effectiveness.
- Acknowledge the importance of context. The organizational culture, the curriculum, teachers and students all contribute to a context that is critical to the effectiveness of educational change.
- Develop more programs that extend over time, to allow for cumulative learning, practice and growth.
- Develop programs that stimulate reflection and learning among participants, raising their awareness of themselves as teachers. This would form the basis for ongoing self-directed development rather than the need to primarily have 'teacher-directed' interventions.
- Re-examine the question of voluntary participation. In many contexts, the requirement to prepare for teaching effectiveness may not be met unless participation is expected and required. Moreover, the voluntary nature of faculty development raises questions about the institutional culture and the values (both explicit and implicit) that it places on teaching and learning.

Implications for future research:

We need to:

- Conduct more rigorous research studies, using control or comparison groups and qualitative methodologies. This requires careful definitions of outcomes, planning for evaluation at the inception of any program, and closer collaboration with research colleagues. We must also find a way to corroborate anecdotal observations and capture faculty members' stories.
- Carry out process-oriented studies in addition to outcome-oriented ones. That is, we need to better understand how change occurs, both as a result of the intervention and within the individual (e.g. how did teachers' beliefs change; did the intervention result in improving teachers' reflective skills). In fact, qualitative methods may be more appropriate here.
- Continue to develop and utilize performance-based measures of change. The use of these methods, which do exist, is an essential and natural next step.
- Use multiple methods and data sources to allow for triangulation of data.
- Assess and report the validity and reliability of instruments used. Further, where appropriate instruments exist, these should be considered in preference to developing new instruments. Using standardized or comparable measures across studies will help to understand the field and improve the quality of research in this area.
- Promote studies in which an intervention is recognized as occurring in a complex environment in which many unforeseen and unpredictable variables play a role. We need to conduct more studies in which the interaction between different factors is investigated, highlighting under what conditions and why an intervention might be successful or not.
- Compare different faculty development methods to enable an analysis of which features of faculty development contribute to changes in teacher performance.
- Assess change over time. This is important both in determining any enduring effects, and in understanding which interventions or factors may be associated with more sustained change. Longitudinal follow-ups may also help us to understand the development of faculty members throughout their careers.
- Develop means of assessing the impact of faculty development on the institution/organization in a more rigorous and systematic fashion.
- Embed our research studies in a theoretical or conceptual framework, and utilize theory in the interpretation of our results.
- Collaborate with colleagues within and outside medicine.

Strengths and limitations of the review

The following strengths and limitations are inherent in this review.

The review process: The review process was 'time-limited' and reflects the literature from 1980 until 2002. It is now time to update this review, based on the methodology developed for this systematic review. Not surprisingly, we would predict an increase in well-designed studies in the first five years of the twenty-first century as well as an increase in behavioral and systems outcomes.

Moreover, while the search process was extensive, it was hampered by the fact that many medical education articles were not indexed in either MEDLINE or ERIC before

2002, and many of the articles had to be found in a hand search. This challenge will probably not be encountered in future searches. It should also be noted that a complex search strategy in a field such as this one, where the terminology is still inconsistent across international and professional boundaries (Freeth *et al.*, 2003), created numerous challenges during the search process. In addition, all of the reviewed studies were found in the English language, with a greater number in the North American literature. As noted in other reviews (Koppel *et al.*, 2001; Freeth *et al.*, 2003), this may reflect a publication bias that prevents a fuller picture of faculty development from an international perspective.

The pilot phase of this review was extensive. Initially, each member of the review team used the same small set of studies to test a prototypical coding sheet. As in other reviews (Freeth *et al.*, 2003), difficulties and differences were discussed and resolved, and led to significant and important changes on the data abstraction sheet. Although lengthy, this iterative process helped to contribute to the rigor of the review. However, inter-rater reliability was a challenge throughout the review process. While the international representation of the TRG was a real strength, and provided an invaluable perspective on faculty development and outcomes research, our ability to meet face-to-face was limited. Such opportunities would have enabled increased reviewer training and more frequent discussions of coding challenges (e.g. level of outcomes; research methods). We should also acknowledge that while we sought to maintain critical reflexivity as individuals and as a research team (Freeth *et al.*, 2003), and we were as vigilant as possible about data coding and quality control, personal biases and misinterpretations of reported data may have led to some errors in the final summary of the studies that we reviewed. We apologize in advance for such errors or inconsistencies and we hope that they will be brought to our attention, to be corrected in the web edition of this review.

