Contents

ARTICLES

Planning and Evaluating Curricula for the Humane Practice of Medicine
Wellesley R. Foshay 3

The Role of Knowledge in Teaching and Assessment of Problem-Solving
Geoffrey R. Norman 7

Simulation in Health Sciences Education
Geoffrey R. Norman, Linda J. Muzzin, Reed G. Williams, and David B. Swanson 11

An Instructional System as Change Agent
Hannelore Vandenscheidt and Ascher J. Segall 18

Organizational Change and the Development of Faculty Evaluation Systems
Deborah A. Orban and Allan J. Abedor 22

Evaluating Short-Term Training Programs:
A Practical Approach
Kent J. Sheets 26

Patterns of Change in Offices of Research in Medical Education: 1979-1984
Paul L. Grover, Douglas U. Smith, and Frank Schimpfhauser 33
Foreword

Craig Locatis, Guest Editor
National Institutes of Health
National Library of Medicine
Bethesda, MD 20209

Instructional developers have common core expertise, but often work in different settings. The health sciences are one of the larger arenas in which developers practice. Most schools of medicine, dentistry, nursing, and allied health have at least one education specialist and many have departments providing instructional services and conducting educational research and development. Unfortunately, information resulting from work in specialized areas may not be transferred, even though relevant and useful to others. For example, a decade ago, researchers in the health sciences assessing teaching effectiveness identified the lack of correspondence between student ratings of highly expressive instructors and the substance of what was taught. This “Dr. Fox effect” spawned a series of studies on educational seduction” that have implications for general teacher evaluation unknown to many outside the health sciences. Journals like JID are uniquely situated to foster information transfer. They are read by individuals confronting similar problems in varied situations and provide a mechanism for sharing information that might not be exchanged otherwise.

Information transfer motivated the DID Health Science Committee to propose this theme issue, one focusing on ID in the health sciences but with broader relevance. The issue does not cover the scope of instructional development in the health sciences—no single issue could—but does highlight critical concerns. The authors have written from a general perspective and have avoided jargon peculiar to the health field. What they say is of interest to developers in professional, higher, and continuing education and those who train specialists to high competence levels in any area.

Fosshay addresses the problem of what constitutes professional competence, describes its dimensions, and presents a strategy for integrating the learning of varied teaching goals. Norman discusses the critical area of problem solving in professional practice, summarizing research about medical problem solving and describing its instructional implications. Norman, Muzzin, Williams, and Swanson review health science simulation research. Since education in the health sciences can involve risks to patients and students alike, simulation is critical in teaching, assessment, and research. Vanderschmidt and Segall describe development of a preventative medicine curriculum that is important because the skills are becoming more crucial in providing health care and because the development strategies have to account for new competencies and engranger change in teaching institutions nationally. Oberan and Akselova continue this ID and innovation theme. They describe a faculty evaluation system that is interesting because the system emphasizes negotiation and its development was managed to insure acceptance. Sheets discusses program evaluation, especially that aimed at short term training. Problems and issues are described concerning data collection and the communication and use of evaluation results. Finally, Grover, Smith, and Schimpfhauser report their survey of medical educational research and development centers, documenting center status and activity trends. Results complement findings of a higher education development center study in an earlier JID issue.

The authors should be congratulated for their contributions as should members of the Health Science Committee who helped bring this issue about. Madeline Berry and Joel Lanphear deserve special thanks for soliciting manuscripts and providing authors with suggestions, as does Norm Higgins who gave the committee helpful advice and insured manuscript submissions received timely, responsive review.
Planning and Evaluating Curricula for the Humane Practice of Medicine

Wellesley R. Foshay
Advanced Systems, Inc.
1602 Tonne Road
Elk Grove Village, IL 60007

There has been growing interest in broadening the content of medical curricula to include humane considerations for the practice of medicine. Contributions from the Humanities and Social Sciences often are called for. While there have been many eloquent statements of the need for such curricula, (e.g., AAMC, 1983) there remains a need to improve methods of development and evaluation in this area. All too often, humane curricula appear to students as disparate experiences drawn from many disciplines, each with vague or misleading learning objectives and an evaluation/feedback strategy which fails to emphasize the most important goals. The critical task of synthesizing this knowledge for practical application in clinical settings is left largely to the student.

The result is that students often master only the technical facts of medicine. Diagnosis of problems and formulation of intervention strategies often occurs in only the narrowest perspective. Lessons learned from humane curricula are of little help beyond the classroom.

The alternative is to develop an integrated curriculum designed from the start to deal with the problem of clinically-oriented synthesis and application. It would focus on the patient. Its central structuring question would be “how should the complete range of factors which comprise the identity of our patients as humans be reflected in the practice of medicine?” rather than, “how can these (social science, humanistic, etc.) disciplines contribute to physician education?”

By focusing on the former question, chances may be improved for achieving the commonly-stated goal of teaching the student clinician to treat the “whole patient.” By posing this central question, an integrated humane medical curriculum may be achieved. This is not a new idea; it has been an ideal of many patient or case centered curricula. The problem has been in implementation: getting faculty to see beyond their disciplines, plan the curriculum and evaluate accordingly. The structure presented here may help to resolve this problem.

The Dimensions of Learning in the Humane Curriculum

To attain integration, it is appropriate to think of the human experience simultaneously in a number of dimensions. For instructional purposes, A. W. Foshay (1975) defines six:

The Intellectual: The symbolic interpretation of experience and symbol manipulation. Many symbol systems are verbal, although others also are common (e.g., mathematics and music). The ability to articulate and manipulate symbol systems is the subject of Bloom’s (1956) Taxonomy of the Cognitive Domain. In medical curricula, most of the content related to the identification and treatment of disease is in this dimension of human experience.

The Emotional: Not only the experience of emotions, but also “the ability to examine (and perhaps rejoice in) one’s own feelings—and those of others.” An extreme (but commonly occurring) example of experience in this dimension is the progression of emotional states in the terminally ill patient, each with corresponding emotional responses by the clinician. A more mundane (and perhaps less commonly studied) example of this dimension might be the clinician’s ability to recognize and deal with the emotional components in the clinician-patient relationship, even in routine outpatient practice. Many now consider emotional states of both patient and clinician significant in the formulation of treatment strategies. Some curricula even include direct instruction in the identification and accurate interpretation of behavioral cues to emotional states, as a basic skill in this dimension.

The Social: The development of social organizations (e.g., the family), and corresponding moral development. An example of clinical skill in this dimension might be the ability to make accurate inferences about roles assumed by the patient in various settings, and to develop treatment plans which are consistent with these roles. Awareness of various professional roles in the health care team also would be an example, as would the identification and description of underlying professional values.

The Physical: “Growth in the realization of one’s self as a physical being.” Examples of curriculum content in this domain might include the development of skills of physical examination or surgical technique. As such skills develop and are integrated, sophisticated pattern perception probably occurs. These perceptions may be seen to merge eventually into what is commonly called “body image,” so that the boundaries become blurred between tools in use and the physician’s own limbs, or between procedures and their effects. In medicine, this dimension might also involve realization of others’ physical being. For example, the physical dimension is familiar when it involves planning the treatment of obese or deformed patients. It may also have application to treatment of athletes and dancers.

The Aesthetic: The range of sensuous, formal, technical, and expressive responses to an object of contemplation (including other humans). When a clinician admires the skill with which a surgical procedure was executed, the response might be viewed as an aesthetic one at the technical level. Additional examples might involve learning appropriate clinical responses to patients with deformities or disfiguring injuries.
Similarly, when teaching the genital examination, clinically appropriate aesthetic responses should be taught. Learning new aesthetic responses also is an appropriate goal in the anatomy dissection laboratory and the surgical residency.

The Spiritual: Relating to the search for ultimate meaning; discovery of a sense of awe and wonder over the phenomena of existence. Examples of this in clinical curricula might include learning to respect human life and the integrity of the patient. Also included might be the development of appreciation for the astonishing complexity and sophistication of the mechanisms of the human body. One might also speculate that the component of clinical medicine described as “the laying on of hands” could be taught as a manifestation of this dimension. Perhaps one reason certain faculty members seem to be charismatic role models is because of their ability to demonstrate knowledge in this dimension.

The Role of the Dimensions in Curriculum Development and Evaluation

If human experience is modeled in these six dimensions, then it follows that any learning experience carries the potential for being taught and evaluated in relation to each. Curricula that fail to address all dimensions probably omit some aspect of humanness. Similarly, curriculum evaluations should be sensitive to learning in all six dimensions, if they are to be complete. Of course, this does not mean that each dimension must carry equal weight in every learning experience. However, if the multidimensional nature of human experience is ignored when planning instruction, then there is a substantial risk that graduates will perceive patients and the provision of health care along only a few dimensions.

Another, more subtle risk exists, however: failure to plan in all dimensions can result in delivery of implicit lessons in the ignored dimensions. These may be inconsistent with the lessons which have been explicitly included in the curriculum. Thus, for example, the conventional medical curriculum risks producing superb technicians who emerge from their training with in appropriate values, attitudes, and social skills. Thus, curricula which fail to include plans in all six dimensions are inhumane (or dehumanizing) because they ignore the full complexity of human experience. Each dimension can be considered a domain in which to specify and assess learning outcomes.

Developing Goals for the Six Domains

In developing goals for curricula in all six domains it may be helpful to keep in mind a progression of “levels” of achievement, much as is now done commonly with the Intellectual (Cognitive) domain. While much developmental work remains to be done in this area, some initial characterizations are presented in Table 1. For the Emotional, Social, Physical, and Spiritual dimensions, the basic progression is from perceiving experience by recognizing experience patterns in others or one’s self, to generalizing patterns beyond single individuals, and finally to evaluating or predicting from such generalizations. This progression corresponds to increasing sophistication in the self/other discrimination. For the Aesthetic domain, Broudy’s (Broudy, Smith, & Burnett, 1964) characterization of four discrete levels of aesthetic response seems useful for instruction. However, in some applications, viewing the four levels of response as a progression seems reasonable.

Writing Objectives for the Six Domains

When reducing the goals of the curriculum to specific objectives, one could use Table 1 to analyze any topic by asking for each of the six domains: (1) “What are the significant cues in relevant clinical situations which apply to this domain?” (2) “What responses to these cues should one expect to observe in the proficient student?” (3) “What criteria for judging a student’s assessment of his/her own response will indicate that the student is attending to the appropriate cues and criteria of judgment?” (4) “What criteria can be used to judge the appropriateness of the student’s inferences about others in this dimension?” These four questions can be answered quite well in many areas by the conventional techniques of behavioral analysis. However, in some cases the student’s responses will not be directly observable (as required by question (2). In such cases, question (2) must be answered by specifying appropriate verbal or nonverbal behavioral responses which are correlates of the response itself. Where such correlates cannot be found, it may be necessary to leave question (2) unanswered, and to specify only the criteria for question (3). This is admittedly an indirect approach to measurement, but it may be worthwhile in some cases to sacrifice some of the reliability of a direct behavioral analysis for improved validity of the objective. Examples of objectives for each dimension are given in Table 2.

An example may help to demonstrate the multidimensional approach. In developing a curriculum to teach medical interviewing, it might be reasonable to develop goals and objectives for gathering the required types of content (chief complaint, present illness, etc.) in the Intellectual dimension. In addition, however, the skilled clinician also is proficient at detecting and interpreting the emotional cues emitted by the patient, and in identifying the cues he/she emits to the patient (the Emotional). Furthermore, the clinician should be able to govern his/her own behavior—both verbal and nonverbal—so as to achieve the goals of the in-

A humane curriculum would train physicians to think of people simultaneously in six dimensions: the intellectual, the emotional, the social, the physical, the aesthetic, and the spiritual.
Table 1
Progressions Within Each Domain

<table>
<thead>
<tr>
<th>Domain</th>
<th>Progression</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Intellectual</td>
<td>1. Predicting impact of overt interventions on social norms, role definitions, and moral codes, of oneself and others</td>
</tr>
<tr>
<td>The Emotional</td>
<td>2. Recognizing emotional states in others and oneself</td>
</tr>
<tr>
<td>The Physical</td>
<td>3. Identifying the mutually causal relationship of emotional states and social experiences, in others and oneself</td>
</tr>
<tr>
<td>The Social</td>
<td>4. Predicting the emotional impact of a social action on others and oneself</td>
</tr>
<tr>
<td>The Aesthetic</td>
<td>5. Recognizing physical self-concept, in oneself and others</td>
</tr>
<tr>
<td>The Spiritual</td>
<td>2. Sensuous experience: awareness of the perceptual process as a separate experience, in one’s self and others</td>
</tr>
<tr>
<td></td>
<td>3. Technical response: Recognizing the skill used to produce that object</td>
</tr>
<tr>
<td></td>
<td>4. Expressive response: Assessing the impact of one’s total reaction to the object</td>
</tr>
</tbody>
</table>

Using the Model for Evaluation

Even where the existing curriculum has not been formulated by the multidimensional model described here, the model may be useful for evaluation. In assessing the learning outcomes of a curriculum, it is common to search for unintended (incidental) learning, as well as achievement of stated objectives. This sometimes takes the form of an attempt to discover the “hidden curriculum.” The scope and precision of such an evaluation can be enhanced by application of the multidimensional model to the design and analysis of questionnaires, observation protocols, and other data-gathering instruments and procedures. Furthermore, the model can be of use in developing and refining the questions or hypotheses which serve to structure the evaluation.

Potential Advantages of the Multidimensional Approach

A medical curriculum planned in six dimensions is likely to include many unfamiliar components. However, the approach has five potential advantages which may facilitate incorporation of these components into a medical curriculum, thus realizing its humane goals.

First, it can help to establish the relationship of the humane concerns to the conventional curriculum content, without slighting either. The multidimensional representation carries to a detailed level the principle that the two areas of content are complementary, and not contradictory.

Second, it can provide a means by which the goals and objectives of the humane curriculum can be specified with sufficient precision to give faculty clear guidelines for implementation.

Third, it can provide a structure by
Table 2
The Six Domains (Dimensions) of Human Experience, With Sample Objectives From Each

<table>
<thead>
<tr>
<th>DOMAIN</th>
<th>DEFINITION</th>
<th>SAMPLE OBJECTIVE*</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Intellectual</td>
<td>similar to Bloom’s cognitive domain</td>
<td>Given a case history of UTI, the student states a rationale for treatment which is justified by relating it to underlying pathophysiological principles.</td>
</tr>
<tr>
<td>The Emotional</td>
<td>feelings, emotions, and emotional development</td>
<td>In an interview, the student identifies the emotional content conveyed by the patient’s facial expressions.</td>
</tr>
<tr>
<td>The Social</td>
<td>development of social organization; moral development</td>
<td>In an interview with a patient who emits cues suggesting needs beyond those normally within the clinical relationship, the student identifies those needs.</td>
</tr>
<tr>
<td>The Physical</td>
<td>psychomotor skills; physical self-concept</td>
<td>Given a case history of a patient who has just undergone a radical mastectomy, the student describes a plan for assisting the patient in modifying appropriately her physical self-concept.</td>
</tr>
<tr>
<td>The Aesthetic</td>
<td>formal, technical, sensuous and expressive response to an object of contemplation</td>
<td>In a physical examination of a disfigured patient, the student evaluates the appropriateness of his/her response to stated professional values.</td>
</tr>
<tr>
<td>The Spiritual</td>
<td>relating to the search for ultimate meaning: awe, wonder</td>
<td>In an interview with a patient who has a chronic condition, the student questions patient to determine the effect of the condition on his/her self-concept.</td>
</tr>
</tbody>
</table>

which faculty from a variety of disciplines can relate the lessons of their specialties to the clinical experience, as in the interpersonal skills example above. In this way, a single, unified framework can be created for curriculum planning across disciplines.

