THE USE OF THE STANDARDIZED PATIENT IN THE MEASUREMENT OF CLINICAL COMPETENCE: THE EVALUATION OF SELECTED MEASUREMENT PROPERTIES

by

Robyn M. Tamblyn

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Epidemiology and Biostatistics
McGill University
Montreal, Quebec

July, 1989

Standardized Patients and the Measurement of Clinical Competence

ABSTRACT

THE USE OF STANDARDIZED PATIENTS IN THE MEASUREMENT OF CLINICAL COMPETENCE: THE EVALUATION OF SELECTED MEASUREMENT PROPERTIES

The standardized patient is one method which can be used in the measurement of clinical competence. The accuracy of reproduction of important features of the patient case by the standardized patient was evaluated in Studies 1 and 2. In 839 encounters reviewed, only 13/89 patients provided an accurate reproduction of the case. Attributes of the patient, training process and evaluation procedure were associated with better patient accuracy. A significant inverse relationship was found between patient accuracy and competence score. In Study 3, the use of standardized patients as raters of behaviour was assessed. There were systematic differences in the scoring by different raters, and overall rater agreement was r=.41.

RESUME

L'UTILISATION DE PATIENTS NORMALISES DANS L'APPRECIATION DE LA COMPETENCE CLINIQUE: EVALUATION DE PROPRIETES CHOISIES DE CETTE METHODE

L'utilisation du patient normalisé constitue une méthode possible pour l'évaluation de la compétence clinique. L'exactitude de la reproduction d'importantes caractéristiques du cas réel par le patient normalisé a été évaluée dans le cadre des Etudes 1 et 2. Dans les 839 interactions cliniques analysées, seuls 13 des 89 patients ont fourni une reproduction exacte du cas. Les attributs du patient, le processus de préparation au rôle à jouer et la technique d'évaluation ont été associés à l'exactitude de reproduction du cas par le patient. On a démontré une relation inverse significative entre l'exactitude de reproduction du cas et le score de competence. Dans le cadre de l'Etude 3, on a évalué l'utilisation de patients normalises comme juges du comportement. Il y avait des differences systématiques dans les scores attribués par les divers patients et la concordance globale parmi ceux-ci était r=0,41.

PREFACE

Contribution of the Author

The idea of conducting a joint standardized patient evaluation of senior medical students was initially discussed by the author and Dr. H. Barrows. Firm plans to investigate the feasibility of such a venture were formulated at a breakfast meeting between Dr. H. Barrows, Dr. D.J. Klass and the author at the American Association of Medical Colleges meeting in 1985. Dr. H. Barrows, Dr.R. Williams and Ms. M. Marcy provided the University of Manitoba with assistance in setting up the evaluation. The joint evaluation was executed for the first time at the University of Manitoba and Southern Illinois University in 1987. In 1987, most of the cases used in the evaluation were developed by faculty at Southern Illinois University. In 1988, most of the cases used in the evaluation were developed by faculty at the University of Manitoba.

The research presented in this thesis was funded by the Manitoba Medical Services Foundation (MMSF), the Medical Research Council (MRC) and the research and travel funds provided as part of an N.H.R.D.P. studentship. The author was responsible for writing the research protocol which was submitted for funding to the MRC and MMSF. Dr. D.J. Klass was the co-investigator for the proposal. Dr. D.J. Klass co-ordinated the execution of the joint evaluation between the University of Manitoba and Southern Illinois University at the University of Manitoba. In Southern Illinois, overall co-ordination of the evaluation was carried out by Dr. R. Williams.

Dr. M. Kopelow was responsible for the challenging job of scheduling students and patients for the evaluation. Organization of videotape sampling of encounters was carried out by Dr. Kopelow. Ms. G. Scnabl was responsible for the recruitment and training of patients at the University of Manitoba. Ms M. Marcy was responsible for the recruitment and training of patients at the Southern Illinois University.

measurement of patient accuracy in accordance with guidelines provided by the author. The author established the protocol for encounter sampling, case tape preparation and the training and pre-testing of accuracy raters. The training and supervision of this activity was carried out by Ms. G. Scnabl in both 1987 and 1988. In 1988, the author developed the checklists for patient accuracy.

Ms M. Marcy and Ms. G. Scnabl co-ordinated the recruitment of standardized patients for the rating of videotaped encounters according to the protocol established by the author.

Patient and trainer information forms used in Study 2 were developed by the author. The author was responsible for overseeing the data collection process in compliance with the established research protocol. This activity was aided by the energy and resourcefulness of the research team in Manitoba (Dr. D. Klass, Dr. M Kopelow, Ms.G Scnabl and Dr.T Hassard).

Data entry and analysis for patient accuracy, predictor data and the results of the rater reliability studies were carried out by the author. The entry of the raw student data and score calculation were co-ordinated by Dr. T. Hassard at the University of Manitoba. Analysis of the relationship between student scores and accuracy was carried out by the author.

Originality

The accuracy of standardized patient presentation has not been evaluated previously. The assumption that patients are, in fact, standardized is one of the assumed methodological advantages of this technique. The research carried out in this thesis permitted this assumption to be evaluated. Factors which may influence patient accuracy and the effect of patient accuracy on competence score have similarly not been the subject of empirical investigation. The findings presented in this thesis represent the first effort to investigate these issues.

The reliability of standardized patient raters has been the subject of

previous investigation. This research used a larger study population to derive reliability estimates and held the performance of students as a constant. Both of these features of the design should improve the precision of the resulting estimates. In addition, this research provides the first estimate of systematic differences which may be present between standardized patients who are presenting and rating the case in different universities.

The results of this research are relevant to investigators who are interested in evaluating competence using standardized patient-based instruments in single and multiple evaluation sites for credentialing or research purposes.

Resume

The author has been involved in the training and use of standardized patients for evaluation, teaching and research purposes since 1974. The use of a standardized patient-based evaluation of competence was first initiated by the author in 1980 using a group of senior nursing students at the University of New Brunswick. Prior to this research project, the author had been the principle or co-author in six papers involving the use of standardized patients for research, evaluation and teaching purposes.

ACKNOWLEDGEMENTS

The author wishes to acknowledge and thank:

The Manitoba Medical Services Foundation, the Medical Research Council and the National Health and Research Development Program for their funding support which permitted the completion of this research.

Dr. John Wade and Dr. Walter Spitzer, who made the possibility of doctoral training a reality.

Dr. Danny Klass for his efforts in getting the project funded, the joint collaboration underway (against all odds) and the data collection completed. His insights and comments on the murky area of competence evaluation as well as the research design and findings have added clarity to this thesis as well as the field in general.

Dr. Murray Kopelow for managing to do the impossible - mount a standardized patient-based evaluation in a novice medical school while simultaneously organizing the collection of data for this project.

Ms. Gail Scnabl for her conscientious attention to all of the details necessary to obtain the data for this research project. Her gift for organization and generous assistance have been invaluable.

Dr.Tom Hassard for providing the student score data and the generous donation of a copy of his kappa program.

Dr. Howard Barrows for his support and assistance in this project and for the development of the standardized patient method to which this project is devoted.

Dr. Reed Williams and Ms. Michelle Marcy for their help in 'getting the data together' in Southern Illinois and their assistance in getting the first evaluation mounted at the University of Manitoba.

Ms. Marielle Olivier for her programming and data analysis assistance.

The members of my thesis committee; Dr. Walter Spitzer, Dr Dale Dauphinee, Dr Renaldo Battista, Dr. Jim Hanley & Dr. Dave Swanson; for their helpful guidance in formulating the ideas which went into this manuscript, aiding in the final polishing and for generally keeping me on track.

Margie Ryan who provided brilliant and timely assistance in getting the many pages of this thesis into presentable shape.

My husband, Richard Menzies, for his encouragement and support during the variety of challenges encountered in completing this thesis.

My daughter Diana for being herself.

INTRODUCTION

This thesis addresses the general issues involved in the definition of clinical competence and it's measurement. The standardized patient is identified as one option which can be used in the measurement instrument. The standardized patient is used to present the clinical situation (the test stimulus) and to rate the clinical actions which are taken in response to the test stimulus. The research questions which are addressed in this thesis relate to these two aspects of standardized patient use.

The thesis is divided into three sections. The first section describes the approaches which have been taken to the complex problem of defining clinical competence. A theoretical model for defining competence in this thesis is presented in Chapter 1. In Chapter 2, the evidence which is available to support the hypothesized relationships depicted in the model is reviewed.

In the second section of this thesis, the approaches which have been taken to the measurement of clinical competence are described. Their relationship to the theoretical model presented in Chapter 1 is identified. A framework for classifying the components common to all measures of competence is provided. Potential sources of systematic and random error in measurement which are associated with each component are discussed.

In the third section of this thesis, one optional aspect of the measurement instrument, the standardized patient is addressed. Literature related to the definition and use of the standardized patient is reviewed. Evidence in support of the reliability and validity of this technique is summarized. Questions which have not been addressed are identified. Two of these questions are the subject of the three studies which are presented in this thesis: the accuracy of the standardized patient's presentation of the clinical situation and the reliability of standardized patients as raters of the clinical encounter. An overview of these three studies and general methods is provided in Chapter 5. The

respective studies are found in Chapter 6, 7 and 8.

The final chapter provides a summary of all three studies and recommends future directions in the research and application of this technique. Investigation of the relationship of clinical competence to performance and health outcome is also recommended.

Abstracts have been provided for each chapter of this thesis. The reader is invited to review all abstracts before focusing on individual chapters in order to gain an overview of the issues which will be addressed in this thesis.

TABLE OF CONTENTS

CHAPTER 2 - DESCRIPTION OF MODEL COMPONENTS AND THEIR RELATIONSHIPS

PART II - THE MEASUREMENT OF CLINICAL COMPETENCE

CHAPTER 3 - THE MEASUREMENT OF CLINICAL COMPETENCE

PART III - THE STANDARDIZED PATIENT

CHAPTER 6 - STUDY 1: THE CONTENT OF STANDARDIZED PATIENT PRESENTATION

CHAPTER 7 - STUDY 2: PREDICTORS OF THE ACCURACY OF STANDARDIZED PATIENT PRESENTATION AND THE IMPACT OF PATIENT ACCURACY ON COMPETENCY SCORE

CHAPTER 8 - STUDY 3: THE RELIABILITY OF STANDARDIZED PATIENTS AS RECORDERS/RATERS

CHAPTER 9 - THE USE OF THE STANDARDZIED PATIENT IN THE EVALUATION OF CLINICAL COMPETENCE: CONCLUSIONS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

PART I – THE MEANING OF CLINICAL COMPETENCE

ABSTRACT

CHAPTER 1

THE MEANING OF CLINICAL COMPETENCE

Standardized patients have been used in the evaluation of clinical competence and performance. To place the advantages and limitations of this method of measurement in perspective, the object of measurement (clinical competence) and its underlying rationale are reviewed.

Clinical competence refers to the ability of a health professional to deliver services to patients. When patient services are delivered by a competent health professional, it is generally believed that the patient's health will be better than if those services were delivered by incompetent health professionals. It is this hypothesized link between the ability to deliver services (competence), actual day to day performance (performance) and patient outcome which provides the basis for public and professional interest in the evaluation of competence.

Job analysis, role delineation and critical incident analysis are the three major methods which are used by professional groups to define the important components of clinical competence. Critical incident analysis is the only method which examines the relationship between competence and outcome. Five groups of abilities are thought to be necessary for competent service delivery: data collection, problem formulation, immediate and continuing management, professional communication and patient communication. The effective development of these abilities is assumed to require certain prerequisites. These prerequisites include: knowledge, skills, judgement and attitudes.

A conceptual model is proposed which provides a method of integrating these hypothesized relationships for future study. In this model, competence is defined as being specific to the clinical situation. The relationship between prerequisites, competence, performance and outcome is described. The mediating influence of other provider and system related determinants of performance is identified and defined. Finally, other socio-cultural and patient factors which could influence health outcome,

and yet are not amenable to control by the health professional, are identified.

The proposed model is compared with other models of competence and performance provided in the educational and health care literature. The advantages and limitations of the proposed model are reviewed. The main limitation of the model is the failure to include the costs of service delivery. The advantage of the model is that the assumed relationships between competence, performance and health outcomes are clarified.

CHAPTER 1

THE MEANING OF CLINICAL COMPETENCE

Clinical competence is a term which is frequently used but rarely defined. A theoretical model for defining clinical competence will be presented in this chapter. It will serve as a structure for this thesis. The literature related to this model will be reviewed in Chapter 2. Methods of measuring clinical competence will be identified in relationship to this model in Chapter 3. The standardized patient is one of the methods which can be used in the measurement of clinical competence. Assumed properties of this method will be evaluated in Chapters 4 through 8.

In this chapter, the rationale for clinical competence evaluation will be initially described. The methods conventionally employed for the definition of competence will be summarized and finally an integrative model for defining competence in the health professions will be presented. The physician will be the primary professional model used for explication and review of the theoretical model proposed since research has been most abundant in this area.

RATIONALE FOR THE EVALUATION OF CLINICAL COMPETENCE

Competence is an abstract attribute or set of attributes which can be used to describe an individual's ability to carry out services associated with a vocation or profession. It is assumed that individuals who are more competent will be able to provide more effective services than those who are not (Popham, 1978). Effective in this sense refers to the expected outcome which should be rendered as a result of the service. Effectiveness also implies the avoidance of harm or a worse outcome than would have resulted if the service had not been provided.

In the health professions we refer to the ability required to deliver services to patients as clinical competence. It is generally believed that when services are effectively delivered by health professionals that the patient's health will be better than if services were not received. It follows then that variation in the level of health professional competence will be associated with variation in the effectiveness of

services delivered which in turn will be associated with patient outcome. Patients who are treated by more competent health professionals should have a better health outcome than those treated by less competent health professionals.

Hence, we are interested in the competence of an individual because we believe it will improve our prediction of those most likely to provide effective services and thereby achieve optimal patient outcomes. It is this assumed causal relationship between provider competence, effectiveness of service delivery and health outcome which provides the primary rationale for the evaluation of clinical competence. To this end society has empowered professional groups with the legislative mandate to control entry and practice within the profession (Holmes, 1986; Mandelbaum, 1987). Through such mechanisms as licensure, certification and recertification society expects to maximize its health outcomes by identifying incompetent providers who are more apt to deliver unsafe or ineffective care.

The costs which may be incurred by the incompetent physician both in resource expenditure and health status loss form an added rationale for the public's interest in clinical competence evaluation. This concern is reflected in recent policy development for the institution of recertification of competence, mandatory continuing education and the creation of Professional Standard Review Organizations (Greene, 1976; Burg, 1982). The institution of mechanisms for assuring safe and effective care is also of interest to third party payment organizations (Greene, 1976). Institutional, regional and national disparities in the use of procedures and health care resources have been documented (Eisenburg, 1986). Differences in patient/population mix account for a small component of practice variation (Eisenburg, 1986). These observations raise two inter-related questions. Is variation in practice associated with a similar variation in health outcomes for patients? Is variation in practice a function of differing levels of provider competence?

The institution of quality assurance mechanisms, in response to escalating costs, presumes that provider competence is a major determinant of

practice variation (Greene, 1976). Donabedian (1982) provides a model for integrating the three components of quality: effectiveness, cost and health outcome. In this model, the ideal provider is conceptualized as the physician who provides the optimal summative gain in health status per unit of time over the natural course of the patient's health problem. Although not explicitly stated, it is assumed that this refers to all patients seen by the physician. When health care resources are restricted, it is assumed that this will attenuate improvements in the health status of patients of the ideal physician. The patients of less competent physicians theoretically would demonstrate a smaller gain in health status or in some instances loss in health status when compared to the natural course of the illness. Resource restriction is again seen as a means of attenuating both the positive gains and negative losses in health status. Incompetent physicians are theoretically expected to increase their probability of causing harm as an increasing number of resources are made available.

In summary, clinical competence is evaluated because it is believed that it will improve our ability to predict those who may do harm as well as those who will have a greater probability of providing safe and effective services. We assume that effective services will more likely result in optimal health benefits. Finally it is theorized that benefits to costs ratio will be maximized if there is a mechanism of identifying and remediating health professionals who are less competent.

METHODS OF DEFINING CLINICAL COMPETENCE

The challenge in defining clinical competence is to identify those components of professional behaviour which are important determinants of effective service and positive health outcomes. There are three general strategies which have been used to identify the important components of clinical competence: job analysis, role delineation and critical incident analysis (D'Costa, 1986). The latter two strategies have been most frequently used in the identification of physician competence.

JOB ANALYSIS

Job analysis involves an itemization and frequency analysis of what a representative sample of working professionals currently do. The importance of itemized tasks may be determined empirically or by rating by an external group of professional judges. The advantage of job analysis is that it provides an objective picture of the current work of the professional, unbiased by theoretical pronouncements of what ought to occur rather than what actually happens. There are a number of disadvantages. Job analysis assumes that what practising professionals currently do is optimally effective in achieving expected health outcomes. Job analysis is also confined to an analysis of the present with no means of accommodating future professional directions.

ROLE DELINEATION STUDIES

Role delineation in contrast is futuristic, defining tasks that professional group members should be responsible for. It typically provides a systematic description of the professional role organized into major and minor functional responsibilities. Expert professional committees are usually responsible for the initial identification of important components of competence. Data from job analysis studies can be employed in this process but the role is not confined to the activities of those in current professional practice.

For example, Young et al. (1983) described the process used to gain consensus on competencies which should be expected of primary care physicians in the management of nutritional problems. Competencies were initially developed to reflect the knowledge and skills required for common nutritional problems by a committee of academic medical practitioners. A delphi survey of faculty chairpersons and local physicians was used to achieve consensus on important competencies.

The validity of this method of competency development has been criticized (D'Costa,1986). Rather than developing competencies to fit knowledge and skills, components of competence should be derived from the services

required in practice. For educational or testing purposes, these competencies can be subsequently broken down into their presumed knowledge and skill requirements. This perspective on the validity of competency development is embodied in the philosophy of the competency-based educational movement. The central belief of this movement is that professional program and evaluation content should be derived from and justified by the competencies required in professional practice (Houston & Warner, 1977).

The principles articulated by Houston & Warner (1977) were applied in the approach taken for the development of the new FLEX (Federal Licensing Examination Program) examination. A sampling frame for examination content was derived from an analysis of the competencies, knowledge and skills required for patient problems in primary care. Important patient problems were identified from an analysis of 102,705 actual patient encounters documented by a sample of physicians from six primary care disciplines (LaDuca et.al.,1984). This study defined general medical competence by using data derived from an analysis of actual practice as well as expert opinion.

Similar approaches have been taken by other health professional societies. Professional consensus has provided the basis for identifying important competencies and prerequisites in family practice (CFPC, 1974, 1981), paediatrics (Ambulatory Paediatric Association, 1984) and obstetrics and gynaecology (CREOG, 1980).

One of the problems with this method is that substantial differences in opinion exist about the importance of various competencies among members within the profession. In a Manitoba study, systematic differences in the rated importance of future components of nursing competence were noted among administrators, educators and practitioners (MARN, 1984). In Young's (1983) study, significant differences in the rating of importance of 30% of the listed competencies existed between academic and non-academic physicians. Wigton (1980) contrasted the ranked importance of components of competence considered to be important for internal medicine residents in three groups: full-time faculty, volunteer faculty and house

officers. Significant differences existed with faculty emphasizing record keeping and physical examination, volunteer physicians emphasizing judgement and data analysis and houseofficers emphasizing patient management and clinical responsibility.

It seems fairly clear that the membership of the 'expert group' will play an important role in the nature of competencies generated. The importance of competencies identified by different expert groups for expected patient outcomes has not been addressed in these studies.

In summary, role delineation uses the opinions of members within the profession to define competence. The advantage of this method is that competence is not confined to the services the majority of professionals are delivering at present. The major disadvantage is that the definition of competence appears to vary among different subgroups in the profession. It is unclear whether or not these differences represent true differences in the competence which may be required in different subgroups within the profession.

CRITICAL INCIDENT ANALYSIS

The critical incident method of defining competence overcomes the problems in the former methods by providing a direct means of ascertaining the importance of certain aspects of competence for patient outcome. This method assumes that provider competence is more critical in some clinical situations than others. These critical situations can be identified by use of professional surveys or by an expert professional group. The aspects of professional competence which are important determinants of the outcome in these critical clinical situations are then identified.

For example, Sanazaro (1968) used critical incident analysis to identify important aspects of competence for physicians in internal medicine. Critical incidents in this study were generated by the respondent and were defined as "a verifiable episode of patient care in which a physician's actions had a clearly beneficial or detrimental effect on the patient". The situations and important aspects of competence were determined by

content analysis of the three beneficial and three detrimental descriptions of performance solicited from each of the 2,449 clinical faculty surveyed. Analysis was restricted to those situations which provided clear evidence of consequence for patient outcome (1,485 descriptions of effective performance and 1,104 descriptions of ineffective performance).

An obvious limitation of this approach is that it is restricted to those clinical actions which, in the opinion of professionals, were causally related to the outcome observed. Secondly the outcomes observed are limited to those which the professional feels are important and will likely be confined to those occurring in the short term. Despite these limitations, this approach provides a means of directly ascertaining the components of competence which, in the opinion of the professional, have been of critical importance in patient outcome.

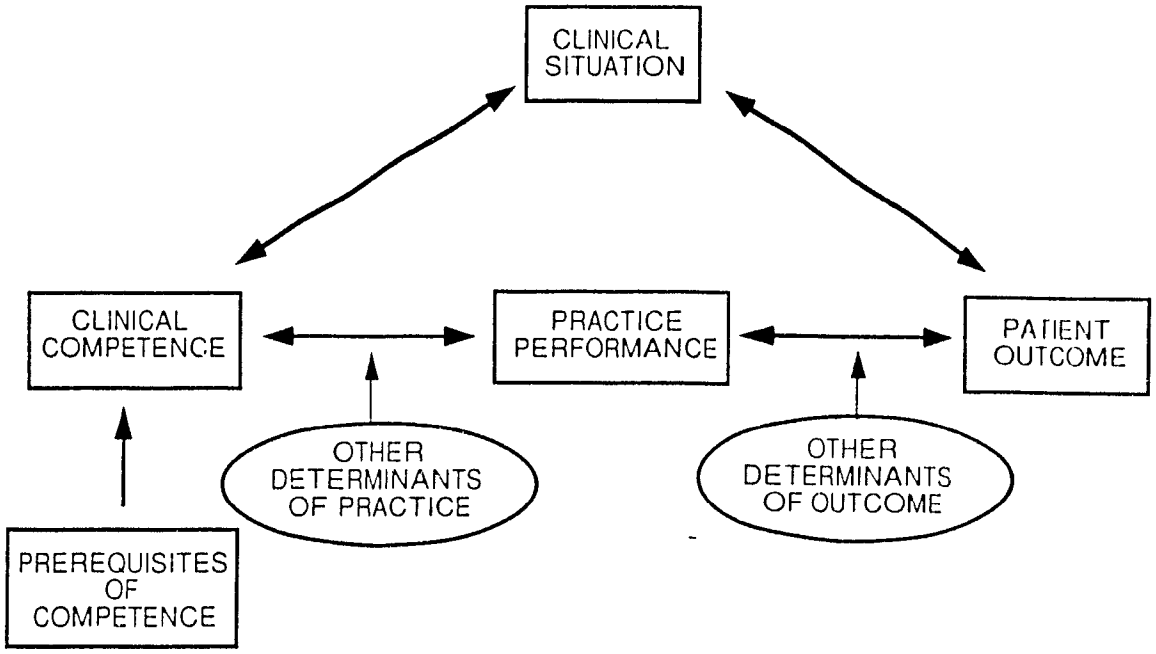
SUMMARY

In summary, job analysis, role delineation and critical incident techniques have been used to identify aspects of competence which are important in the delivery of effective services and the optimization of patient outcome. Of all current methods, critical incident analysis is most compatible with the stated rationale for clinical competence evaluation. It provides the most rational basis for identifying individuals who are more apt to render unsafe, costly or ineffective services to the public. The results of all of these methods should be viewed as hypothesis generating. The validation of these hypotheses would require empirical evidence of the presence of a causal relationship between components of competence generated, effectiveness of provider services and health outcome. The next section provides a theoretical model for integrating the important dimensions of this relationship.