The BEME Coding Sheet was both a strength and a limitation. While it provided a coherent structure to the review, considerable time was spent in adapting the form to our review and piloting it to ensure that it would work, as definitions of concepts were needed to ensure inter-rater agreement. Some reviewers have argued that the BEME Coding Sheet puts too much emphasis on methodological issues and too little emphasis on theoretical issues (Dolmans, 2003). However, this observation may be more reflective of the state of the literature than the nature of the form. Study quality and strength of findings should also be elaborated in a future iteration.

The nature of the articles reviewed: The nature of the articles reviewed presented a number of challenges. As stated earlier, the study designs were often limited. As well, authors frequently did not report on response rates or statistical methods used and, as a result, it was difficult to perform certain analyses of the available data. Basic background information (e.g. discipline; duration), critical to understanding the context of the intervention, was also lacking in many reports and the reviewers often had difficulty ascertaining key aspects of the study (e.g. methods; results). In addition, an inconsistent use of terminology (e.g. to describe program types) often led to different interpretations of the same information. Finally, it is worth noting that negative results were rarely reported. This may be due to a possible publication bias towards positive results, which is often a challenge for those engaged in a systematic review.

Next steps

As mentioned at the outset, this review was limited to faculty development designed to improve teaching effectiveness in medicine. It would now be worthwhile to update this review and to conduct a similar review of faculty development targeting other faculty roles (e.g. research; administration). It would also be worth examining the literature on faculty development for other health professionals and for residents-in-training. Interestingly, the majority of randomized controlled trials in this area can be found in studies addressing faculty development for residents.

The aim of Best Evidence Medical Education is to encourage teachers to think more clearly about the actions they are taking as teachers and to utilize evidence where it is relevant and available to inform their decisions (Harden *et al.*, 1999). The goal of this review has been to assess the evidence on the impact of faculty development activities on teachers' knowledge, attitudes and skills and on the organizations in which they work. The breadth and depth of faculty development programs offered (and described in the literature) is impressive. We must now develop new methodologies to assess impact over time and collaborate more systematically across programs and institutions to achieve our goals.

Acknowledgements

The authors would like to extend their appreciation and gratitude to the following:

Aventis Pharma, for educational grants to support this initiative; Dr Carole Bland and Dr Dave Davis, for initial contributions to the review process; Dr Terri Collin, for statistical consultation; Ms Jocelyn Godolphin, for systematically reviewing and coding all of the articles and for providing insightful comments and suggestions; Dr Marilyn Hammick, for invaluable feedback on earlier versions of this manuscript; Dr Ron Harden and Dr Ian Hart, for initiating and supporting this BEME review; Ms Susan Huculak, for retrieving all of the articles and entering the references into *Reference Manager*; Ms Angella Lambrou, for conducting the comprehensive literature search; Ms Maureen Leaman, for invaluable help in coordinating and administering this review; Mrs Pat Lilley, for administrative support throughout this BEME process; and Ms Susan Lu, for assistance with the pilot review.

Notes on contributors

YVONNE STEINERT, PhD, is a Professor of Family Medicine, Associate Dean for Faculty Development and Director of the Center for Medical Education in the Faculty of Medicine at McGill University, Montreal, Canada. Her special interests relate to teaching and learning in medicine, the impact of faculty development on the individual and the organization, and the continuing professional development of faculty members.

KAREN MANN, PhD, is Professor and Director of the Division of Medical Education in the Faculty of Medicine of Dalhousie University, Halifax,

Canada. Her special research and practice interests lie in teaching and learning in medicine, faculty/staff development, reflective practice and the preparation of residents as teachers.

Angel Centeno, MD PhD, is Professor of Internal Medicine and Director of the Department of Biomedical Education at the Faculty of Biomedical Sciences, Austral University in Buenos Aires, Argentina. His special interests are faculty development and the academic development of medical education as a science.

Diana Dolmans, PhD, is Associate Professor at the Department of Educational Development and Research of Maastricht University in The Netherlands. Her special interests relate to research on teaching and learning in innovative learning environments and especially teaching and learning in work-based learning environments.

John Spencer, FRCGP, is Professor of Medical Education in the School of Medical Education Development, University of Newcastle upon Tyne, UK, and Sub-Dean for Primary and Community Care. His special interests include teaching and learning in clinical settings, community-based education, faculty development, and the role of the patient in healthcare education.