Fourth, it provides a framework for evaluating humane curricula and providing feedback to students about their achievement. Existing curricula can be evaluated for the lessons taught in each dimension, and new curricula can improve the precision of their program and student evaluation procedures by basing them on objectives formulated according to the six-dimensional model.

Fifth, it can provide a structure which recognizes the contributions of all health care professionals. For example, recent reforms of some nursing curricula place considerable emphasis on goals which might be best interpreted within a multidimensional model such as this one.

Conclusion
The model described here implies very little about the details of topics to be included in a humane curriculum of medicine. Such decisions are at base value judgments, and consequently they lie beyond the scope of this model. Instead, the model is useful for operationalizing such value judgments by providing a framework for specification of goals, objectives, and evaluation procedures. Thus, the model could be applied productively to the development and/or evaluation of many different curricula, each with different aims. Its utility is as an analytic tool, not a prescriptive one.

AUTHOR NOTE. This work was done while the author was at the School of Clinical Medicine, University of Illinois, Urbana, Illinois. Craig Locatis, Reed Williams, and DeLayne Hudspeth made many thoughtful comments and suggests about earlier drafts of this paper. Their help is greatly appreciated.

References
AAMC Project on General Professional Education of the Physician and College Preparation for Medicine. (1983). Emerging perspectives on the
The Role of Knowledge in Teaching and Assessment of Problem-Solving

Geoffrey R. Norman
Associate Professor
Department of Clinical Epidemiology and Biostatistics
McMaster University
Hamilton, Ontario, L8N 3Z5

Recent research has resulted in a theory of medical problem-solving that is radically different from the theory which prevailed for nearly two decades. The new theory has not yet led to revisions in medical school curricula or evaluation methods, but potentially may engender change based upon a deeper understanding of the cognitive processes of expert clinicians. In this article, two theories of clinical problem-solving are contrasted. Studies of medical problem-solving conducted by the author and others are described. Evidence is presented to highlight the central role of experience in the development of competence. Finally, implications for the teaching and assessment of problem solving, especially in medicine, are suggested.

In the early 1960s medical education moved away from emphasizing the acquisition of rote knowledge towards the development of problem-solving strategies. Signaling this change were the “critical incident” study commissioned by the National Board of Medical Examiners, which clearly highlighted the multiple dimensions of clinical competence, and the appearance of objective measures of clinical problem-solving using card or paper based patient simulations (Rimoldi, 1961; McGuire and Babbott, 1967). These evaluation methods implicitly viewed “problem-solving” as a general skill or strategy applicable to different problems. Findings show moderate correlations between performance on these problems and measures of content knowledge were interpreted as evidence for the presence of a general problem-solving strategy (McGuire and Babbott, 1967), although equally low correlations were observed when scores on different problems were compared (Berner, 1977).

Curriculum efforts focusing on the development of problem-solving skills inevitably followed. Problem-based learning (Barrows and Tamblyn, 1981) in which basic science material was learned in the course of solving clinical problems, usually in a small group setting, exemplified the growing interest of educators in clinical problem-solving. The philosophy of this approach was simple—if students were to become effective problem-solvers, the faculty had a responsibility to teach problem-solving, and problem-based learning emerged as a natural approach to achieve this goal. However, since little was known about problem-solving, it was difficult to know what to teach. In response to this dilemma, studies were undertaken at McMaster University and Michigan State University in the early 1970s (Barrows et al., 1982; Neufeld et al., 1981; Elstein et al., 1978) to describe the problem-solving process of clinicians and its development in medical students. The results of these studies were inconsistent with the expectations of the investigators. They found a general process—the now classic “hypothetico-deductive method,” but no evidence that some clinicians were consistently better or worse problem-solvers than others or that the problem-solving process improved from entry into medical school through clinical practice. Perhaps most disquieting was the finding that the best predictor of successful problem-solving, at least as assessed by obtaining the correct diagnosis, was the content of the diagnostic hypothesis.

As evidence of the fundamental role of early hypothesis generation, over 80% of the diagnostic hypotheses generated by clinicians occurred in the first quarter of the encounter. Of those encounters where the clinician arrived at the correct diagnoses 95% had thought of it within the first five minutes. Conversely, when physicians did not reach
the correct diagnosis, 95% had not entertained it as an hypothesis. If clinicians thought of it in the first 5 minutes, they got it; otherwise they never did.

The conclusion from these data is that experts are experts because they think of better, more correct or more appropriate hypotheses very early on. The search for a general problem-solving process supported the central role of knowledge, as it is reflected in the accurate hypotheses generated by experts.

Studies in other domains, including chess (deGroot, 1965; Chase & Simon, 1973), the oriental game of GO (Reitman, 1976), physics (Larkin, et al., 1980), and other domains (Schoenfeld & Herrmann, 1982), have arrived at similar conclusions. As an example, chess masters did not differ from novices in their thought processes—the number of moves considered, search strategies, depth of moves—but they thought of better moves (deGroot, 1965). As Glaser (1984) noted in his review article:

"Our interpretation is that the problem-solving difficulties of novices can be attributed largely to the inadequacies of their knowledge base and not to limitations in their problem-solving strategies." (p. 99)

One feature of memory which distinguishes experts from novices in chess is their ability to recall legal chess positions. After a five second viewing of a chess board, experts could remember better than 90% of the board; however, if pieces were placed at random on the board, no differences in recall between experts and novices were observed (deGroot, 1965).

These studies have been replicated in medicine (Norman et al., 1979). Four case protocols were developed, two typical cases of gall bladder disease and lung cancer and two "random" protocols, and were given to second year students, first and third year residents in family medicine, and family physicians. Each subject read the case through once at his own speed, and then recalled what he read. The amount recalled divided by time spent reading, for the typical and atypical protocols yielded results consistent with the chess findings. Typical cases showed a consistent increase in recall with increasing expertise, but no differences in recall were present with random protocols. One intriguing feature of this study was that the gradient of performance on typical cases began in residency education, suggesting that differences reflected a different kind of knowledge than the formal basic and clinical science knowledge taught in medical school.

Another intriguing finding emerged from a secondary analysis. One of the "random" protocols had been developed from an actual patient and might more correctly be called an "atypical" or unusual case. This seemed to be handled differently than the "true random" protocol. Items recalled for this case followed a gradient with education level similar to the typical cases (Figure 1). Time spent reading the protocol (Figure 2) was very different from either typical or random protocols, in that expert family physicians spent an inordinate amount of time puzzling over the case. This raises the issue of differences in expert behavior when dealing with typical and atypical cases, and what is it about a particular problem which makes it typical or atypical.

Typical Cases and Pattern Recognition

Typical cases probably represent 75 to 95% of the problems seen by a physician. One distinguishing feature of experts in dealing with these problems is their ability to see patterns or "chunks" of data, as evidenced by the natural grouping of phrases in free recall tasks (Figure 3). This can be measured by the time taken to recall an individual item, and studies of experts and novices in cardiology and rheumatology have demonstrated that this time is inversely related to expertise.

Two other phenomena occur in experts faced with this recall task. First, they tend to make large inductive inferences from the data, so that, for example, a productive cough, chronic and steadily worsening, over a period of several months or years, without hemoptysis, is labelled a "bronchitic cough." This appears to occur at a perceptual level so that experts are unable to recall individual features. Because of this phenomenon, studies of subspecialists could no longer demonstrate differences in total amount of recall with expertise (Muzzin et al., 1982, 1983); however experts could recall more important features—what Patel (1983) calls "frame" information, or what others have called "deep structure" (Chi et al., 1981). Finally, there is the rapidity or automaticity with which experts can appreciate patterns in data, best illustrated by visual tasks such as radiologic diagnosis. A study of chest film interpretation showed that experts were able to correctly identify abnormal films in an average of 7 seconds, and normal films in about 10 seconds (Norman, et al., 1984).

Automaticity is further supported by studies of disease prototypes (Bordage, 1984). A prototypical disease is one which best exemplifies the category, for example tuberculosis is a prototype of infectious pulmonary disease. Prototypes are not necessarily the most common or serious diseases within a category; rather, they are most representative in the sense that they have most features which distinguish that category.
from others. Disease prototypes are recalled earlier and more frequently by clinicians, are judged to have a greater number of overlapping features with the category, and are more rapidly and correctly classified as true or false when they are presented to clinicians in statements such as “diabetes is an example of an endocrine disease.”

To summarize, experts process common disorders almost automatically at a perceptual level, so that the expert literally “sees” patterns which are not available to the novice. This processing, strongly influenced by prior experience and based on a pattern-matching process, shows little evidence of general problem-solving strategies which are independent of experience or knowledge, and may explain why studies of clinical problem solving have had difficulty revealing correlations in performance across problems.

Studies of Atypical Cases

As demonstrated in the earlier study of expert-novice memory, experts treated the single atypical case in the study very differently—recalling about as much data as in a typical case, but taking much more time. Similar results arose from two studies of visual processing—the radiology study, also discussed earlier, (Norman et al., 1985) and a study of recognition of skin lesions by students and clinicians (Norman et al., 1985). In both studies, correct interpretation of the stimuli by experts occurred more rapidly than by students. However, incorrect identification showed a positive relationship to expertise, so that expert dermatologists took an average of 25 seconds to incorrectly label a lesion, versus 14 seconds for clinical clerks and 11 seconds for first year students. It is significant that errors were associated with long processing times—which reflect a much less automatic form of processing and probably more genuine “problem-solving.” Chase and Simon (1973) arrived at a similar conclusion in studying chess masters.

None of these studies explicitly deals with what makes an atypical problem atypical. Bordage and Allen (1982) have, however, dealt explicitly with this issue by devising four types of written case protocols, the first corresponding to a textbook or prototypical presentation, and the remaining three containing either (a) a single critical cue necessary to support the correct diagnosis; (b) a single distracting cue unrelated to the diagnosis or (c) a misfitting cue which suggests an alternative, correct diagnosis. He showed that each atypical presentation leads to a large increase in error rates among both experts and novices, presumably because they demand a logical, cue-by-cue inference rather than pattern recognition. Placing these demands on the expert clinician results in a high error rate. Further research is required as to whether the differences in error rates and response times are in fact a reflection of qualitatively different processes, such as pattern recognition and problem-solving.

Implications for Education

Medical education in the twentieth century has been guided by two successive paradigms. The first rooted medicine firmly in the basic sciences and viewed clinical expertise as a mastery of a defined body of scientific facts and principles. The more recent paradigm has viewed the clinician as a problem-solver, distinguished by unique set of skills, but able to marshall and apply facts as needed to solve problems. To the extent that the interpretations made on the data here are correct, clinical expertise appears to be quite different from either characterization. Certainly, in viewing the diagnostic process, the distinguishing characteristic of the expert clinician is the availability of vast clinical experience against which new presenting situations can be matched in an almost intuitive and automatic manner, complemented by an occasional venture into a more psychologically or scientifically “pure” form of problem-solving behavior when the occasion arises. In this latter instance, one may presume that the clinician uses basic principles of anatomy, physiology and so on to arrive at a solution. No data have yet been collected on this aspect of problem-solving.

Any extrapolation from the limited data presented thus far to the educational domain must be viewed as conjectural. However, if clinical experience is a major determinant of expertise, it follows that the more clinical problems a student can encounter, the better. Further, given the significant role of context in the ability to subsequently recall and recognize information (Riccio & Reitman, 1984), it is reasonable to suggest that since clinical and basic science “facts” will be used in a patient context, learning these principles in a patient situation should enhance recall. There is some evidence to support this conclusion from a study by Claessen and Boshuizen (1985) at Maastricht (a problem-based curriculum) and Utrecht (a “subject-based” curriculum). Using a recall task similar to those already described, there was consistent improvement in recall of the Maastricht students compared to Utrecht students as a function of educational level.

The implications for evaluation are also significant, if one accepts the theory that clinical problem-solving usually
represents an interaction between the presenting situation and previous representations in memory. Since there is no reason to suppose that availability of appropriate representations for one problem would be related to a similar availability for a second, it may be that problem-solving is "content-specific," but not in the way it is commonly assessed on multiple choice tests. Rather, the "content" is more likely to represent complex, configurational, and perhaps unconscious representations of signs and symptoms. This theory of expert memory would also account for the frequently observed lack of relationship between thoroughness in data gathering from patients experience. To the extent that experience leads to automatic associations between initial patterns and categories in memory, the expert will be seen to make apparently inductive "leaps" from minimal data, and the search for additional data is only confirmatory, or even irrelevant.

The study of general problem-solving strategies, if they exist, will require the careful selection of challenging problems similar to those of Bordage and Allen (1982). Under these circumstances, one should be able to observe the logical inference based on scientific principles which might be called "problem-solving." But in more common situations, these activities may well be absent.

Since the access of relevant knowledge in a patient context is the hallmark of expertise, the expectation that such knowledge should be available in other contexts, like multiple choice tests, stacks the cards against the expert. That increasing experience results in generally poorer performance on conventional tests is not surprising and should not be interpreted as an indicator of diminished clinical competence. The studies discussed may be a first step towards the development of new testing methodologies. The methods are efficient, objective, and have shown high reliability across repeated cases—in contrast to many problem-solving tests. However, much work needs to be done before these approaches can be used as valid measures of competence. As yet, all that is certain is that scores are positively correlated with experience, and it remains to be demonstrated that they are related to competence.

**Conclusions**

Recent evidence from studies of the psychology of clinical reasoning suggests that expertise in medicine is characterized by the gradual accumulation of experience in memory which results in an ability to perceive, organize, and recall the extensive information presented by the patient. The experienced clinician is a better problem-solver by virtue of this accumulated experience, not as a result of any innate or learned problem-solving skill. These results suggest curriculum innovations focused on the provision of an environment rich with clinical experience, and the learning of concepts in the context of clinical problems. Further research is required to identify the strategies used by experts in dealing with unusual or atypical problems.

**References**


![Figure 3: Examples of Transcribed Recall from an Experienced Clinician and from a Year II Student](image-url)

**Figure 3. Examples of Transcribed Recall from an Experienced Clinician and from a Year II Student.**

From an experienced clinician...

- My memory of these cases is based on actual exposure, but doesn't seem to have been particularly solving strategies. I feel it's a good case about the role of the patient in the actual exposure, but the patient is not the patient. The problem is not how the patient should be treated, but how the patient should be treated... and how the patient should be treated.