A THEORETICAL MODEL OF CLINICAL COMPETENCE

A model for conceptualizing the relationship between clinical competence service delivery and patient outcome is displayed in Figure 1.1.

FIGURE 1.1 A CONCEPTUAL MODEL OF CLINICAL COMPETENCE



DEFINITION OF MODEL COMPONENTS

A brief glossary of terms is provided to aid the reader in the subsequent discussion.

Prerequisites of Competence: The abilities that are necessary to demonstrate competence in a clinical situation. These abilities include:

knowledge; skills (interpersonal, problem-solving and technical); clinical judgement; self-study and self-evaluation abilities; attitudes and professional habits (Bloom, 1956; Ho Ping Kong, 1987; ABIM, 1979).

Clinical Competence: The ability to carry out relevant tasks, services and responsibilities in the provision of care to patients. Relevance refers to those activities which are thought to be causally related to the achievement of optimal patient outcomes or the avoidance of adverse patient effects. Professional services which are considered to be important determinants of the outcome include:

1) Data Collection [e.g. history, physical examination, diagnostic studies]

2) Problem Formulation [e.g. diagnosis, differential diagnosis]

3) Immediate and Continuing Medical Management [e.g. pharmaceutical, biological, consultative, counselling]

4) Professional Communication [e.g. record-keeping, intra and inter-disciplinary communication]

5) Doctor-Patient Relationship

(Coulehan, 1984; Sanazaro, 1968; ABIM, 1979; Platt, 1979)

Performance: The actual performance of relevant tasks, services and responsibilities in the day to day provision of care to patients under usual practice conditions (Senior, 1976; Cronbach, 1970). It includes the same classification of activities as outlined above. It is the circumstances in which they are measured which differs.

Health Outcome: A measure of the individual's health state at a point(s) in time when the impact of provider intervention should be evident. Measures of health state classically include the five D's: death, disease, discomfort, disability and dissatisfaction (Fletcher et al., 1982). Cost to the patient and society is an additional outcome which has received increasing attention (Donabedian, 1982; Fletcher, 1982; Evans, 1984). A composite index, reflecting functional status in the physical, psychological and social domains, has also been used as a method of characterizing

health status (Donabedian, 1982; Tugwell, 1979; Ware et al, 1987). The outcome which can be expected in a clinical situation is usually specified by professionals who are knowledgeable in the field (Palmer, 1976). The outcome which is considered to be desirable or worthy of the costs and risks is specified by society and/or the individual (Donabedian, 1982).

Clinical Situation: The characteristics of the situation in which the provider is required to render services. Characteristics refer to those attributes of the clinical situations which are predictive of what will be required of the provider in the clinical situation and/or of the expected health outcome. Characteristics used for classification include:

1) Health Problem: diagnostic category, presenting symptom, severity, complexity, physiological system, prototypicality, disease mechanism

2) Individual: age, gender, socio-economic status, race

3) Management Required: level of management-eg.primary, secondary; scope of management-eg.preventive, rehabilitative; type of management-eg.acute, chronic

4) Location of Management: remote, rural, urban

(Chong, 1984; Hennan, 1984; Laduca, 1984; Bordage, 1982; Wilson-Pessano, 1984; Piscano, 1986; Wakeford, 1986).

Treatability and prevalence are additional characteristics which are used to earmark situations where provider competence is of greater importance to the general public welfare (Chong, 1984; Hennan, 1984).

Other Provider and System Related Determinants of Performance: Other factors which have been associated with systematic differences in the resources used by providers or the quality of care delivered. They include attributes of the provider, practice setting, institution, administrative systems, economic policies and work situation.

Individual/Disease Factors: Determinants of health outcome which are particular to the individual and disease. They include factors listed under the classification of the clinical situation (health problem-type, severity, complexity and individual characteristics (eg. age, race,

socio-economic status, gender). Health related values, response to the illness, health seeking behaviour and the presence of barriers to access for health care are additional factors which have been associated with the individual's probability of benefiting from relevant therapy.

Other Socio-Cultural Determinants: Other determinants of health status in the socio-cultural milieu within which care is delivered which are beyond the control of the individual provider. They include such factors as living conditions, nutrition, employment, sanitation, environmental hazards and pollution, availability of social assistance and funding and availability of health services.

OVERVIEW

The horizontal axis of the model depicts the relationship which forms the primary rationale for the evaluation of clinical competence. Competence is seen as being predictive of practice performance which in turn is predictive of patient outcome.

The clinical situation is positioned at the apex of this relationship having a pivotal role in the identification of both the components of competence which are considered to be important and the patient outcomes which can reasonably be expected as a result of effective service delivery.

The model incorporates the contribution of additional determinants of patient outcome and provider performance other than clinical competence. Individual and disease factors are positioned on the direct axis between the clinical situation and health outcome. This acknowledges the influence of these factors with or without entry into the health care system. A second group of determinants has been identified as influential in both the quality of provider performance and the quantity of health resources used in care episodes. The relationship of these factors to clinical competence is unclear but all appear to be associated with systematic differences in provider performance. Finally a host of system related and socio-cultural factors appear to influence health status in addition to

those explained by the characteristics of the individual and disease state. These are entered on the axis between performance and health outcome although a legitimate case could be made for their position at other parts of the model.

The inclusion of prerequisites of competence in the model may not be intuitively obvious from the previous discussion of the rationale of clinical competence evaluation. The dissection of important components of competence into their assumed prerequisite abilities has traditionally formed the basis for professional program content and the target of competency evaluation procedures. It has been assumed that the possession of specified prerequisite abilities is predictive of clinical competence which in turn is predictive of practice performance and health outcome.

In this model clinical competence is defined as the individual's ability to select and deliver services which have a probability of improving the natural outcome of the clinical situation presented. The definition is situation specific. It follows that individuals may be competent in only some of the clinical situations in a domain such as internal medicine. No optimal method of grouping clinical situations into homogeneous groups which require similar abilities has been identified in this model or in the literature.

The Relationship of the Model to Other Theories of Competence/ Performance

Theories of competence and performance arise from two sources: the educational literature and the literature on quality of care. In the educational literature, competence is often defined in terms of prerequisite abilities; knowledge, skill, judgement and attitude (General Medical Council, 1984; Ho Ping Kong, 1987). This perspective is consistent with the conventional paradigm of the professional (Moline, 1986). In this paradigm the possession of specialized knowledge, skills and judgement is viewed as being synonymous with professional expertise. The model presented expands on the conventional definitions to provide explicit recognition of the situation in which professional expertise must be

rendered and the end point which effective service delivery is expected to achieve.

The inclusion of the situation and outcome in this model is more consistent with the deliberate action theory of professional expertise - (Kennedy, 1987). In this theory, the major task of the professional is assumed to be the analysis of situations from the perspective of required action. The perception of the problem, goals, means and ends interact to influence the professional's performance in any situation.

Bashook and Lloyd (1985) have developed a topological paradigm of physician competence and performance. It includes four dimensions of competence: clinical care activities, interpersonal relationships and teaching, management of the health care system and humanistic concerns in patient care. It is a useful paradigm for defining the scope of measurement and predicting the relationship which can be expected among different clinical situations which tap varying proportions of the four domains. Differences among medical disciplines and levels of training are characterized by differences in the specific content sampled in each domain.

There are no theoretical grounds inherent in Bashook's model for justifying the domains identified or their content. In the model presented, competence is theoretically defined as those aspects of the providers' process which are important determinants of patient outcome. Patient outcome is not considered in Bashook's model. The distinction between competence and performance is also unclear. In Bashook's model, competence is an abstract attribute which is inferred from performance under both optimal and usual practice circumstances. This conceptualization fails to consider other determinants of provider performance which have been identified. The scope of competence defined by Bashook's model is broader than that of the model presented. The model presented fails to include competence in student teaching and management of the health care system in its conceptualization of important components of competence. Although the inclusion of student teaching as an important component of clinical competence could be debated, competence in the

management of the health care system may have an important relationship to patient outcome. Clearly the physician must have the ability to mobilize resources in the health care system to provide effective management in many clinical situations. Unfortunately, this ability has not been evaluated as an aspect of competence or performance in any of the studies reported. It should be theoretically considered as an important component of competence and examined empirically.

Pierce and Downing (1982) employed a combination of two theoretical models; the clinical reasoning model (Elstein et al., 1978) and the conflict model of decision-making (Janis & Mann, 1977) to develop a method of measuring evolving competence in internal medicine. The clinical reasoning model identifies theoretically important determinants of diagnostic accuracy and optimal management decision-making in the physician's cognitive process. Conflict theory identifies important situational determinants which might influence the quality of diagnostic and therapeutic decisions. These latter determinants have been incorporated, in the model presented, in the category of other determinants of performance. They would be included under the heading of practice setting characteristics. The elements of the physician's cognitive process which are theoretical determinants of the quality of diagnosis and management from the clinical reasoning model are included as prerequisites of competence in this model. Pierce and Downing (1982) also describe theoretical prerequisites for the development and maintenance of competence. The prerequisites include the individual's motivation and responsiveness to learning opportunities in the clinical situation. In the model presented in this chapter, these provider characteristics are included under prerequisites of competence as self-evaluation and self-study abilities.

In the quality of care literature, Donabedian's (1982) integrative model of quality of performance, costs and health outcome is related to the model presented. Although, clinical competence is not an explicit component of Donabedian's model it could be seen as the attributes of performance characteristic of the optimal provider. The optimal provider in Donabedian's model is the physician who produces the greatest possible

cumulative improvement in health status given current knowledge and technology and within existing resource restraints. The dichotomy of the conventional classification of competence (competence/incompetence) is intended to separate those who have a greater probability of being safe and effective from those who do not. In Donabedian's model the unsafe and ineffective are clearly depicted by those who render a worse patient outcome than would be realized through the natural course of the illness. Effectiveness on the other hand lies on a continuum between no effect/no harm to the maximum effectiveness possible by the theoretically optimal provider. It could be argued that maximally effective provider represents excellence in performance. Competence on the other hand falls somewhere in the range from no effect to optimal effectiveness. Neither model provides a means of establishing what might be considered a sufficiently acceptable level of performance to be considered competent by conventional classification.

Finally, it is interesting to note that a similar model is proposed for the definition of competence outside of the health professions. Haertal (1986) describes a model for defining teacher competence which is based on the relationship between competence, performance and outcome. Important components of competence for teachers are seen to be theoretically related to classroom performance and to the learning gains that can be made by their respective students. Teacher competence is then defined as those behaviours which have an impact on learning outcome. Similar consideration must be made for other determinants of learning outcome which are unrelated to the performance of the provider.

Limitations

There are a number of potentially important factors which have not been adequately addressed in the model presented in this chapter. The relationship of the costs incurred by the provider in the delivery of care to competence, performance or outcome has not been explicitly included in the model proposed. Costs are of course intimately related to resource availability,the patient's ability to pay and various obstacles to the use of resources on the part of the patient and provider. Resource

availability and obstacles to access should certainly be included as additional determinants of performance and outcome. What is not known is whether or not competence and costs are related when studied in homogeneous groups of patients and practice settings. Do more competent providers incur greater costs, less costs or are costs per se unrelated to the effectiveness of care?

The issue of the professional's moral, social and fiduciary responsibilities in patient care have similarly received inadequate attention in this model. The relationship of these responsibilities to competence, performance and outcome requires clarification. It has been difficult to articulate and gain consensus on responsibilities which should be expected of the professional particularly in the area of social responsibility (Eisenburg, 1986). It may well be that some of these responsibilities are being indirectly quantified by current measurement methods.

For example, Sheehan et al. (1980) found a relationship between level of moral development and supervisor's ratings of clinical performance in 244 residents evaluated. The results of this study suggest that moral development may be related to the quality of clinical performance. Alternatively, attending supervisors may be evaluating a different attribute of performance which is influenced by moral development. Clearly the relationship of these responsibilities to clinical competence, performance and outcome requires further study.

ABSTRACT

CHAPTER 2

DESCRIPTION OF MODEL COMPONENTS AND THEIR RELATIONSHIPS

In Chapter 2, the literature related to the model of clinical competence described in Chapter 1 is reviewed. Evidence which supports and negates the hypothesized relationships between competence, prerequisites of competence, clinical performance, the clinical situation and health outcome is summarized. Limitations in the available literature are identified.

Of the four identified prerequisites of competence, knowledge is the only prerequisite where sufficient evidence is available to draw any conclusions. Medical knowledge has a modest and positive relationship to competence in clinical diagnosis and management. A weaker but positive relationship is present between medical knowledge and clinical performance. No relationship has been demonstrated between knowledge and patient outcome in the one study in which this was investigated. Prerequisites of competence are associated with the clinical situation with different knowledge and skills being required in different situations.

There is a relationship between competence and performance when it is measured in the same type of clinical situation. Other factors which influence performance include characteristics of the provider and practice setting, institutional policies and affiliation and remuneration practices. The relationship between competence and these factors has not been studied. The competence and performance of the provider vary in different types of clinical situations in the same practice discipline. Factors which contribute to this observed variation are unclear. They may include differences in the knowledge and skill required in different situations and/or be attributable to measurement error.

The relationship between competence, performance and health outcome has not hitherto been studied. There has been a relationship found between provider performance and selected health outcomes in some studies. The

quality of the provider's performance appears to be more critical at certain points in the patient's clinical course.

CHAPTER 2

DESCRIPTION OF MODEL COMPONENTS AND THEIR RELATIONSHIPS

This chapter examines the literature related to the components of the model presented in Chapter 1 and their respective relationships. At the completion of this chapter, the reader should have an understanding of clinical competence and its relationship to theoretical prerequisites, the clinical situation, practice performance, and patient outcomes. Limitations of current research and gaps in our understanding of these relationships will be identified as areas for future research. This will provide the reader with a context for reviewing the potential use of standardized patients and other evaluation methods for research in this area.

PREREQUISITES OF COMPETENCE

Variants of Bloom's taxonomy of instructional objectives have been used to group prerequisite abilities into four categories: knowledge, skill, judgement and attitudes (ABIM, 1979; General Medical Council, 1984; Ho Ping Kong, 1987; CREOG, 1980; CFPC, 1974). Self-Evaluation and self-study skills are two additional prerequisite abilities. They have been identified as being important for the development and maintenance of competence (Barrows & Tamblyn,1980; Pierce & Downing,1982).

It is assumed that if the individual possesses these abilities, relevant to the situations he/she will encounter, he/she will be capable of rendering a competent performance. It is similarly assumed that if an individual demonstrates a persistent predilection for substandard performance that deficits in knowledge, skill,judgement or attitudes will be present. These assumptions form the rationale for the evaluation of knowledge as a predictor of competence on licensing examinations and the employment of continuing medical education and recertification as methods of remedying or assuring continuing competence in practice. Empirical evidence of these assumed relationships will be discussed in the next three sections.

THE RELATIONSHIP OF PREREQUISITES TO OTHER COMPONENTS

THE RELATIONSHIP OF PREREQUISITES TO COMPETENCE

Overview

The relationship of knowledge to competence has been the major focus of study. The components of competence evaluated include data collection, problem formulation, short term medical management and doctor-patient relationship. These components are usually assessed through patient simulation formats (patient management problems (PMP), computerized management problems, and standardized patients). Criteria of performance are usually established by medical faculty and scores are based on a summation of criteria met.

The correlation coefficient is customarily used to estimate the strength of the relationship between components of competence and knowledge. With one exception, a weak to modest relationship has consistently been found between knowledge and competence in diagnosis and management. In the one study reported, no relationship was found between knowledge and competence in the doctor-patient relationship.

The Evidence

Wolf et.al. (1983) estimated the association between scores on 15 P.M.P.'s given to 175 medical students and their scores on Part 1 of the National Board examination. Correlation coefficients of .21 to .53 were found. It is impossible to interpret the meaning of this relationship since the domain of knowledge sampled by the National Board examination and that implicitly sampled by the 15 Patient Management Problems (P.M.P.) are not reported.

To be conceptually meaningful, knowledge presumed necessary for competent performance in a clinical situation needs to be articulated and its relationship to a competent performance estimated in the same domain.

Dawson-Saunders et.al. (1984) attended to this problem by studying the relationship between the possession of knowledge required for three surgical problems and the student's ability to identify relevant data and generate likely hypotheses of causation. The apriori possession of knowledge about the area did not seem to be an exclusive predictor of variation in hypotheses generation. Correlations across problems among students possessing adequate knowledge in all areas were -.03 to .18 for diagnostic accuracy.

Maatsch (1983) examined the relationship between knowledge considered to be necessary for the management of emergency medicine problems and competence in the diagnosis and management of emergency medicine problems presented by simulated patients. Among 182 emergency medicine certification candidates, a correlation of .38 was found between scores on multiple choice tests and patient simulation.

Norcini et al. (1986,1987) have studied the relationship between knowledge measured by multiple-choice formats and competence in diagnosis and management among candidates for the American Board of Internal Medicine (ABIM) certification and re-certification exam. Competence in diagnosis and management was measured in the certification sample with 3-6 PMPs and with both computer and PMP formats in the re-certification sample. Correlations between composite multiple-choice scores and PMP scores were consistently high over the three reported years of certification examination administration (r=.75,.71,.76). For the re-certification sample, correlations of .49 and .61 were found between PMP performance and scores for two types of multiple choice questions. Correlations of .41 and .42 were found between computerized simulation and multiple choice scores.

The magnitude of the associations found by Norcini et.al.(1986,1987), in contrast to those reported in other studies, may be partially explained by improvements in the reliability of the measures. An equally plausible explanation is the difference in content measured by the multiple-choice items used by Norcini to that conventionally sampled in other studies. The multiple-choice items used in the ABIM exams measure both knowledge and

the appropriateness of actions selected in the diagnosis and management of patient problems presented (the proportionate mix is unspecified). The PMP used in the ABIM examination are different from those customarily studied. Competence in the ability to collect relevant items on history and physical examination is not included in the overall score. These data are presented to the examinee at the beginning of the problem. The PMP case score is based on the appropriate selection of investigations and medical management. In essence the two formats are measuring essentially the same thing, competence in diagnosis and management when relevant clinical data are provided. The relationship of knowledge to competence cannot be determined from the information provided in these studies.

Bordage (1982) evaluated the relationship between medical knowledge and diagnostic errors in a sample of medical students, nurses and general practitioners. Problem related knowledge did not seem to be a major determinant of diagnostic accuracy. In 57% of the cases which were incorrectly diagnosed, subjects were reported to have adequate knowledge. Although the likelihood of making an incorrect diagnosis increased with greater deficiencies in knowledge, accurate diagnoses were still made by subjects whose knowledge was classified as being inadequate.

Norman et al. (1983) evaluated the relationship between knowledge relevant to eight clinical problems in two subspecialty areas and competence in data collection, diagnosis and management in 30 subjects. There was no significant relationship between scores on multiple-choice tests of knowledge and performance with the 8 standardized patient problems.

The relationship of knowledge to competence in communication and the doctor-patient relationship was estimated by Lamont and Hennen (1972) in their examination of 117 family practice certification candidates. Scores on the multiple-choice examination of knowledge correlated at .09 and -.09 with measures of communication and doctor-patient relationship carried out in a simulated office situation.

Summary

In summary, medical knowledge appears to have a modest and positive relationship with competence in clinical diagnosis and management. There is currently no evidence that knowledge is related to competence in communication, doctor-patient relationship or data collection.

Limitations

There are a number of limitations in the studies presented which could inflate or attenuate the magnitude of the estimated relationship. Both range truncation and unreliability of measures may contribute to a lower estimate of the correlation coefficient. Maatsch (1983) was able to estimate the impact of range truncation on the relationship between measures of knowledge and competence in emergency medicine. When the score range was truncated by the application of a minimum required knowledge test score prior to acceptance for competency testing, the original estimate of the relationship between knowledge and competence fell from .75 to .39. An estimate of the true correlation of measures, corrected for measurement unreliability was provided by Norcini (1986). After correcting for measurement unreliability, correlations of .61 and .49 for PMP,s were increased to .80 and .61 and the estimates for computer simulation performance were increased from .42 and .41 to .62 and .58.

The magnitude of the true relationship between knowledge and competence will be similarly attenuated if content is sampled from unrelated domains. For example the knowledge required for the diagnosis and management of clinical situations in neurology may be completely unrelated to the knowledge required to diagnose and manage infectious disease or endocrine problems. No meaningful interpretation of the relationship can be gained by the customary practice of comparing omnibus, multi-disciplinary tests of knowledge with performance in a few clinical problems. If a relationship exists it will surely be diluted by inadequate sampling of relevant items (unreliability) and performance on other unrelated aspects of knowledge tested.

The ability to apply discipline related theory and general principles to

clinical situations is hypothesized to be predictive of clinical competence (Kennedy, 1987). The use of multiple-choice tests for evaluating this kind of knowledge has been challenged; the criticism being that they tend to evaluate the recall of isolated irrelevant facts rather than the comprehension and appropriate application of general principles (Neufeld, 1985; McGuire, 1962). An analysis of the level of questions used in the reported studies is not provided. Their ability to measure theoretically relevant knowledge is therefore unknown.

Biases towards the null hypothesis have been suspected as being more likely than biases which would create a stronger estimate of the relationship than actually existed. The most likely bias in the latter category would be introduced by an absence of blinding. Raters scoring competence may be aware of the individual's scores on knowledge tests introducing the potential for bias in the scoring procedure. The issue of rater blinding has not been discussed in any of these studies.

Finally, the measures employed to estimate competence have their own limitations which could act to create biases both away from and towards the null hypothesis. Biases away from the null are created by measures of competence which emphasize the purely cognitive aspects of care delivery particularly when measured in situations which demand people-oriented and psychomotor skills. Biases towards the null are created by low reliability, range truncation, measurement of extraneous attributes and dilution of effects.

THE RELATIONSHIP OF PREREQUISITES TO PERFORMANCE

Overview

The relationship of knowledge to performance has again been the focus of study. The understanding of this relationship comes from two types of studies. The first group of studies examine the relationship between scores on multiple-choice tests of knowledge with supervisor's or colleagues ratings of performance in practice. The second group have studied the impact of increasing medical knowledge on practice perfor-

mance.

The Evidence

Veloski et al. (1979) examined the relationship between performance on Part 3 of the National Board examination with supervisor's ratings of performance at the end of their first year of postgraduate training for 1,866 medical students. The examination included both measures of knowledge and competence in diagnosis and management in paper case simulations (mix unspecified). Test scores correlated at .24 with supervisor's ratings of knowledge, r=.21 for data gathering, r=.21 for clinical judgement, and r=.13 for professional attitudes.

Maatsch (1983), in his study of emergency medicine candidates, compared their performance on a multiple-choice examination of knowledge with ratings of diagnosis and management on stimulated recall (mini-oral) of charts completed by the resident in daily practice. A correlation of .23 was found. When corrected for range truncation and attenuation,the estimated correlation was r=.39.