Mark Gelula, PhD, is Assistant Dean for Faculty Development at the University of Illinois at Chicago College of Medicine and Research Assistant Professor of Medical Education in the Department of Medical Education, USA. His research interests focus on the use of standardized students and on alternative approaches to establish lasting transfer of learning among faculty involved in faculty development workshops, courses, and programs.

David Prideaux, PhD, is Professor and Head, Department of Medical Education in the School of Medicine at Flinders University in South Australia. His interests focus on educational change and innovation and the role of faculty development in promoting and facilitating change.

References

Andriole, D.A., Evans, S.R., Foy, H.M., Atnip, R.G. & Mancino, A.T. (1998) Can a one-day workshop alter surgeons' teaching practices? *American Journal of Surgery*, 175(6), pp. 518–520.

Bandura, A. (1986) *Social Foundations of Thought and Action: A Social Cognitive Theory* (Englewood Cliffs, NJ, Prentice-Hall).

Baroffio, A., Kayser, B., Vermeulen, B., Jacquet, J. & Vu, N.V. (1999) Improvement of tutorial skills: an effect of workshops or experience? *Academic Medicine*, 74(10 Suppl.), pp. S75–S77.

Baxley, E.G., Probst, J.C., Schell, B.J., Bogdewic, S.P. & Cleghorn, G.D. (1999) Program-centered education: a new model for faculty development, *Teaching and Learning in Medicine*, 11(2), pp. 94–99.

Belfield, C., Thomas, H., Bullock, A., Eynon, R. & Wall, D. (2001) Measuring effectiveness for best evidence medical education: a discussion, *Medical Teacher*, 23(2), pp. 164–170.

Bing-You, R.G., Renfrew, R.A. & Hampton, S.H. (1999) Faculty development of community-based preceptors through a collegial site-visit program, *Teaching and Learning in Medicine*, 11(2), pp. 100–104.

Bird, J., Hall, A., Maguire, P. & Heavy, A. (1993) Workshops for consultants on the teaching of clinical communication skills, *Medical Education*, 27(2), pp. 181–185.

Bland, C.J. & Froberg, D.G. (1982) A systematic approach to faculty development for family practice faculty, *Journal of Family Practice*, 14(3), pp. 537–543.

Bland, C.J. & Schmitz, C.C. (1986) Characteristics of the successful researcher and implications for faculty development, *Journal of Medical Education*, 61, pp. 22–31.

Bland, C.J., Schmitz, C.C., Stritter, F.T., Henry, R.C. & Aluise, J.J. (1990) *Successful Faculty in Academic Medicine* (New York, Springer Publishing).

Bligh, J. (2005) Faculty development, *Medical Education*, 39(2), pp. 120–121.

Blumberg, P. & Deveau, E.J. (1995) Using a practical program evaluation model to chart the outcomes of an educational initiative: problem-based learning, *Medical Teacher*, 17(2), pp. 205–213.

Boud, D. & Middleton, H. (2003) Learning from others at work: communities of practice and informal learning, *Journal of Workplace Learning*, 15(5), pp. 194–202.

Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A.L., Sandercock, P., Speigelhalter, D. & Tyrer, P. (2000) Framework for design and evaluation of complex interventions to improve health, *British Medical Journal*, 321(7262), pp. 694–696.

Centra, J.A. (1978) Types of faculty development programs, *Journal of Higher Education*, 49, pp. 151–162.

Clark, J.M., Houston, T.K., Kolodner, K., Branch Jr, W.T., Levine, R.B. & Kern, D.E. (2004) Teaching the teachers: national survey of faculty development in departments of medicine of US teaching hospitals, *Journal of General Internal Medicine*, 19(3), pp. 205–214.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (New York, Academic Press).

Coles, C.R. & Tomlinson, J.M. (1994) Teaching student-centred educational approaches to general practice teachers, *Medical Education*, 28(3), pp. 234–238.

Creswell, J. (2002) *Educational Research: Planning, Conducting and Evaluating Quantitative and Qualitative Research* (Upper Saddle River, NJ, Merrill Prentice Hall).

Creswell, J. (2003) *Research Design: Qualitative, Quantitative and Mixed Methods Approaches* (Thousand Oaks, CA, Sage Publications).

DaRosa, D.A., Folse, J.R., Reznick, R.K., Dunnington, G.L. & Sachdeva, A.K. (1996) Description and evaluation of the Surgeons as Educators course, *Journal of the American College of Surgeons*, 183(5), pp. 499–505.