From a Year II student...

- Oh, this is very different... that patient's not the same patient, but the patient who answers the question about the patient's not the same patient... and the patient should be treated... and the patient should be treated... and the patient should be treated...
Simulation in Health Sciences Education

Geoffrey R. Norman  
McMaster University

Linda J. Muzzin  
McMaster University

Reed C. Williams  
Southern Illinois University

David B. Swanson  
American Board of Internal Medicine

Simulation has had an increasing role in medical and health science education for over two decades due both to new technology and the movement from content-based curricula toward approaches emphasizing problem solving and clinical reasoning. Various simulation methods have been widely researched as a result and pertinent findings are reviewed here. The review focuses on those techniques involving an encounter between a health provider and client. Excluded are simulations of laboratory experiments, physiological functions, or triage exercises involving the treatment of many patients in simulated emergencies. Five types of simulation methods, oral examinations, live simulated patients, mechanical or electronic simulations, and written and computer-based simulation are reviewed. Each class of simulation will be discussed under the following headings:

1. Fidelity—To what degree does the simulation method resemble a real life experience?
2. Reliability—How consistently can performance be assessed with each simulation type?
3. Validity—Does the simulation method have construct, concurrent or predictive validity?
4. Learning—Does the simulation result in student learning?
5. Feasibility—Is the method affordable and feasible to implement?

The questions concern dimensions along which different methods can be compared, however it may not always be possible to find one method which ranks high on all dimensions. Improving a method on one dimension often reduces its effectiveness on others. In the abstract, many simulation methods will have a near “zero sum” and their value depends on the specific educational applications.

Oral Examinations

Oral examinations include a variety of techniques that provoke the examinee to simulate clinical reasoning in response to examiners’ questions, either at the bedside or in a direct one-on-one encounter. Skills involved include data gathering, interpretation, differential diagnosis, and management.

Fidelity. Claims about the usefulness of oral exams in judging ability to apply knowledge, to problem solve, to respond to dynamic situations, to demonstrate interpersonal skills and professional attitudes are typically based on unsubstantiated impressions or on reports that the exam seems life like (e.g. Van Wart, 1974), rather than on a more systematic content analysis (Levine & McGuire 1970a, 1970b). The issue of fidelity commonly concerns the extent to which the “reasoning” displayed by candidates in an oral exam reflects how they would actually reason when confronted with a clinical case. Oral exam scores may be based partially on behavior unrelated to clinical competence, such as anxiety level (Pokorney & Frazier, 1966; Waugh & Moyse, 1969), the percentage of words contributed by examinees (Evans, Ingersoll & Smith, 1966), examinee visual impressions (Holloway, Collins & Start, 1968) and examinee self confidence (Wigton, 1980). Whether oral examiners are able to use these cues appropriately remains an issue.

Reliability. Most research has focused on lack of agreement among examiners (Bull, 1956). But the lack of agreement typically found between totally independent ratings of an examinee in two different situations (Hubbard et al., 1973) might have other explanations. The behavior observed may
be different or the prerequisite skills for the solution of each problem may differ. This interpretation is consistent with the finding that pairs of raters observing the same encounter have reliability coefficients of .75 to .89, while examiners observing different sessions have reliabilities of .25 to .45 (Carter, 1962; Wilson et al., 1969; O'Donohue & Wergin, 1978). While the former may be partially due to consultation between examiners, it does not explain why consensus is often unpredictable among three or more raters (Colton & Peterson, 1967). The problem may result from nonconformists on rating teams whose removal might improve consistency (Newble, Hoare & Sheldrake, 1980; Lloyd, 1983). When checklists and rating forms are used, inter-rater reliability has ranged from .59 to .72 (Maatsch, 1980; Littlefield, Harrington & Garman, 1977) and in general, more O'Donohue & Wergin, 1978). These results are often interpreted as evidence that orals and written examinations measure different aspects of clinical competence. However, test unreliability could also reduce these correlations (Levine & McGuire, 1970a; Meskauskas, 1975).

Learning. The oral exam has often been cited as a flexible technique allowing direct feedback between teacher and student and it is institutionalized in most medical schools at the bedside in the form of medical rounds and evaluative clinical exercises in residency programs and clerkships (Fletcher, Sanderson & Pusler, 1977). With a few exceptions however, (Vu, Johnson & Mertz, 1981; Powles et al., 1981) the learning value of these exercises has not been reported in the literature.

Feasibility. Oral exams are not feasible for large groups of learners because of logistical problems and costs resulting from having examinees at one location examining the same patients and being rated by the same examiners, now the widely accepted basis for standardization. Both the National Board of Medical Examiners and the American Board of Internal Medicine have had to discontinue orals because of their unreliability and expense.

Validity. Studies of construct validity of medical orals (Miller, 1968; Maatsch, 1980), show differences in performance within levels of training greater than differences between levels. Oral exam validation has tended to rely exclusively on how scores achieved on multiple-choice tests and orals intercorrelate. Although use of written tests as a criterion for validating oral exams is somewhat anomalous, multiple-choice tests are highly reliable, widely administered, known to measure factual knowledge and are more familiar than other measures. Most studies show small positive correlations between multiple-choice tests and oral exams (e.g., Bull, 1956; Ludbrook & Marshall, 1974; of logistical problems and costs resulting from having examinees at one location examining the same patients and being rated by the same examiners, now the widely accepted basis for standardization. Both the National Board of Medical Examiners and the American Board of Internal Medicine have had to discontinue orals because of their unreliability and expense.

Discussion. There may be trade-offs between reliability and content validity. As Perinumer remarked, "flexibility is perhaps inversely proportional to standardization" (American Board of Medical Specialties, 1983, p.64). As examinations are made uniform to achieve consistency, the flexibility, uniqueness and spontaneity of the oral exam is lost. High reliability resulting from standardization may only allow sampling a very small part of overall clinical competence, while longer and less structured orals might measure more skills but with lower reliability (O'Donohue & Wergin, 1978). Another issue concerns examiner ability to judge whether an examinee will make a good doctor. Although training and checklists can improve

Five simulation methods are reviewed: oral examinations, live simulated patients, mechanical or electronic simulations, written, and computer-based simulation.

Simulated Patients

Although healthy individuals acting as patients have always had a role in teaching the clinical examination, Barrows (1971) was the first to formally train people to simulate all aspects of a patient problem—history, physical findings and emotions. Simulated patients are purported to have several advantages over actual patients: they can be scheduled and chosen for their appropriateness to learning objectives; they can simulate acute, serious or rare conditions which might not usually be encountered; free discussion can be conducted in their presence that might be viewed as unethical or impolite with real patients; junior students can approach and examine them without concern for physical or emotional harm; and simulated patients can provide accurate and objective feedback without fear of consequences. This approach has been extended to real patients with chronic conditions trained to simulate their manner when they first presented their problem to a health professional (Stillman et al., 1980).

Fidelity. Trained simulated patients have been introduced into family physicians' offices without detection (Burri, McLaughan & Barrows, 1976; Owen & Winkler, 1974). However, other studies have resulted in a detection rate of 20% (Neufeld et al., 1983). Focussing more directly on behavior with simulated and real patients. Norman and Tugwell (1982) found no difference in history-taking, physical examination and diagnosis.

Reliability. The reliability of examinee performance in encounters with simulated patients depends on the scoring method. Subjective ratings using behavior categories and five point scales yield inter-rater reliability coefficients in the range 0.5 to 0.7 (Sinkel & Norman, 1973). Other approaches have much higher reliability. Because the patient problem is standardized, a criterion group can determine what options on history, physical, and so on constitute an appropriate workup of the patient's problem, and agreement of better than 95 percent (kappa = .86) between recall of the simulated patient and observer rating on these checklists has been demonstrated (Neufeld et al., 1983).
Other studies report similarly high reliability, with phi coefficients ranging from .77 to .89 (Stillman et al., 1980). However, several studies have shown low correlations of performance on different problems. In one study, the correlation across seven gastrointestinal problems was only .16 (Killer et al., 1983). Another study found correlations ranging from .26 to .42 for presentations of the same problem to .01 for problems in different specialties (Norman et al., 1983). Similar results, with correlations from -.26 to .42, were found by Rutala et al., (1980). Thus, content specificity once more appears to be a problem common to nearly all evaluations based on a single patient encounter.

Validity. Validity studies have examined how individual performance in simulated patient encounters correlates with performance on other measures. Content-specificity places an upper limit on validity since a stable estimate of performance based on simulated patients requires a large number of encounters. One study (Wakefield & Norman, 1977) compared performance of family medicine residents in an average of twenty-two observed encounters with their supervisor performance reports and their later performance on the certification examination. Scores on data gathering were uncorrelated with either criterion. Clinical judgment correlated .51 with the same competency assessed by supervisors, and .54 with examiners' ratings of treatment plans in the certification oral. Interpersonal and interview skills correlated .3 to .4 with similar categories assessed by supervisor report and oral examinations. Positive correlations have been found between scores of content assessed by patients trained as simulators and peer ratings in data acquisition, problem recognition, attitudes, and competence, supporting concurrent validity. But two studies (Killer et al., 1983; Norman et al., 1983) show no significant correlation between performance on simulated patients and multiple choice test scores. Scores from simulated patient encounters might primarily reflect interpersonal skill unrelated to content knowledge.

Learning. In contrast to reliability and validity, numerous studies show simulated patients have positive educational effects. Several studies have contrasted learning from simulated and real patients (Tinning, 1975; Livingstone & Ostrow, 1978; Holzman et al., 1977; Johnson, Murchison & Reiter, 1976) and students taught with simulated patients have consistently out-performed those taught with real patients.

Feasibility. A major barrier to feasibility is the cost of training and using simulated patients, which requires use of faculty resources. Training and application costs are consistently overestimated, however. At McMaster University, simulated patients are trained in two to three hours at a cost of $10 per hour, and cost of application is also $10 per hour, which compares favorably with that of computer simulations. The real cost is faculty time required to observe and evaluate simulated patient encounters, but there has been significant progress towards using simulated patients without the requirement for faculty participation.

Discussion. It is clear from this review that simulated patients have a strong, positive effect on learning. Further, the approach has demonstrated validity in comparison to real patients which is unequaled by other simulation methods. Both these features suggest a useful role for live simulated patients in teaching and evaluation which is not, as yet, fully realized.

All methods of evaluation must deal with the problem of "content specificity," the tendency for mastery of one skill domain to be unrelated to mastery of another.

Manikins

Perhaps the most widespread but least discussed simulation devices in health sciences education are manikins such as RESUSCI-ANNIE (for cardiopulmonary resuscitation) and GINIE (for pelvic examination). The term "manikin" is an over-simplification. Although some are the essence of simplicity, others, like the SIM-1 computer-controlled manikin (Abrahamson, Denson & Wolf, 1969) or the cardiac simulator HARVEY (Gordon, 1981), are quite complex. Most are mechanical or electronic, but also included are animal "models" such as anesthetized cats for teaching pediatrics intubation (Thompson, 1979). Manikins are designed universally to teach and evaluate a defined procedural or technical skill. In most applications, the focus is on learning rather than assessment. Some controlled experiments have been conducted with manikins. Evaluation using simulators can be process oriented, by using detailed checklists, or outcome oriented, by comparing the student's formulation (e.g. split third heart sound, retinal exudates) with a known abnormality. Either or both orientations may be appropriate.

Fidelity. Fidelity depends on the specific device. SIM-1, is highly realistic, even to blinking eyes, but other manikins are less realistic. Little empirical evidence of fidelity exists. Thompson (1979) stated that intubation of kittens was highly realistic, which only proves that beauty is in the eye of the beholder, or throat of the beholder. Seventy-five to 85% of physicians found HARVEY realistic (Gordon, 1981). However, fidelity of simulations need only be consistent with the educational purpose (Simon, 1981). For example, airplane simulators need a complete array of instruments and authentic responses to pilot actions but do not need wings to serve their purpose.

Reliability. Performance is usually assessed with detailed behavioral checklists and reliability is usually high.

Inter-rater agreements of 95% have been reported with the cardiac simulator HARVEY (Gordon, 1981) while other studies using checklists, without manikins, report reliabilities in the range of .75 to .94 (Liu et al., 1980; Andrews, 1977). Very high reliability is achievable through the use of manikins and behavioral checklists.

Validity. Little empirical evidence of validity exists. Abrahamson et al. (1969) showed that residents trained on SIM-1 performed better than others in the clinical setting, implying some concurrent validity. However, a study of a less life-like simulator found that a group trained on a heart sound simulator (Sajid, Magro & Feinheimer, 1977) showed large gains in recognizing taped heart sounds, but there was no dif-
ference in performance with actual patients. Clearly, further research is required.

Learning. Thompson (1979) indicated that residents found practicing intubations on cats “useful”. Penta and Kofman (1973) conducted a randomized trial, and showed significant differences between groups using the Iowa Ophthalmoscope model and the Strabismus Cover Test Simulator, but no differences for the Barter eye model or a heart sound simulator. As already noted, students trained on SIM-1 and on a heart sound simulator showed superior performance, but only SIM-1 training transferred to clinical practice.

Feasibility. Widespread use of simple models like RESUSCI-ANNIE and heart sound simulators attest to their feasibility. Conversely, the enormous development costs of computer controlled manikins like SIM-1 has undoubtedly led to few adoptions. Given their variety, a global judgment about the feasibility of manikins cannot be made.

Discussion. Manikins, or technical simulators have achieved a limited, but important, role in teaching clinical skills. When defined technical or motor skills are to be acquired, these simulation devices can be useful. However, it is discouraging to see limited rigorous evaluation of their validity and effectiveness.

Written and Computer-Based Simulations

Written and computer-based simulations have been used to assess problem solving since the late 1960s and have been investigated in many studies. Findings from research on both kinds of simulation are deliberately integrated since most work with computer simulation has simply replaced paper and pencil with a computer terminal, with little evidence indicating that computer use has important measurement consequences. Recent applications of computer simulation, using microcomputer and videodisc technology does represent significant progress but as yet no studies of these methods have been reported.

Fidelity. Several studies have investigated whether actions taken on these simulations are similar to those which occur with real patients. Three different research paradigms have been used—introspection/self report, chart audit, and live simulated patients. Introspection/self report studies (e.g. McGuire & Babcock, 1967; Harless et al., 1978; Finchman et al., 1976; National Board of Medical Examiners and American Board of Internal Medicine, 1981) indicate that examinees consider the simulations realistic. Audit studies, comparing behavior on the simulations with similar real life cases, show performance similarity, at least in selection of important diagnostic and therapeutic actions (Goran, Williamson & Gonnella, 1973; Harless et al., 1978; National Board of Medical Examiners and American Board of Internal Medicine, 1981), but more data is collected using formats that cue the examinee by delineating specific options. Comparisons of behavior on written and computer-based simulations with identical cases using simulated patients have also demonstrated cueing effects (Feighn & Norman, 1978; Norman & Feighn, 1981; Page & Fielding, 1980), which lead to higher scores than in uncued formats or real life. Cueing problems may be worse in the written than computer formats because examinees may read ahead and use choices listed in later sections to determine which earlier options are appropriate.