Norcini (1987) estimated the relationship between scores on the ABIM multiple choice examination and program director's rating of competence among candidates for internal medicine certification. Correlations of .32 to .33 were found in three examination cohorts. The proportion of variance in supervisor's ratings explained by all aspects of the testing procedure was 12.2-14.7%, the majority of which was explained by multiple-choice test performance. A similar study was conducted on 289 physicians applying for recertification in internal medicine. The ratings of performance were generated by the chief of service and two of the subject's peers. Performance on the multiple-choice test correlated at .27 and .30 with performance ratings (Norcini,1986). Correlations of .27 and .30 were found with PMP and computer simulation scores. All measures combined accounted for 12.3% of the variance in performance rating with computer case simulation performance accounting for the largest unique contribution to variance.

Gonnella and Hojat (1983) hypothesized that the relationship between measures of academic achievement on knowledge tests and subsequent clinical performance would vary by specialty. This hypothesis was confirmed in a study of 441 graduates entering three specialty areas. Supervisor's ratings of medical knowledge, data collection, clinical judgement and professional attitudes in the first postgraduate year were unrelated to undergraduate grades in paediatrics. The highest correlation in internal medicine was found between senior grades and clinical judgement (r=.36). Stronger relationships were found in obstetrics and gynaecology with a correlation of r=.49 between ratings of medical knowledge and junior grades; 24% of the variation in rating being accounted for by undergraduate grades. It was concluded that relationships between undergraduate achievement and postgraduate performance may be masked by pooling across specialty areas and across various levels of performance.

The effects of continuing medical education on medical knowledge and performance have been studied by a number of authors. Sibley et al. (1982) used a randomized trial design to estimate the impact of continuing medical education on change in knowledge and performance among 16 family physicians. Improvements in medical knowledge were demonstrated in the treatment group for both preferred and unpreferred topics. A similar gain in medical knowledge was also demonstrated in the control group for preferred topics although no formal intervention was provided. Quality of care provided improved modestly in both the experimental and control groups by 5% and 2% respectively. The only significant differences in quality of care were in low preference conditions where the experimental group improved by 10% while the control group remained the same.

Jennett (1988) used a randomized controlled trial to study the effect of continuing medical education (CME) on change in the quality of care provided by 31 volunteer family physicians. Quality of care was studied by chart audit pre-intervention and at 6 and 12 months post-intervention. Significant changes in the use of targeted preventive strategies were demonstrated in both intervention groups; 43.8% when compared to a 10.7% improvement in the control group for cancer detection

and 44.7% in contrast to a 4% improvement in the control group for hypertension control. These differences persisted for 12 months only in the hypertension group only.

Cohen et al. (1985) used a less individualized approach to CME among 85 internal medicine residents. Change in the use of selected preventive practices was the outcome measured by chart review; the intervention consisting of selected readings. Although there was a change in knowledge demonstrated as a result of reading there was no change in the use of these preventive practices despite the stated intention to do so.

Greenberg and Jewett (1985) report a difference in knowledge gain with two formats of continuing education among 23 participants of a paediatric continuing medical education course. The lecture method was associated with a 29% gain; the case presentation method with a 64% gain. Standardized patients were subsequently used to evaluate the impact of CME formats on patient care and the relationship between knowledge and performance. Knowledge was not associated with the quality of diagnosis or treatment plans in the four conditions studied.

Other methods of educational intervention which include feedback and corrective information on performance have demonstrated more consistent short term gains in performance (Winicoff et al., 1984; Putnam & Curry, 1985; Kroencke et al., 1987). Since the target of these interventions is not a change in knowledge per se, they have not been included in this review.

Summary

Medical knowledge has a weak,positive relationship to the quality of performance rated by supervisors or colleagues. Estimated relationships between knowledge and performance are not as large as those between knowledge and competence. This may be a function of the reliability of the performance measures or the influence of more powerful determinants on performance. These will be reviewed subsequently. A change in medical knowledge has an impact on the overall quality of practice

performance. This would suggest that deficits in relevant medical knowledge are associated with a greater probability of substandard practice when compared to the prevailing practice norms.

Limitations

The limitations reviewed for the relationship between prerequisites and competence are all applicable to the studies reviewed in this section. The outcome measures used to evaluate performance in these studies; ratings by supervisors and colleagues, and chart review have their own limitations. For the purposes of understanding the knowledge performance relationship, it is important to note that the use of supervisor's ratings as the outcome measure may attenuate the estimate of the true relationship by poor reliability, the measurement of attributes unrelated to performance and the constriction of the range of true performance variation. Chart audit is flawed by errors in omission and commission. The former has been estimated at 25% in similar study populations. (Norman et al., 1985; Page & Fielding, 1980). The potential bias created by this problem could be in either direction.

THE RELATIONSHIP OF PREREQUISITES TO THE CLINICAL SITUATION

Overview

Competence and performance are situation dependent. What is required for a competent performance in one situation may be quite different from that required in a different situation. This is a fundamental premise of the theoretical model described in Chapter 1 and is consistent with the models and definitions of competence proposed by others (D'Costa, 1986; ABIM, 1979; Broski et al., 1977; LaDuca et al., 1984).

Knowledge, skills, judgement and attitudes required for the demonstration of competence in different clinical situations would be expected to vary. The evidence related to this supposition is meagre. The results of four studies which addressed this issue will be reviewed.

The Evidence

LaDuca et al. (1984),through an intensive study of primary physician-patient encounters in the United States, constructed a framework of clinical situations and applied it to the analysis of prerequisites of competence in the design of a national licensing examination. The 102,705 records of patient encounters of physicians practising in six disciplines and 8 setting-care classifications were grouped by diagnosis (ICDA-8). A total of 137 unique medical problems were identified.

These problems were subsequently grouped into 13 clinical management paradigms which in the opinion of experts required similar provider competence. These paradigms represented service location, problem acuity, problem type and type of service delivered. An analysis of the components of competence and knowledge required for situations within these paradigms led to a further sub-classification of the paradigms into 40 subgroups. Other presumed prerequisites (skills, judgement and attitudes) were not addressed.

The Manitoba Association of Registered Nurses (1984) conducted a study of the expected competence required for nurses practising in 7 major clinical areas. Expected competence was determined by a random sample survey of provincial nurses. The identification of common and important clinical situations, specific competence required and related prerequisite abilities (knowledge, skill and experience) were identified by expert nursing panels in each of the clinical areas. A content analysis of prerequisite abilities identified independently in each of the seven clinical areas produced groups of abilities common to all clinical situations along with those which were specific to specified clinical areas or situations. Similar to LaDuca, (1984) clinical competence required of the provider varied by situation. Approximately 50% of the knowledge required was specific to clinical areas or situations. The interpersonal, cognitive and psychomotor skills required depended to a much greater extent on the clinical area or situation (approximately 80%). Experience, as a proxy index of judgement and attitude, was similarly sensitive to the clinical area and situation.

The study of Sibley et al. (1982) provides an interesting insight into the specificity of knowledge required for different clinical situations. Knowledge supplementation targeted for specific clinical problems did not have a spill-over effect on performance for a hidden tracer condition used in the study.

Summary

The knowledge and skills required for a competent performance appear to be sensitive to the clinical situation in both medicine and nursing. The end result of judgement, competence in diagnostic and management decisions, has been the usual avenue for study of this prerequisite. The relationship of the latter to the clinical situation will be reviewed subsequently. Prerequisite attitudes have been difficult to articulate and hence hard to measure. They are usually translated into desirable behaviour which should be demonstrated in patient care (eg. responsibility, interpersonal relationships). They are usually measured in a manner which would not allow their relationship to the clinical situation to be elucidated.

Limitations

Expert opinion is the basis for the majority of evidence presented in support of a relationship between prerequisite abilities and the clinical situation. Prevailing beliefs about the abilities required for practice may bias the expert's opinion about the knowledge and skills required in certain clinical situations.

There needs to be a better understanding of the specificity of prerequisite abilities to the clinical situation. This issue is fundamental to the development of the content of curriculum for professional formation and for the remediation of substandard performance. Must the knowledge and skills for all clinical situations the provider will encounter be covered in curricula or tested in licensure/certification examinations or covered in targeted remedial activities? Before addressing these issues we must be confident that

important components of competence have been identified, that they are predictive of performance and that performance when rendered according to optimal professional standards is an important determinant of health outcome.

CLINICAL COMPETENCE AND PERFORMANCE

Competence and performance are differentiated by the context rather than the content of measurement. Competence is an estimate of the provider's actual ability to perform. Performance refers to what the provider does in day to day practice. Typically performance is the target of quality assurance mechanisms where as competence traditionally is the target of licensure and certification (D'Costa, 1986; Holmes, 1986; Senior, 1976; Eisenberg, 1986).

Components of clinical competence have been identified and are listed in the definition (Chapter 1). They describe features of the process of care that the provider should demonstrate in clinical situations. They are believed to be important determinants of optimal patient outcome.

The relationship among components of clinical competence has been examined by a number of authors (Maatsch, 1983; Arnold et al., 1984; Verhulst et al., 1986; Klass et al., 1988). The objective of these studies has been to determine whether one attribute or several independent attributes of provider behaviour are being measured? This issue is of relevance to those who are required to make decisions about academic progress, licensure and certification.

Although the number of attributes identified has varied from one to four, the grouping of components has been rather consistent; technical skills in data collection and management, intellectual abilities relevant to diagnostic and management decision-making, professional relationships with colleagues and communication skills with patients and families. The number of factors identified is related to the number of components actually measured. The accurate measurement of all components considered to be important is not a feature of any of the studies reported.

The likelihood that competence represents several underlying and independent behaviourial attributes poses a problem for those required to render a decision about competence or incompetence for practice. In order to protect the public from unsafe, costly or ineffective service, groups charged with the responsibility of certification and licensure decisions will need to know which components of competence are predictive of safe, effective performance and optimal patient outcome.

The relationship of competence to performance and outcome will be reviewed in the next sections. It is hypothesized that the provider's ability to perform will be predictive of actual performance and that the quality of performance will be associated with the probability of achieving better health outcomes. It is also hypothesized that the clinical situation will modify required competence, performance and expected outcome (i.e. competence and performance will vary across clinical situations and the relationship of provider competence to performance and outcome will also be situation dependent).

THE RELATIONSHIP BETWEEN COMPETENCE AND PERFORMANCE

Overview

A number of studies have examined the relationship between the competence of the provider in standardized clinical situations and their subsequent or contemporaneous performance in practice. Clinical simulation methods or observed performance with real patients have been the methods used to evaluate components of competence. Ratings by colleagues/supervisors, chart audit, billing data, and blind evaluation using standardized patients have been the methods used to measure performance in practice. In order to understand the relationship between competence and performance, additional provider and system-related determinants of performance must be taken into consideration. These will be reviewed subsequently.

The Evidence

The Relationship Between Competence and Performance

Competence and performance relationships have been most frequently studied in populations of medical students and residents. Supervisor's ratings have been the most common measurement of performance. Clinical competence has been evaluated by a number of methods.

The Clinical Examination Exercise (CEX) was recommended by the American Board of Internal Medicine (ABIM) for the evaluation of competence in data collection, diagnosis, proposed management and patient communication. Direct observation and scoring of a patient work-up and subsequent record documentation are used to measure competence. The relationship of scores achieved in this exercise with supervisory ratings has been evaluated in two studies. Kroboth et al. (1985) found a correlation of r=.30 between competence and performance scores among 27 house officers examined. The largest correlation was .36 for ratings of medical history, the lowest was for scores on physical examination (r=.14). The authors suspected that this low correlation represented the absence of rigorous evaluation of physical examination ability in practice. The study by Wooliscraft et al. (1984) provides credence to this explanation. They found that the 120 medical house officers examined demonstrated recurrent inadequacies in the social and family history, mental status and neurological exam. Furthermore the technical quality of the interview/exam correlated with the accuracy of findings elicited (criterion=faculty). Despite assiduous attention to inter-rater reliability, a similar modest correlation was found between competence and performance measures.

Competence as measured in clinical simulation formats has been compared with peer and supervisors ratings of performance in both house officer and practising physician populations. Norcini (1986) found a correlation of r=.26 for the relationship between performance on Patient Management Problems (PMP) and peer ratings of performance among practising physicians. A slightly higher correlation was noted (r=.28) when competence was measured by an extended computer case simulation. The largest correlation in this latter format was between diagnosis and peer

ratings (r=.92 when corrected for reliability) followed by competence in therapy (r=.57,corrected). Low associations were found for all other components (history, physical exam and lab investigations). Similar results were found in the house officer group with correlations of r=.30 to .36 found between PMP performance and supervisory ratings (Norcini, 1987).

Maatsch et al. (1983) used a different measure of performance, stimulated chart recall, and compared the resulting scores with those achieved on measures of competence with standardized patients. In a sample of 182 candidate for emergency medicine certification, he found a similar magnitude of relationship to that found by others (r=.33, corrected r=.45).

Goran et al. (1973) improved on the design of the previous studies by comparing competence scores achieved on a PMP exercise with performance on the same problem in the ambulatory clinic in 22 clinic teams (n=35). Chart review was used to derive measures of performance on 33 clinic patients. A 12% difference in the number of actions performed on the PMP versus real patients was noted. Although this difference could be accounted for by errors of omission in charting, a dramatic difference in the ordering of urinary cultures was noted (46%) which would not be accounted for by this explanation. The cuing inherent in the PMP format has been offered as the most likely explanation of this observation.

Rethans and Boven (1987) used the same clinical problem as Goran in an uncued form to study the relationship between competence and performance. Potential differences in patient mix as contributors to performance variation were controlled through the use of standardized patients sent blind into practice settings. In this study, minor differences between competence and performance scores were noted in the 48 physicians studied. The differences which were noted may be explained by the artificialities of using a written format to measure competence in patient communication.

Page and Fielding (1980) used a similar design to that of Rethans and

Boven (1987) to study the relationship of competence, assessed by PMP performance to practice performance measured by standardized patients. The subject population was 16 practising pharmacists and the relationship was analyzed in four common practice situations. In this study, performance with a cued PMP was substantially better than that in practice, predicting less than a third of practice behaviours. Correlations ranged from .26 to .68. The artificialities of the PMP format probably explain some of these differences. Since other determinants of performance were not taken into account in this study, it is not possible to determine the extent to which competence versus other determinants contributed to the observed variation in practice performance.

The Relationship Between Performance and Competence

A bi-directional arrow between competence and performance is evident in the model presented in Chapter 1. It represents the hypothesis that competence is influenced by the opportunity to perform in actual practice. Although a provider may be competent to perform at one point in time, it is hypothesized that the absence of opportunities to exercise this ability will have negative consequences for the continued competence of the provider. The evidence to evaluate this hypothesis is scant. There is a relationship between patient volume, problem-related experience and performance which would lend support to the existence of this relationship (to be discussed in the next section). Stross (1983) found that in 132 physicians tested one year after advanced cardiac life support (ACLS) certification that only 39.4% could still maintain adequate ventilation of the mannequin and 47% could maintain adequate cardiac compression. This deterioration in performance was not influenced by selective educational strategies aimed at improving retention. Although the interim experience of physicians varied re:opportunities to perform ACLS on the job,the relationship of experience to maintenance of skills was not evaluated.

Summary

Measures of competence appear to be predictive of performance when it is evaluated in the same clinical situation, using the same criteria with a format which does not artificially cue the provider about clinical actions and decisions. The generalizability of this observation to all clinical problems and other subpopulations of physicians is unknown. A very modest relationship is consistently found between various measures of competence (PMP, computer and standardized patients) and supervisor's or stimulated chart recall ratings of performance.

There is little evidence to support or disprove the hypothesis that opportunities to perform in practice influences the development or maintenance of continuing competence.

Limitations

There are limitations in both the content and reliability of supervisor's ratings of performance. They, provide no means of controlling for differences in case mix for different individuals, a problem which may confound the estimated relationship between competence and performance. Stimulated chart recall has the same problem along with the associated difficulties of reliability with oral evaluations (Muzzin & Hart, 1985). The study by Rethans & Boven (1987) and Page & Fielding (1980) standardized both patient mix and the content of the evaluation in the study design. The limitation in these studies is one of generalizability and potential biases created by the competence measurement format. None of the studies reported examined the relationship of competence, performance and other provider and system related determinants of performance variation.

THE RELATIONSHIP OF COMPETENCE AND PERFORMANCE WITH OTHER COMPONENTS

THE RELATIONSHIP OF PROVIDER CHARACTERISTICS AND SYSTEM RELATED DETERMINANTS TO PERFORMANCE

Overview

The observed institutional, regional and national variation in physician practice has led to a number of studies which have evaluated potential determinants of performance. This research has been motivated by a concern for the costs incurred in health care delivery and the quality of care rendered. For these reasons variation in resource utilization and measures of quality of care have been the usual endpoints studied. No study to date has included the evaluation of all determinants identified. The relative and independent contribution of each of the determinants reviewed is therefore unknown. In addition, the competence of the provider to deliver care has not been studied in relationship to these determinants. The relationship between competence and provider and system related determinants is therefore unknown.

The Evidence: Provider Characteristics

Of the provider related determinants identified; three serve as proxy indices of competence at the completion of training (length of clinical training, certification status, and location/quality of the training program); four serve as proxies for continuing competence (age/years since graduation, continuing medical education (CME) involvement, medical knowledge, and health problem experience).

Personal characteristics of the provider (gender, practice philosophy, risk taking behaviour, socio-economic background, religion, political persuasion and ethnic background) and the practice setting have also been studied. The impact of these factors on the providers ability to perform (competence) has not been studied.

Proxy Indices of Competence at the Completion of Training

The impact of the length of clinical training and certification status on quality of care has been examined in a number of studies (Ramsay & Beminoff, 1981; Butterworth & Reppart, 1960; Clute, 1963; Morehead, 1958; Peterson et al., 1956; Rhee, 1976, 1977; Arnold, 1970; Trussel et al., 1962; Committee V.A., 1977). The length of clinical training appears to only be of benefit when physicians are practicing in the domain of their specialty (Morehead, 1958; Rhee, 1976). Two early studies found that the length of clinical training was only associated with better quality of care when carried out in approved training programs (Morehead, 1958; Clute, 1963).

The relationship of clinical training to diagnostic accuracy was evaluated by Berwick and Thibodeau (1983). They estimated the impact of the length of clinical training on the accuracy of the clinician's predictions of the results of chest x-ray and throat cultures ordered in the emergency room. Length of training was only associated with the accuracy of prediction of x-rays results. The range of training studied was quite narrow which may have attenuated the estimate of the true relationship.

Length of clinical training also appears to be an important determinant of the quantity of resources used in patient management (Cherkin et al., 1987; Young et al., 1987). Geeritsma & Smal's (1986) study of the clinical reasoning process of 16 family physicians and 16 internists in the Netherlands provides some insight into the potential sources of these differences. Differences between internists and family physicians were predominantly observed in the first patient encounter where internists tended to spend more time, ask more questions, carried out a more extensive physical examination, and ordered three times as many lab tests for the same problem. Despite these differences in process, no differences in patient management were found. The complexity of the case appeared to be an important factor in determining the extent of the work-up. Family physicians tended to carry out a more extensive work-up in complex cases

than internists; these differences were presumed to be due to differences in experience or confidence in the physician's ability. This observation is consistent with that of Young et al. (1987) who studied threshold differences in ordering coronary angiography between family physicians and cardiologists. Family physicians had a greater tendency to order angiography and were less confident than internists about the probability of cardiac disease. Differences in confidence about the probability of disease may provide one explanation for variation in test-ordering among physicians with different types and lengths of clinical training.

Certification status has been associated with better quality of care in three studies (Arnold, 1970; Morehead, 1958; Trussel et al., 1962). No relationship was found in three studies, all of which were more recent (Committee V.A., 1977; Peterson & Barsamian, 1976; Rhee, 1976). The Stanford study of post-operative morbidity and mortality in 17 hospitals in fact found board certification to be associated with worse outcomes after adjustment for pre-operative status (1974).

It has been hypothesized that training may act as a determinant of the content of practice, with physicians selectively carrying out services in areas where they feel they are more competent. Variation in service would then be explained by the provider's efforts to reach a target income by providing more services in areas in which they were competent. Curry (1985) compared the practice content of general practice and family practice trained clinicians using provincial billing data and found no differences in practice content between the two groups.

The relationship of the quality of the training program with the quality of care delivered by its graduates has been primarily studied in comparisons of U.S./Canadian and foreign medical graduates (it is assumed that foreign programs are of inferior quality). The foreign medical graduates studied are limited to the subset who were able to gain access to the U.S. system. Rhee's (1986) study of the quality of care delivered by 1,156 physicians in 14,203 patient encounters found no differences in the quality of care delivered by foreign medical graduates. In some tracer conditions, foreign medical graduates tended to do better. This may be

attributed to the observation that U.S. graduates were more likely to practice outside of the domain of their specialty. Saywell et.al (1980) studied differences in the documented quality of the history and physical examination among 556 U.S. and 342 foreign graduate house officers in 14 hospital settings. No main effect for training location was found. Hospital and diagnostic category were the main contributors to variation in quality.

In contrast, Stillman et al. (1986) in a study of the clinical skills of internal medicine residents from 14 New England programs did find that the program's academic reputation was associated with both the average performance score obtained and the range of scores observed. Better and more homogeneous scores were achieved by residents from programs with a strong academic reputation.

Indices of Continuing Competence

The impact of medical knowledge on variation in the quality of practice performance has been assessed in the evaluation of CME programs. This literature was reviewed in the section dealing with prerequisite, performance relationships. From these studies it can be concluded that medical knowledge acts as a significant but minor contributor to variation in the quality of care. The number of CME credits accumulated, on the other hand, has had no demonstratable relationship to variation in the quality of care. (Clute, 1963; Lewis & Hassanein, 1970). CME attendance is obviously a particularly crude index of practice related learning.

Age or years since graduation could conceptually be viewed as an index of either the potential for improved competence resulting from practice experience or deterioration in capability to perform consistent with state of the art standards of care. In keeping with the latter supposition, most studies have found that younger physicians provide better quality of care than older physicians (Butterworth & Reppart, 1960; Clute, 1963; Hulka et al., 1976; Evans et al., 1986). In addition, Rosenblatt & Moscovice (1984) found that older physicians were more likely to admit patients to hospital when patient characteristics, occupancy rate, clinical training and

practice setting were taken into account.

The results of a number of studies suggest that this relationship is more complex. Payne et al. (1984) studied the quality of care delivered by 1,156 physicians in 5 practice settings. Although there was an overall tendency for physicians with less than 10 years of practice experience to do better, those with 10-19 and >20 years of experience did better in certain conditions. Rhee (1976), in contrast, found that physicians with less than 6 years of experience provided the worst care, the best care being provided by those with 6-15 years of experience with quality of care declining thereafter.

This disparity in results could be accounted for by differences in the population range and characteristics of physicians sampled, modelling assumptions or by the tracer conditions used in the evaluation. It is plausible that experience may be of benefit in some clinical situations particularly where major changes in treatment have not occurred. On the other hand, in situations where management standards have changed (eg. hypertension management, preventive health practices), the older physician may be handicapped if they have been unable to keep up to date. Lomas and Haynes (1987), for example, report that the use of blood pressure screening (a relatively new practice) is inversely related to age. Battista et al. (1986) found cervical cancer detection scores to be higher in younger physicians but the same relationship did not hold for breast cancer, colorectal or lung cancer screening practices.