Davis, D., Evans, M., Jadad, A., Perrier, L., Rath, D., Ryan, D., Sibbald, G., Straus, S., Rappolt, S., Wowk, M. & Zwarenstein, M. (2003) The case for knowledge translation: shortening the journey from evidence to effect, *British Medical Journal*, 327(7405), pp. 33–35.

Davis, D.A., Thomson, M.A., Oxman, A.D. & Haynes, R.B. (1995) Changing physician performance: a systematic review of the effect of continuing medical education strategies, *Journal of the American Medical Association*, 274(9), pp. 700–705.

Dawson, B. & Trapp, R.G. (2001) *Basic and Clinical Biostatistics* (New York, Lange Medical Books).

Dennick, R. (1998) Teaching medical educators to teach: the structure and participant evaluation of the Teaching Improvement Project, *Medical Teacher*, 20(6), pp. 598–601.

D'eon, M., Overgaard, V. & Harding, S.R. (2000) Teaching as a social practice: implications for faculty development, *Advances in Health Sciences Education*, 5(2), pp. 151–162.

DeWitt, T.G., Goldberg, R.L. & Roberts, K.B. (1993) Developing community faculty: principles, practice, and evaluation, *American Journal of Diseases of Children*, 147(1), pp. 49–53.

Dolmans, D. (2003) The effectiveness of PBL: the debate continues. Some concerns about the BEME movement, *Medical Education*, 37(12), pp. 1129–1130.

Drescher, U., Warren, F. & Norton, K. (2004) Towards evidence-based practice in medical training: making evaluations more meaningful, *Medical Education*, 38(12), pp. 1288–1294.

Elliot, D.L., Skeff, K.M. & Stratos, G.A. (1999) How do you get to the improvement of teaching? A longitudinal faculty development program for medical educators, *Teaching & Learning in Medicine*, 11(1), pp. 52–57.

Flynn, S.P., Bedinghaus, J., Snyder, C. & Hekelman, F. (1994) Peer coaching in clinical teaching: a case report, *Educational Research and Methods*, 26(9), pp. 569–570.

Freeman, J., Sharf, B.F. & Benson, J.L. (1992) Observational data in faculty development: an underused method, *Teaching and Learning in Medicine*, 4(2), pp. 103–109.

Freeth, D., Hammick, M., Koppel, I., Reeves, S. & Barr, H. (2003) *A Critical Review of Evaluations of Interprofessional Education* (London, Higher Education Academy Learning and Teaching Support Network for Health Sciences and Practice) [Available at: http://www.health.ltsn.ac.uk/publications/occasionalpaper/occasionalpaper02.pdf].

Freeth, D., Hammick, M., Koppel, I., Reeves, S. & Barr, H. (2005) *Evaluating Interprofessional Education: A Self-Help Guide* (London: Higher Education Academy Learning and Teaching Support Network for Health Sciences and Practice) [Available at: http://www.health.ltsn.ac.uk/publications/occasionalpaper/occp5.pdf].

Gibbs, G. & Coffey, M. (2004) The impact of training of university teachers on their teaching skills, their approach to teaching and the approach to learning of their students, *Active Learning in Higher Education*, 5(1), pp. 87–100.

Gordon, G.H. & Levinson, W. (1990) Attitudes toward learner-centered learning at a faculty development course, *Teaching and Learning in Medicine*, 2(2), pp. 106–109.

Harden, R.M., Grant, J., Buckley, G. & Hart, I.R. (1999) BEME Guide No. 1: Best Evidence Medical Education, *Medical Teacher*, 21(6), pp. 553–562.

Hewson, M.G. (2000) A theory-based faculty development program for clinician-educators, *Academic Medicine*, 75(5), pp. 498–501.

Hewson, M.G. & Copeland, H.L. (1999) Outcomes assessment of a faculty development program in medicine and pediatrics, *Academic Medicine*, 74(10 Suppl.), pp. S68–S71.

Hewson, M.G., Copeland, H.L. & Fishleder, A.J. (2001) What's the use of faculty development? Program evaluation using retrospective self-assessments and independent performance ratings, *Teaching and Learning in Medicine*, 13(3), pp. 153–160.