Reliability. Three sources of measurement error have been investigated—variation introduced by the scoring procedure used, inter-rater variation in developing scoring keys and intercase variation in performance. Most scoring schemes for written and computer simulations involve categorizing each option (e.g. essential, indicated, contraindicated and risky) and assigning a numerical weight to each category (e.g. essential = 2, indicated = 1, contraindicated = 0, and risky = 0). Changes in number of categories and moderate changes in the weights assigned categories make little difference in resulting scores (Bligh, 1980, Norcini et al., 1983a). Correlations between score variants are so high that reliability and validity are unaffected. These results mirror research on weighted versus unweighted composite scores (Lord & Novick, 1968) and prediction accuracy using linear composites with different regression weights (Dawes & Corrigan, 1974).

Poor inter-rater agreement on appropriate (and inappropriate) options adversely affects scoring clinical simulations. There can be substantial variation in workup and treatment recommendations among skilled physicians on real and simulated cases (Barrows et al., 1978; Elstein, Shulman & Spreafco, 1978). As a result, consensus judgment by an expert panel seems necessary and appropriate for constructing scoring keys (Mazzucco & Cohen, 1982), but it is unclear if consensus scoring adequately rewards or penalizes approaches to patient management that reflecting differences in quality of care versus differences in style and opinion.

Previous studies have shown that clinical problem solving is not a general trait but is specific to particular content domains (e.g. diabetes, coma, emotional problems). Physicians have abilities which vary from specialty, and indeed, from case to case within a specialty (Elstein, Shulman & Spreafco, 1978; Barrows et al., 1978; Swanson et al., 1982). This is reflected psychometrically in low intercorrelations between scores on different cases and poor reliability of composite test scores. Most correlations are from 0.1 to 0.4 (Marshall, 1977; Berner, Bligh & Guerin, 1977; Mast et al., 1982) indicating that performance on one case is a poor predictor of performance on other cases. Therefore, a large number of cases are needed for acceptable levels of intercase reliability. If intercase correlations average 0.2, and a total test reliability of 0.8 is desired (a minimum value for assessment of individual performance), then about 16 cases are required. Most tests using written or computer based simulations are significantly shorter than is psychometrically reasonable, and yield scores with marginal reliability.

Validity. Construct validity studies have examined the relationship between different subscores on the simulation using medical problem solving theories to guide interpretation (Juhl, Noe & Nereberg, 1979; Berner, Bligh & Guerin, 1977; Harasym et al., 1980). Results
have varied since each uses somewhat different subjects, measures, and analytic procedures, but most use factor analysis as a "discovery procedure" to identify underlying dimensions of problem solving performance. Investigators have found 1) a single, weak underlying factor, 2) two underlying factors typically labeled data-gathering and decision-making, or 3) no interpretable factor. Exploratory factor analysis may be inappropriate, given usually small sample sizes and low intercase correlations which affect factor stability.

Correlations between written or computer simulation performance and other written clinical competence measures, particularly multiple choice test scores, are low to moderate ranging from 0.2 to 0.5 (Case, 1981; Langdon et al., 1978; McGuire & Babbott, 1967). As with oral exams, this is interpreted to mean that simulations measure something different, but given the low intercase reliability of the typical simulation-based exam, low intercorrelations may simply reflect attenuation due to measurement error. A study of the daylong certifying exam in internal medicine using three years of data from five thousand subjects per year (Norcini, Swanson & Webster, 1983) found correlations ranging from 0.71 to 0.76 between the multiple-choice and simulation components and true correlations greater than 0.9, strongly supporting the attenuation interpretation. Thus, no conclusions regarding construct validity can be drawn because of wide variations in results and attenuation of correlations. More research and better analytic procedures are badly needed.

Criterion-referenced validity has been investigated using the relative performance of groups which should differ in ability and correlations with some performance-based measure of ability, usually ratings from clinical instructors. Written simulations can distinguish performance of medical students and residents at different training stages (McGuire & Babbott, 1967; Schumacher, 1974; Schumacher et al., 1974; National Board of Medical Examiners and American Board of Internal Medicine, 1981; Grosso & Webster, 1983). Given that residents have a year or two of additional education and clinical experience, this is not surprising. A well-constructed twenty-item multiple choice test could probably demonstrate similar differences so only mild support for validity is implied by these results.

Studies relating simulation scores to real-world criterion measures of clinical ability such as clinical instructor ratings have found correlations in the 0.1 to 0.4 range (Langdon et al., 1978; Norcini, Swanson & Webster, 1983; Schumacher, Burg & Taylor, 1974), approximating correlations between multiple-choice tests and similar real world criterion measures. Studies of incremental validity—whether predictions of the criterion improve when written simulations are added to a multiple-choice test battery—typically show modest improvements (Schumacher, 1974; Langdon et al., 1979; Norcini, Swanson & Webster, 1983), suggesting that multiple choice tests and simulations measure similar competencies. Again, the low reliability of simulation-based exams make these results difficult to interpret and lack of a good criterion measure of real world clinical performance obscures interpretation further, particularly since instructor ratings are typically plagued by reliability problems.

Manikins have an important role in teaching clinical skills, but there has been only limited rigorous evaluation of their validity and effectiveness.

It is often assumed that written and computer based simulations possess a high degree of content validity because of their high fidelity. Although the evidence is not overwhelming a number of investigators have asserted that the similarity of the problem-solving tasks in patient care and written simulations is sufficiently great that the simulations are content valid (McGuire & Babbott, 1967; Swanson et al., 1982). Evidence of consistency of behavior on similar real or simulated cases is, indeed, relevant to an assessment of content validity, but equating content validity with fidelity is incorrect. For a test to be content valid, it is important that the simulated cases be sampled from the real world clinical domain in such a way that they are representative. Judgment enters into the choice and poor case selections can easily be made. Also, the discouraging intercase reliability findings reviewed previously indicate that a large sample of cases is necessary to develop an accurate ability estimate. Tests using small numbers of cases cannot be content valid, because they do not sample clinical situations sufficiently.

Feasibility. The widespread use of written simulations by specialty boards, licensing bodies, and medical schools attests to their feasibility. They are fairly inexpensive to design, can be administered to large numbers of examinees and machine-scored at relatively low cost. Conversely, computer simulations have not proceeded beyond pilot testing or small scale implementation despite years of development. Estimates of development costs range from a few thousand to $50,000 per problem, and costs of administration have been estimated at $50 to $1,000 per examinee (Diamond & Weiler, 1974; Senator, 1976).

It must, however, be recognized that two factors may alter these figures. First, computer equipment costs have decreased dramatically since the mid-1970s and are not reflected in these estimates, and second, computer simu-
measure a distinct and central component of competence; however, such evidence is lacking. The value of computer simulations is more problematic, since early developments apparently possessed all of the problems of written simulations and the additional handicap of high development costs, but new developments such as videodisc technology may represent a real advantage in evaluation which remains to be demonstrated.

Conclusions

Content specificity seems to be a fundamental problem in assessing clinical problem solving ability with simulations. Even with very high fidelity computer simulations, it can be anticipated that correlations between performance on different cases will be low. This seems to be a characteristic of problem solving in real clinical life, so it can be expected on simulations as well. It is necessary to use large numbers of cases to adequately assess problem solving ability. Measuring performance on a single case with more fidelity and accuracy can well result in a less valid test because more testing time is usually required. Certainly, short simulation-based exams, consisting of only a few cases, should not be used for either testing or research purposes. One line of inquiry which may reduce the demands for a large number of cases in testing situations is directed at a better understanding of the characteristics of physician knowledge about medical problems. Bordage & Allen (1982) have made an important step in this direction, and other recent papers address the issue (Borner, 1984).

Despite shortcomings, simulations have a significant role to play in health sciences curricula. Live simulations are a powerful method to teach data gathering, interpersonal skills and technical skills, and can be usefully employed to evaluate these skills provided sufficient numbers of patient problems are employed. Some types of manikins have been shown to be useful for both teaching and evaluation specific technical or procedural skills, although further research is necessary to determine those characteristics of the simulation which result in effective transfer to the clinical setting. Finally, computer simulations have had an uncertain role in the past, but the evolution of new technology holds great promise for the future.

References


Langdon, L.O., Maskauskas, J.A., Nordin, J.J., & Webster, G.D. ABIN MERIT: Recertification na-
nations, on research in medical education, Washington, DC.


1985, VOL. 8, NO. 1
An Instructional System as Change Agent

Hannelore Vanderschmidt
Associate Professor
Ascher J. Segall
Professor
Center for Educational Development in Health
Boston University
67 Bay State Road
Boston, MA 02215

Since its inception in the mid-1970s, the Project on Curriculum Development in Preventive Medicine has sought to change teaching of preventive medicine in professional and continuing education. Project development staff have learned lessons that go beyond the domain of educational technology. They have had to solve questions of authorship, academic responsibility, and limitations in the state of preventive health care delivery systems. But most of all, they have had to cope with problems inherent in institutional change.

In this article, background information about preventive health services is discussed first to provide a context for understanding the project and its development, but its main focus is on three aspects of the project's development process that are distinctive: 1) the identification of skills to be taught that are presently seldom employed in professional practice, and for which knowledge is limited or consensus on acceptable procedure is lacking; 2) the use and management of large numbers of experts from many disciplines to author instruction; and 3) dissemination strategies involving material design, "entrepreneurship" and retrieval systems for material access.

Background

Preventive health services focus on risk reduction and health promotion. For example, the risk of cardiovascular disease is reduced by smoking cessation, weight/cholesterol reduction, exercise, and hypertension control. Other preventable conditions include cancer of the lung, cirrhosis of the liver, many infectious diseases, and occupational disorders. Health promotion includes proper nutrition, exercise, lifestyle enhancement, and stress reduction. Some examples of preventive health services are immunization, blood pressure measurement, and pap smear analysis.

At the turn of the century, clinical preventive medicine was widely practiced in this country for infectious diseases like typhoid, tuberculosis, diphtheria, cholera, and tetanus. The fight against infectious diseases was aided by the availability of effective countermeasures, including immunization. With the decline of infectious disease, longevity increased and chronic diseases such as heart disease, cancer, and stroke became leading causes of death. Since infectious disease was no longer a problem and chronic disease was considered part of the natural aging process, the teaching of clinical prevention declined.

However, several decades of epidemiologic studies have demonstrated that primary and secondary chronic disease prevention measures are effective (United States Department of Health, Education, and Welfare, 1979). Primary prevention is any intervention with the purpose of reducing risk of disease occurrence, while secondary prevention is any intervention with the purpose of detecting asymptomatic remediable disease or reducing risk of disease recurrence (Stokes et al., 1982). Even though a scientific basis for clinical prevention now exists, the associated knowledge and skills are not emphasized in medical schools, nursing programs or primary care residencies. Only three percent of the medical school curriculum is currently devoted to preventive medicine (Bishop, 1983; AAMC, 1983; AMA, 1983; NLN, 1981).

To strengthen preventive medicine teaching, the Project on Curriculum Development in Preventive Medicine was undertaken by the Association of Teachers of Preventive Medicine and the Center for Educational Development in Health with the Kellogg Foundation support. The curriculum resulting from the project, entitled Health Maintenance in Clinical Practice, consists of three modules. Its contents and components are summarized in Figure 1. A systematic, competency-based approach was used to design the curriculum (Segall et al., 1975.)

Defining Outcomes

The first development task was to delineate the scope of professional performance in an area where skills are new, evolving, and seldom practiced. This was accomplished by first listing both generic responsibilities (those cutting across specific conditions or risks) and categorical responsibilities (those limited to a specific condition or risk). Outlining these responsibilities made it possible to identify a process for planning, implementing, and evaluating preventive services. An algorithm depicting the process, shown in Figure 2, has been found applicable to a range of services, including those associated with the cardiovascular risk factors of smoking, hypertension, and nutrition, as well as occupational health, breast cancer, alcohol abuse, immunization, and other responsibility categories. Developing a general procedure for performing professional responsibilities and several specific categorical examples was useful for conceptualizing complex, interrelated skills, and providing a framework for identifying these skills and organizing materials.

To verify responsibilities of physicians and nurses, two modified Delphi surveys (Linstone & Turoff, 1975) were conducted: the first with 50 practicing physicians; the second with 35 nurses. They were either engaged in clinical practice or were teaching clinical prevention. Panels of experts, composed
of physicians and nurses serving on the project's advisory and steering committees, also were consulted. Both the Delphi results and the expert panel suggestions were used to determine the scope of the curriculum and the three major areas of module development.

Practitioners could respond to the Delphi artifically only in areas where they had some experience. Further, some performance discrepancies were identified that were not due to lack of skill or knowledge. For example, provision of mammograms to women over 40 is now recommended by the American Cancer Society (1980) but infrequently practiced because of organizational policy or lack of compensation by third party payers such as health insurance and maintenance organizations. Because of the inability of any one group of respondents to see the "total picture," verifying assumptions about professional performance was accomplished using complementary groups. Moreover, instructional materials were developed that addressed skill/knowledge deficits, not policy issues beyond the control of the developers and their sponsors.

In deriving educational objectives from responsibilities there was a major behavioral/attitudinal component. Since some people smoke, eat high cholesterol foods, drink too much alcohol, and refrain from exercise because this helps them cope with everyday problems and frustrations, behavior modification skills had to be taught. Experiences of colleagues in behavioral medicine, especially those involved in major intervention trials such as the Multiple Risk Factor Intervention Trial, MRFIT, (Davidson et al., 1980; Benfari, 1981) were used to develop the training objectives. Because of the difficulties inherent in effecting behavior modification, some skills were defined for which there were few well established techniques compared to other aspects of medicine. Altering patient attitudes, however, is a critical part of preventive care, and it was essential to teach these skills and the behavior modification approaches that are available even though knowledge in the area is relatively limited. The credibility of the curriculum would be compromised if difficult-to-teach but important subjects were avoided.

Authoring

When curriculum projects are large scale and multidisciplinary, it is often necessary to enlist the aid of many authors — professional and scientific writers, educational technologists, and subject experts — with varied teaching backgrounds. In this project, authors who knew the subject matter, who could apply educational technology principles, and who were experienced teachers performed best. They did not have to read themselves into familiarity with the subject, and they had a clear sense about how to communicate content and how to create resources that instructors could use. Finally, most successful work was done by authors who worked throughout the development project, insuring continuity through different iterations of materials tryout and revision.

In addition to authors, other subject experts were needed to set objectives, identify readings and review draft materials. An overall editor checked format consistency, standardized bibliographies, developed cross references, and caught omissions and redundancy. Outside reviewers insured content accuracy, while student readers identified what was confusing, wordy, or difficult. Where disagreement arose, project directors refereed.
Testing and Packaging

Initial tests in nine medical schools and a school of osteopathic medicine revealed that the user group required more than an instructor's manuals, the only resource originally developed. They asked for cases/exercises, up-to-date readings, and a job aid (Lineberry & Bullock, 1980) for performing preventive procedures. As a result, a Learning Resources book, an Anthology of readings, and a Guidebook were added. Instructors required precise instructions as well as model answers to study questions so they could use less trained teaching staff not conversant with the details of a given preventive strategy. Because of the rapidly changing nature of preventive practices, senior faculty also welcomed the additional materials.