One of the difficulties with using age as an index of practice experience is that it is a crude measure of the provider's actual experience with the tracer conditions being studied. If competence is situation specific and improves with experience, the relationship would be masked by classification errors in the measurement of the independent variable. A few studies have estimated the relationship between volume of practice experience with the tracer condition studied and provider performance. Eisele et al. (1956) found a significant association between case volume and the audited quality of diabetic management. Graham & Paloucek (1963) noted that physicians seeing fewer than 25 patients/annum had higher case

fatality rates for cancer of the cervix. Thibodeau and Berwick (1980) noted variation in the diagnosis of otitis media among interns and residents rotating for 6 weeks in the emergency room. The frequency of diagnosis was associated with the number of weeks spent in the emergency room with more errors of commission being made early in the course of the rotation. Norman et al. (1988) found a similar relationship between experience and diagnostic accuracy in a study of physician competence in the diagnosis of dermatological conditions. Although all studies point to the presence of a relationship between volume of experience with the condition studied and the quality of performance, we do not know whether this is true for all clinical situations.

Personal Characteristics of the Provider

The gender of the physician has been associated with variation in breast cancer detection services (Battista et al., 1986), the use of hysterectomy and prescription of diazepam (Lomas & Haynes, 1987). Female physicians are better represented in the younger age groups however the tendency to provide a different approach to female health problems persists after age has been taken into account in the analysis.

Verhaak (1986) examined the influence of practice philosophy on variation in psychosocial diagnosis and the elicitation of psychosocial findings in the patient encounter. Physicians who were classified by questionnaire as being science, cure and intervention oriented were less apt to elicit psychosocial complaints and entertain psychosocial hypotheses of symptom etiology.

In a related area of study, Hull (1979) examined physician's self-reports of psychiatric referral in relationship to their age, residential background (urban/rural), religion, ethnicity, socio-economic background, and political persuasion. He found that older physicians, those of the Jewish faith, those of Eastern European background and those coming from mid to upper socio-economic backgrounds were less apt to report the use of psychiatric referral. He hypothesized that these factors represented a difference in cultural and social mores which conditioned the physician's

response to mental health conditions.

Nightengale (1988) examined the influence of the physician's predisposition to take risks in treatment on admitting patterns in the emergency room. Physicians who were more willing to take risks to improve the possible outcome of a clinical problem were more apt to admit patients to hospital in situations which would be considered more discretionary (2.3 patients per shift vs. 1.38/shift in the low risk group).

Practice Setting Characteristics

Consistent and systematic differences have been demonstrated in the quality of care provided by group and solo fee-for-service practices, with the quality of care being better in the former (Eisenberg et al., 1974; Kahn et al., 1977; Brook & Williams, 1976; Morehead et al., 1971; Payne & Lyons, 1972; Peterson et al., 1956; Clute, 1963; Roemer & Gartside, 1973). Williamson (1975) found that physicians in group practices tend to adopt drug innovations more quickly than those in solo practice. It has been hypothesized that physicians in group practice benefit from the availability of peer review and collegial input, both potentially positive factors in improving quality of care. A related hypothesis is suggested by the work of Rosenblatt and Moscovice (1984). They found that solo practice physicians were less apt to admit patients to hospital,a finding they attributed to the difficulty solo practitioners may have in covering both inpatients and their usual office practice. The quality of care which can be provided by solo practitioners may be limited by inadequacies in manpower for on-call and hospital coverage.

In the same study, Rosenblatt and Moscovice (1984) examined the impact of practice volume and scope of services on hospital admissions. Two week log diaries were used to ascertain the frequency and reason for admission among 287 physicians. Increased practice volume decreased the probability of admission. A broad range of services was associated with increased admissions even when obstetrical services were removed from the data base. The relationship of resource use to quality of physician performance was not addressed in this study. It is reasonable to hypothesize however

that there is an optimal range of practice volume. Insufficient volume (as discussed earlier) appears to lower the quality of performance, likely by a direct effect on the providers competence to perform. Excess volume could similarly lower the quality of care but for different reasons (eg. fatigue or inadequate provider time). A study by Engel et al. (1987) provides some data on the effect of fatigue on the quality of care. The quality of care delivered by fatigued and rested interns was examined in two standardized patient situations. A 12% difference in performance scores was present although not significant due to the small sample size. Cohn (1985) provides support for the effect of practice conditions on errors in performance. He comments that while errors which were made early in residency training were attributable to incompetence, those in later training were result of time constraints.

Summary

Three proxy indices of competence, length of clinical training, change in medical knowledge, and physician experience (age, health problem experience) appear to be consistently associated with variation in practice performance. An inverse linear relationship between age and quality of performance seems to be present in clinical situations requiring the application of newer innovations in medicine. A curvilinear relationship between quality of care and age probably exists for those clinical situations where provider experience, rather than application of recent practice innovation, would be expected to improve diagnostic and management decision-making. A volume threshold for problem related experience appears to be a necessary prerequisite for maintaining an adequate standard of care, particularly in situations requiring the use of manual intervention. Fatigue and time constraints are factors which have been adversely associated with quality of care. This would suggest that their is an upper threshold on the beneficial effects of patient volume on quality.

Personal characteristics of the physician influence resource utilization and quality of care in certain types of clinical situations. These include physician gender in female health problems; practice philosophy,

ethnicity, age, socio-economic background, and religion in the diagnosis and management of psycho-social problems; and risk taking behaviour in the utilization of patient resources. The relationship of these characteristics to competence is unclear. They probably influence career choice and the selective development of abilities required for the demonstration of competence in certain clinical situations.

The physician's practice setting influences the quality of care delivered and the resources used. The lower quality of care demonstrated by physicians in solo practice may be due to self-selection (less competent physicians choose to practice alone), a conort effect (older physicians are more common in solo practice), inadequacies in manpower for on-call and hospital coverage or the relative poverty of resources available for reviewing their own performance and adopting relevant practice innovation.

The Evidence: System Related Determinants

Certain attributes of the health care system are associated with variation in resource utilization and quality of care. They can be organized into three groups: factors which influence continuing competence, economic factors which influence medical decision-making, and other factors which influence quality of care.

Factors Which Influence Continuing Competence

There are a number of factors which probably exercise their effect on performance variation by providing a means of selecting for or improving provider competence. They include the organization, administration, volume and content of care delivered by institutions with which the provider is associated and the employment of systems which provide performance monitoring and corrective feedback.

University affiliated hospitals and those providing medical teaching have demonstrated better quality of care in a number of studies (Stapleton & Zwerneman, 1965; Committee V.A., 1977; Morehead & Donaldson, 1964; Peterson & Barsamian, 1976; Sparling, 1962; Trussel et al., 1962; Yankauer

& Allaway, 1958; Duff et al., 1972; Rhee, 1977; Rosenfeld, 1957). This effect is likely due to a combination of factors: physician selection, the learning effects of medical teaching and the organization and availability of a number of avenues for performance review.

Specialized services and patient volume have also been associated with better quality of care (Morehead et al., 1971; Staff Stanford, 1974;Graham & Paloucek, 1963; Yankauer, 1958; Committee V.A.,1977). These factors are probably inter-related. The availability of specialized services likely attracts an increase in patient volume which in turn serves to provide individual physicians with a sufficient number of patients to maintain their competence. Specialized services also provide a structure for the organization of ancillary health services which have their own independent effect on quality of care (Georgopoulus & Mann, 1962). Both specialization and adequate patient volume seem to be necessary conditions for improved quality of care. For example, Bloom & Peterson (1973) found that coronary care units with less than 6 beds had a worse mortality experience.

A number of administrative methods for structuring and/or correcting deficiencies in provider performance have been studied. The problem-oriented medical record (POMR) was introduced as a means of enhancing provider competence in diagnostic and management decisions and improving professional communication and review (Weed, 1969). Fernow et.al. (1978) studied the impact of the POMR on quality of care in three London teaching hospitals. Most of the variation in performance was accounted for by differences among the 28 medical and surgical firms employed in the study. Significant improvements in quality of care were only noted in the surgical firms and only in one of the conditions studied. Although compliance with the format was not studied, it seems unlikely that use of POMR, per se, has an appreciable impact on the quality of care.

The use of performance audit does appear to have an impact on care delivery when the provider is personally involved in either the specification of audit standards or the feedback provided. Winickoff et al. (1984), for example, found that providers improved their employment of

cancer detection strategies only when the audit provided them with a comparison of their performance with peers. Putnam and Curry (1985) found that physician involvement in the specification of performance standards improved their performance in those conditions with no change in a hidden condition audited during and after the intervention. Both of these findings are consistent with the most powerful determinants of learning identified by Bloom (1985): individualized corrective feedback, group involvement and relevance of the learning task.

In one study, administrative systems which required justification for the use of health resources have had a positive effect on reducing the number of unnecessary tests ordered. Kroenke et al. (1987) indicate that this provided an additional 17-19% improvement in the number of appropriate tests ordered over gains made by an educational and audit intervention alone. Administrative systems which require the provider to think about the rationale and priority of care delivery (eg. chart format, order form structure, resource limitations) probably act to increase the provider's reflection on the quality of their diagnostic and management decisions. This would be consistent with the deliberate action model of expertise development (Kennedy, 1987).

Economic Factors Which Influence Medical Decision-Making

Eisenberg (1986) describes the three roles of the physician; self-fulfilling practitioner, patient agent and guarantor of the social good. The role of economic factors in medical decision-making is related to the role of self-fulfilling practitioner and guarantor of the social good. In contrast, clinical competence is relevant to the physician's ability to act effectively as the patient's agent. All three roles have a potential to influence variation in practice performance. In relationship to the role of self-fulfilling practitioner, Evans (1984) postulates that the physician sets a target income and manipulates the number and type of services rendered in order to reach the target. If this were the case then differences in services rendered would be partly explained by the remuneration of those services (the fee-for-service schedule), the degree of local competition for available patients and the source of referral.

The impact of physician density as a proxy index of competition was studied by Hemenway and Fallow (1985) in the management of three paper case simulations. More aggressive treatment and potentially remunerative services were used by physicians from high density areas and those with a reported income of over $100,000.

Differences in the reported use of preventive services between salaried and fee-for-service physicians has been reported by Battista et al. (1986) in a study of primary care physicians in Quebec. The absence of remuneration for preventive services has been offered as a partial explanation of these findings. The salaried method of remunerating physicians in health maintenance organizations (HMO) has been cited as one of the reasons why HMO's are able to provide less costly care for similar diagnostic related groupings (Luft, 1981). Lomas and Haynes (1988) report that manipulation of the remuneration provided for obstetrical services is being used by the Ontario government to entice physicians into providing a greater number of services in this area. Differences between salaried and fee-for-service physicians in the management of tension headache were reported by Renaud et al. (1980). Fee-for-service physicians spent a shorter time with patients, provided less patient explanation and were more apt to use medications in headache management. These studies suggest that the quality of care may be adversely affected by the need to generate a certain proportion of highly remunerative services or patient volume to meet the target income goals of the provider.

Rhee et al. (1980) hypothesized that the source of the provider's practice base would influence the length of patient stay in hospital. Those who were dependent on patient self-referral for their practice base had longer lengths of stay than physicians dependent on colleague referral. The classifications employed were confounded by differences in the length of provider training.

Other Factors Which Influence Quality of Care

The availability of other health care professionals (eg. nursing, physio-

therapy) and services (eg. lab, diagnostic) are additional factors which could influence the quality of care to patients. Although no association could be found between the quality of blood-bank and pathology services on surgical morbidity and mortality, the ratio of registered nurses and nurses to patients has been associated with the quality of care delivered (Georgopoulus & Mann, 1962; Staff Stanford, 1974). A better understanding of the impact of these additional determinants on the quality of performance and outcome is required.

Summary

Factors which influence the selection and facilitation of continuing competence of the provider include: practice in a university affiliated hospital or one with teaching responsibilities, specialization of services, an adequate patient volume, and administrative methods which facilitate a review and rationalization of performance (e.g. individualized chart audit, policy for resource use). The contribution of these factors to variation in physician performance is likely realized through a direct impact on the physician's competence to perform as the patient's agent.

Additional economic determinants of performance are associated with the physician's role as self-fulfilling practitioner. These factors may act to alter the quality of physician performance particularly in circumstances where standards of management are more ambiguous (Eisenberg, 1986; Palmer & Reilly, 1979). The remuneration method, fee schedule and competition for patients are all factors which can contribute to variation in performance unrelated to the competence of the provider.

The availability and quality of other resources in the health care system likely influence the quality of care. The quantity and quality of nursing resources have been the only factors identified to date, although few have been systematically studied.

Limitations

The variables associated with provider performance have not been studied simultaneously in any study. Their relationship to each other is therefore unknown (Palmer & Reilly, 1979). Patient mix provides an important determinant of variation in provider performance. The adequacy of methods used in adjusting for differences in patient mix has been criticized by Tugwell (1979) and in some studies the contribution of patient mix to practice variation was not estimated. Standardized simulation formats were used in a number of studies to control for the contribution of patient mix to practice performance. Direct measures of competence in patient situations have not been a feature of any of the studies reviewed. The actual contribution that the practice setting, institution, administrative and economic policy makes to provider competence is not clear. The identification of effective strategies for producing a sustained improvement in performance will require an understanding of the mechanism by which these factors produce performance differences. At present it is not clear whether or not they represent physician self-selection, motivation to perform or determinants of actual competence.

THE RELATIONSHIP BETWEEN COMPETENCE, PERFORMANCE AND THE CLINICAL SITUATION

Overview

The relationship between competence and performance can be obscured by measurement in different clinical situations. The relationship between prerequisite abilities and the clinical situation has been reviewed. To the extent that professionals know what abilities are required, the clinical situation would appear to influence knowledge, skill, judgement and attitude requirements. This section will review the evidence for theoretically proposing that the clinical situation influences the competence and performance required of the provider. This premise seems so obvious that the reader might wonder why a review is relevant. The issue is one of measurement. We want to predict the ability of physicians to

perform in clinical situations they will encounter .r different performance capabilities are required in different situations, then we need to know the features of the clinical situation which will influence performance requirements to establish a frame for sampling. We similarly need to know how to define the domain of clinical situations to which we wish to draw inferences. The issue which will be addressed in this review is the evidence that the question to be addressed in measurement is not just how many problems but how many and of what kind for a stable estimate of provider competence of performance.

The next section will also review the evidence for the hypothesizing that the competence of providers influences their perception of the clinical situation and as a result their diagnostic and management decisions.

The Evidence

Standard Protocol Application VS. Situation Specific Data Collection

Two landmark studies on the clinical reasoning process of the physician negated the commonly held belief that the application of a standard and rigorous protocol of data collection across all clinical problems would result in appropriate diagnosis and management. The work of Elstein et al. (1978) and Barrows et al. (1978) found that early hypothesis generation was characteristic of most physician/patient encounters. They found that it was the quality of the hypotheses generated rather than the quantity of data collected which was predictive of the accuracy of diagnosis. In fact, in the study of Barrows et.al.(1978), physicians on average only collected 60% of relevant patient data; the items included varying from physician to physician.

The importance of hypothesis generation for patient management can be inferred from the work of Starfield and Sheff (1972). They found that the detection, diagnosis and management of abnormal haemoglobin in children was associated with the reason for ordering the test. For those ordered on a routine basis (the application of standard data collection protocols), 35% of the abnormal haemoglobin results were recognized in contrast to 72%

when the test was ordered to evaluate a symptom or diagnostic hypothesis. Group differences persisted even when haemoglobin values were less than 9gm/100ml. Two of the undetected cases resulted in hospital admission, one for a bleeding ulcer, the other for sickle cell anaemia.

The generation of diagnostic hypotheses early in the patient encounter provides a structure for selecting and interpreting relevant patient data (Fredericksen, 1984). It is hypothesized that the early generation of diagnostic hypotheses in the patient encounter is a more important determinant of diagnostic accuracy and appropriate management than the application of a standard protocol for data collection across all situations. The appropriateness of hypotheses generated will of course vary by the situation presented. Scherger et al. (1980) also found that the number of hypotheses considered varied by the situation presented. Since we wish to predict competence to perform in a situation we need to know whether the appropriateness of hypotheses generated by the individual provider along with related diagnosis and management will vary from one situation to another and if so the factors associated with that variation.

Variation in Competence/Performance by the Same Provider Across Clinical Situations

The 'case effect' has been a well recognized phenomenon in all instances where components of competence have been tested with standardized patient formats (Norman et al, 1983). Typically the performance of students or practitioners will vary from one case to another with correlations across cases being generally low. Although part of this phenomenon may be due to differences in the difficulty among cases, the usual presence of a case-subject interaction suggests that different subjects do better or worse with different cases. The relative contribution of these factors to explained variance in scores for diagnosis and patient management was estimated by Swanson et al.(in press) using generalizability theory. He found that most of the variance was explained by a case-subject interaction, followed by cases and subjects. Similar findings were noted by Tamblyn et al. (1985) in the evaluation of nursing students.

Different components of competence appear to be more or less sensitive to features of the clinical situation. Diagnostic and management competence seem to be more sensitive to the nature of the clinical situation than data collection and communication skills. It has been estimated that 25-40 clinical situations may be necessary to reliably estimate competence in diagnosis and management (in a specialty domain of practice) whereas 6-10 would be needed for data collection and 15 for communication skills (Stillman et al., 1986). Harasym et al. (1980) similarly found that hypothesis refinement, physical examination, lab investigations, and final diagnosis were problem dependent when analyzed in 71 medical students and three clinical problems. They did not however find that history-taking or hypothesis generation were context dependent. This may be a function of the lack of clinical sophistication in the population studied or the scoring system employed for these components.

Measures of the quality of performance in diagnosis and management show the same sensitivity to the nature of the clinical situation when measured. Erviti et al. (1980) used chart audit to evaluate the performance of paediatric residents on 5 tracer conditions and found low correlations across conditions studied. When analysis was limited to those subjects who had a sufficient number of care episodes evaluated for a stable estimate of performance on each tracer condition (10/condition), higher correlations were found within two groups of tracer conditions; health screening and acute medical.

Factors Contributing to Variation in Competence/Performance Across Clinical Situation

Two explanations for variation in performance across clinical situations have been considered; true effects and sampling error. The study of Erviti et al.'s (1980) would indicate that both are contributing factors to observed variation. Features of the clinical situation which contribute to these true effects have been identified in the quality of care literature.

From Erviti's (1980) study it is evident that the type of service which the provider will need to render in a clinical situation (screening

vs. acute medical management) may influence the quality of their performance.

Rhee et al. (1976) have noted that the quality of performance was a function of the specialty domain of the problem with better performance by providers in the problems congruent with their domain of training than outside of that domain. A similar finding was reported by Norman et al. (1983) with higher correlations across problems more commonly presented within one specialty than across specialties. These observations may be due to a commonality in the underlying knowledge and skill prerequisites.

Demographic and clinical attributes of the patient and their situation have also been noted to contribute to variation in performance. The patient's race, socio-economic status, gender, risk of disease or complications, living situation, likelihood of compliance, and availability of social supports have been associated with differences in the quality of care delivered and resources utilized (Rhee, 1979; Mushlin & Appel, 1976; Yergan et al., 1987; Kuder et al., 1987; Epstein & McNeil, 1985; Larsson et al., 1987; Bergman & Beck, 1986).

The structure of the clinical situation, as a potential contributor to variation in diagnostic accuracy, has been an additional attribute studied. It has been hypothesized that diagnostic accuracy improves when the problem is a more typical presentation of a disorder. Bordage & Allen (1982) found that prototypicality provided a good explanation for errors in diagnosis among physicians, medical students, and nurses. Norman et al. (1988) failed to demonstrate evidence of this relationship in the diagnosis of dermatological conditions among practicing physicians.

Variation in Perceptions of the Clinical Situation as a Function of Competence

It is hypothesized that the competence of the provider will influence their perception of the situation and as a result contribute to variation in diagnosis and management. This relationship has been examined in a few studies.

Verhaak (1986) rated the communication skills of 30 Dutch general practitioners by videotaping their performance with 50 real patients. They found that the physician's communication skills were related to both the number of psychosocial complaints elicited and the likelihood that they would identify psychosocial aspects of the problem. Differences among the populations of patients seen by the practitioners could, however, be an alternate explanation for this observation.

Held et al. (1984) found that clinical training was related to the likelihood of hearing a third heart sound generated by an artificial simulator. Cardiologists possessed the greatest sensitivity on auscultation followed by residents and medical students.

Summary

Variation in competence and performance across clinical situations has been consistently found in both the educational and quality of care literature. Low correlations across clinical situations are a function of sampling error (in achieving a reliable estimate of performance in a given condition) and true differences in competence in different clinical situations. Factors contributing to these true differences in performance include demographic and clinical attributes of the patient situation as well as presumed differences in the knowledge, skill and judgement required for the management of problems common to certain specialties. The structure of the problem may also contribute to performance variation although the evidence to date is conflicting.

Limitations

Variation in competence across clinical situations has been studied with clinical simulation methods. These methods require the test developer to recreate the factors which are of relevance in the clinical situation. These methods are limited therefore by our understanding of important situational determinants of provider competence. In addition it is customary practice to use a representative sample of situations from

the practice domain, with one sample of performance per situation. As a result it is not possible to partition the error term into sources of variance attributable to sampling error and true differences in competence across situations.

The analysis of factors influencing performance in practice is limited by the small number of physicians which are evaluated in a sufficient number of instances across all tracer conditions. Since the impact of other factors is usually the objective of study,the analysis is conducted either within tracer conditions or with a summary score across all conditions. The study by Erviti et al. (1980) provides an exception to this practice and as such contributes valuable insights into the potential contribution of sampling error and the nature of the clinical situation to score variance.

All studies reported have confined their evaluation to episode specific management. Situational factors which contribute to variation in the quality of continuing management have not been studied. In addition, little is known about situational factors which may contribute to variation in patient communication skills. Kagan et al. (1967) identified four types of patient attributes which were difficult for health professionals to manage (hostility, seductiveness, silence and talkativeness). The influence of these situational characteristics on patient communication and other components of competence has not been studied.

THE RELATIONSHIP BETWEEN COMPETENCE, PERFORMANCE AND HEALTH OUTCOME

Overview

A fundamental assumption which is made in the evaluation of competence and performance is that professionals are aware of the components of the provider's process of care which are important determinants of the outcome. This assumption has been challenged by those involved in the measurement of quality of care. A number of studies have been conducted which evaluated the assumed relationship between the quality of care and indices of health outcome. Quality of care has usually been evaluated by

chart audit using standards established by a group of experts. No studies have examined the relationship between competence, performance and outcome.

The model of competence presented in Chapter 1 proposes that the outcomes experienced by the provider's patients may influence their subsequent diagnostic and management practices with other patients. The scant evidence which is available to evaluate this premise will also be reviewed.

The Evidence

Performance to Outcome

Nobrega et al. (1977) evaluated the relationship between provider performance and outcome in 138 patients who were newly diagnosed as hypertensive. Eighty-three process criteria were identified by hypertension specialists and general internists. The components of performance studied included: data collection (history & physical-49 items); lab and diagnostic procedures (22 items); medical management (18 items). No relationship was found between the number of process criteria met and diastolic blood pressure ascertained by nurse follow-up. This finding was consistent across the three classifications of patient severity and within all categories of criteria measured.

