Figure 1

Conceptual Framework

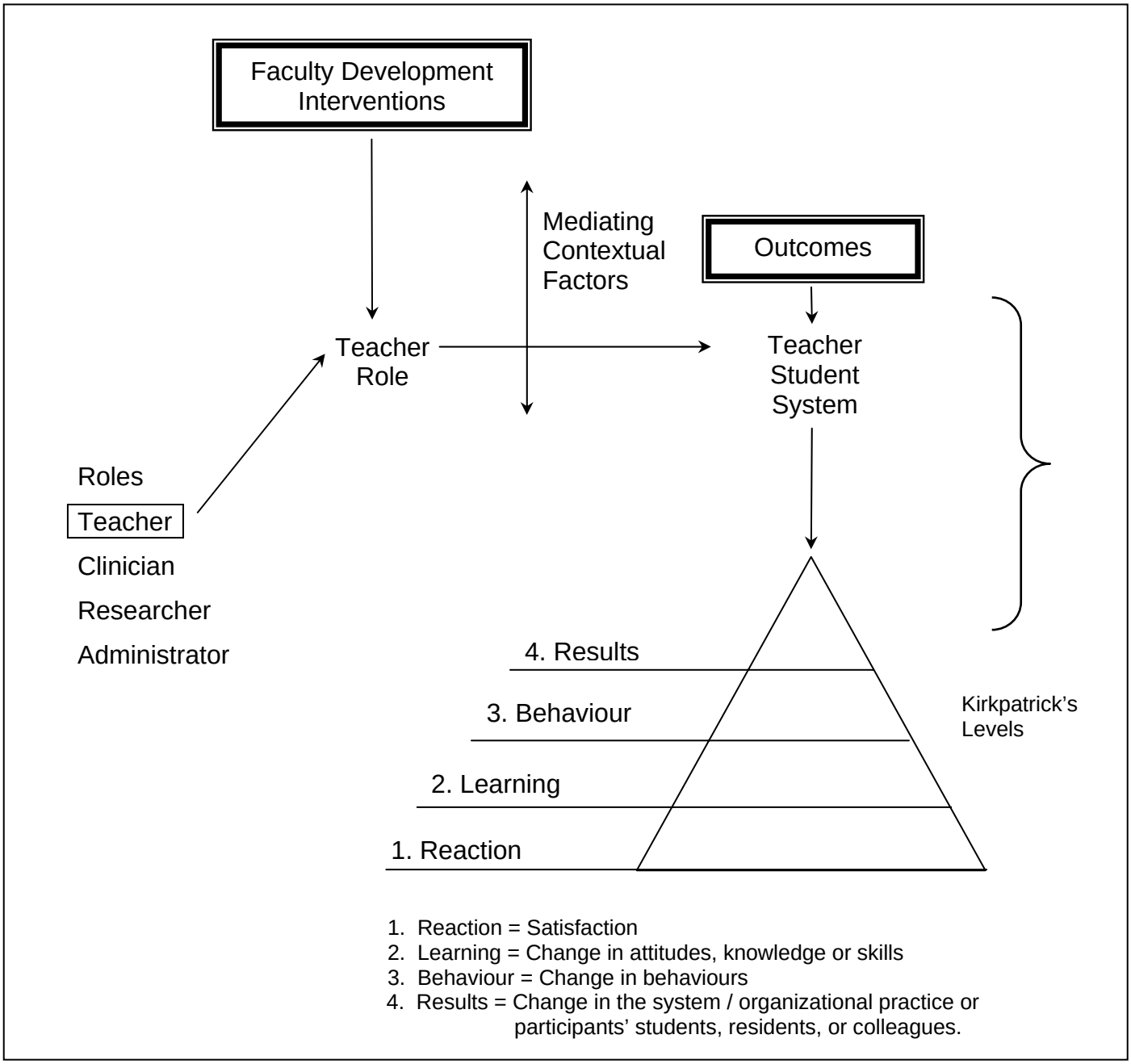

Figure 2

Kirkpatrick's Model for Evaluating Educational Outcomes*

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	**BEHAVIOUR** - Change in behaviours	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.

* 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.