In its early revisions, the entire curriculum system was presented in a loose-leaf notebook with tabs for separate components. The volume was found "heavy and forbidding." Users took out what they wanted from the notebook and left the rest. Some used articles only, in order to bring their standard course in preventive medicine up to date. Some used the exercises and case studies only, thus introducing interactive teaching into what would otherwise have been a lecture course. Some users adopted all offerings but changed the sequence.

Materials were repackaged accordingly, placed in twelve thinner books and bound and formatted for maximum flexibility. Further subdivision is likely. In some other settings, instructors may follow procedures slavishly and demand tight structure—indeed, it may be mandated by an organization. Such an approach would be rejected by this target instructor group working in varied institutional settings.

When making revisions, designers work down from the most general design specification (the job description and instructional objectives) to the specific aspects of instructional design. This top down approach, moving from the general to the specific, is a logical way of designing instruction and revising it. It is also cost effective, because testing the course when it is at a fairly high level of generality can indicate what additional components are needed. Initially creating many specific materials may result in a curriculum with many components that are redundant or unnecessary, and deletions are often more difficult to identify than additions.

Dissemination

In disseminating instructional materials to groups such as physicians, nurses, nutritionists, and physician assistants, the instructional designer must be aware of the political and social contexts of education. Some users will be sensitive to the role of women or to that of non-physicians as health care providers. Examples cited should reflect a range of user groups. Titles and terms included should be acceptable for all users, based on feedback from the intended users. Plans for dissemination should be considered early in designing and packaging materials.

A problem any instructional innovation faces is how to win converts. There are many reasons given for not adopting a new instructional program:

- I'm doing it already.
- I have little time for teaching. I have a
private practice and I do research on the side.
• In order to be promoted I need research publications. Good teaching
does not help.
• I'll get around to it in a year or two.
• It was invented by you guys in the East. This is Miracle Falls, Idaho. We'll
invent our own, thank you.
• Our curriculum is already crammed full. We can't add another thing.

To overcome resistance to our pro-
gram, efforts were made to gain con-
fidence by repeated visits, phone calls,
and letters. Instructors had to be urged
on. Many used the resources only after
they got to know project staff. In-
dividuals who do use the materials are
contacted frequently. Once they have
"bought into the system," they often call
back to ask for further consultation, re-
quest additional material, and begin to
structure new courses. Once a preven-
tive health program is implemented, in-
structors often want to see for
themselves how materials can be effec-
tively used. In this project, over sixty
such inquiries have been made by in-
structors in health science institutions
throughout the world, and a network of
demonstration centers is being estab-
lished. The demonstration centers will:
• Offer courses for continuing educa-
tion of physicians and nurses in clinical
prevention.
• Introduce preventive initiatives into
residency training and graduate nursing
programs where clinical facilities are
available for delivering preventive ser-
vice. Here residents (graduate physi-
cians) and nurses can obtain "hands-on"
learning experience.
• Include the materials in clinical
clerkships and rotations of third and
fourth year medical and nursing
students, where students spend one to
several weeks at a time developing dif-
ferent specialized skills.
• Include the materials in first and sec-
ond year courses where the more didac-
tic portions of the three modules can be
taught.

This integrated, comprehensive ap-
proach to the teaching of preventive
health care should serve as a stimulus to
the would-be user.

By the time the project is complete in
1986, there will be upward of 1,000 cur-
rriculum components. Because of its
complexity, a plan has been developed
for computer retrieval of materials. The
data base and automated access system
depicted in Figure 3 will enable faculty
to pick and choose among holdings and
enhance flexibility in adoption and use
of materials. The system allows retrieval
not only by content, but by selected in-
structional design attributes.

Conclusion
What seems most attractive to the in-
structors who use the program materials
is adapting them to their own instruc-
tional setting, rather than wholesale
adaptation. Allowing adaptation engenders change in successive approx-
imations. This approach is particularly
appropriate in a time of scarce resources, as it leads to gradual, con-
cordant progress toward measurable
goals.

The project started out by applying a
systematic approach to instructional
design to the development of instruc-
tional materials in preventive medicine.
Both the approach and the products
have become more flexible. There
should be rigor in curriculum design but
other factors must be addressed to effect
institutional change. These include
developing a meaningful and credible
method of identifying new skills,
developing and packaging materials for
adaptation to local needs, establishing
retrieval tools that facilitate adaptation,
and decision-making centers to show
how material can be applied, and
building interpersonal relationships be-
tween producers and users of materials.
Large scale projects also require the
identification of appropriate par-
ticipants and coordinating their efforts
to effectively achieve desired outcomes.

References
Association of American Medical Colleges. (1983). Direct-
ory of American medical education. Washington, DC: Au-
thor.
directory of residency training programs. Chicago, IL: Au-
thor.
American Cancer Society. (1980). Guidelines for
the cancer related checkup: Recommendations
and rationale. CA: A Cancer Journal for Clin-
icians, 30(4), 294-242.
Benfari, R. (1981). The multiple risk factor inter-
vention trial, MRFIT. The model for interven-
tion. Preventive Medicine, 10, 420-442.
Bishop, F.M. (1983). Departments, sections, and
programs of preventive/community medicine
and public health in US medical schools. Journal of
Community Health, 8(4), 269-273.
Behavioral medicine: Changing health lifestyles.
instructional design library. Vol. 25. Englewood
method: Techniques and applications. Reading
MA: Addison-Wesley.
National League for Nursing. (1981). Bu-
calculator education in nursing: Key to a profes-
sional career in nursing. New York, NY: Na-
tional League for Nursing.
Segall, A.J., Vanderschmidt, H.F., Burglass, R., &
Frostman, T. (1975). Systematic course design
for the health fields. New York: John Wiley.
Stokes, J., Noren, J., & Tindell, S. (1962). Defin-
tion of terms and concepts applicable to clinical
preventive medicine. Journal of Community
Health, 8(1), 33-41.
and disease prevention. Washington, DC: USGPO.
Organizational Change and the Development of Faculty Evaluation Systems

Deborah A. Orban, Ph.D.
and
Allan J. Abedor, Ph.D.
University of Texas Health Science Center
School of Allied Health Sciences
Division of Instructional Development
G. 703 Freeman Building
Houston, TX 77025

Problems in Performance Appraisal

The evaluation of faculty performance is not uniformly conducted and, if done, usually consists of overly simplistic, quantitative general effectiveness ratings. Minimal data result for administrative decision making, the support of promotion and tenure, and professional growth and development. These inadequacies have led some institutions to adopt a more complex performance appraisal approach. The resulting faculty evaluation systems are not without problems, but they appear to provide more relevant and timely information.

Developing and implementing a performance appraisal system raises several substantive problems. One is that the decisions to be served by the evaluation are unclear. Since performance evaluations are multi-purpose by nature, the co-mingling of administrative personnel decisions and faculty development uses of the same data is common and often destructive. A second problem is relevance. To be effective, evaluation must be a continuous, systematic process related to the functions and goals of individuals and the organization. More precise definitions of job duties than typical faculty position descriptions are required. A third problem concerns systems management. To maximize probability of accurate and useful data, the evaluation program must be well received and supported.

New evaluation systems can be devised to incorporate advances in performance appraisal while remaining compatible with existing organizational structures. These systems combine characteristics of management by objective systems (Holbrook, 1975), portfolio evaluation (Berquist & Phillips, 1977), and the congruence evaluation approach (Stake, 1973). They involve developing a portfolio of documents to guide objective setting, negotiation of workload and criteria, data collection, performance rating, and review. They include descriptions of each evaluation activity, a set of instructions to use during implementation, and rating criteria. Instruments used to obtain measures for each activity may vary among programs, although standardization is desirable.

System Assumptions

There are at least five fundamental assumptions about faculty members and the purpose of performance assessments underlying system development and use.

Assumption 1: Faculty should be evaluated on their performance of activities for which they are presently employed.

Faculty should be evaluated on their productivity in areas of teaching, service, scholarly activity, and professional development according to the percentage distributions specified by their employment agreement. This is of primary importance to Allied Health educators whose workload usually differs from other academic faculties and, therefore, is frequently incompatible with typical promotion and tenure requirements. An important outcome is to bring workload and promotion requirements into close alignment.

Assumption 2: Faculty are autonomous professionals who ought to have a significant input into their evaluation process.

The grounds for evaluation should be mutually agreed upon by the faculty and the evaluator before the evaluation begins. This should include what is evaluated (activities performed), and how performance is to be evaluated (instruments and procedures used).

Assumption 3: Evaluation should accommodate individual differences in faculty as well as differences between departments.

Faculty members often possess specialized expertise which permit them to make unique contributions to the

Introduction

While most institutions of higher education support the concept of faculty evaluation, such appraisal systems are seldom well received. These negative reactions are often dismissed as resistance to accountability, but they can be based on knowledge of the many real problems associated with performance evaluation programs (Baird, Beatty & Schneier, 1982) and on policies affecting their implementation (Centra, 1980; Millman, 1981). Addressing these problems through application of organizational change theory and diffusion strategies which are participatory permit faculty to construct evaluation systems which combine their practical knowledge and experience with expert suggestions so as to fit existing empirically derived evaluation principles. The resulting evaluation systems are more likely to reflect the needs and context of the institution.

In this article, problems in faculty evaluation and performance appraisal are described. Assumptions and features of newer alternative faculty evaluation systems are presented. A strategy for developing systems involving use of committees and structured phases for faculty input based on organizational change theory is discussed. The recommendations here are based in part on a review of relevant literature and the authors' experiences devising a system for the School of Allied Health Sciences at the University of Texas—Houston.

22

JOURNAL OF INSTRUCTIONAL DEVELOPMENT
achievement of departmental goals and these should be accounted. Moreover, academic departments often are very different in distribution of teaching, research, and service responsibilities; curricular emphasis on didactic and clinical instruction, student practicums and internships; and the division of responsibilities for course development, coordination, and actual delivery of instruction. A school-wide program must accommodate between-department differences while providing within-department consistency.

Assumption 4: Faculty evaluation should be used as a means for improving faculty productivity and effectiveness as well as for determining promotion, tenure, pay increases, and/or continued appointment.

This assumption raises the thorny issue of mixing developmental and administrative uses of the same evaluation process. Addressing both can meet the needs of faculty and administrators, providing systems are designed that reward the faculty members for involvement in self-improvement activities, without requiring them to chronicle specific areas of weakness in any detail. Professional growth should be treated as a desirable pursuit expected of all faculty members and reported in a manner similar to descriptions of teaching, research, and service productivity.

Assumption 5: The developmental function of the evaluation requires supervisors to provide performance feedback based on systematically collected data presented in a frequent and timely manner.

Accurate and timely feedback is prerequisite to performance improvement. A global assessment of accomplishments on an annual basis is unlikely to be expressed in specific, performance terms which facilitate growth. In contrast, frequent, specific reviews assist in developing new skills and accomplishing negotiated goals.

System Features

The first feature of newer faculty evaluation systems is the requirement of a negotiated agreement between the faculty member and supervisor-evaluator concerning activities, evidence, and criteria for each evaluation period. The two parties meet and the faculty member’s workload for the academic year is discussed and agreed upon. Teaching, research, service, and professional development activities are proposed by the faculty member, discussed, and more fully defined. The amount of effort expected in each area, the outcomes to be achieved, and the criteria to be satisfied are determined. It is the responsibility of the supervisor to negotiate workloads consistent with departmental requirements and goals, while faculty members negotiate for workloads consistent with their professional interests.

A second feature is the integration and demarcation of developmental and administrative uses of evaluation results. A system should provide data gathering mechanisms addressing both purposes. Three mechanisms address faculty professional development. First, development activities are included as an evaluation category and are considered a professional responsibility. Second, faculty members and supervisors meet regularly (usually quarterly) to review progress, providing “coaching” opportunities for supervisors and renegotiation of any unrealistic or inappropriate performance goals. Such flexibility is crucial since duty reassignments and resource shortages occasionally occur, making it impossible for faculty to perform as negotiated. Third, the faculty member and the supervisor must specify both the individual’s and the department’s action to remediate any identified deficiencies in performance at least once annually. Administrative functions, in contrast, are satisfied during initial negotiation and annual reviews concerning workload distribution and overall performance. The combined emphasis on development and administration issues is often considered to be the most difficult aspect of performance appraisal to implement (Baird, Beatty, & Schneier, 1982).

A third feature is use of annual portfolios as performance evidence (Berquist & Phillips, 1977) which are rated on standard forms. Portfolio construction is a faculty responsibility that is cumulative and expected to be integrated with the dossier preparation required by promotion and tenure committees.

The fourth feature is a management plan which includes a set of critical events and a timeline. The evaluation period is usually defined as one academic year for administrative purposes, with quarterly meetings to accomplish developmental functions. The events and activities are depicted in Figure 2.

System Development

Faculty evaluation is often an administrative mandate imposed on employees and their supervisors. Such autocratic implementation risks user resistance and sabotage (Havelock, 1973; Rogers & Shoemaker, 1971).

Alternatives to autocratic implementation are systems based on what is known about successful organizational change. This approach treats faculty evaluation as an innovation which is adapted and diffused throughout the organization in a progressive manner.
characterized by a series of activities keyed to psychological stages of change. It encourages faculty and administration involvement in designing and developing the innovation. Faculty and administration involvement will result in a more relevant evaluation system and increased user commitment to the system that is developed.

An innovation is any “idea, practice, or object which is perceived as new by an individual” (Rogers & Shoemaker, 1971). Whether the idea is actually new or just newly discovered is unimportant. It is unfamiliarity and the inherent challenge and uncertainty “newness” represents, that change theory and diffusion models address. Although a number of schemas exist describing the psychological processes of accepting or rejecting new ideas (Zaltman, Duncan & Holbek, 1973), most suggest change originates with some awareness of a gap between a current situation and a desired situation which might be bridged by an innovation (Zaltman, Florio & Sikorski, 1977). This initial awareness stage is followed by acquisition of additional knowledge, the formation of attitudes about the innovation, a mental or an actual trial, and evaluation supporting decisions to adopt or reject. Although these change stages are not strictly linear, they have sufficient accuracy and descriptive power that strategies facilitating an individual’s progression toward change at each stage can be applied. Such strategies are essential if change is to be successful and enduring. Change strategies associated with the development of faculty evaluation systems are summarized in Figure 3.

The participatory process in developing faculty evaluation systems can be initiated with the formation of a committee to obtain direct faculty participation since committees are a common and familiar means of governance and policy setting in higher education. Faculty opinion leaders from each program in a school and instructional developers might comprise the committee. This structure creates a good working relationship between experts and clients, provides a reasonable likelihood of identifying needs, and helps insure the innovation will not only be generated, but also eventually accepted.