Lindsay et al. (1977) studied the relationship between process and outcome in patients discharged after their first myocardial infarction. Process was evaluated using eight care standards along with standards specific to the management of complications and new events. The presence of cardiac symptoms, compliance and hospitalization were the outcome measures employed. A relationship between the process of care and outcome was found only in the quality of the process at the first visit post discharge and survival. The two year follow-up performance score was unrelated to the outcomes measured. This study suggests that the quality of provider performance may be of greatest importance at certain critical points in clinical course of the disease. This conclusion is supported by Sanazaro's

(1978) multi-centre study of process outcome relationships in 7 tracer conditions. The provider's adherence to treatment standards in the first 48 hours of care was predictive of the outcomes studied in acute MI's and bacterial pneumonia.

Lindsay et al. (1976) also studied process outcome relationships in the management of acute bacterial cystitis. Symptoms 6 months after the index episode and urinary culture were the patient outcomes studied in 42 patients. Performance was scored by retrospective chart audit. No association was found. Since only 2/42 urine cultures were positive, the power of this study to detect clinically important differences was inadequate. Potential differences in case mix were also not taken into account in the estimate.

Starfield and Scheff (1972) found an association between provider performance and outcome in 52 children who had low haemoglobin. Improvement in follow-up haemoglobin values was associated with providers who recognized, appropriately diagnosed, and managed the problem.

Greenfield et al. (1977) used a criteria mapping approach to assess performance in the management of patients presenting with chest pain in the emergency room. The outcomes assessed included death and subsequent hospitalization. Associations between the quality of the process and outcome were found. This has been attributed to the criteria mapping approach which allows the relationship to be studied in homogeneous subsets of patients with performance criteria which are of particular relevance to their situation.

Patient satisfaction is another outcome of the provider patient relationship which has been studied. Although its inclusion as an outcome index has been appropriately challenged, it will be reviewed in this section as one of the commonly employed intermediary outcomes of both health care programs and provider services (Palmer, 1976). DiMatteo (1979, 1980) has studied the attributes of provider performance which influence patient satisfaction in hospital and ambulatory patients. The amount of provider contact time and the ratings of the socio-emotional quality of

the encounter were the two most important predictors of patient satisfaction and patient ratings of the competence of the provider. The socio-emotional aspects of the provider's performance were similar to those identified by Falvo and Smith (1983). They included the physician's ability to meet the patient's expectations, inform the patient about their treatment and problem, listen to their concerns, and take their needs into consideration in the construction of a treatment plan. Gender and socio-economic status were also associated with differences in the rating of patient satisfaction.

The study by Evans et al. (1986) provides the only data on the relationship between the possession of prerequisite knowledge and patient outcome. In an innovative design which permitted control of differences in patient mix through random allocation, physicians knowledge of hypertensive management was evaluated in relationship to their patients' diastolic blood pressure. No association was found between knowledge and outcome although knowledge was inversely related to the year of graduation.

Cohn (1985), in an anecdotal recount of his surgical residency experience, reports that deficiencies in competence early in his residency were associated with adverse patient outcomes (morbidity and mortality). Deficiencies in time were more likely to account for adverse outcomes in the latter parts of his training. No empirical investigations of this relationship have been reported.

Outcome to Performance

Lockyer et al. (1985), in their survey of specialists and general practitioners in Alberta, identified that the most frequent precipitant in the physician's adoption of a new clinical policy was perceived benefit for patients. Specifically, a search for alternate clinical strategies was initiated when patient's were having problems making progress with current therapy.

What is not known is the extent to which the adverse and beneficial

patient outcomes which have been experienced by the provider alter their subsequent performance with other patients. The influence of the physician's recent experiences with clinical situations in their practice has been addressed in two studies. Norman et al. (1988) noted that they influenced the diagnosis of dermatological conditions (diagnostic bias) and Lomas and Haynes (1988) comment on their influence on management policy.

Summary

A relationship between the quality of the provider's performance and selected patient outcomes has been found in some studies. The outcomes studied to date include mortality and morbidity (in terms of lab data, symptoms and physical examination findings). The quality of the provider's process appears to be of critical importance at certain points in the patient's clinical course. A failure to find a relationship in some studies may be explained by this observation, differences in impact which could be expected in different clinical situations or a result of problems reviewed in the subsequent discussion of limitations.

In the one study reviewed, knowledge was not associated with patient outcome. The socio-emotional aspects of the provider's perfonmance do have a relationship with patient satisfaction. The relationship of patient satisfaction to health status is not clear.

Little information is available on the relationship of outcomes to performance. Although outcomes are considered to be theoretically important determinants of practice performance, the existence and nature of the relationship requires systematic investigation.

Limitations

Tugwell (1979) and McAuliffe (1978) provide a critical review of the work to date. Tugwell (1979) criticizes the narrowness of outcome measures selected suggesting that adequate measurement of health status should involve the evaluation of physical, social, and emotional function. From a

methodological perspective, the research design must provide the means of controlling for differences in patient mix. To avoid problems of selection bias, patient eligibility should be by presenting symptom rather than diagnosis.

McAuliffe (1978) raises two important issues. The process standards and outcomes selected for evaluation should be conceptually related. The weight placed on standards for data collection in some studies would not be expected to have a conceptually plausible or direct relationship to morbidity and mortality. He comments that performance scores should be constructed to reflect the quantity of those performance attributes which we believe are meaningfully related to the outcome (i.e. a greater emphasis on the measurement of management provided).

Secondly, he comments that the relationship of quality of performance can not be understood if studied in situations where effective standards of care are unknown. If the efficacy of management standards have not been demonstrated then the absence of a performance outcome relationship could well mean that the standards of performance themselves have no relationship to health status.

All studies have used chart audit as the means of evaluating the extent of provider compliance with expected care standards. Performance outcome relationships are likely attenuated by the problem of under reporting, failure to measure and adjust for other important covariates and inadequate power.

CONCLUSIONS

This chapter has reviewed the components of the model presented in Chapter 1 along with the evidence for the theoretically proposed relationships. The bulk of the evidence has been derived from observational studies. The major limitations which must be considered in interpretation have been addressed in each section. This section will summarize what is known about these relationships, review the implications for clinical competence evaluation, and identify further areas of relevant research.

SUMMARY OF THE EVIDENCE

The relationship of prerequisite abilities to competence has been confined to a study of the relationship of medical knowledge to competence, provider performance and patient outcome. A weak but consistent relationship is found between medical knowledge and competence. Change in medical knowledge is similarly associated with small gains in provider performance which may not be sustained over time. No relationship was found between knowledge and patient outcome in the one study in which this was examined. The relationship between knowledge, competence, performance and outcome may have been attenuated by unreliability of competence, performance and outcome measures, the measurement of knowledge unrelated (both in level assessed and content) to the tracer condition studied or the truncation of scores in the populations available for study.

Competence appears to have a strong association with performance when examined in the same providers in a common clinical situation. The relationship between measures of competence and supervisor's ratings is consistently positive but of a low order of magnitude. Differences in the domain of clinical situations in which competence and performance were measured may provide one explanation for this phenomenon. The reliability of supervisor's ratings along with differences in the actual attributes being measured may provide additional explanations for these findings.

The contribution of other determinants of performance has not been taken into consideration in those studies which have measured competence. The relationship between these determinants and the competence of the provider remains unclear. Specific provider and system determinants seem to be consistently associated with variation in performance. Some are likely related to the competence or continuing competence of the provider. They include the length of clinical training, the age of the provider or years since graduation, the provider's institutional affiliation (university, teaching hospital), patient volume and the availability of avenues in the practice setting for individualized review and corrective feedback on performance. Additional personal attributes of the provider are associated with variation in performance in specific clinical situations. Their

relationship to provider competence is unclear but they probably predispose the provider to selectively develop competence relevant to certain situations. They include the provider's gender, practice philosophy, religion, ethnicity and socio-economic background.

Other determinants are probably unrelated to provider competence. They represent environmental conditions or incentives which can influence performance. They include the adverse effects of fatigue and time constraints which may be associated with excess patient volume and a busy practice setting. These factors in turn may be related to institutional policy or practice setting characteristics which induce excessive demands for manpower coverage. Alternately they may be due to economic policies which create competing incentives and determinants of provider performance. Included in this latter list are the methods of provider remuneration, fee schedule structure and policies aimed at resource restraint.

The clinical situation is a modifier of provider performance. Different prerequisite abilities and components of competence/performance are required in different clinical situations. Variation in competence and performance for the same provider across clinical situations has been found in all studies. This is likely attributable to both sampling error and true effects. The attributes of the clinical situation which contribute to variation in competence and performance required further study. Those which have been identified include the underlying medical condition (its etiology,severity and related risk factors); the age, sex, race, living situation and socio-economic status of the patient and the type of management.

The quality of provider performance appears to be an important determinant of certain patient outcomes in some but not likely all clinical situations. The impact of the provider on patient outcome appears to be more critical at certain points in the course of the illness. The components of clinical competence or performance which are important determinants of the social,psychological and physical aspects of health status have not been studied. Adequate control for differences in patient mix, the evaluation

of performance with conceptually and efficaciously supportable standards and broader indices of health status are important design considerations for future study in this area.

THE IMPLICATIONS FOR THE EVALUATION OF CLINICAL COMPETENCE

The rationale for evaluating clinical competence was to predict those who were more likely to provide safe and effective services and in turn be more likely to render improved health outcomes. The reason for doing so is to protect the public from the unnecessary adverse effects of incompetent care and to maximize their potential health status through effective service delivery.

The available evidence would suggest that the customary practice of assessing the knowledge necessary for practice is a weak predictor of subsequent practice performance. Assessment of competence in a standardized clinical situation has the potential to be a better predictor of performance. However the question of how many situations and of what kind for reliable estimates to a prespecified domain of practice remains unanswered. In addition, the components of competence which are of particular importance in improving health status are poorly understood. This issue is of particular relevance to the decision of not just what to measure but how important observed deficiencies might be for the patient's welfare.

Of equal consideration is the importance and relative contribution of additional provider and system related determinants to performance. We do not know how these factors influence the provider's competence to perform. Their relative contribution to variation in performance among homogeneous groups or providers is similarly poorly understood. There is a real possibility that economic policy and structure within the practice setting may be considerably more important determinants of performance variation. If this is the case, then professional licensing and standard setting bodies might better address their attention to these factors rather than the current preoccupation with entry level competence.

ISSUES FOR FURTHER RESEARCH

The relationship of competence to performance along with its various additional determinants will be an important area to address. This will provide an empirical basis for determining the relative priority of various mechanisms of assuring the public of the probability of effective service.

We need to understand what components of competence/performance are important for improving various dimensions of health status and the relative value that society places on those dimensions if we are to make meaningful decisions about competence to enter or remain in the profession. For the purposes of measurement, we need to know what attributes of the situation contribute to differences in the competence required so that a sound framework for sampling may be established.

Finally, we need a method which will allow us to measure the components of competence and performance that we theoretically feel might be important. It is this issue which is the subject of this thesis. The next chapter will review the current methods of measurement, the components of competence which they evaluate, and related measurement issues.

PART II - THE MEASUREMENT OF CLINICAL COMPETENCE

ABSTRACT

CHAPTER 3

THE MEASUREMENT OF CLINICAL COMPETENCE

This chapter reviews the general properties of measurement instruments. Reliability and validity are identified as general properties which are of concern with all measurement instruments. Reliability is defined as the degree to which the empirical measure is free of random error, (a condition necessary for the detection of associations between concepts of interest). Validity is defined as the strength of the relationship between the empirical measure and the conceptual entity, in this instance clinical competence. The absence of systematic bias and random error in the empirical measure is a necessary condition for instrument validity.

The measurement of clinical competence is complex. Eight components of the procedure are identified, four relating to the measurement process and four relating to the instrument selected. The four components of the measurement process include: Domain Definition, Sampling Method, Measurement Process and Score Classification. The four components of the instrument include: The Test Stimulus, Establishment of Performance Criteria, Rating/Recording of Behaviour and The Assignment of Numerical Values to Performance Criteria.

In selecting a measurement method there are several feasible options for each of the eight components. Each option will be associated with specific sources of random error and/or systematic bias. Poor reliability in scores derived from a specific measurement method and instrument may be the result of random errors contributed by one or more of these components.

The validity of a clinical competence measure rests on the assumption that the scores produced will be predictive of the quality of day to day clinical performance. In turn, clinical performance is assumed to be associated with patient outcome. The evidence needed to support the validity of a specific procedure must be tailored to address these assumptions. Systematic sources of bias which may invalidate these assumptions are identified in relationship to the various approaches which

can be taken with each component of the procedure.

The standardized patient is one method that is used for the presentation of the clinical situation (the test stimulus) and the rating/recording of actions taken. The evaluation of random error and systematic bias in the presentation of the test stimulus will be addressed in Study 1 and 2 of this thesis. In Study 3, random error and systematic bias and/or error in recording by the standardized patient raters will be evaluated.

CHAPTER 3

THE MEASUREMENT OF CLINICAL COMPETENCE

In Chapter 1, a theoretical construct of clinical competence was introduced. The components of clinical competence, which have been derived from role delineation and critical incident analysis studies, were defined. The assumptions which form the basic rationale for the measurement and interpretation of competency scores were identified. In Chapter 2, the literature related to these assumptions was reviewed. Current limitations in our understanding of these relationships which would require further study were identified.

The objective of this chapter is to provide a context for reviewing the role that standardized patients play in the measurement of clinical competence and their potential contribution to measurement error. The first section will address general issues in scientific measurement. The next section will identify the four groups of approaches which have been used to measure competence and performance in relationship to the theoretical model of competence presented in Chapter 1. The final section will identify the common components of all clinical competence measures. The options, which may be used within each of these components, will be identified along with their potential contribution to measurement error.

GENERAL ISSUES IN MEASUREMENT

THE DEFINITION OF MEASUREMENT

In order to gain a better understanding of clinical competence and its assumed relationships to educational prerequisites, performance and patient outcome, a method of measuring the extent to which it is possessed by an individual is required. Stevens (1951) has defined measurement as "the assignment of numbers to objects or events according to rules". A similar definition is provided by Nunnally (1978): "measurement consists of rules for assigning numbers to objects in such a way as to represent quantities of attributes".

Carmines and Zeller (1979) comment that these definitions do not adequately describe the process of measurement for more abstract phenomena. These are phenomena which cannot be accurately classified as an object (can be seen or touched) or an event (result, consequence or outcome). Clinical competence, empathy and problem-solving are all examples of such abstract phenomena. Carmines and Zeller (1979) comment that measurement of abstract phenomena is better described as the "process of linking abstract concepts to empirical indicants". This includes an explicit organized plan for classifying and quantifying the empirical indicants to represent the abstract concept.

Miettinen (1985) describes a similar perspective in the conceptualization of measurement in epidemiology. The measurement of outcomes, modifiers and determinants involves the specification of the conceptual entity, the conceptual scale and the corresponding empirical scale which is used to classify or quantify the conceptual entity.

All three authors comment on the importance of adequately linking the conceptual entity with the empirical scale. If the link between the conceptual entity and empirical indicant is poor then little can been learned about the relationship among different concepts, or worse, erroneous conclusions about their relationships may be drawn.

It is this link between the conceptual entity and the empirical indicant which is one of the most challenging problems in the measurement of clinical competence. There are two properties of empirical indicants which are of concern: validity and reliability. These will be reviewed.

RELIABILITY

Reliability is concerned with the extent to which a measurement procedure yields the same results on repeated trials (Carmines & Zeller, 1979). A measure is never perfectly reliable. There are two types of errors which influence measurements, systematic bias and random error (Nunnally, 1978). Systematic bias has also been referred to as systematic errors. Systematic bias is used to refer to both terms in this thesis. The

reliability of a measurement is concerned with the extent to which random errors are influencing the repeated measurement of the same phenomenon. Since random errors are non-systematic, they should be normally distributed on repeated measurements of the same phenomenon. The standard deviation of a distribution derived from repeated measurements provides an index of the extent to which random errors are present in the measurement procedure. The magnitude of the standard deviation will be inversely related to the reliability of the measurement. The precision or reproducibility of a measurement procedure are other terms used to describe reliability (Colton, 1974; Carmines & Zeller, 1979).

The reliability of measurement is important because it influences the extent to which the true nature of relationships among concepts and phenomena can be identified. Unreliable measurements contribute unwanted variability to the measurement of phenomena. The strength of a relationship between two phenomena will be limited by the reliability of measurement. For example, a linear relationship could exist between knowledge and performance. The measurement of performance is typically unreliable. Variation in performance may be as much due to different raters and conditions as due to true differences in subjects. Because variation attributable to the subjects being measured accounts for only a small proportion of performance score, the true relationship between knowledge and performance may be obscured.

It is for this reason that the reliability of measurement is seen as a necessary prerequisite for validity (Nunnally, 1978). The magnitude of the validity coefficient between two measures will be limited by the reliability of those measures. However, an instrument which is reliable is not necessarily valid. If systematic biases are influencing the measurement or one is measuring a conceptually unrelated domain, then a reliable measurement will not be valid. In epidemiology, the accuracy of an instrument is a term which may be used to describe the extent to which an instrument is reliable and valid for the purpose intended. In Colton's definition (1974), accuracy is influenced by both systematic bias and random errors in measurement. This is only one of the many methods of conceptualizing and defining accuracy.

VALIDITY

Validity refers to the evidence that an association exists between the empirical indicant and the conceptual entity (Carmine & Zeller, 1979). More simply stated, does the empirical scale measure what we want it to measure? Whereas reliability relates to random error in measurement, validity is related to the degree to which systematic bias may influence the measurement of a phenomenon.

The validity of an instrument is conditional on the interpretations, inferences, and decisions one wishes to make on the basis of the measurement results (Kane, 1987; Carmine and Zeller, 1979). For example, an instrument which may be relatively valid for predicting present clinical performance may not be valid for predicting future clinical performance.

Because the 'truth' about what one is intending to measure is often not known, the validity of an instrument is relative. It rests on the mobilization of evidence to support the interpretations or inferences one wishes to draw on the basis of measurement results. The evidence to support the validity of an instrument has been conventionally categorized into three groups: content validity, construct validity and criterion validity. These groups refer to different approaches and types of evidence which are used to support instrument validity. They have been recently relabelled as content-related evidence, construct-related evidence and criterion-related evidence by the American Psychological Association (Kane, 1987).

Content-Related Evidence

Content-related evidence is used to demonstrate that the content of the measurement provides adequate representation of the conceptual domain to which one wishes to draw inferences. In this approach to the establishment of the validity the domain to which inferences are to be drawn must be specified and the representativeness of content measured assessed in relationship to that domain. It also includes a systematic investigation

of the adequacy and representativeness of sampling methods and item quality (Kane, 1987). An example of this approach is provided in the study of LaDuca et.al. (1984). The clinical problems, knowledge and abilities required for the general practice of medicine were initially specified. The subsequent test was developed to include a representative sample of items from the defined domain. In this study, required skills and attitudes were systematically excluded in test construction since they were less amenable to reliable measurement. These exclusions would limit the validity of the test instrument for the intended domain of inference.

Criterion-Related Evidence

Criterion-related evidence is used to demonstrate that scores arising from the measurement instrument are predictive of "some important form of behaviour that is external to the measurement instrument itself, the latter being referred to as the criterion" (Nunnally, 1978). Generally the criterion is a measure which is accepted by the scientific community as providing an accurate approximation of the 'truth'. Criteria, when they exist, are usually labelled as the 'gold standard' for the measurement of a particular phenomenon. Criterion validity includes both concurrent and predictive validity. The former refers to the association between the measure and the criterion at the same point in time. Predictive validity refers to the association between the measure and the criterion, the latter being measured at some later point in the future.

There is no accepted criterion measure for the evaluation of clinical competence. Nevertheless, new measures are often compared with ratings of clinical performance in actual practice. There are several important limitations in this approach to instrument validation. Clinical performance ratings, when carried out by supervisors or colleagues, tend to be unreliable, limiting the strength of association which can be detected. Chart audit approaches to performance measurement are limited in the components of competence which can be evaluated. Selection of charts by diagnostic category may also bias the inferences which are made about performance (Tugwell, 1979). Finally, clinical performance is influenced by a variety of factors other than clinical competence (see Chapter 2). If

no association is found, it does not necessarily mean that the measure of competence is invalid since different content domains may be measured, other factors may be influencing performance or the criterion may be unreliable.

Construct-Related Evidence

Construct-related evidence is used to demonstrate that scores arising from the measurement instrument behave in a manner consistent with the underlying theoretical construct of the phenomena being measured (Carmine & Zeller, 1979). The theoretical construct contains assumptions about the relationships of the concept being measured. If the instrument is measuring the theoretical construct, then scores derived from the instrument should behave in a manner consistent with the hypothesized relationships in the theoretical construct. Construct-related evidence of validity is, by necessity, frequently employed in the validation of abstract phenomena (Nunnally, 1978). Clinical competence is an example of an abstract phenomenon. A construct of this phenomenon is presented in Chapter 1. Hypotheses of the relationships between clinical competence, the clinical situation, prerequisites, performance, and outcome could be used to investigate the construct validity of scores resulting from a clinical competency measure. For example, it is hypothesized that for a homogeneous group of patients, those who are more competent would have better patient outcomes than those who are incompetent. If providers with worse patient outcomes had significantly lower scores on an instrument developed to measure competence than evidence in support of the construct validity of the instrument would be provided.

Interpretive Validity

Kane (1987) points out that these three approaches to measurement validation are not mutually exclusive. From his perspective, content validity and criterion validity are two subcategories of construct validity. Kane defines an alternate, more encompassing form of validity: interpretive validity. He argues that instrument validation efforts are often misguided by efforts to provide evidence using the

conventional categories of methods. In interpretive validity, the researcher tailors the evidence for validation to the specific interpretation which is to be made from measurement results. In order to do so, the researcher is required to articulate the assumptions which are being made about test scores in relationship to the conceptual entity for which the measure is intended. Literature review and selected evaluation efforts are then used to examine each of the articulated assumptions. Emphasis is placed on those assumptions which are the most suspect. The resulting evidence is then reviewed to assess the overall plausibility of the intended interpretation. This general view is shared by Messick (1989). He points out that it is not just the interpretation of scores which must be considered in a validity argument but also the social consequences of decisions arising from those interpretations.

This approach to instrument validation is particularly well suited to the measurement of clinical competence. It provides a framework to elucidate the rather muddy area of clinical competence and the assumptions which are being made in using a specific instrument to draw inferences about it's presence or absence.

Kane (1982) extends his argument, for interpretive validity, to the particular types of evidence that are required to support the interpretation of scores from professional licensure examinations. The assumption which is made in licensure examinations is that they measure certain "critical abilities that are necessary, although not sufficient, for effective performance in practice" (Kane, 1982). The inference which would be drawn from this assumption is that persons possessing low scores on the examination of critical abilities would perform inadequately in practice.

In terms of the model presented in Chapter 1, critical abilities may refer to either competence or prerequisites of competence. Similar to the model presented in Chapter 1, Kane (1982) defines critical abilities as those which would have a significant influence on client outcome, a relationship which should be made explicit in instrument development. Critical abilities which are required frequently in practice and/or would have a

high impact on client outcome would be weighted more heavily. This principle is similar to the use of clinical situations in health professional licensing examinations which are either prevalent or important by virtue of their treatability. It also forms the basis for an experimental format which is under current development for the Canadian licensing examination of physicians (Page, 1988).

Finally, Kane (1982) sees the identification of critical abilities as the responsibility of the relevant departments of research in the profession. In the health professions, the bulk of this evidence would come from departments of clinical research and epidemiology or their equivalent. These departments are usually involved in the evaluation of the efficacy of various provider services.