Hitchcock, M.A., Lamkin, B.D., Mygdal, W.K., Clarke, C.M. & Clarke, S.O. (1986) Affective changes in faculty development fellows in family medicine, *Journal of Medical Education*, 61(5), pp. 394–403.

Hitchcock, M.A., Stritter, F.T. & Bland, C.J. (1992) Faculty development in the health professions: conclusions and recommendations, *Medical Teacher*, 14(4), pp. 295–309.

Irby, D.M. (1996) Models of faculty development for problem-based learning, *Advances in Health Sciences Education*, 1, pp. 69–81.

Irby, D.M., Vontver, L.A. & Stenchever, M.A. (1982) Improving teaching in a multisite clerkship: faculty-development workshops, *Journal of Reproductive Medicine*, 27(6), pp. 307–310.

Issenberg, S.B., McGaghie, W.C., Petrusa, E., Gordon, D.L. & Scalese, R.J. (2005) Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. BEME Guide No. 4 (Dundee, Scotland, Association for Medical Education in Europe).

Johnson, D.H. & Zammit–Montebello, A. (1990) A new faculty training model for countries developing academic family medicine programs, *Academic Medicine*, 65(5), pp. 346–348.

Kirkpatrick, D.L. (1967) Evaluation of training, in: R. Craig & I. Mittel (Eds) *Training and Development Handbook*, pp. 87–112 (New York, McGraw-Hill).

Kirkpatrick, D.L. (1994) *Evaluating Training Programs: The Four Levels* (San Francisco, CA, Berrett-Koehler Publishers).

Knowles, M.S. (1988) *The Modern Practice of Adult Education: From Pedagogy to Androgogy* (New York, Cambridge Books).

Kolb, D.A. (1984) *Experiential Learning: Experience as the Source of Learning and Development* (Englewood Cliffs, NJ, Prentice-Hall).

Koppel, I., Barr, H., Reeves, S., Freeth, D. & Hammick, M. (2001) Establishing a systematic approach to evaluating the effectiveness of interprofessional education, *Issues in Interdisciplinary Care*, 3(1), pp. 41–49.

Lave, J. & Wenger, E. (1991) *Situated Learning: Legitimate Peripheral Participation* (Cambridge, Cambridge University Press).

Litzelman, D.K., Stratos, G.A., Marriott, D.J., Lazaridis, E.N. & Skeff, K.M. (1998) Beneficial and harmful effects of augmented feedback on physicians' clinical-teaching performances, *Academic Medicine*, 73(3), pp. 324–332.

Lye, P.S., Simpson, D.E., Wendelberger, K.J. & Bragg, D.S. (1998) Clinical teaching rounds: a case-oriented faculty development program, *Archives of Pediatrics and Adolescent Medicine*, 152(3), pp. 293–295.

Mahler, S. & Benor, D.E. (1984) Short and long term effects of a teacher-training workshop in medical school, *Higher Education*, 13(3), pp. 265–273.

Mahler, S. & Neumann, L. (1987) Faculty development in medical instruction: the cognitive dimension, *Israel Journal of Medical Sciences*, 23(12), pp. 1247–1251.

Mann, K.V. (2002) Thinking about learning: implications for principle-based professional education, *Journal of Continuing Education in the Health Professions*, 22(2), pp. 69–76.

Marvel, M.K. (1991) Improving clinical teaching skills using the parallel process model, *Family Medicine*, 23(4), pp. 279–284.

McDermott Jr, J.F. & Anderson, A.S. (1991) Retraining faculty for the problem-based curriculum at the University of Hawaii 1989–1991, *Academic Medicine*, 66(12), pp. 778–779.

Moon, J. (1999) *Reflection in Learning and Professional Development: Theory and Practice* (London, Kogan Page).

Morzinski, J.A., Diehr, S., Bower, D.J. & Simpson, D.E. (1996) A descriptive, cross-sectional study of formal mentoring for faculty, *Family Medicine*, 28(6), pp. 434–438.

Nasmith, L. & Steinert, Y. (2001) The evaluation of a workshop to promote interactive lecturing, *Teaching and Learning in Medicine*, 13(1), pp. 43–48.

Nasmith, L., Steinert, Y., Saroyan, A., Daigle, N. & Franco, E. (1997) Assessing the impact of a faculty development workshop: a methodological study, *Teaching & Learning in Medicine*, 9(3), pp. 209–214.

Nathan, R.G. & Smith, M.F. (1992) Students' evaluations of faculty members' teaching before and after a teacher-training workshop, *Academic Medicine*, 67(2), pp. 134–135.