Initial committee deliberations should center on developing the evaluation plan and operating procedures through a consensus process. Operating procedures consistent with change theory (Zaltman, Florio & Sikorski, 1977; Havelock, 1973; Rogers & Shoemaker, 1971) allow individual faculty and program concerns to be brought forth as discussion topics and full faculty review to be solicited at key junctures. Each component of the plan can be revised to reflect these concerns and to insure “ownership.” A “sunset provision” requiring formal review after one year, helps to insure a genuine trial and protect against adopting a flawed program. Finally, mechanisms to communicate the entire plan to the faculty should be specified, including written descriptions, open discussion meetings, and assessment procedure demonstrations, as appropriate. These strategies permit the committee to make increasingly specific determinations of faculty needs, to design an innovation which reflects the organizational context, and to provide sufficient data supporting adoption/rejection decisions.

It is possible to develop plans in three general phases, each representing a key juncture requiring a written report distributed to the entire faculty for comment and approval. During the first phase, the committee specifies a set of basic assumptions concerning functions the evaluation program would serve and the nature of the work to be evaluated. During the second phase, a conceptual evaluation model is defined based on the assumptions approved. Phase three involves operationalizing the evaluation model by developing procedures and instruments. A draft of the full plan is presented to the entire faculty, and feedback obtained to fine-tune the system. Once this done, training can be provided.

Conclusion

Newer, more responsive, relevant, and equitable faculty evaluation systems are beginning to emerge. Although they share many common assumptions and features described here, how they are articulated in a given school depends on local circumstances. In a way, every institution must develop its own system, tailoring general features to specific requirements. Development is as much a change problem as a technical one. Meaningful implementation requires development approaches incorporating strategies concerning diffusion and adoption of innovation. The development process described here is predicated on the notion that change at the organizational level requires a large number of adoption decisions by individual members. There are “early,” “middle,” and “late adopters” who must be provided sufficient lead time to assess innovations. Change strategies need to be incorporated into a development approach that address their individual differences. One is building a “critical mass” of early supporters who can pro-

![Figure 2. Annual Timeline for Evaluation Activities](image-url)
provide leadership. Another is providing information about the innovation to all potential adopters, including factual data, descriptions, and actual demonstrations of the innovation at appropriate stages. Third is gradually expanding involvement to include the full adopting community during the trial stage. The innovation should be either newly designed or specifically adapted within the organization to insure suitability to group culture and operating procedures. Ownership promotes endurance.

References

Figure 3. Strategies for Developing and Implementing Faculty Evaluation System

Awareness
- Form a critical mass of opinion leaders
- Insure representation from all departments
- Adopt a problem solving-orientation
- Communicate committee charge to full faculty
- Provide factual data, descriptions, illustrations
- Inform/insure support of committee for diffusion strategies

Acquisition of Additional Knowledge
- Include ID/resource person on development team
- Engage in literature and existing evaluation program reviews
- Present reports on technical matters to committee as relevant
- Encourage members to report progress at department meetings
- Solicit input about needs/values/attitudes of full faculty

Formation of Attitude
- Distribute reports of committee progress to the full faculty
- Solicit full faculty review & input at key junctures
- Revise documents to reflect faculty comments & language

Evaluation/Trial of Innovation
- Involve full user community
- Conduct open meetings with full faculty
- Verbally present the plan using visuals to support key points
- Conduct role play demonstrations of major activities in plan
- Engage faculty in small group discussions/critiques of plan

Adopt/Reject Decision
- Include a “sunset provision” mandating formal program review
- Examine faculty identified problems & revise accordingly
- Initiate use only after overwhelming consensus is expressed

Implementation/Integration
- Maintain sufficient flexibility in plan to accomodate department & individual faculty differences
- Present an orientation & demonstration
- Rely on committee members from each department for troubleshooting
- Provide full training to users
- Include examples of accomplishment documentation
- Insure continuing verbalized support by Dean
- Establish ongoing dialogue with promotion & tenure committee

Review/Renewal
- Survey faculty & department chairmen after 6 months & 1 year
- Use third parties to conduct interviews
- Listen/engage in hallway talk with users
- Review system for possible revisions on a biannual basis
- Establish links with institutional goals wherever possible
- Communicate revisions & goals to faculty regularly
Evaluating Short-Term Training Programs

A Practical Approach

Kent I. Sheets, Ph.D.
Department of Family Practice
University of Michigan Medical School
Ann Arbor, Michigan 48109

Abstract. Short-term training programs are frequently used to implement faculty development activities in post-secondary education. However, rarely are these programs systematically evaluated to assess their impact on program participants. This article describes an evaluation framework that was designed, developed, and implemented as an evaluation study of an existing faculty development program for family physicians. The study demonstrated the effectiveness of the faculty development program and a metaevaluation of the evaluation study identified the strengths and weaknesses of the evaluation framework. The original framework was modified according to the metaevaluation results and presented for possible adoption by other instructional developers and evaluators.

Introduction

Short-term training programs are conducted regularly by schools, corporations, hospitals, businesses, churches, the military, and others. In post-secondary education, short-term training programs are frequently used in faculty development programs directed toward the improvement of instruction and teaching. This article describes an evaluation framework designed for use with short-term training programs.

Considerable time, effort, and resources have been expended on the design and implementation of short-term training programs, but little is known about their impact because they rarely are systematically evaluated (Centra, 1976; Gaff, 1979; and Littlefield, Hendrickson, Kleffner and Burns, 1979). A review of the research on improving college teaching reported more studies than expected, but of lower quality than hoped (Levinson and Menge, 1979). A subsequent review reported virtually no adequate studies of workshop impact (Menge and Levinson-Rose, 1980).

If indeed the major objective of faculty development programs is to change instructor behavior (Davis, 1979), the evidence supporting successful change is mostly based on participant satisfaction measures (Caldwell, 1981; Donnelly, Ware, Wolkon and Naftulin, 1972; and Stephens, 1981). The problem is that faculty give a generally positive assessment of their experiences, but do not teach differently (Gaff and Morstein, 1978). In addition, most evaluation efforts stop when the program ends and there are few attempts at follow-up observation of participants' behavior in actual practice.

In light of these reports, serious attempts should be undertaken to evaluate the impact of short-term training used within faculty development programs. These efforts should be designed to provide more rigorous data than the mere tabulation of participant opinions. Such an evaluation study was conducted during 1981 and 1982 at Michigan State University as part of the author's doctoral dissertation. This attempt to rigorously evaluate the impact of a faculty development program is detailed in this paper.

The evaluation approach used during the study, its strengths and weaknesses, and its applicability to other faculty development programs are presented and discussed. Although the approach was used to evaluate a single program, a number of issues are discussed based on that single evaluation study, including recommendations for further use.

The Program

The Family Medicine Faculty Development Program (FMFDP) was conducted by the Office of Medical Education Research and Development (OMERAD) at Michigan State University (MSU). The FMFDP began operation in July 1978 with support from the Bureau of Health Manpower, United States Public Health Service. It had two major objectives. One was to identify and train new physician teaching faculty for family medicine training programs; the other was to help current family medicine faculty develop or refine their teaching skills. One component of the FMFDP was a teaching fellowship offered to allopathic (M.D.) and osteopathic (D.O.) physicians who had completed or were near completion of a family medicine residency program and to family medicine physicians with one year or less of academic teaching experience. The fellowship experience began in September 1981 as a two week session at MSU. The fellows returned in January, March, and May for three additional one week sessions. The subject matter presented in these sessions focused on teaching principles and techniques used in medical schools and residency training programs. The content was presented in a variety of instructional formats including workshops, seminars, and simulations. Sample topics presented during the sessions included: elements of group development, presentation skills, principles of learning and motivation, and clinical teaching.

Fourteen fellows (thirteen males and one female) participated in the program; four were D.O.'s and ten were M.D.'s. Each fellow received a stipend to defray program participation costs.

An Evaluation Framework

The evaluation framework was designed and developed to provide a mechanism for evaluating the impact of short-term training programs on program participants. It is a set of conceptual components and guidelines based on review of the literature. Factors considered during the review and the
Table 1
Evaluation Questions Asked During Evaluation of the FMFDP

1. How satisfied were participants in the September session with the content, instructors, and activities of that session?
2. How satisfied were the program directors with the content, participants, and activities of the September session?
3. How satisfied were the supervisors with the program and the participants?
4. What was the participants' level of cognitive knowledge of the September session at the beginning of the session?
5. Was there a significant retention of the cognitive knowledge of the content of the session by the participants six months following the completion of the session?
6. Was there a significant retention of the cognitive knowledge of the content of the session by the participants six months following the completion of the session?
7. How much additional study of the content of the session was undertaken by the participants between the end of the session and the administration of the delayed posttest?
8. How did the participants perceive their own expertise in each of the content areas of the September session both before and after the session?
9. What types of skills or techniques did the participants use following the completion of the September session?
10. What types of skills or techniques did the participants expect to use in the next six months?
11. How did the supervisors perceive the participants' ability to apply the content of the session?
12. How did the participants rate their own performance in a series of three completed simulations or presentations?
13. How did the participants rate their expected performance in a repeat of the series of three simulations or presentations?
14. How well did the participants utilize the content of the session related specifically to presentation skills in two different presentations?
15. How did the program directors perceive the participants' presentation skills?
16. Did the supervisors perceive any change in the participants' teaching due to the session?
17. Did the participants perceive any change in their own role or function in their home institutions due to the session?
18. Did the supervisors perceive any change in the participants' role or function in their home institutions due to the session?

The framework's design included the importance of assessing the program impact, the focus on providing information to decision makers, and the need for flexibility. One assumption was that evaluators would not have access to participants before training starts and, therefore, data could be gathered only during and after the program.

There are five major components of the evaluation framework: type of data collected, who or what is assessed, data sources, data gathering methods, and questions.

Type of Data Gathered
The three types of data to be collected when using the framework are:
1. Reaction (satisfaction) data.
2. Cognitive (learning) data.
3. Behavioral (performance) data.
Most research addresses only reaction data with some additional self-reporting or cognitive or behavioral change, but objective measures of cognitive or behavioral change are seldom used.

Who or What is Assessed
One of the purposes of the evaluation framework is to assess the impact of a short-term training program on its participants. This component is concerned with identifying who or what is assessed in order to determine program impact. For example, the program content is assessed, as are the participants and the program faculty.

Sources of Data
There are numerous potential data sources and, indeed, all those in a position to comment on participant changes should be considered. The participants, program faculty, and the participants' supervisors should at least be consulted. Other possible data sources include the participants' subordinates, peers, students, clients, family members, or others with whom the participants interact. If the program teaches the creation of certain products or materials, it is also possible to use those developed by the participants as data sources. Multiple sources of information are recommended when conducting evaluation studies (Cronbach, et al., 1980; Patton, 1980).

Methods of Gathering Data
Data gathering methods may be quantitative, qualitative, or both and need not be the same for each data source. Appropriate methods have been suggested by several authors (Baron & Baron, 1980; Bryk, 1978; Cronbach, et al., 1980; Patton, 1980; and Posavac & Carey, 1980) and include, but are not limited to, the following: interviews, questionnaires, tests, direct observation (live, videotapes, films, audiotapes), and participant and staff self-reports.

Evaluation Questions
General evaluation questions are part of the framework as they are in other evaluation approaches (e.g., Grotelueschen, 1980). The eighteen evaluation questions developed for the evaluation of the FMFDP are presented in Table 1.

The five components of the evaluation framework form a grid, Figure 1, designed for use as a guide when developing specific evaluation plans. Information that is placed in the cells of the grid becomes the basis for evaluation plans like the one used in this study.

The FMFDP and the fourteen participants were assessed during the study. Data sources included the participants, their supervisors, and the directors of the FMFDP. Data gathering methods included questionnaires, interviews, debriefings, tests, and videotapes. The methods addressed the evaluation questions in Table 1.

Instruments
Several different instruments were used during the evaluation study. Some of these instruments were already being used by the FMFDP staff. Other instruments were developed and used for the first time during the study. The five types of instruments included: End-of-Week Evaluation Forms, Pretest, Posttest, Delayed Posttest, Videotape Rating Scale, Interview Protocols, and Final Debriefing Questionnaire.

End-of-Week Evaluation Forms
These were designed and administered...
by program directors at the end of both the first and second weeks. They focused on logistics, instructors, facilities, curriculum, and other programmatic elements and measured participant satisfaction and their self-reports of competence on selected topics taught during the session. Likert-scale and open-ended items were included on these instruments.

Pretest, Posttest, Delayed Posttest
A pretest, posttest, and delayed posttest were administered to assess entry level cognitive knowledge, changes in this knowledge after the session, and its retention six months later. The tests were developed using items submitted by FMFDP faculty. Only the session segments directly concerned with skills, techniques, and theories related to teaching and learning were tested. The same 40-item short-answer essay test was administered as a pretest, posttest, and delayed posttest. All fourteen fellows completed the tests.

All three sets of tests were scored independently by trained raters. The scores showed sufficient inter-rater reliability (pretest: .77; posttest: .83; and delayed posttest: .81) that ratings could be averaged into a single score for each participant.

Videotape Rating Scale
The participants were required to give several presentations to the program directors and fellows on skills taught during the session. These presentations were videotaped allowing participants to evaluate themselves. A 16-item rating scale was developed for two of the presentations given by each participant. Each item was rated using a Likert-type scale.

Two trained raters scored the tapes according to criteria established during the training session. Due to unequal length, only the first ten minutes of each presentation were rated. The inter-rater reliability coefficients were not as high as those coefficients obtained from the written tests (January presentation: .48; May presentation: .59), but were still of sufficient reliability to be averaged.

Interview Protocols
To gather additional information a series of interview protocols was developed. Participants, their supervisors, and two program directors were interviewed. Open-ended and Likert-type items were used on the protocols.

Final Debriefing Questionnaire
As part of the regular evaluation system of the FMFDP, a final debriefing was conducted by the program directors. In addition to the open-ended written questions included on the questionnaire administered during this debriefing session, a question concerning the evaluation activities was inserted to ascertain the fellow's reactions to the various evaluation procedures. After the fellows completed the questionnaire, a discussion was conducted by the program directors.

Analysis Procedures
Quantitative data analysis was required for the test data, rating of presentations, and quantitative items on the interview protocols and end-of-week evaluations. The rest of the data was qualitative, notably the open-ended comments made during interviews, end-of-week evaluations, and the final debriefing.

For those items which were quantitative, frequencies were determined and where appropriate, descriptive statistics were calculated. For the ratings of the two videotaped presentations, descriptive statistics were computed for the individual items, each rater, and the average of the two raters. Descriptive statistics were computed for the three tests and a univariate analysis of variance (ANOVA) was conducted comparing the results of both the pretest and posttest with those of the delayed posttest.

Qualitative analysis of the open-ended comments was conducted by reading and grouping similar comments for each question and instrument. Comments made two or more times were reported.

Metaevaluation of the Evaluation Study
A metaevaluation was designed and conducted to answer the following questions about the evaluation:
1. What specific problems were encountered during the evaluation study?
2. Was the evaluation framework practical in its use of resources?
3. Was the evaluation framework useful in providing information to the decision makers?
4. Were the methods and instruments used during the evaluation study technically adequate?
5. Were the methods and instruments used during the evaluation study conducted in an ethical manner?