GENERAL CATEGORIES OF INSTRUMENTS USED IN THE MEASUREMENT OF COMPETENCE AND PERFORMANCE

In this section, the major categories of methods which have been used in the evaluation of competence and performance will be reviewed. The general characteristics of instruments within each category will be described. A detailed review of each instrument, it's limitations and the related evidence of reliability and validity is beyond the scope of this thesis. Salient issues relevant to the reliability and validity of these instruments was reviewed in relationship to the interpretation of studies in Chapter 2. A review of instruments in each category can be found in the text by Neufeld and Norman (1985).

Four groups of instruments can be identified by the content of measurement and the assumptions involved in the interpretation of scores arising from their use. They have been located in the model of clinical competence described in Chapter 1 and are illustrated in Figure 3.1.

The four groups of instruments are denoted by the letters A to D in the theoretical model. Each group and the related assumptions will be described subsequently. In the next section a more detailed review of the measurement issues in the B category of instruments will be provided.

GROUP A: INSTRUMENTS WHICH MEASURE PREREQUISITES OF COMPETENCE

The Instruments

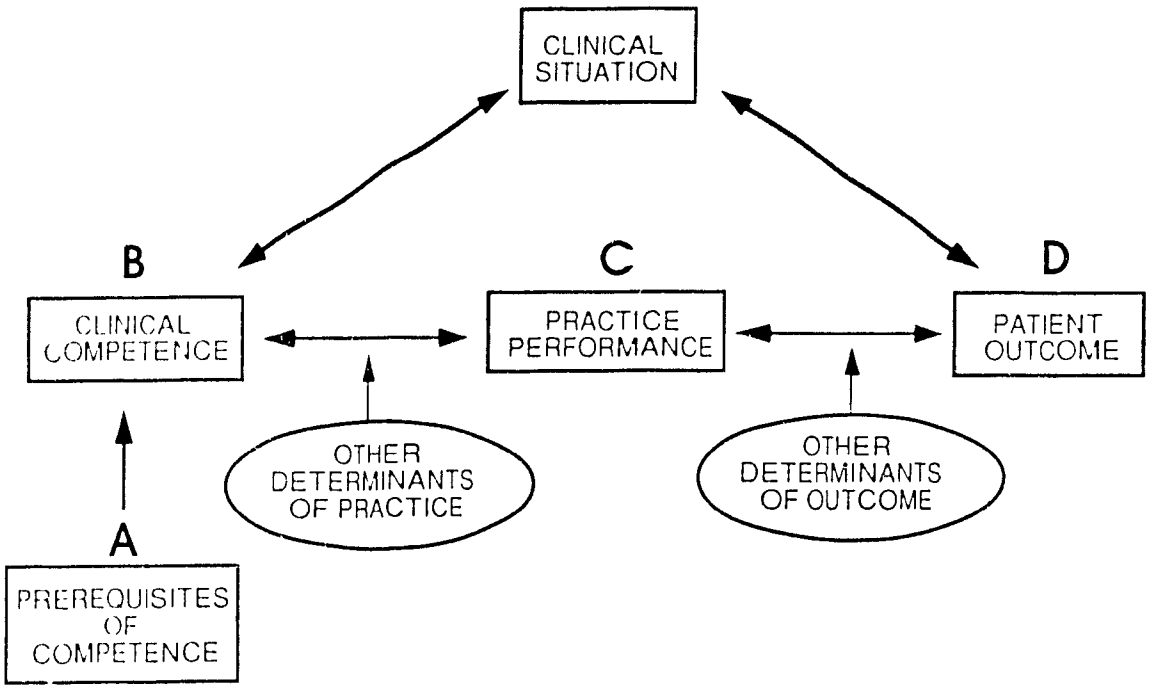
Four major prerequisites of competence were defined in Chapters 1 and 2. They include knowledge, skills, judgement and attitudes. Test instruments have been mainly developed for the measurement of knowledge and skills. Knowledge tests may also attest that they are measuring problem-solving, an attribute of judgement.

The most common form of knowledge test is the multiple-choice test. This test form is structured to provide the subject with a statement, summary of information or question. The subject is then asked to: a) select the correct response(s) from a list of alternatives or b) indicate whether the statement is true or false or c) fill in a blank with the correct response. Alternate methods of testing knowledge include the oral examination and the short or long answer essay. These latter methods are believed to provide a better measurement of the subject's understanding of a given content area. For a given test length they sample a narrower area of content than multiple-choice formats. Raters are an additional source of measurement error in essay and oral exam formats.

Three types of skills have been generally tested: interpersonal skills, data collection and interpretation skills, and technical intervention skills (eg. suturing). A variety of human models, animal models and simulations/mannequins have been used to test skills. Typically, direct observation and rating by a trained or expert observer is used to rate behaviour. These skills may be evaluated in the B category of instruments. The difference between measurement of these skills in the A category is that they are measured outside of a specific clinical context. For example, the ability of a subject to carry out a neurological examination may be evaluated by observing the subject carry out an examination on another subject. The content and technical quality of the examination would be typically rated. In the B category of instruments, the subject's ability to choose and carry out appropriate aspects of the

neurological examination for the clinical situation presented would be evaluated.

FIGURE 3.1 INSTRUMENTS USED IN THE MEASUREMENT OF CLINICAL COMPETENCE: THEIR RELATIONSHIP TO THE THEORETICAL MODEL OF CLINICAL COMPETENCE



The objective structured clinical exam (OSCE), developed by Harden (1975), is a method which has been used to structure and standardize an approach to the sampling and evaluation of knowledge and skills. The general principles espoused in this method have also been employed for B category instruments and mixtures of A and B. The OSCE dictates an approach to measurement rather than its content. OSCEs operate by first specifying a domain of knowledge and skills (for A category instruments) or clinical situations and components of competence (for B category instruments). A representative sample is drawn from the domain. A testing station is

developed to evaluate each item sampled. The requirements of a station are that it have a standardized stimulus (eg. X-Ray, oral examiner, patient) and a prospectively defined and standardized set of criteria which can be used to rate performance. Subjects rotate in random sequence through all stations. The number and length of stations is variable.

The Assumptions

When A category instruments are used to measure clinical competence, it is assumed that:

1) the knowledge, skills, judgement and attitudes required for clinical competence in a specified domain are known and that the instrument selected is able to measure these abilities

2) a subject who does not possess adequate amounts of the specified prerequisites will not be able to demonstrate clinical competence in the clinical situations included in the defined domain

3) a subject who does not possess adequate amounts of the specified prerequisites will not be able to perform effectively in day to day practice

4) the patients seen by subjects who do not possess adequate amounts of the specified prerequisites will have worse outcomes than patients who are seen by subjects who do possess the specified prerequisites.

The evidence to support these assumptions was reviewed in Chapter 2.

GROUP B: INSTRUMENTS WHICH MEASURE CLINICAL COMPETENCE

The Instruments

The next section will provide a detailed analysis of the instruments used to measure clinical competence. The general characteristics of this group

of instruments in relation to the other groups will be identified in this section along with the related assumptions in score interpretation.

This group of instruments is developed to measure some or all of the components of clinical competence (eg. data collection, diagnosis, management) with a defined patient problem. This differentiates them from A category instruments where measurement is carried out without a specific clinical problem context. Unlike C category instruments, components of clinical competence/performance are not measured in day to day practice. Rather the subject's ability to perform in a defined clinical situation within a defined time period and setting is assessed. There is usually some effort to standardize the clinical situation and other factors which could influence the quality of their performance (eg. delays in the return of lab data). This is usually done by simulating the clinical situation and standardizing the measurement procedure. In theory, this group of instruments evaluates the subject's ability under optimal practice circumstances whereas the C group of instruments evaluates their performance under usual practice conditions (which may or may not be optimal) (Cronbach, 1970).

The Assumptions

When clinical competence is evaluated using this group of instruments, it is assumed that:

1) components of clinical competence which are important determinants of patient outcome have been identified

2) that the actions required for optimal patient outcome in each clinical situation evaluated have been identified

3) the instrument is able to accurately measure the identified components of competence

4) the subject who does not possess adequate amounts of each of these components will be unable to perform effectively in day to

day practice in the defined domain of inference

5) the subject who does not possess adequate amounts of these components will have major deficiencies in one or more of the prerequisite abilities: knowledge, skills, judgement and/or attitudes

6) within equivalent groups of patients, that the patients of subjects who do not possess adequate amounts of each of these components will have worse health outcomes than the patients of a subject who possesses these components of competence.

The evidence to evaluate these assumptions was reviewed in Chapter 2.

GROUP C: INSTRUMENTS WHICH MEASURE PERFORMANCE

The Instruments

The instruments in this category measure the quality of day to day performance in clinical situations. One or more of the following components of performance are measured : data collection, diagnosis, test selection and interpretation, management, professional communication, professional responsibility, working knowledge, and patient communication. Professional responsibility is the one component of performance which is not amenable to measurement with B category instruments. It is usually defined as the individual's ability to demonstrate responsibility for patients in daily practice.

The quality of performance is most commonly ascertained by one of two methods: clinical supervisor/colleague ratings of performance in day to day practice and chart audit. When performance is rated by colleagues or supervisors, a standardized form is generally employed. It lists and in some instances defines the components of interest. A five to seven category Likert scale is generally employed for rating the quality of performance for each component. Behaviourially anchored rating scales have also been used. The source of data used to assess performance is

unstructured. Stillman (1987) comments that most of the data used to evaluate performance are derived from oral presentation and discussion of patient cases rather than direct observation or chart review.

Chart audit is a more structured approach to the ascertainment of performance quality. Prospectively defined criteria are used to audit a sample of provider charts over a defined time period. The criteria may be specific to certain types of clinical situations (defined as tracer or indicator conditions). Alternatively, criteria may relate to the general requirements for the process of care for all clinical situations within a specific domain of practice. A quality of performance score is generated by calculating the number of times the provider complied with the defined criteria in the charts sampled. This method of evaluation is not suitable for the evaluation of some areas of competence, namely patient communication, professional communication, and professional responsibility.

The Assumptions

This category of instruments is used to evaluate clinical performance. The resulting scores are often used to draw inferences about clinical competence and the possession of prerequisite abilities. For example, a subject who has low scores on a clinical performance measure is assumed to be less competent. It is also assumed that the subject will have major deficiencies in one or more prerequisite abilities: knowledge, skills, judgement and/or attitudes. When clinical performance measures are being used to measure clinical competence, the following assumptions are being made:

1) components of clinical performance which are important determinants of patient outcome have been identified

2) the actions required in each component for optimal patient outcome in each clinical situation evaluated have been identified

3) the instrument is able to accurately measure the identified

components of performance

4) subjects who have low scores on performance evaluation measures are less competent than subjects with high scores

5) other system and provider related determinants which may influence performance are either associated with subject competence or are not associated with performance evaluation score

6) subjects with low scores on performance evaluation measures have major deficiencies in one or more of the prerequisite abilities: knowledge, skill, judgement and/or attitude

7) the outcomes for the patients of subjects with low scores on performance evaluation measures will be worse than for subjects with high scores (within similar groups of patients)

The evidence which is available to support these assumptions was reviewed in Chapter 2.

GROUP D: INSTRUMENTS WHICH MEASURE PATIENT OUTCOME

The Instruments

The most frequently measured patient outcomes are mortality, morbidity, satisfaction, and compliance with therapy. Chart audit and follow-up patient interviews, examinations/lab data and/or questionnaires have been used to measure outcome.

The Assumptions

When patient outcome is used to draw inferences about the quality of provider performance or competence, the following assumptions are made:

1) the outcome being measured is of importance to the patient or society

2) the outcome of interest is accurately measured using the selected method

3) other factors which may influence patient outcome, other than the performance of the provider, have been taken into account

4) the performance evaluation score of providers delivering care to patients with poorer outcomes will be less than for the providers of patients with better outcomes (after other patient, disease and health care system factors have been taken into account).

5) when other patient, disease and health care system factors have been taken into account, patients with poor outcomes will have had their care delivered by less competent providers than patients with good outcomes

6) when other patient, disease and health care system factors have been taken into account, patients with poor outcomes will have had their care delivered by providers who have major deficiencies in knowledge, skill, judgement and/or attitudes than patients with good outcomes

The evidence to support these assumptions is meagre. With the exception of surgical problems, it is not common to find patient outcome measures used as a means of drawing inferences about provider competence or performance.

INSTRUMENTS USED IN THE MEASUREMENT OF CLINICAL COMPETENCE

The measurement of clinical competence requires the selection of an instrument and the specification of the method of measurement and score interpretation. A variety of approaches can be taken in the specification of the measurement method and instrument selection. In the literature, an array of eponyms are used to differentiate these various approaches and

instruments. This section will review the components which are common to all instruments and methods. The options which are available for each component identified will then be described. Finally the potential sources of measurement error, both random and systematic, associated with each option will be identified.

THE COMPONENTS OF THE MEASUREMENT METHOD

Clinical competence is usually measured with the objective of classifying subjects as competent or incompetent in a defined domain of clinical practice. There are four components of the measurement method which must be defined in relationship to this objective:

1) The clinical practice domain to which inferences will be drawn needs to be defined (Domain Definition)

2) The method of sampling from the domain needs to be specified (Sampling Method)

3) The method of measuring competence needs to be identified including instruments, setting and time frame (Measurement Process)

4) The method of classifying subjects into competent and incompetent categories needs to be specified (Score Classification Method)

The optional approaches which have been used to address each of these components of the measurement method will be reviewed along with the sources of measurement error which are associated with each option.

Component 1: Domain Definition

The definition of the clinical practice domain characteristically includes the identification of clinical situations which would be included in the domain, the location and level of practice expected (eg. primary, secondary, urban ambulatory) and the components of competence to be

measured (eg. data collection, management). The domain definition represents the sample frame for component 2, sampling method. It is usually organized in a multi-way grid of clinical situation by location of practice by level of practice by components of competence. Decisions must be made about how variables within the grid are defined or classified and how they are generated.

The Options

In Chapter 1, the three major methods of identifying and classifying components of competence were identified: job analysis, role delineation, and critical incident analysis. The advantages and limitations of each of these methods were identified. Critical incident analysis was identified as the method which was most compatible with the assumptions and inferences which are usually drawn about scores arising from clinical competence measures.

The major methods of classifying the clinical situation were also identified in Chapter 1. In Chapter 2, the limitations in our understanding of the attributes of the clinical situation which influence performance were discussed. The major methods of clinical situation classification include: by prevalence (of the diagnosis or symptom), by treatability, by age and gender, by severity/chronicity, and by type of management required. Although prevalence, age and gender can be generated empirically, the remaining attributes are often classified by panel of experts or representatives from the clinical domain of interest. The clinical situations to be included and excluded from the clinical practice domain are also customarily determined by a combination of expert panel and empirical data on disease and symptom frequency.

Location and level of practice are defined and classified by expert or representative panels. Levels of practice include: primary, secondary and tertiary. Location of practice includes: hospital (general and critical care), ambulatory (office and clinic), urban, rural and remote. These categories are not mutually exclusive. There is no clear differentiation of the types of services which would be expected in each level and

location of practice.

Sources of Measurement Error

A certain amount of random error in the specification of the clinical practice domain is to be expected when expert or representative panels of professionals are used. The magnitude of this problem would be influenced by the size of the expert/representative group, method of selection, and variance in professional opinion about the clinical practice domain.

Systematic bias in domain definition is probably a greater problem when expert or representative professional panels are used. As described in Chapter 1, different components of competence are generated by different subgroups within the profession. The approach taken by LaDuca et.al.(1984) in the definition of a general practice domain provides an example of control of systematic bias in domain definition. Empirical data from 102,705 physician patient encounters were content analyzed to produce the definition of the domain of clinical practice to which inferences were to be drawn. Expert panels were then used to identify important components of clinical competence required for each of the 40 patient management paradigms included in the domain.

Component 2: Sampling Method

The Options

Two decisions have to be made with respect to sampling method. The first decision relates to how many clinical situations must be sampled and how many times components of competence must be measured to derive a stable estimate of subject performance. The literature and principles in relationship to this question have been recently reviewed by Swanson and Norcini (1987). The variance in performance within the domain, the type and number of components of competence tested and the instrument selected will all influence sample size requirements. Feasibility in test administration ultimately determines the number of clinical situations which will be sampled.

The second decision relates to how sampling is to be carried out. The question of how and what to sample is influenced by the instrument which has been selected to measure competence. Various formats are available to present the clinical situation (see components of the instrument). With patient formats, it is usually not possible to present all clinical situations included in the clinical practice domain. With paper and computer formats, it is not possible to measure all components of competence considered to be of importance. Depending on the instrument selected, sampling from the defined domain must be limited to those clinical situations and/or components of competence which can be evaluated. Given these restrictions, a stratified random sample of situations by location of practice and components of competence may be selected. More commonly, a sample of clinical situations is selected to represent the various disciplines included in the domain and practice locations. The components of competence which are most relevant to these clinical situations and locations of practice are then identified by an expert or representative panel.

Sources of Measurement Error

If an insufficient number of clinical situations are sampled, random error in the estimation of clinical competence is a potential problem. Restrictions introduced by instrument format are a potential source of systematic bias since only a select subset of situations or components of competence can be evaluated. Systematic bias may also be introduced by sampling methods which are non-random. Different selection groups may place emphasis on different clinical situations or components of competence in the practice domain. Inferences about competence would be biased by these selection practices.

Component 3: The Measurement Process

Three decisions are usually required in the specification of the measurement process: the selection of the instrument(s), the specification of the test setting and the specification of the time frame and conditions

of measurement.

The Options

The options which are available in the selection of the four components of the measurement instrument are detailed in the next section. Decisions in relationship to the test setting, in part, relate to the type of instrument selected. For paper and computer formats, a classroom or office setting is usually employed. For patient formats, the issue of test setting is probably more critical. The three options usually employed are a hospital or clinic setting, a simulated office/hospital setting or a classroom. It is hypothesized that the subject's behaviour in response to the clinical situation (for patient formats) may be influenced by the characteristics of the setting in which the clinical problem is presented. For example, if a patient with acute chest pain is seen by the subject in a classroom setting, the subject may confine his/her behaviour to the solicitation of relevant history since facilities for patient examination are not provided.

Test time per subject is determined by how many components of competence are to be tested within each clinical situation and the total number of clinical situations included in the evaluation. The length of time per clinical situation can be as short as 30 seconds, for multiple choice formats of presentation, to one hour for patient formats of presentation. For shorter time periods, the subject's response to the clinical situation is usually structured within a defined set of parameters.

Finally, the conditions of the measurement process must be specified. These include a statement about the purpose of the evaluation, the method of measurement and scoring, the decisions which will be made as a result of test performance and the consequences of an adverse performance outcome. It is hypothesized that subjects, even in the absence of information, will draw conclusions about each of these issues which in turn will influence their performance.

Sources of Measurement Error

Measurement error in relationship to the choice of instrument will be covered in the next section. Systematic bias is the major potential source of error in relationship to the other attributes of the measurement process. With patient formats, artificialities of the test location may bias the resulting estimates of competence. It is hypothesized that the potential for this form of bias would be greater for clinical situations which would normally be seen and managed in resource intensive settings (eg. emergency problems).

All measurement methods provide some limitation on the time spent per clinical situation. This is a recognized artificiality which is necessary for large scale administration of the measurement method. With shorter test times, the subject's approach to the clinical situation must be structured. It is hypothesized that the more structured the approach, the greater the likelihood for bias in the estimation of the subject's true ability.

Biases introduced by the conditions of the measurement method have given the general label of 'test-wiseness' (Cronbach, 1970). For example, subjects may conclude that bonus points can be accrued by the completion of an exhaustive history and the avoidance of costly investigations. If this were the case, the subject may respond to the clinical situation in a different manner during the measurement process than in actual practice.

Component 4: Score Classification Method

A measurement process will typically generate a score for competence in each of the clinical situations tested. In multiple-choice formats, this score is generally dichotomous (correct, incorrect). In other paper and patient formats, the score generated is continuous, representing the number and importance of performance criteria which were met by the subject. In most situations, each clinical situation is given equal weight

and an overall competence score is calculated by averaging the scores for each clinical situation.

This score may be interpreted in two major ways: as an estimate of the subject's competence in the defined practice domain or as an estimate of the subject's competence relative to other subjects who have completed the same measurement procedure. When competence is being measured to determine academic progress, licensure or certification, the continuous score must be collapsed into a binary classification of competent and incompetent.

The Options

One of two approaches is generally used. In the criterion approach, an expert or representative panel of professionals prospectively specifies the minimum score which is required to be considered competent. This may take a number a forms. For example, the minimum score per clinical situation may be specified and a specified number of clinical situations must be above that minimum score for the individual to be considered competent. Alternately, certain performance criteria may be identified as being critical and all must be carried out in the clinical situations tested for the subject to be considered competent.

The second approach is the normative classification method. Subjects who fall below a specified standardized score on a normal distribution of scores generated by all subjects taking the test (present and past) are considered to be incompetent. This approach is described in detail by Cronbach (1970).

Sources of Measurement Error

As might be expected, random error in the classification of competence would be greatest for subjects who are near the cutpoint on the continuous scale. Random errors in the classification of subjects who are two standard deviations above or below the cutpoint are less likely. For normative classification methods, random errors in classification are problematic when a small sample of subjects has been used to establish the

normal score distribution.

Systematic bias in classification is a potential problem for the criterion approach. As indicated in the previous sections, the composition of the expert or representative professional group can bias the classification of competence. Different types of groups could generate different classification criteria. In the absence of empirical data which would identify the critical components of competence, both approaches are at risk of measuring, scoring and ultimately classifying subjects with criteria which are irrelevant to performance and patient outcome. Systematic errors in classification could be in either direction.

Systematic bias is also a potentially important problem when a normative classification approach is being used to draw inferences to a clinical practice domain. Cohort effects may clearly lead to classification errors in either direction.

THE COMPONENTS OF THE INSTRUMENT USED TO MEASURE CLINICAL COMPETENCE

Instruments which are used to measure competence have four components:

1) a method of presenting the clinical situation in which competence is to be measured (the test stimulus)

2) a method of establishing what behaviour is to be measured in response to the clinical situation (performance criteria)

3) a method of recording or rating behaviour which occurs in response to the clinical situation (rating method)

4) a method of assigning a numerical value to the behaviour measured (scoring)

In the development or selection of an instrument, decisions must be made about each of these components. The potential sources of measurement error and their consequences for instrument reliability and validity will vary

depending on the option selected for each component. These options and their consequences for measurement will be reviewed.

Component 1: Presenting the Clinical Situation

There are four major ways of presenting the clinical situation: by paper, computer, oral examiner and by patient.

Paper Formats

a) The Options

This is the most common method of presenting the clinical situation in clinical competence evaluation. Included within this category are the Patient Management Problem (PMP) (McGuire, 1967), the Sequential Management Problem (SMP) (Barrows, 1975), the Multiple-Choice Question format (MCQ) and the Portable Patient Problem Pack (P4) (Barrows and Tamblyn, 1980).

The various options within this category differ by the amount of relevant patient information which is provided to the subject before they are required to select an action, the number of actions which are available for the subject to choose from and the components of competence which can be measured.

The MCQ format generally provides the most information to the subject and the fewest options to choose from. The components of competence it measures are usually limited to diagnosis, test selection and interpretation and management decisions.

The P4, PMP and SMP all provide an initial statement of the patient's problem. All further data and actions must be determined by the subject. The PMP provides a branching format with 5 to 10 possible actions the subject may choose from at each branch. Data are provided according to the actions selected.

The SMP provides no options, the subjects must generate their own actions at each step in the process of managing the patient's problem. After the subject generates actions, the actions and resulting patient data generated in the actual patient situation are provided. This process continues in a sequential fashion.