Nayer, M. (1995) Faculty development for problem-based learning programs, *Teaching and Learning in Medicine*, 7(3), pp. 138–148.

Olmesdahl, P.J. & Manning, D.M. (1999) Impact of training on PBL facilitators, *Medical Education*, 33(10), pp. 753–755.

Olukoya, A. (1986) Teaching medical teachers how to teach in Lagos, Nigeria, *Medical Teacher*, 8(2), pp. 145–148.

Oxman, A.D., Thomson, M.A., Davis, D.A. & Haynes, R.B. (1995) No magic bullets: a systematic review of 102 trials of interventions to improve professional practice, *Canadian Medical Association Journal*, 153(10), pp. 1423–1431.

Pinheiro, S.O., Rohrer, J.D. & Heimann, C.F.L. (1998) *Assessing Change in the Teaching Practice of Faculty in a Faculty Development Fellowship Program for Primary Care Physicians: Toward a Mixed Method Evaluation Approach*. Report No. TM028458 (East Lansing, MI, College of Osteopathic Medicine). [ERIC Document Reproduction Service No. ED421499.]

Pololi, L., Clay, M.C., Lipkin M., Jr, Hewson, M., Kaplan, C. & Frankel, R.M. (2001) Reflections on integrating theories of adult education into a medical school faculty development course, *Medical Teacher*, 23(3), pp. 276–283.

Quirk, M.E., Dewitt, T., Lasser, D., Huppert, M. & Hunniwell, E. (1998) Evaluation of primary care futures: a faculty development program for community health center preceptors, *Academic Medicine*, 73(6), pp. 705–707.

Rayner, H., Morton, A., McCulloch, R., Heyes, L. & Ryan, J. (1997) Delivering training in teaching skills to hospital doctors, *Medical Teacher*, 19(3), pp. 209–211.

Reeves, S., Koppel, I., Barr, H., Freeth, D. & Hammick, M. (2002) Twelve tips for undertaking a systematic review, *Medical Teacher*, 24(4), pp. 358–363.

Regehr, G. & Norman, G.R. (1996) Issues in cognitive psychology: implications for professional education, *Academic Medicine*, 71(9), pp. 988–1001.

Reid, A., Stritter, F.T. & Arndt, J.E. (1997) Assessment of faculty development program outcomes, *Family Medicine*, 29(4), pp. 242–247.

Sachdeva, A.K. & Kelliher, G.J. (1994) Impact of a faculty development workshop on three institutions, in: A.I. Rothman & R. Cohen (Eds) *Proceedings of the Sixth Ottawa Conference on Medical Education*, pp. 75–79 (Ontario, Canada, University of Toronto Bookstore Custom Publishing).

Schmidt, H.G., Bouhuijs, P.A.J., Khattab, T. & Makladi, F. (1989) Attitude change among medical teachers: effects of a workshop on tutorials, in: H.G. Schmidt, M. Lipkin, M. de Vries & J. Greep (Eds) *New Directions for Medical Education*, pp. 243–247 (New York, Springer Verlag).

Schön, D.A. (1987) *Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions* (San Francisco, CA, Jossey-Bass).

Sheets, K.J. (1985) Evaluating short-term training programs: a practical approach, *Journal of Instructional Development*, 8(1), pp. 26–32.

Sheets, K.J. & Henry, R.C. (1984) Assessing the impact of faculty development programs in medical education, *Journal of Medical Education*, 59(9), pp. 746–748.

Sheets, K.J. & Henry, R.C. (1988) Evaluation of a faculty development program for family physicians, *Medical Teacher*, 10(1), pp. 75–83.

Sheets, K.J. & Schwenk, T.L. (1990) Faculty development for family medicine educators: an agenda for future activities, *Teaching and Learning in Medicine*, 2, pp. 141–148.

Simpson, D., Lawrence, S.L. & Krogull, S.R. (1992) Using standardized ambulatory teaching situations for faculty development, *Teaching and Learning in Medicine*, 4(1), pp. 58–61.

Skeff, K.M. (1983) Evaluation of a method for improving the teaching performance of attending physicians, *American Journal of Medicine*, 75(3), pp. 465–470.

Skeff, K.M., Campbell, M., Stratos, G.A., Jones III, H.W. & Cooke, M. (1984) Assessment by attending physicians of a seminar method to improve clinical teaching, *Journal of Medical Education*, 59(12), pp. 944–950.