The metaevaluation examined the quality of the process and product of the evaluation study. Three different metaevaluation activities were conducted, self-report, program director interview, and analysis of evaluation procedures.

The evaluator prepared a self-report that addressed the problems encountered, including difficulties related to the development, administration, scoring, and analysis of evaluation instruments. The self-report provided a problem overview and answered the first question.

The second stage of the metaevaluation was approached in two steps. The first step was to identify evaluation standards related to each of the metaevaluation questions. The source of the standards used was the 1981 publication, Standards for Evaluations of Educational Programs, Projects, and Materials. After one or more standards were identified for each metaevaluation question, specific questions were formulated to address the concepts within each of the standards. The specific questions were then arranged as an open-ended questionnaire.

Two directors of the FMFDP were
terviewed by telephone to answer the final four questions. Additional questions not specifically related to the questions, yet concerned with the effectiveness of the evaluation, were also asked during the interview.

The final stage of the metaevaluation consisted of an analysis of the evaluation procedures used. With the assistance of the program directors and another individual experienced in conducting and evaluating short-term training programs, the evaluator identified five effectiveness factors (Table 2).

Each of the procedures was rated by the evaluator and program directors independently in relation to each of the five factors. A scale from one to three was used to rate each procedure, with one meaning low, two medium, and three high. The procedures were rated according to their actual performance during the study.

Results

The results of the evaluation study were extremely encouraging and indicated that the two-week session was successful and had an impact. The strengths and weaknesses of the content, instructors, and activities were identified and mostly agreed upon by participants and program directors. The supervisor reactions were also favorable although their comments were limited. Moderate to high satisfaction was reported by the fellows, program directors, and supervisors.

Among the results of the cognitive measures, there was an apparent relationship between participant test performance, self-report of expertise, and self-report of handout use. There was also evidence that suggested this relationship extended to the degree of satisfaction with the topics.

There was a meaningful change in participant cognitive knowledge. The ANOVA results indicated that this change occurred during the first two weeks and was stable six months later.

Behavioral data indicated the participants were using many skills and techniques taught. Support for this finding was provided by data originating from all three sources of information and from the different data gathering methods.

Across the three types of data there were some notable trends that deserve identification. The practical, concrete, skill-oriented presentations were most highly enjoyed and the content most used. In contrast, a session on learning theory was viewed as theoretical and irrelevant.

In most circumstances, the information provided by participants was consistent to the point of redundancy. There was sufficient reason to believe the data provided by the participants on the reaction measure and self-reports were reliable because they were supported by the supervisors’ and program directors’ data. The data provided by the participants, program directors, and supervisors were also in agreement in most circumstances.

There was an apparent relationship between the performance of the participants on the three cognitive tests and the videotape ratings. Those individuals ranked at the top of the test results were likely to be highly ranked for the videotaped presentations. Similar results were found for those individuals who scored poorly on the cognitive tests. Scores and ranks for each of the fellows are presented in Table 3.

Results of Metaevaluation

Table 4 contains the five metaevaluation questions and a brief answer to each.

The additional data gathered during the metaevaluation helped identify those procedures and data that were most useful to the program directors and the three preferred procedures were the end-of-week evaluations, final debriefings, and videotape ratings. The cognitive tests were least preferred due to their low validity and poor ability to discriminate. The interviews with the fellows and the supervisors were also rated low, primarily due to the time and cost associated with conducting the interviews.

Program directors identified behavioral data as being more useful to them than cognitive and reaction data. However, the program directors indicated that a strength of the evaluation study was that reaction, cognitive, and behavioral data were all collected using several methods. Much of the information collected was redundant and, at times, overwhelming.

The redundancy of information was noteworthy in that the reliability and validity of the data gathered during the end-of-week evaluations were strengthened by the data collected by the evaluator during later interviews and were also notable in terms of their relationship with other study outcomes. There was a relationship between participant reactions to topics and presentations and their subsequent behavior and performance.

The terms “relevant” and “applicable” were frequently used by the fellows in their comments. The presentations the fellows enjoyed and perceived as most relevant or applicable to their present or future activities were also those on which they had the highest cognitive test scores. The reported use of handouts

<table>
<thead>
<tr>
<th>Table 2 Evaluation Factors</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Direct utility of information for decision making</td>
<td>Ease of applying information to decision making.</td>
</tr>
<tr>
<td>2. Time efficiency</td>
<td>Time required for developing, administering, scoring, and analyzing procedures and data.</td>
</tr>
<tr>
<td>3. Resource efficiency</td>
<td>Personnel, materials, and equipment required for developing, administering, scoring, and analyzing procedures and data.</td>
</tr>
<tr>
<td>4. Credibility</td>
<td>Reliability and validity of procedures and data.</td>
</tr>
<tr>
<td>5. Data manageability</td>
<td>Ease of data representation, summarizing, and analysis.</td>
</tr>
</tbody>
</table>
Table 3
Composite Scores and Ranks on Tests and Videotapes

<table>
<thead>
<tr>
<th>Fellow</th>
<th>Pretest*</th>
<th>Rank</th>
<th>Posttest 1*</th>
<th>Rank</th>
<th>Posttest 2*</th>
<th>Rank</th>
<th>Video 1**</th>
<th>Rank</th>
<th>Video 2**</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23.0</td>
<td>14</td>
<td>30.0</td>
<td>11</td>
<td>34.0</td>
<td>14</td>
<td>21.0</td>
<td>13</td>
<td>40.0</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>45.0</td>
<td>10</td>
<td>46.5</td>
<td>14</td>
<td>46.0</td>
<td>12</td>
<td>41.0</td>
<td>8</td>
<td>52.0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>48.0</td>
<td>6</td>
<td>69.5</td>
<td>8</td>
<td>61.5</td>
<td>7</td>
<td>48.5</td>
<td>7</td>
<td>41.5</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>39.0</td>
<td>13</td>
<td>73.0</td>
<td>6</td>
<td>75.0</td>
<td>1</td>
<td>41.5</td>
<td>6</td>
<td>34.5</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>52.0</td>
<td>2</td>
<td>81.5</td>
<td>2</td>
<td>73.0</td>
<td>2</td>
<td>41.5</td>
<td>6</td>
<td>52.0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>39.5</td>
<td>12</td>
<td>79.5</td>
<td>3</td>
<td>63.0</td>
<td>6</td>
<td>32.5</td>
<td>12</td>
<td>25.0</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>45.5</td>
<td>8</td>
<td>57.0</td>
<td>12</td>
<td>56.5</td>
<td>11</td>
<td>39.5</td>
<td>10</td>
<td>29.5</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>31.0</td>
<td>3</td>
<td>78.5</td>
<td>4</td>
<td>61.0</td>
<td>9</td>
<td>36.0</td>
<td>11</td>
<td>42.5</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>46.0</td>
<td>7</td>
<td>72.0</td>
<td>2</td>
<td>66.0</td>
<td>4</td>
<td>48.0</td>
<td>3</td>
<td>44.0</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>41.0</td>
<td>10</td>
<td>68.0</td>
<td>9</td>
<td>61.0</td>
<td>8</td>
<td>65.5</td>
<td>4</td>
<td>40.0</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>59.0</td>
<td>1</td>
<td>63.5</td>
<td>1</td>
<td>65.5</td>
<td>5</td>
<td>51.0</td>
<td>1</td>
<td>47.0</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>64.0</td>
<td>9</td>
<td>59.5</td>
<td>10</td>
<td>55.0</td>
<td>12</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>40.0</td>
</tr>
<tr>
<td>13</td>
<td>46.5</td>
<td>6</td>
<td>57.0</td>
<td>12</td>
<td>57.5</td>
<td>3</td>
<td>41.0</td>
<td>8</td>
<td>48.0</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>33.0</td>
<td>3</td>
<td>74.5</td>
<td>5</td>
<td>57.5</td>
<td>10</td>
<td>42.5</td>
<td>3</td>
<td>47.0</td>
<td>4</td>
</tr>
</tbody>
</table>

*120 points possible  **80 points possible  ***did not participate

from these presentations and the reported use of skills and techniques taught were also highest.

Conclusions

The strengths and weaknesses of the evaluation framework for short-term training programs are summarized in Table 3.

Since a major focus of the evaluation framework is on determining impact, much data collection must be conducted after program completion. Data were collected as late as eight months following the session. Although no timelines are prescribed by the framework, follow-up has to be delayed to allow participants time to integrate new skills and techniques into their daily routines. Final evaluation results and their use for making programmatic changes are delayed accordingly.

There is no conclusive evidence concerning the necessity of collecting all three types of data. While the reaction and behavioral data were easily used by the program directors, it was difficult for them to use the cognitive test data because of its low validity and poor ability to discriminate, and its utility could not be determined.

The lack of explicit criteria for the cognitive test and videotape rating results constitutes a weakness in the evaluation when results are examined by individuals external to the program. Without explicit criteria and cutoff scores, percentage scores have little or no meaning to other individuals examining the results of the evaluation; a weakness in the procedures used, not in the evaluation framework.

Table 4
Summary of Responses to Metaevaluation

Metaevaluation Question  Response
1. What specific problems were encountered during the evaluation?  The major problems related to the collection of behavioral data and to the development, administration, and validation of the cognitive tests.
2. Was the evaluation framework practical in its use of resources?  Yes, the program directors felt justified in committing the resources required to conduct the field test.
3. Was the evaluation framework useful in providing information to the decision-makers?  Yes, the data were comprehensive and confirmed the program directors' subjective assessments of the program's quality and impact.
4. Were the methods and instruments used during the field test of the evaluation framework technically adequate?  Yes, with the exception of the cognitive tests. There were reasons to question the validity of the cognitive test results.
5. Were the methods and instruments used during the field test of the evaluation framework conducted in an ethical manner?  Yes, the methods and instruments were conducted in an ethical manner. The evaluator was candid in his interactions with people during the field test.
The framework was designed specifically for one program type, short-term training, and was used with a program for training family physicians interested in academic medicine. To fully assess the value of the evaluation framework, it should be used with short-term training programs of varying length, content, and audience.

The evaluation framework might be as effective or more effective with programs that are not short-term. No conclusions can be made based solely on the results of the evaluation and metaevaluation.

Implications for Educational Practice

The ultimate value of this study is based on the products, processes, and procedures which can be used by evaluators, educators, trainers, administrators, and other individuals responsible for short-term training. The procedures and instruments used during this study could be adapted and modified for use in a number of situations. The conceptual approach to short-term training presented throughout the paper could be useful to practitioners.

The primary purpose of the study was to develop an evaluation approach to assess the impact of short-term training programs. The evaluation framework was designed, developed, and field tested to serve that purpose. One particular short-term training program, the September 1981 session of the FMFDP, was evaluated using an evaluation design based on the evaluation framework. The results of the evaluation and metaevaluation were presented with the expectation that the results would be of interest and value to potential users of the framework. Several conditions related to the FMFDP require discussion because these conditions limit the general applicability of the results of the evaluation. First, the group of fourteen fellows were paid participants using release time to participate in the fellowship. It is more often the case that the participants in short-term training programs or the participants organizations pay to attend such programs. Frequently, participants must use weekends, evening hours, or vacation time to attend.

Second, the two week segment evaluated during the study was part of a longer, continuing relationship between the fellows and the program directors. This continuing relationship facilitated certain data collection activities that might have been more difficult otherwise.

The framework used to collect the evaluation data was designed eclectically, an approach recommended by Baron & Baron (1980), Patton (1980), and Steele (1973). The implication of their suggestions to follow an eclectic approach to evaluation was that the existing models were not functioning satisfactorily. The framework was operationalized and the evaluation was successful. The question of whether the evaluation framework is superior to other approaches remains unanswered and will remain unanswered until the evaluation framework is used by others.

To fully assess the value of the evaluation framework, additional short-term training programs in other settings with different content, length, and populations must be evaluated. Only then can more definite conclusions be drawn. A revised matrix of the evaluation framework is outlined in Table 6. The revised matrix incorporates the major findings of this study into its design and offers a prescriptive version of the evaluation framework.

References


### Table 6
Revised Matrix of the Evaluation Framework

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Assessment of Data</th>
<th>Source of Data</th>
<th>Method of Gathering Data</th>
<th>Evaluation Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaction</td>
<td>STP</td>
<td>P</td>
<td>EOW, FD</td>
<td>How satisfied with the program were the participants?</td>
</tr>
<tr>
<td>(Satisfaction)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive (Learning)</td>
<td>P</td>
<td>P</td>
<td>TESTS</td>
<td>How much of the content of the program did the participants learn and retain?</td>
</tr>
<tr>
<td>Behavioral (Performance)</td>
<td>P</td>
<td>P</td>
<td>S-R</td>
<td>How did the participants perceive their own learning and retention of the content of the program?</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>P</td>
<td>DO, VT</td>
<td>How well did the participants apply to the content of the program in simulated or actual performance settings?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S-R</td>
<td>How did the participants perceive their ability to apply the content of the program in simulated or actual performance settings?</td>
</tr>
</tbody>
</table>

**Key:**
- STP—Short-term training program
- EOW—End-of-week evaluations
- S-R—Self-reports
- P—Participants
- DO—Final debriefing
- VT—Videotape of participant performance in simulated or actual setting

---


Patterns of Change in Offices of Research in Medical Education: 1979-1984

Paul L. Grover, Ph.D.
Douglas U. Smith, Ph.D.
and
Frank Schimpfhauser, Ph.D.

Abstract. Directors of offices of research in medical education were surveyed in July of 1983 in order to assess changes within their offices occurring since a similar survey conducted in 1979. The goals, organizational relationships, effort devoted to various educationally-related activities, and current research endeavors were obtained from 39 offices. Descriptive statistics are presented which summarize the responses to the survey, including a detailed content analysis of 175 reported active research projects. Comparison data for 24 offices of research in medical education participating in both the 1979 and 1983 studies are also presented. Findings indicate that despite only modest increases in budget since 1979, offices have expanded both their research efforts as well as their involvement in faculty development. Other trends are discussed and compared to the findings of a survey of instructional improvement centers in higher education.

Perspective
The first office of research in medical education in a U.S. college of medicine was established nearly twenty-five years ago and by 1977 offices were reported operating in over half of the 121 medical schools in the United States and Canada (Miller, 1977). In general, these offices provide the educational support necessary to allow medical schools to train more physicians, while maintaining high educational standards. Moreover, they have sought to coalesce physician, educator, and administrator efforts in such diverse areas as (a) educational and institutional research, (b) curriculum development and evaluation, (c) faculty development, and (d) development and implementation of grant proposals for curricular improvement. The structure, function, and staffing patterns of the offices have usually reflected these roles.

However, recent national studies (American Association of Medical Colleges, 1983; American Medical Association, 1982) indicate increasing the number of graduating physicians is less crucial than improving the quality of learning and instruction, increasing faculty teaching commitment, and expanding the role of undergraduate medical education in the physician's professional development. These changes may be difficult because new economic conditions and governmental priorities threaten the viability and direction of medical education by diverting faculty attention away from teaching and other educationally related activities (Petersdorf, 1983). Pressures on the clinical faculty to generate income through patient care and basic science faculty to produce more grant proposals simply to maintain existing support reduce time for curricular planning and evaluation.