The P4 provides the same array of actions at each step in the management of the patient's problem. The actions are grouped by clinical category (eg. history, physical, tests) and include all options normally available in a health care setting. As the subject selects each action, the information which would be gained from the patient with that action is provided.

These three formats can measure data collection (in terms of appropriateness of items selected), diagnosis, test selection and interpretation, and the selection of management options. They do not measure the ability to carry out data collection or patient communication.

Potential Sources of Measurement Error

Paper formats have the advantage of standardizing the presentation of the clinical situation for all subjects. This eliminates sources of random error in the presentation of the clinical situation. There are two potential sources of systematic bias. The first is cueing bias. Competence may be overestimated if the subject is cued to select actions by virtue of the choices available. It has been assumed that the more limited the choice, the more likely the subject is to be cued by options available and hence the more biased the estimate of competence.

The second source of potential systematic bias is format bias. It has been hypothesized that the more artificial the representation of the clinical situation, the more likely that behaviour produced will be different than actual practice. For example, the MCQ format is the most artificial representation of the clinical situation. Competence in diagnosis measured with this format may under or overestimate the subject's ability to competently diagnose situations from the same domain in practice.

Computer Formats

The Options

Two types of computer formats have been developed. The first provides a format similar to the Patient Management Problem which is presented by computer rather than on paper. The second is still in the process of development. It provides an uncued, natural language format. The presenting problem is provided and the subject must request all subsequent data by entering a request from the keyboard. It is conceivable that in the future, patient response could be provided by interactive video disc. The components of competence measured by computer formats are the same as for paper formats.

Sources of Measurement Error

Similar to paper formats, random error in the presentation of the problem is eliminated by use of computer presentation. The PMP type of computer format has the same cueing bias potential as the PMP. Format bias is a potential problem with all forms of simulation.

In addition, both types of computer programs have a potential computerphobe bias. Subjects who are less comfortable with computer technology may be apt to select fewer actions than those who are more adept. This bias was demonstrated in an early study by Feightner and Norman (1978). More actions were selected by subjects in a paper presentation of a patient management problem than a computer presentation.

Oral Examiner Format

The Options

The clinical situation can be presented to the subject by an oral

examiner. This is done using one of two methods. In the first method the examiner presents some initial clinical data to the subject and asks the subject for his/her impressions or plan of action. Depending on the subject's response, additional questions or clinical data may be provided. In the second method, the oral examiner assumes the role of the patient. The subject is expected to carry out actions they would take in an actual situation. The examiner provides the data which would be obtained for each of the clinical actions selected.

Sources of Measurement Error

In this format, both random error and systematic bias may influence the content of the clinical problem presented. Random errors may be present in both the content of the presentation (eg. some data are provided for some subjects but omitted for others) and the manner in which the data are provided (eg. emphasis is placed on one aspect of the problem for one subject but not for another).

Systematic bias in the content and manner of clinical problem presentation is probably more commonplace. One of the characteristics of this method of presenting the problem is that the examiner can vary the difficulty of the problem and provide explanation when an aspect of the problem is unclear. This is perceived to be advantageous in that the extent and limits of the subject's ability to manage the clinical situation presented can be effectively explored. The difficulty is that each subject is essentially given a different test and is measured on a different scale. In addition, attributes other than clinical competence may influence both the subject's response to the situation and examiner behaviour.

Patient Formats

The Options

The two methods of presenting the clinical situation with a patient format are with standardized patients and with unstandardized real patients. The

standardized patient format includes real patients and healthy individuals who have been trained to simulate a particular clinical problem. They are differentiated from unstandardized real patients by being trained to provide a reproducible portrayal of an actual clinical problem at a specified point in time. When real patients are used, they present the actual problem which brought them to the health care system. The content of the problem may vary depending on the point in time when they are seen by the subject.

When standardized patients are used as the method of presenting the clinical situation, the same situation is used to measure all subjects evaluated. When real patients are used as the test stimulus, different real patients with a similar clinical problem are used to evaluate all subjects. All components of competence except long term management and professional responsibility can be measured with both patient formats.

Sources of Measurement Error

Unlike paper and computer formats, random error is a potential problem in the presentation of the clinical situation with both types of patient formats. It has been assumed that standardized patient training would eliminate random errors in the presentation of the clinical situation. This assumption will be evaluated in Study 1 of this thesis. Random error in the unstandardized real patient's presentation of their clinical problem is often anecdotally reported. The patient may neglect to recall certain items on history for one subject but remember for another. Subtle physical findings may be present on some occasions but not on others.

Two types of systematic bias may operate in both types of patient formats. Errors of omission and commission in the recall of relevant patient data may alter the content of the clinical situation presented and as a consequence, the subject's choice of actions in response to the clinical problem. Secondly, true changes in the real patient's problem over time may alter the content of the presentation with different clinical actions selected by subjects seeing the patient in different time periods. The time period bias may also influence the standardized

patient's presentation. Fatigue may influence the content of the presentation for subjects seeing the patient in the late afternoon vs. the morning. With the standardized patient format, these two forms of systematic bias are assumed to be eliminated by training. This assumption will be evaluated in Study 1 and 2 of this thesis.

Component 2: Establishing Performance Criteria

The decision of what behaviour to measure in a clinical situation is probably one of the most important for instrument validity. The choice of performance criteria relates to assumptions 1 and 2 of the B category of instruments. It is assumed that the components of competence which are selected for measurement and the clinical actions specified within each of those components are important determinants of patient outcome. If irrelevant performance criteria are specified, then subjects who are incompetent (i.e. they may do harm to patients) may not be detected.

There are two decisions which must be made with respect to the specification of performance criteria: who should specify performance criteria and how should it be done.

The Options

Two types of groups are customarily used to establish performance criteria: the expert group and the representative professional group. The expert group is composed of individuals who are recognized for their expertise in the clinical problems used in the evaluation. The representative professional group is composed of practitioners who represent the subcategory within the profession to which one wishes to draw inferences about competence. For example, if one wishes to draw inferences about a subject's ability to be competent in general practice situations, a representative professional group would consist of general practitioners from various practice types and locations.

Those who advocate the use of the expert group hypothesize that experts are more knowledgeable about those clinical actions which are most

critical to patient outcome. Advocates of the representative professional group hypothesize that criteria established by experts are tailored to the select subset of difficult clinical problems which are referred to their practice setting. As a result, criteria set by experts are both unrealistic and inappropriate for the most common types of problem presentation.

Two methods are used for the generation of performance criteria: group development and rating and performance-based development. The first method relies on the group to generate performance criteria and rate or reach consensus on the final set of criteria to be employed. The second method is less common. The performance of the designated group in the clinical situation of interest is used to establish expected performance criteria. All actions taken by the group may be included or only those which occurred most frequently.

The generation of performance criteria for the MCQ format of problem presentation is a special subset of the group development and rating method. The available options to choose from and the most correct response is generated by the group.

Sources of Measurement Error

Random error in criteria specification is a potential problem. Different membership in the two types of criteria specification groups could result in different performance criteria. This is particularly true when group size is small or when the optimal approach to the clinical problem is controversial.

Systematic bias, as indicated, could arise by the use of different types of groups: expert vs. representative professional. In addition, regional, national and cultural differences in performance expectations with certain clinical problems could result in the establishment of different performance criteria.

Component 3: Recording/Rating Behaviour Using Performance Criteria

Two decisions about the rating/recording of behaviour produced in response to a clinical situation must be made. The first decision is who will be responsible for rating behaviour. The second decision is how the rating will be done.

The Options

In relationship to the first decision, three types of raters/recorders may be selected: the subjects themselves, an expert rater or a trained non-expert rater. With computer and paper presentations of the clinical problem, the subjects themselves are acting as recorders of the actions selected in response to the clinical situation. With oral examiner and patient formats of presentation, either an expert rater or a trained non-expert rater is used to rate/record clinical actions. In the measurement of competence, medical faculty are traditionally considered to be expert raters. They may or not be trained to rate/record the presence of clinical actions specified by the performance criteria. Standardized patients are one example of a trained non-expert recorder. They are trained to record the presence/absence of specified performance criteria and rate the quality of patient communication.

With respect to the second decision, the rating/recording of clinical actions may be done by:

1) the selection of actions by the subject from a predefined list,

2) the direct observation of the subject interacting with the patient or oral examiner,

3) review of the written record which results from the interaction with the patient or

4) by the subjects oral response to the clinical situation

presented.

In addition, for options 2, 3, and 4, actions may be recorded as being present or absent or the relative quality of the action may be rated.

Sources of Measurement Error

When the subject is used to record the actions they have selected from a predefined list, random errors in the recording of subject behaviour are virtually eliminated. When experts or trained non-experts are used to rate/record actions, a certain amount of random error in measurement is to be expected. It is hypothesized that random error in measurement may be greater for the direct observation and oral response forms of clinical action ascertainment. In these two methods, the rater does not usually have the time to reflect on the content or quality of the actions undertaken by the subject.

The type of rater used to record/rate clinical behaviour may systematically influence the performance criteria documented as being present or absent. The bias could operate in either direction. With respect to how the ratings are done, written and oral forms of response may under or over estimate the subject's ability to actually comply with performance criteria. Underestimates of the subject's ability would occur if the subject was unable to adequately report the actions they would take in an oral or written fashion. Overestimates of ability would occur if the subject was unable or disinclined to carry out the stated clinical actions in the clinical situation being evaluated.

Component 4: Assigning a Numerical Value to Performance Criteria

Establishing performance criteria was identified as one of the most important decisions in relationship to instrument validity. The method used to assign numerical values to specified performance criteria is of similar importance. Theoretically, the numerical values assigned to performance criteria should reflect the relative importance of each clinical action for patient outcome. If this is not the case, then undue weight is given to actions which may be of marginal importance. The

resulting score for clinical competence would then not represent those who are more likely to achieve better or worse patient outcomes.

The Options

The methods used to assign numerical values to performance criteria are the same as those used to generate criteria. It is typical for the group who establishes the criteria to also assign numerical values.

Numerical values are assigned by group consensus or by frequency weight. When established by group consensus, a value of 1 can be assigned to all criteria generated or some may be given greater weight by assigning values of greater than 1. The latter method uses the number of times a criterion is rated as being important by different group members as the assigned weight (in the instances of performance generated criteria, the weight is the number of times the action is performed).

Sources of Measurement Error

The same sources of measurement error which were discussed in the establishment of performance criteria apply to the assignment of numerical values.

SUMMARY

In this chapter the measurement properties of the reliability and validity were described. Reliability was identified as a property of the empirical measure. It reflects the degree to which the empirical measure is free of random error, a condition necessary for the detection of associations between concepts of interest. Validity was identified as the strength of the relationship between the empirical measure and the conceptual entity, in this instance clinical competence. The absence of systematic bias in the empirical measure is a necessary condition for instrument validity.

The procedure used to measure clinical competence is complex. Eight components of the procedure were identified, four relating to the measurement process and four relating to the instrument selected. A number of approaches can be taken with each of the identified components. Each approach may be influenced by potential sources of random error and/or systematic bias. Poor reliability in scores derived from a specific procedure and instrument may be the result of random errors contributed by one or more of these components. The evidence needed to support the validity of a specific procedure must be tailored to address the identified assumptions inherent in score interpretation. Systematic sources of bias which may invalidate these assumptions were identified in relationship to the various approaches which can be taken with each component of the procedure.

The standardized patient was identified as one approach which can be taken in the presentation of the clinical situation (the test stimulus) and the rating/recording of actions taken in response to the test stimulus. The evaluation of random error and systematic bias in the presentation of the test stimulus will be addressed in Study 1 and 2 of this thesis. In Study 3, random error in recording and selected sources of systematic bias associated with standardized patient raters in two university sites will be evaluated.

PART III - THE STANDARDIZED PATIENT

ABSTRACT

CHAPTER 4

THE STANDARDIZED PATIENT: A REVIEW OF THE METHOD

The standardized patient is defined as an individual who has been trained to provide a consistent presentation of the history, physical findings and affect of a real patient case. Individuals who are trained as standardized patients may be healthy people or real patients who have the problem they are being trained to present.

Six characteristics of individuals which are important to consider in patient selection have been identified by authors experienced in this technique. They include patient age, stamina, motivation, experience, employment and current health status. The Barrows method of training standardized patients is described. It is the method employed by most authors reporting the use of this technique in the literature.

Studies which have reported on the use of the standardized patient technique for educational, evaluation and research purposes are summarized. The standardized patient has been used to teach interviewing, history-taking, physical examination and patient education skills. This method has been used for both formative and summative evaluation of clinical competence. Consideration is being given to its use for licensure examinations. The standardized patient has also been used in quality of care research. The validity of the standardized patient as a method of presenting the clinical situation is supported by evidence from six studies which found that the standardized patient could not be differentiated from a real patient by clinicians. In addition, scores on standardized patient-based evaluations of competence were no different from scores based on the evaluation of performance with real patients in most instances. When compared to supervisor's ratings of performance, standardized patient-based evaluations were more strongly correlated than paper problem-based evaluations. Direct estimation of the standardized patient's accuracy in the presentation of the case has not been reported. Errors in the presentation of the clinical situation were implicated as being responsible for differences in competence score in two

studies.

The reliability of standardized patients as raters of the clinical encounter has been studied in a number of ways by different authors. The sample size used in most studies is small limiting the precision and generalizability of estimates. From the data available, agreement between standardized patients and faculty appears to be no better or worse than the agreement observed between faculty. Poorer reliability in the rating of communication is characteristic of both types of raters. In one study it was estimated that standardized patient raters contributed 2% to variance in data collection scores and 9% to variance in communication skills. Comparisons between standardized patients who are presenting the same case is limited. Systematic differences between patients were noted in one study.

It was concluded that standardized patients represent a potential source of measurement error which is confounded with case in the measurement process used to evaluate competence. Estimation of the contribution of random and systematic errors made by the patient in case presentation and encounter rating was recommended. If problematic, control may provide a means of reducing the number of cases required to produce a stable estimate of competence and avoid bias when multiple centres are being used in competency evaluation.

CHAPTER 4

THE STANDARDIZED PATIENT: A REVIEW OF THE METHOD

INTRODUCTION

The standardized patient is one of several methods which have been used by investigators to develop and evaluate clinical competence. Specifically, the standardized patient has been used as:

1) a means of presenting the clinical problem for evaluation or learning

2) a recorder who documents the actions taken by the provider during the patient encounter and

3) a rater who judges the ability of the provider to establish a patient relationship and communicate with the patient and family.

4) as a source of corrective feedback and instruction to students on interviewing, history-taking and physical examination skills

This chapter will provide a detailed review of the use of the standardized patient as an educational, evaluation and research tool. Research related to the reliability and validity of the standardized patient will be reviewed. Areas for further investigation will be identified and addressed in subsequent chapters.

THE STANDARDIZED PATIENT: DEFINITION, CHARACTERISTICS AND TRAINING

DEFINITION

The simulated patient was first described in the literature by Barrows and Abrahamson in 1964. The simulated patient is defined as a person "who has been trained to accurately recreate the history, personality, emotional structure, responses and physical findings of an actual patient" (Barrows, 1971). The term 'standardized patient' has subsequently been employed to

characterize persons (both real patients and those simulating a clinical problem) who have been trained to present a clinical problem in a reproducible manner (Barrows, 1987; Stillman, 1987).

CHARACTERISTICS

Socio-Demographic Characteristics of Standardized Patients

The characteristics of individuals who can be effectively trained to become standardized patients have not been the subject of empirical study. Factors important in standardized patient selection have been identified by Barrows (1971) and Stillman (1986) on the basis of their experience with this technique. These factors can be grouped into six categories: age, employment, experience, motivation, stamina, current health status and physical & psychological attributes. Experiential knowledge which has been gained about these factors is described in Figure 4.1.

Generally, persons trained to be standardized patients are between the ages of 5 and 70. They may or may not have had experience with the problem they are simulating. Students, health professionals, actors/actresses and individuals with flexible working hours are the most common occupational groups trained. Interest in contributing to health professional education or financial need are the most common motivations for becoming a standardized patient. Intelligence and sufficient stamina to withstand repeated examinations are necessary prerequisites for successful training and use.

Naftulin & Andrew (1975) provide the only reported study of characteristics of standardized patients. This descriptive study evaluated the psychiatric profile, emotional and physical disorders of 9 standardized patients in comparison to 10 comparable community controls. No difference in MMPI scores, physical or emotional problems were identified.

FIGURE 4.1 SOCIODEMOGRAPHIC CHARACTERISTICS OF STANDARDIZED PATIENTS

Age:

The most commonly reported age interval for standardized patients is 20-50 years(Stillman,1982; Page & Fielding,1980; Burri 1974; Renaud,1980; Rethans & Bovin, 1987; Barnes,1978; Lincoln,1978; Jason,1971; Behrans,1979). Children have been trained, the youngest reported being 5 years of age. It has been difficult to train children to be consistent in their performance. This problem coupled with their lack of availability during school hours are obstacles which impede extensive use (Barrows, 1986). An upper age limit for patient recruitment has not been identified. The training time required for those over the age of 70 is reported to be longer (Marcy,1987; Schnabel,1987).

Employment:

The employment groups most commonly used as standardized patients include: professional and amateur actors and actresses (Coggan,1980; Jason,1971; Meadow & Hewitt,1972; Werner and Schneider, 1974; Lichstein & Newman,1985), college students(Renaud,1980; Page & Fielding,1980; Behrans,1979; Lincoln,1978;), health professionals and students(Hulzman,1977; Barnes,1978; Hanway,1980), retirees,housewives and those with flexible hours of employment(eg. shift workers, salespersons, clergy, university faculty (Owen & Winkler,1974; Barrows & Tamblyn,1979; Nowotony & Grove,1982; Kerr,1977). Certain employment groups are preferred because they self-select individuals who would have a greater aptitude/interest in playing the role of a patient(actors), have expertise in managing patient situations and a familiarity with the needs of health professional students (health professionals and students) or because they will be more readily available for training and use(eg. retirees).

Experience:

Standardized patient trainers have reported that it is easier to train an individual who has had similar experiences to the patient role for which they are being trained(Marcy,1987; Gliva,1980). For example, a person who has experienced lower back pain can be more easily trained to provide a realistic presentation of a herniated disc. Experience with the health problem was one of the criteria used in selecting standardized patients in Carroll & Hutchins(1978) report of standardized patient use in a medical interviewing course. Both Barrows (1971) and Coggan(1980) comment however on the potential hazards of training an individual with a problem which is too close to the reality of their own situation. The emotional response to their own situation may adversely influence their ability to accurately portray the patient role for which they are being trained.

Motivation & Intelligence:

Barrows(1971),Stillman(1986) and Coggan(1980) comment that motivation and intelligence are two of the most important factors in selecting a standardized patient. An initial screening interview and training session are customarily used to evaluate these attributes. It is recommended that individuals who are using the standardized patient role to seek health care assistance or vocalize their grievances against the health care system be avoided (Barrows,1971; Coggan,1980).

Stamina:

Stillman(1986) and Burri(1974) comment that standardized patients need to be able to withstand repeated examinations in a short period. This is particularly true when patients are being trained for large scale evaluations. Elderly individuals and those with chronic health problems appear to be at greater risk for fatigue related inconsistencies in the portrayal of their role. For example, Woodward(1985) found that the three real patients with chronic health problems who were being trained

for a quality of care evaluation project had to be dropped because of flare-ups in their disease process and inconsistencies in problem presentation.

Current Health Status,Physical and Psychological Attributes:

Two additional issues are considered in standardized patient selection. Physical and psychological attributes need to be considered in recruitment if they are important to the clinical problem which has been selected. For example, the height, weight and appearance of the individual may be important for a case of hyperthyroidism, delayed growth or obesity. Type-casting is a helpful training adjunct in certain physical and mental health problems(eg. an extrovert finds the repeated portrayal of a mania easier than an introvert).

When certain non-simulatable physical findings are required for the accurate portrayal of a clinical problem, individuals who possess those findings need to be recruited. Stillman(1980) was the first to report the active recruitment and use of individuals as standardized patients who possessed hard physical findings. Individuals with hard findings have been trained as both patient instructors (Stillman,1980) and as standardized patients(Woodward,1985; Stillman,1986; Klass, 1987).

Characteristics of Standardized Patients in Relationship to Role-Play, Real Patients and Patient Models/Instructors

Standardized patients are often confused with role-playing and patient models and instructors. A description of the key attributes of these various methods is provided in Figure 4.2. The central differentiating characteristic of the standardized patient is that they are assumed to provide an accurate reproduction of a real patient problem.

FIGURE 4.2 CHARACTERISTICS OF STANDARDIZED PATIENTS IN RELATIONSHIP TO ROLE-PLAY, REAL PATIENTS AND PATIENT MODELS/INSTRUCTORS

Role-Play:

Standardized patients are always used to provide an experience with a selected type of patient problem for evaluation, teaching or research purposes. The accurate and consistent presentation of the clinical problem is critical to the evaluation, teaching or research goal. Role-playing, in contrast, is used for teaching purposes. The student is given a role to play either to gain an understanding of a patient's experience or to allow for peer practice as the care-giver. Role-playing and standardized patient experience is often combined in interviewing courses with different educational goals for each exercise (Froelich, 1969; Quirk & Letendre, 1986).

Real Patients:

The main difference between real patients and standardized patients is that the latter have been trained to be reproducible in their presentation of the clinical problem. Barrows(1964), Norman(1982) and Newble (1980) have outlined some of the problems in using real patients for evaluation purposes

Patient Models/Instructors:

Patient models are usually lay individuals who are used to provide health professional students with experience in practising physical examination and interviewing skills (Godkins, 1974; Barrows, 1968). They are not used or trained to present a clinical problem. Patient Instructors are individuals who have been trained to provide feedback and instruction on interviewing or physical examination skills(Frazer & Miller,1977; Holzman,1977; Behrans,1979; Godkins,1974). In some instances they are selected because they have hard physical findings which can be used for teaching (Stillman,1980; Anderson & Meyer,1978). Lay persons, health professionals and patients have been used. The essential difference between a standardized patient and a patient instructor is that the latter has not necessarily been trained to present a particular patient problem, he/she are acting as a proxy for the faculty preceptor. The two roles may be combined and Stillman(1980) provides an example of this combination for undergraduate teaching.

In contrast to real patients, the standardized patient provides a reproducible presentation of a clinical problem at a fixed point in time

in it's clinical course. In role-playing, accurate reproduction of a real clinical problem is not a necessary requirement for it's effective use. Patient models and instructors are differentiated from standardized patients in that they are not used to present a clinical problem.

TRAINING

Presenting the Clinical Problem

The training methods outlined by Barrows (1971) are the most common methods reported for training standardized patients (Jason, 1971; Werner & Scneider, 1974; Lichstein & Neiman, 1985; Anderson, 1979; Vayda, 1976; Norman, 1985; Burri, 1976; Lamont & Hennen, 1972; Coggan, 1980; Stillman, 1987; Nowotony & Grove, 1982; Page & Fielding, 1980). The main features of the method have been outlined in Figure 4.3.

Using Barrows method, training time for a patient case is on average 2-4 hours in length (usually broken down into 2-3 sessions). Complex cases or those involving the simulation of physical findings take 4-5 hours. Prior standardized patient experience, type-casting, age and intelligence are factors which influence the length of the training time (Barrows, 1971; Marcy, 1986; Schnabel, 1986).

FIGURE 4.3 THE KEY FEATURES OF BARROW'S METHOD OF STANDARDIZED PATIENT TRAINING

1) The standardized patient is trained to recreate an actual patient case at a selected point in it's clinical course.