Figure 3

Literature Review and Selection of Articles for Review

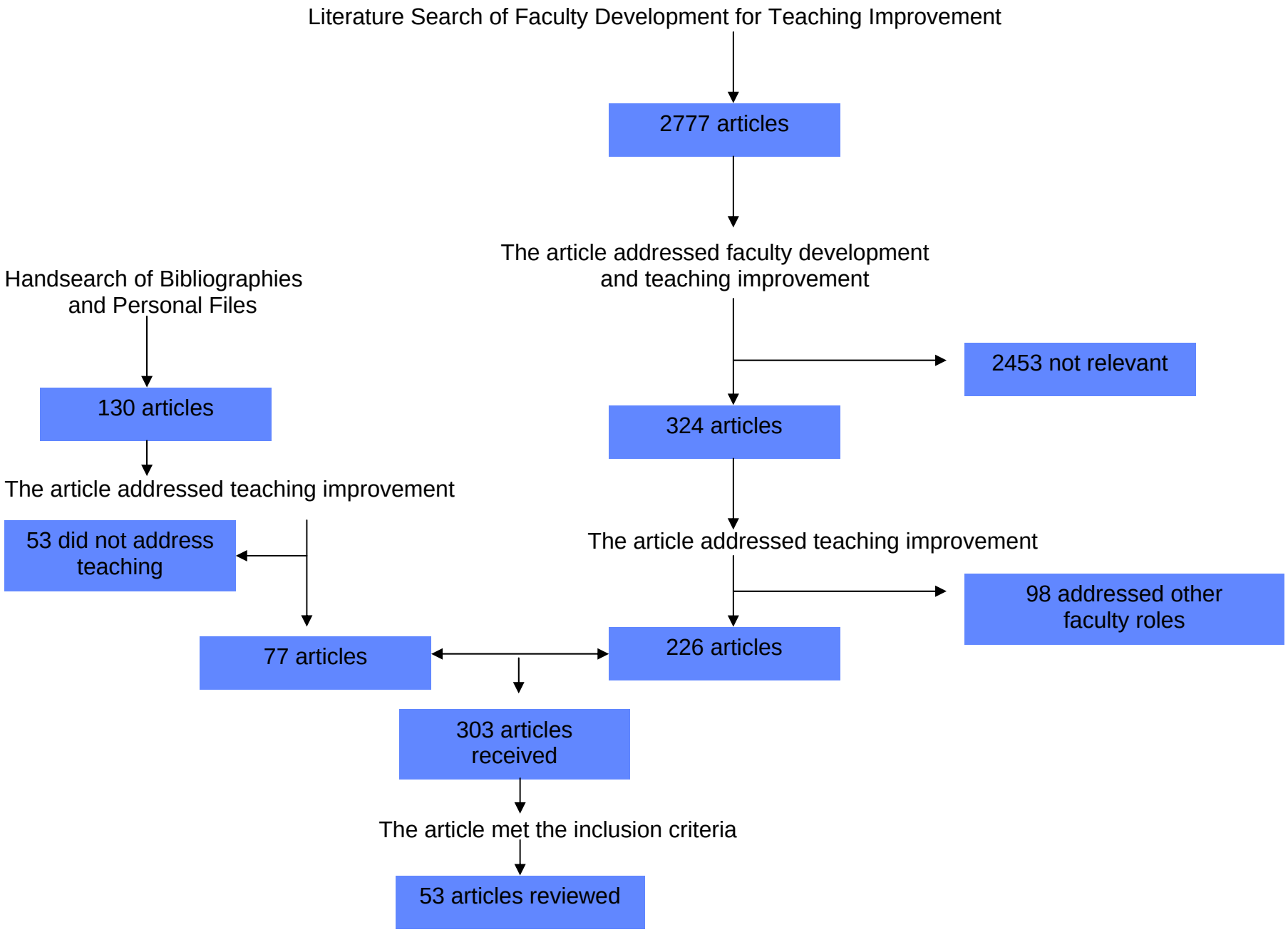

Table 2

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

Behaviour 72%
13/53 assessed self-reported changes in behaviour
25/53 assessed observed changes in behaviour

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.

Table 3

Summary of the 8 Most Highly Rated Studies

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 mid-way through their 4-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 many need to be tailored for different needs.
Barroffio *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 in either 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 are particularly beneficially for this group of individuals.
Mahler and 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 behaviour 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 (approximatly 3 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 and Neumann (1987) examined the effects of the above workshop (Mahler and 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.

Skeff (1983) evaluated the effect of intensive feedback. 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 than compared to 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 behaviour. Changes in teacher attitudes and ratings of teacher impact significantly favoured the experimental group; specifically, changes in the teachers' behaviour 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 core 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 pretest-posttest ratings. At the system level, 20 of 45 (44%) clinical teaching recommendations for improvement were judged by facilitators 6 months later as having made significant progress made toward implementation.

Marvel (1991) conducted an evaluation of an intervention to improve teaching skills in a family practice residency program. 16 faculty physicians participated. The intervention consisted of individuals viewing videotapes of their teaching, using a check list 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 behaviours improved following the intervention. Individualized feedback was provided to each faculty member following baseline data collection. Patient ratings of residents increased, but not significantly.