Since these forces may also have influenced the form, funding, and function of medical education research, the Directors of Research in Medical Education voted in 1983 to survey its membership. This paper describes the results of that survey, compares these findings to those of an earlier survey in 1979, and discusses trends in relation to those found in another recent survey of instructional improvement centers in higher education.

Methods
A structured survey instrument was developed based on earlier surveys (Schimpfhauser, 1980; Miller, 1980) and was mailed to the 53 member offices listed in the 1983 Directors of Research in Medical Education directory. Thirty-nine (74%) responded with 32 (62%) representing U.S. medical schools, and the rest (28%) Canadian institutions. All but 7 (18%) were state-supported. Each office director was asked to provide the following information:
1. Unit characteristics, including staffing and the office's position within the institution's organizational structure.
2. Percent funding from both internal and external sources.
3. Current mission, activities, and any changes in these over the last three years.
4. Present and estimated future percent of effort in the following areas: educational research; program evaluation; instructional development; faculty development; direct teaching; educational support services; other activities
5. Current research activities including estimates of changes in effort expected for 1984.

Descriptive statistics (means, standard deviations, and frequency distributions for each response category) were derived for all questions. Two of the authors independently analyzed the content of questions requiring open-ended responses. Key words appearing in answers were used to develop categories for grouping these responses.

Results
Survey data indicate the present status of offices of medical education and, when compared to those of an earlier study (Schimpfhauser, 1979), document trends.

Staffing
Thirty-two offices (82%) were headed by Ph.D.'s or Ed.D.'s, six (15%) by M.D.'s, and one by a director with Ph.D. and M.D. degrees. The mean number of professional staff involved in educational research and evaluation was 6.5 with a majority of the offices (67%) reporting from one to five full time equivalent professional staff in this area.

Structure
Offices are either independent organizational entities (e.g., an academic department), or affiliated with
another unit (e.g., Office of the Dean). Of the thirty-nine offices responding, twenty-two (56%) are affiliated with another unit, while sixteen (41%) are independent. Most directors report to either a dean (36%) or an associate dean (30%).

**Funding**

Table 1 shows funding sources for offices grouped by institution type. Whereas the level of institutional support within both public and private institutions accounts for approximately 70% of total budget, the most noticeable difference is the percentage of funding derived from grants and contracts. Offices at private institutions exceed their public-supported counterparts in this respect by almost 10 percent.

Although external funding constitutes a substantial portion of many offices' budgets, twelve (39%) receive no outside support. A total of eighty-six grants and contracts were reported: thirty-eight (44%) from federal sources, twenty-six grants (30%) from private sources, ten (11%) from institutional sources, and twelve (15%) from state and local sources. Two or more externally funded projects were reported by 51% of the offices.

Budgetary increases for most offices are modest at best. Nineteen (56%) anticipated some budget increase in 1984, but only 29% of those offices anticipated increases over 5%. Eleven (32%) indicated no anticipated budget change and four (12%) anticipated decline.

**Mission and Activities**

Fifty-one percent of the offices cited educational research and 44% cited faculty development in their mission statements. Other frequently stated missions included general educational consultation (36%), program evaluation (36%), curriculum development (31%), service/technical support (21%), teaching (8%), innovation (5%), and enhancing the institution's image (5%). Activities mentioned only once included attracting outside funding, improving student retention, and testing.

Thirty-seven of the 39 offices responded to the question concerning mission changes in the last three years. 57% cited changes in specific effort areas. These included increases in educational research (19%), attempts to attract external funding (12%), curriculum involvement (14%), computer related instruction (14%), program evaluation (14%), teaching (10%), workshops (10%), and institutional studies (10%). Three offices (14%) cited a shift from research to service activities. Other areas of increased activity were student counseling and problem-based learning, each reported by a single office. Areas of decreased activity cited by single offices included course evaluation and testing. Overall, reports of increased efforts far exceeded decreased efforts. Whether the effort report differences were due to underreporting could not be determined.

**Activity Changes**

Estimates of the percent of total office activity devoted to educational research, program evaluation, faculty development, direct teaching, and educational service by twenty-four offices participating in both a survey conducted by Schimpfhauser in 1979 and the current one are summarized in Table 2 and the data are represented in Figure 1. Not all respondents in this comparison sample responded to each question.

Figure 1 reveals both changing patterns of activity as well as a great diversity of programmatic emphasis among offices. More than half (56%) of the offices responding to both surveys indicated increases in educational research, while 33% cited no change and
12% a decline. In spite of the decline in the median percent program evaluation activity, 50% reported increases while 22% noted a decline and 28% reported no change. Only 37% reported increases in instructional development, while 59% noted decline and 11% no change. Faculty development activity increased in 53% of the offices, decreased in only 11%, and remained unchanged in 37%. Only 11% reported increases in teaching activity, while 47% noted a decline and 42% no change. There is a wide range in commitment to educational support services with a third of the offices in the comparison sample reporting increases, 47% reporting a decline, and 17% no change.

Research Activity

The level of research activity at offices responding to the 1983 survey is substantial. 175 projects (funded both by grants and local resources) were reported, an average of 4.5 projects per office. The results of a content analysis of the project titles in which each was assigned to one of twelve major areas appear in Table 3. Areas are ranked by number of offices reporting and number of projects reported.

Research on medical student characteristics includes development of clinical reasoning/problem solving ability, professional socialization, and analysis of the effects of prior academic achievement, learning styles, and personality factors on achievement. Research on program/course/clinical evaluation includes studies of evaluation research methodologies, clinical performance evaluations, the technology of student assessment, and longitudinal studies of graduate performance. Instrucional method and media studies involved techniques of clinical instruction, small group teaching, and CAI/microcomputer applications. Social or humanistic issue studies include patient education or disease prevention, physician manpower supply, humanistic health care, and health services delivery. While faculty evaluation and development studies concern distinguished teacher characteristics, lecture improvement techniques, and reliability/validity of student ratings of instructors and courses. Institutional studies included accreditation self-studies and evaluating projects specific to the home institution, while testing and measurement studies include test validation, criterion-referenced testing, testing procedures, and faculty review of

<table>
<thead>
<tr>
<th>Areas of Study</th>
<th>Total Offices Reporting (N=39)</th>
<th>Total Projects Reported (N=175)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F (%) Rank</td>
<td>F (%) Rank</td>
</tr>
<tr>
<td>Characteristics of Medical Students</td>
<td>24 62 1</td>
<td>42 24 1</td>
</tr>
<tr>
<td>Program/Course/Clincal Evaluation</td>
<td>20 51 2</td>
<td>24 14 2</td>
</tr>
<tr>
<td>Instructional Methods &amp; Media</td>
<td>15 38 3</td>
<td>23 13 3</td>
</tr>
<tr>
<td>Social/Humanistic Topics</td>
<td>14 38 4</td>
<td>22 13 4</td>
</tr>
<tr>
<td>Faculty Evaluation/Development</td>
<td>10 26 5</td>
<td>12 7 5</td>
</tr>
<tr>
<td>Testing and Measurement</td>
<td>9 23 6</td>
<td>9 5 7</td>
</tr>
<tr>
<td>Institutional Studies</td>
<td>9 23 6</td>
<td>11 6 6</td>
</tr>
<tr>
<td>Post Grad. Training &amp; CME</td>
<td>8 20 7</td>
<td>8 4 8</td>
</tr>
<tr>
<td>Curriculum Development</td>
<td>7 19 8</td>
<td>8 4 8</td>
</tr>
<tr>
<td>Specialty/Career Choice</td>
<td>7 18 8</td>
<td>7 4 9</td>
</tr>
<tr>
<td>Medical School Admissions</td>
<td>5 13 9</td>
<td>5 3 10</td>
</tr>
<tr>
<td>Allied Health Professions</td>
<td>3 8 10</td>
<td>4 2 11</td>
</tr>
</tbody>
</table>

1Percentages will not total 100% due to multiple projects conducted by offices

2A heterogeneous category including: health manpower analyses, humanistic and social issues, patient education and health promotion.

Conclusions

Based on the information obtained from both the 1979 and 1983 surveys of medical education offices as well as a 1983 survey of instructional improvement centers (IICs) at colleges and universities (Gustafson & Bratton, 1978), the authors feel there is evidence to support a number of inferences. First, it appears that the age of expansion both in number and size of offices of research in medical education has ended and that there is a similar trend in university instructional improvement centers. In fact, there has been a slight reduction in average number of staff as well as in the total number of units in both medical and higher education. Based on a random sample of 72 of the 275 instructional improvement centers existing in 1975, Gustafson and Bratton estimated that 28% had ceased operation by 1983.
Miller's 1977 survey of offices of research in medical education identified 65 units. Forty-four of these units are still in operation and an additional 11 were established since 1977 for a net loss of 9 (14%).

The pattern of reported changes in activities within offices of research in medical education reflects those social and economic forces enumerated by Petersdorf (1983). There was a decline in percent of effort devoted by professional staff to direct teaching, instructional development, educational support services, and program evaluation, and a simultaneous increase in research activity. This indicates offices are responding to the same forces that have purportedly caused faculty to shift attention from education to research. Although an increase in faculty development activity runs counter to this analysis, much of the faculty development activity has a research and evaluation focus. Fortunately for the educational enterprise, the thrust of research activity is not outside medical education, but rather in applied areas such as clinical reasoning/problem solving ability which allow office staff to blend their personal needs for intellectual inquiry and research productivity with instructional improvement.

Offices of research in medical education are changing. Much of this change may be due to their institutionalization. This institutionalization may be a positive, evolutionary development. But, as research offices have become part of their colleges they have also become subject to the forces (both internal and external) that affect medical colleges. With this “maturity,” the offices have lost their special status as sheltered, innovative enterprises, and must now compete under the same conditions and criteria as other, more established, departments. It remains to be seen if these offices can demonstrate the research productivity this competition demands while maintaining the educational functions which have been their unique contribution.

Author Note. Dr. Grover is Professor and Director, Dr. Smith is an Assistant Professor with the Division of Educational Communications and Studies, SUNY, Upstate Medical Center, Syracuse, New York. Dr. Schimpfhauser is Professor and Director of the Educational Evaluation, Development and Research Unit, College of Medicine, SUNY, Buffalo.

References

New Associate Managing Editor
David F. Salisbury

We are happy to welcome Dr. David F. Salisbury as the new Associate Managing Editor of JID. David Salisbury is a Research Associate at the Center for Educational Technology and Assistant Professor of Instructional Design at Florida State University.
<table>
<thead>
<tr>
<th>Year Established</th>
<th>Affiliation (S-State)(P-Private)</th>
<th>Region</th>
<th>FTE</th>
<th>Grants</th>
<th>% of Total Office Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ER</td>
</tr>
<tr>
<td>1959</td>
<td>S</td>
<td>MW</td>
<td>28</td>
<td>(20.5)</td>
<td>10 (4)</td>
</tr>
<tr>
<td>1963</td>
<td>P</td>
<td>W</td>
<td>5</td>
<td>(15)</td>
<td>55 (75)</td>
</tr>
<tr>
<td>1965</td>
<td>S</td>
<td>W</td>
<td>1</td>
<td>(3)</td>
<td>40 (25)</td>
</tr>
<tr>
<td>1967</td>
<td>S</td>
<td>MW</td>
<td>13</td>
<td>(12)</td>
<td>66 (27)</td>
</tr>
<tr>
<td>1967</td>
<td>S</td>
<td>CDN</td>
<td>3</td>
<td>(2.6)</td>
<td>40 (20)</td>
</tr>
<tr>
<td>1967</td>
<td>S</td>
<td>NE</td>
<td>4</td>
<td>(3.4)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>1968</td>
<td>S</td>
<td>NE</td>
<td>4</td>
<td>(7)</td>
<td>50 (47)</td>
</tr>
<tr>
<td>1968</td>
<td>S</td>
<td>NW</td>
<td>12</td>
<td>(14.5)</td>
<td>75 (92)</td>
</tr>
<tr>
<td>1970</td>
<td>P</td>
<td>NE</td>
<td>3</td>
<td>(2)</td>
<td>2 (5)</td>
</tr>
<tr>
<td>1970</td>
<td>S</td>
<td>SE</td>
<td>6</td>
<td>(8)</td>
<td>39 (48)</td>
</tr>
<tr>
<td>1971</td>
<td>S</td>
<td>SW</td>
<td>15</td>
<td>(5.25)</td>
<td>0 (20)</td>
</tr>
<tr>
<td>1972</td>
<td>S</td>
<td>E</td>
<td>4</td>
<td>(4)</td>
<td>15 (20)</td>
</tr>
<tr>
<td>1972</td>
<td>S</td>
<td>SW</td>
<td>5</td>
<td>(5.5)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>1972</td>
<td>A</td>
<td>CDN</td>
<td>5</td>
<td>(3)</td>
<td>50 (40)</td>
</tr>
<tr>
<td>1973</td>
<td>A</td>
<td>E</td>
<td>20</td>
<td>(10)</td>
<td>60 (36)</td>
</tr>
<tr>
<td>1973</td>
<td>A</td>
<td>CDN</td>
<td>2</td>
<td>(2.3)</td>
<td>10 (3)</td>
</tr>
<tr>
<td>1974</td>
<td>S</td>
<td>NW</td>
<td>0</td>
<td>(1)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>1974</td>
<td>S</td>
<td>MV</td>
<td>6</td>
<td>(6)</td>
<td>10 (10)</td>
</tr>
<tr>
<td>1975</td>
<td>P</td>
<td>MW</td>
<td>7</td>
<td>(10)</td>
<td>90 (65)</td>
</tr>
<tr>
<td>1975</td>
<td>S</td>
<td>SE</td>
<td>4</td>
<td>(4)</td>
<td>0 (5)</td>
</tr>
<tr>
<td>1975</td>
<td>S</td>
<td>S</td>
<td>10</td>
<td>(13)</td>
<td>--(40)</td>
</tr>
<tr>
<td>1976</td>
<td>S</td>
<td>NE</td>
<td>4</td>
<td>(5)</td>
<td>20 (30)</td>
</tr>
<tr>
<td>1978</td>
<td>P</td>
<td>SE</td>
<td>4</td>
<td>(3)</td>
<td>20 (33)</td>
</tr>
</tbody>
</table>

(Footnotes:)

1. 1979 survey results are indicated in parentheses; unreported data are represented by '-'s.
2. number of FTE's involved in research & evaluation
3. percent of office support from grants/contracts
4. 1979 survey results are combined staff development and other categories
5. *25% combined ER & EV

<table>
<thead>
<tr>
<th>RR: Educational Research</th>
<th>EV: Educational Evaluation</th>
<th>FD: Faculty Development</th>
<th>T: Direct Teaching</th>
<th>ES: Educational Service</th>
<th>Other</th>
</tr>
</thead>
</table>