Skeff, K.M. & Stratos, G.A. (1985) *Issues in the Improvement of Clinical Instruction*. Report No. HE018321 (Chicago, IL, American Education Research Association). [ERIC Document Reproduction Service No. ED256257.]

Skeff, K.M., Stratos, G.A. & Bergen, M.R. (1992a) Evaluation of a medical faculty development program: a comparison of traditional pre/post and retrospective pre/post self-assessment ratings, *Evaluation and the Health Professions*, 15(3), pp. 350–366.

Skeff, K.M., Stratos, G.A., Bergen, M.R. & Regula, D.P., Jr (1998) A pilot study of faculty development for basic science teachers, *Academic Medicine*, 73(6), pp. 701–704.

Skeff, K.M., Stratos, G.A., Bergen, M.R., Albright, C.L., Berman, J., Farquhar, J.W. & Sox Jr, H.C. (1992b) The Stanford faculty development program: a dissemination approach to faculty development for medical teachers, *Teaching and Learning in Medicine*, 4(3), pp. 180–187.

Skeff, K.M., Stratos, G.A., Bergen, M.R., Sampson, K. & Deutsch, S.L. (1999) Regional teaching improvement programs for community-based teachers, *American Journal of Medicine*, 106(1), pp. 76–80.

Skeff, K.M., Stratos, G.A., Berman, J. & Bergen, M.R. (1992c) Improving clinical teaching: evaluation of a national dissemination program, *Archives of Internal Medicine*, 152(6), pp. 1156–1161.

Skeff, K.M., Stratos, G., Campbell, M., Cooke, M. & Jones, H.W., III (1986) Evaluation of the seminar method to improve clinical teaching, *Journal of General Internal Medicine*, 1(5), pp. 315–322.

Skeff, K.M., Stratos, W., Mygdal, T.A., Dewitt, T.A., Manfred, L., Quirk, M., Roberts, K., Greenberg, L. & Bland, C.J. (1997) Faculty development: a resource for clinical teachers, *Journal of General Internal Medicine*, 12(Suppl. 2), pp. 56–63.

Snyder, S. (2001) A program to teach curriculum development to junior faculty, *Family Medicine*, 33(5), pp. 382–387.

Steinert, Y. (2000) Faculty development in the new millennium: key challenges and future directions, *Medical Teacher*, 22, pp. 44–50.

Steinert, Y., Nasmith, L., Daigle, N. & Franco, E.D. (2001) Improving teachers' skills in working with 'problem' residents: a workshop description and evaluation, *Medical Teacher*, 23(3), pp. 284–288.

Stratos, G.A., Bergen, M.R., Albright, C.L., Skeff, K.M. & Owens, D.K. (1997) Use of faculty development to improve ambulatory-care education, *Medical Teacher*, 19(4), pp. 285–292.

Stritter, F.T. (1983) Faculty evaluation and development, in: C.H. McGuire, R.P. Foley, A. Gorr & R.W. Richards (Eds) *Handbook of Health Professions Education*, pp. 294–318 (San Francisco, CA, Jossey-Bass).

Stuart, M.R., Orzano, A.J. & Eidus, R. (1980) Preceptor development in residency training through a faculty facilitator, *Journal of Family Practice*, 11(4), pp. 591–595.

Ullian, J.A. & Stritter, F.J. (1997) Types of faculty development programs, *Family Medicine*, 29, pp. 237–241.

Valdiserri, R.O., Koziol, S.M., Korth, W.W. & Haynes, L.T. (1986) A workshop to improve the teaching skills of physician faculty members, *Journal of Medical Education*, 61(6), pp. 469–471.

Wilkerson, L. & Irby, D.M. (1998) Strategies for improving teaching practices: a comprehensive approach to faculty development, *Academic Medicine*, 73(4), pp. 387–396.

Wilkerson, L. & Sarkin, R.T. (1998) Arrows in the quiver: evaluation of a workshop on ambulatory teaching, *Academic Medicine*, 73(10 Suppl. 1), pp. 67–69.

Zabar, S., Hanley, K., Stevens, D.L., Kalet, A., Schwartz, M.D., Pearlman, E., Brenner, J., Kachur, E.K. & Lipkin, M. (2004) Measuring the competence of residents as teachers, *Journal of General Internal Medicine*, 19(5 Pt 2), pp. 530–533.