2) A protocol of the important clinical features of the clinical case is created and used in training. (Norman (1982) comments that historical details of the case should be based on the actual patient's recall since this may be quite different from the data documented in the medical record).

3) The provider who cared for the actual patient(or who has cared for similar patients) is involved in the training process.

4) The standardized patient is coached to understand the symptoms and emotional response to the problem from the patient's point of view. (This feature of the training process is thought to be essential for the patient to provide a realistic portrayal of the case. In addition it enables the standardized patient to provide a response to unanticipated inquiry in a manner which would be consistent with the actual patient) (Videotapes or an interview with the actual patient have been strategies used to accomplish this aspect of the training process (Nowotony & Grove, 1982)

5) Education about the illness or use of medical terminology is avoided in the training process unless used by the actual patient.

6) The standardized patient's own social-medical background is used to provide responses to inquiry which is unrelated to the actual problem being created. This eliminates the need to train the standardized patient for all possible 'negatives' or unrelated medical information which may be requested by the clinician. This practice shortens the required training time and maintains the reality of the simulation by providing flexibility in the patient's ability to respond to any line of inquiry which might be pursued. The standardized patient's own socio-medical background is of course reviewed by the trainer-clinician to ensure that it does not introduce conflicting data or 'red herrings' into the case being presented.

7) The standardized patient is coached on the information which would be provided spontaneously by the actual patient and that which would be provided only to certain types of inquiry. The emotional response of the standardized patient to various types of provider actions is coached in a similar manner. (This feature of the training process is important for both consistency and validity of the presentation.)

8) The standardized patient has the opportunity to practice the role and receive corrective feedback from the trainer.

9) An independent, experienced clinician 'works-up' the case at the completion of training and provides feedback on the credibility of the presentation.

10) The standardized patient presentation is reviewed each time before he/she is used for evaluation, research or educational purposes.

Providing Feedback and Instruction

Training standardized patients to provide feedback and instruction is relevant when they are being used for educational purposes. Meier et al. (1982) report on a self-instructional program used to train standardized patients to provide feedback to students on their interview process. The program took approximately 6 hours to complete. It consisted of training tapes, written instructions, self-assessment tests and a practice run with corrective feedback. Standardized patients who completed the training (n=22) received better ratings on the quality of their feedback from the patient trainer than non-randomized controls (n=19). Modest, non-significant differences were noted in the quality of patient feedback when rated by students and faculty, both favouring the self-instructional group.

Stillman (1980, 1986), Frazer & Miller (1977) and Behrans (1979) provide a detailed outline of the process used to train individuals to provide instructional feedback on selected aspects of the physical examination. Stillman (1986) estimates that 25-30 hrs. of training time is required to train each individual, 15 of which were carried out in groups. All authors trained individuals on the correct method of physical examination, some including instruction on related anatomy and physiology. The subjective feelings associated with correct and incorrect technique were covered in all programs through faculty demonstration. The ability to detect errors in technique and provide appropriate corrective feedback is pre-tested by faculty.

Recording Clinical Actions (history and physical) and Rating Satisfaction

Training the patient to rate or record events which took place in the patient-provider encounter is relevant when standardized patients are being used for evaluation or research purposes. The documented reliability of standardized patient recording/rating will be reviewed in a subsequent section. The training process for recording the content of the history and physical examination was reported by Stillman (1980). Stillman (1980)

employed a cyclical group training process to gain acceptable levels of reliability among recorders. Standardized patients observed an encounter and provided independent recording of actions taken. Inter-rater reliability was calculated and feedback/discussion used to explore areas of disagreement. The process was then repeated.

Carroll & Hutchins (1978) comment that standardized patients were trained to rate the doctor-patient relationship using a standardized rating form. No description of the training process is provided in this study or others who have used standardized patients for this purpose (Hannay, 1980; Stillman, 1986; Klass, 1987; Williams, 1987).

THE STANDARDIZED PATIENT: APPLICATION OF THE METHOD

The standardized patient has been used for educational, evaluation and research purposes. The reported applications of this technique in these areas will be summarized.

EDUCATION

Barrows (1968, 1971) has enumerated the advantages of using standardized patients in health professional curricula. The use of this method has been reported in the United Kingdom, Europe, Canada, Australia and the United States. It has been used to accomplish a variety of educational objectives including the teaching of:

Interviewing Skills (Werner & Scneider, 1974; Carroll & Hutchins, 1978; Jason, 1971; Froelich, 1969; Meadow & Hewitt, 1972; Wolf, 1987; Quirk & Letendre, 1986; Evans & Curtis, 1983; Carroll, 1981; Hannay, 1980; Lichstein & Neiman, 1985)

Assessment Skills (Whatling & Wodale, 1979; Barnes, 1978; LaSor, 1979; Vayda, 1976; Lincoln, 1978; Kerr, 1977)

History-Taking Skills (Engel, 1976; MacGuire, 1976, 1977)

Physical Exam Skills (Rubenstein, 1979; Frazer & Miller, 1977; Stillman, 1980; Holzman, 1977; Behrans, 1979; Godkins, 1974; Anderson & Meyer, 1978)

Patient Education Skills (Callaway, 1977; Anderson, 1979)

The standardized patient has been used for educational purposes with students in medicine(undergraduate & postgraduate), nursing, social work and pharmacy. Generally standardized patients are used to provide an opportunity to develop and practice skills prior to the assumption of responsibility for real patients. They are also used in clinical situations where supervised instruction is difficult to provide (emotionally sensitive, emergency or uncommon situations). Instruction typically takes place in small group sessions allowing opportunity for student practice and feedback.

EVALUATION

There is a growing trend to utilize standardized patients in the evaluation of clinical competence and quality of care. This has been done on a formative basis with the aim being to provide providers/students with feedback on performance strengths and deficiencies (Barrows & Tamblyn, 1976, Stillman, 1980; Mumford, 1987; Carroll & Hutchins, 1978; Lichstein & Neiman, 1985; Bishop, 1981; Stuart, 1980)

Recent efforts have focused on the use of the standardized patient in summative evaluation methods. The objective of these methods is to make decisions on student progress, licensure or certification on the basis of the documented presence of an acceptable level of clinical competence (Stillman, 1982, 1986, 1987; Klass, 1987; Williams, 1987; Lamont & Hennen, 1972; Hannay, 1979; Coggan, 1980; Nowotony & Grove, 1982).

The use of standardized patients has provided evaluators with the opportunity to evaluate dimensions of competence which cannot be assessed with written or computerized tests. Standardized patients are believed to have several advantages over the use of real patients. They provide the

opportunity for evaluators to select the specific case which will be used for evaluation in advance. Case-specific performance criteria can then be established prospectively. Selected emergency, psychiatric and emotionally sensitive problems can be used in the evaluation, a practice which would not be ethical or feasible with real patients. Finally, the contribution of real patient variation to measures of provider competence can theoretically be eliminated since the same problem is used to evaluate all students and the patient presentation is standardized.

Similar rationale has been provided by those who advocate or have employed standardized patients for the purposes of evaluating quality of care (Abernathy & Crowder, 1979; Burri, 1976; Renaud, 1980; Norman, 1985; Owen & Winkler, 1974; Woodward, 1985). Norman (1985), in a study of 17 Ontario physicians, found that 48% of the criteria used to evaluate performance were documented in the medical record. In contrast, the standardized patient's report of the same encounter indicated that 73% of criteria were performed. Owen & Winkler (1974) found little association in the physician's reported methods for managing a case of depression and what he/she actually did with the same patient problem when seen undetected in practice. Both authors conclude that the standardized patient provides a means of gaining information about the quality of practice which could not be ascertained by conventional methods.

RESEARCH

The appeal of the standardized patient for research purposes is based on the ability of the investigator to gain greater experimental control over the clinical situation. Standardized patients have been used to:

Evaluate changes in clinical performance as a result of educational interventions (DaRosa, 1982; Love, 1978; Barrows & Tamblyn, 1978; Greenburg, 1984)

Evaluate the concurrent validity of other performance measures (Page & Fielding, 1980; Tamblyn, 1979; Rethans & Boven, 1987; Norman, 1985)

Evaluate the nature and determinants of the clinical reasoning process (Elstein, 1978; Barrows, 1978; Ekwo, 1979)

Evaluate determinants of quality of care (Owen & Winkler, 1974; Renaud, 1980)

In addition Adler (1977) outlines the potential for utilizing standardized patients in research on the doctor-patient relationship.

The utilization of the standardized patient for evaluation and research has raised a host of questions about the reliability and validity of the method. The evidence currently available will be reviewed in the next section.

THE STANDARDIZED PATIENT: A REVIEW OF MEASUREMENT PROPERTIES

This review will focus on the use of standardized patients in the evaluation of clinical competence and performance. In Chapter 3, the components of clinical competence measures were reviewed. It was noted that the accuracy or validity of the numerical score or classification (eg. competent/incompetent) which results from a clinical competence measure will depend on the absence of random and systematic errors in each of the components of the measurement procedure and instrument. The standardized patient has the potential to contribute random and systematic errors in measurement in two of the components of the measurement instrument: in the presentation of the test stimulus (the clinical problem) and in the recording/rating of behaviour which results from the test stimulus. The research which has been carried out in relationship to each of these components will be reviewed.

COMPONENT 1: RANDOM AND SYSTEMATIC ERRORS IN THE PRESENTATION OF THE CLINICAL PROBLEM BY THE STANDARDIZED PATIENT

The Relationship of Measurement Properties Examined to the Theoretical Model of Clinical Competence

Measurement issues relevant to the presentation of the problem relate to three aspects of the theoretical model presented in Chapter 1: the clinical situation, competence and performance. The specific assumptions which have been evaluated are:

1) the standardized patient's presentation of a clinical problem is a valid representation of a real patient's presentation of a clinical problem

2) clinical actions taken in response to a standardized patient problem in an evaluation setting are valid indicators of clinical actions which would be taken in response to a real patient problem in the same evaluation setting and

3) numerical scores of clinical competence which are derived from the evaluation of actions taken with standardized patients in an evaluation setting are valid indicators of performance with real patients in a clinical setting.

The Evidence

Assumption #1: Validity of Problem Presentation

The validity of the standardized patient's presentation of a clinical problem requires the absence of random and systematic errors in problem presentation when compared with a real patient. In this sense the real patient serves as the gold standard against which the standardized patient's presentation is judged. Direct quantification of random and systematic errors in standardized patient presentation has not been reported in the literature.

An indirect method of evaluating the assumption of validity has been used. The approach which has been taken has been to evaluate a construct about the behaviour of clinicians in response to the measurement stimulus. The construct which has been evaluated most frequently is: if the standardized patient is a valid representative of a real patient, clinicians will be unable to detect a standardized patient who a) enters their practice or b) who is being used in conjunction with real patients to evaluate clinical competence. Papers reporting on the evaluation of this construct have been summarized in Table 4.1.

Detection rates vary among studies from 0%-67%. Higher detection rates are noted in situations where the clinicians knew the type of case being used in the evaluation (Norman, 1985) or were guessing from a limited number of choices (Norman, 1982). The evidence to date provides consistent support for the construct with no reported detection rate exceeding chance.

TABLE 4.1 PAPERS REPORTING ON DETECTION RATES FOR STANDARDIZED PATIENTS

Author	Year	#Patients	#Physicians	Method	Detection Rate
Owen & Winkler	1974	1 problem 10 stand pt 5=female 5=male	25	no knowledge of stand pt entry into practice saw 1 female and 1 male version of the same problem	detection=0/50
Burri et.al.	1976	1 stand pt	7	consent to see stand pt at unknown time	detection=1/7
Norman et al.	1982	real=4 simulated=4	10	residents saw 8 problems(4 real 4 simulated)in random order	guessed wrong= 33% guessed right= 67%(chance=50%)
Nowotony & Grove	1982	6 problems real=6 simulated=12 paired 1:2	109	each student saw 3 problems-blind to patient status	average wrong guesses=16.5 (sd=13.5) average right
Norman et al.	1985	7	17	all consent to see stand pt-specific time unknown 10 involved criteria development 7 no involvement	criteria grp= 11/28(39%) non-criteria grp=10/49(21%)
Rethans & Boven	1987	1 problem > 1 patient	48	all consent to see stand pt at unknown time in practice	detection=2/48 (4%) guesses=29.8 (sd=17.7)

Legend: # Patients: the number of patients studied (real patients and/or simulated patients)
Physicians: the number of physicians (or medical students) included in each study
Method: the method used to determine whether standardized patients were detected
Detection Rate: the number of times the standardized patient was correctly identified

Assumption #2: Validity of Clinical Behaviour Produced

For standardized patients to be a valid stimulus for measuring clinical competence, the clinical behaviour produced by a standardized patient must be predictive of that which would be produced by a real patient. In the evaluation of this assumption, differences may exist between real and standardized patient problem presentations but differences are assumed to be irrelevant if the resulting behaviour being measured is the same. The chief concern in using standardized patients as the measurement stimulus is the potential for systematic errors in competency estimates. This problem has been noted with paper problem simulations. Generally more clinical actions are taken and higher scores are achieved on paper versus real problem situations (Page & Fielding, 1980; Goran et. al., 1973). The exception which has been noted is in the area of patient teaching where fewer actions are taken with paper problems (Rethans & Boven, 1987).

Two methods have been used to evaluate this assumption: a direct comparison of the competency scores achieved by residents/students with real and standardized patient presentations of the same or equivalent case and an analysis of the clinical content which can be measured with each format of problem presentation.

Three studies have reported results using the direct comparison method. Norman et.al.(1982) compared the performance of 10 residents on 4 cases. Each case was seen twice by each resident in real and standardized patient formats. No differences were found in the amount of data collected on history and physical examination. Significant differences in format were found in the number of critical findings elicited but not for scores achieved on diagnosis or planned investigation. For these latter scores, differences as large as 15% were found between formats in one case however the power was insufficient to detect differences of this magnitude. The authors comment that the primary reason for observed format differences was a result of discrepancies in the data provided by the real

and standardized patient in one case. The scoring system was based on information in the real patient's record however the real patient failed to recall and provide this information during the encounter.

Nowotony and Grove compared the performance of 54 medical students on real and standardized patient presentations of 6 cases. Students were randomly allocated to format for each case and scores on history taking compared. No differences were found in 5 cases. Format differences were found in one case, a finding attributed to systematic errors in case presentation by one standardized patient.

Finally, Sanson and Poole (1980), using a repeated measures design, compared empathy scores for students rated in standardized patient and equivalent real patient situations. No significant differences in scores were found.

The second approach which has been taken is to identify the components of clinical competence which are important to measure in a real clinical situation and evaluate whether all components of interest can be measured in a standardized patient situation (a form of content-related evidence). Barrows and Tamblyn (1980) provide a summary of this form of analysis for standardized patients and other methods of problem simulation. Although the standardized patient measurement stimulus provides more comprehensive coverage of important components of competence when compared to paper and computer simulation methods, it does not provide a means of assessing long term management, or most technical/surgical skills.

Assumption #3: Validity of Clinical Competence Estimates

Comparison of scores from standardized patient-based tests with supervisor's ratings of clinical performance is the most common form of evidence provided to evaluate instrument validity. There are a number of limitations in this type of evidence.

Components of the measurement instrument and procedure, above and beyond

the use of standardized patients as the measurement stimulus, have the potential to contribute random and systematic errors to the reported relationships. For the measurement instrument, these components include sources of systematic and random error in: the specification of performance criteria, rating of performance criteria, and the assignment of a numerical score to criteria specified. These sources of measurement error are applicable to both the clinical performance measure and standardized patient measure of competence. For the measurement procedure, relevant components sources of error include: the comparability of clinical problem domains sampled, the sample size (number of cases sampled to derive overall estimates) and attributes of the measurement procedure (case length and evaluation setting). Finally, the relationship between competence and performance measures is being examined with the assumption that other determinants of performance in the practice setting, besides clinical competence, are not operative.

Swanson (1989) provides an excellent summary of studies which have examined this relationship along with data on a number of the component sources of error in standardized patient measurement. Table 4.2 summarizes the correlations found between standardized patient measures and clinical performance measures.

TABLE 4.2 PEARSON PRODUCT MOMENT CORRELATIONS BETWEEN STANDARDIZED PATIENT-BASED AND PAPER SIMULATION BASED TESTS OF COMPETENCE AND CLINICAL PERFORMANCE MEASURES

Author	Year	Format	Observed Correlation	Corrected Correlation
Norcini et.al.	1986	6 PMPs: practising physicians	.26	.27
Stillman et.al.	1987	30 min S.P. encounter:	.42	.52
Petrusa et.al.	1987	5 min. S.P. encounter: residents	.46	.73
Williams et.al.	1987	15-30 min S.P. encounter: med. students	.65	.75
Klass et.al.	1987	15-30 min S.P. encounter: med. students	.44-.52	
Petrusa et.al.	1988	5 min S.P. encounter: med. students residents	.37	.56
Webster et.al.	1988	16 PMPs: internal med. certification candidates	.30	
Stillman (in press) et.al.		10-15 min S.P. encounter: med. students	.25	.31

For the purposes of comparison, correlations found in two large studies of paper problem methods of presentation are also provided (Norcini, 1986; Webster, 1987). Both observed correlations and those statistically corrected for reliability are provided. Estimates of the component sources of error associated with clinical supervisor's ratings are unknown. Observed correlations of .25 to .67 have been found between standardized patient-based measures and clinical performance ratings. When the former is corrected for disattenuation, the true relationship is estimated to be in a range of .31 to .75.. In contrast, observed correlations of .27 to .30 were found between competency scores derived from paper problem presentations and clinical ratings of performance.

Component sources of error which were estimated in standardized patient-based ratings included those attributable to the sample size of cases & case length (a composite index) and those attributable to the recording of clinical behaviour by raters. Both components contributed measurement error to the estimates(ABIM,1988). It should be noted that neither of these two components are specific to standardized patient-based estimates. Comparability of clinical problem domains and case sample size are relevant sources of error with all performance-based measures. Raters are a potential source of error in all evaluation methods which use observers to record clinical behaviour.

COMPONENT 2 SYSTEMATIC AND RANDOM ERRORS IN THE RATING/RECORDING OF CLINICAL BEHAVIOUR RESULTING FROM THE MEASUREMENT STIMULUS

The Relationship of the Measurement Properties Examined to the Theoretical Model of Clinical Competence

In the evaluation of rater effects, the clinical situation is presented by a standardized patient. Competence tends to be defined on a case by case basis, usually by one or more medical faculty who are proficient in the content area of the case. The clinical behaviours which would be required for a competent performance on history, physical examination, communication, diagnosis and management are identified. Case-specific

checklists of identified criteria are created for data collection, communication and some immediate management items. The remaining components of competence are usually scored by inspection of the written record resulting from the patient encounter.

For checklist items, actions taken by the clinician are rated/recorded by either 1) the standardized patient who presented the case at the completion of the clinician-patient encounter or 2) a faculty member on the basis of direct observation of each encounter. It is the measurement of these aspects of competence where measurement error attributable to standardized patient rating/recording would be expected to have an impact on the resulting estimates of competence.

The assumptions which have been evaluated to date include:

1) the standardized patient's recall of actions taken by the clinician is equivalent to the documentation of actions recorded by direct observation

2) the standardized patient's rating of actions taken by the clinician will be equivalent when the same encounter is rated on two occasions

3) the actions rated by two standardized patients or two faculty observers will be equivalent

4) the actions rated by standardized patients by direct observation will be equivalent to those rated by faculty observers.

The Evidence

The evidence for the four assumptions evaluated has been summarized in Tables 4.3 and 4.4. In most studies the equivalency of rating has been summarized by using a Pearson product moment correlation coefficient. It should be noted that this statistic is not an appropriate index of agreement (Bland & Altman,1986). Substantial systematic differences in the

magnitude of scores assigned by different raters may exist and yet the correlation could still be 1. Indices which reflect observed agreement are more appropriate measures of equivalency when systematic differences in the magnitude of the score assigned by different raters is of concern. The studies reviewed provide less evidence about this potential source of measurement error. The generalizability and precision of the reported estimates of rater effects for the studies reviewed is limited by the small number of raters and encounters used in the respective estimations.

TABLE 4.3 STUDIES ESTIMATING THE RELIABILITY OF STANDARDIZED PATIENT RATING

Author	Year	#stand pts.	#episode per pt.	Tot	Pair	Content	Reliability Estimate % agree	K	R
Stillman et.al.	1980	CV=3 Pulm=4	?	?	M.D. + Pts	CV PE actions Pulmon PE actions			phi=.89 phi=.77
Carroll	1981	3	2	6	St.Pt	Communic Rating			r=.6 to .85
Norman et.al.	1985	7	2	14	res. asst.	Hist,PE, Manage actions	93%	.86	
Stillman et.al.	1986	?12	1	48	M.D.	Hist,PE actions			r=.7
						Communic rating			r=.52
		?12	1	37	res. asst.	Hist act			r=.82
						PE actions			r=.86
						Communic rating			r=.67
Stillman et.al.	1986	8	?1	?8	M.D.	PE actions	95%		
Rethans & Boven	1987	>1	3		M.D.	Hist,PE Manage actions	89-100%		
		>1	3		self	same	85-96%		
Stillman et.al.	1987	14	?	?	res. asst	Hist,PE actions			r=.93
						Communic rating			r=.77
Dawson-Saunders	1987	14	?	?	St.Pt	Data Coll. Communication Diagnosis Management	[difference in mean student scores for 2 patients presenting the same case-signif. on 5/7 cases]		
Williams et al.	1987	34	?	40	res asst.	Dat Coll Communic.	80%		
Petrusa et al.	1987	17	?	?	M.D.	Data Coll		.8	

Author	Year	#stand pts.	#episode per pt.	Tot	Pair	Content	Reliability Estimate % agree	K	R
Nieman et al.	1988	7	?	49	res assist.	Interview	indices used are not comparable		

Legend: Tot: total encounters rated (number of patients * number of encounters/patient)

Pair: standardized patient ratings were compared with other standardized patients, a research assistant or medical faculty

Content: the content of the items being rated specified as history (hist), physical examination (PE), communication (Communic), Cardiovascular (CV), pulmonary (pulm), data collection (data coll), diagnosis or management.

% Agree: the observed agreement for total score or the items rated

K: Kappa

R: Pearson product moment correlation unless otherwise indicated

Assumption#1 Direct Vs. Recalled Recording of Actions Taken

Two studies have examined the standardized patient's ability to recall and record actions taken in the encounter in comparison to actions recorded by direct observation. Observed agreement of 93% (kappa=.86) was found in Norman's (1985) study and 89%-100% agreement was found in the study by Rethans & Boven (1987). Nieman et al. (1988) reported on the ability of 7 geriatric standardized patients to recall and rate the interview behaviour of 49 medical students. The 'gold standard' used in the comparison was a coder's rating of the audiotape of the interview. Better recall was demonstrated for actions which were taken than for those which were not taken.

Assumption#2 Test-Retest Reliability of Standardized Patient Recording

The rater's ability to record actions in an equivalent manner on two separate occasions for the same encounter was evaluated by Rethans & Boven (1987) for standardized patient raters and by Mumford (1987) for faculty

raters. Similar results were obtained with observed agreement of 85% or greater for standardized patients and an intraclass correlation coefficient of .75 for faculty observers.

Assumption#3 Inter-rater Reliability-Standardized Patients & Faculty

The equivalency in rating by two or more standardized patients rating actions taken by the clinician on the same case has been evaluated in two studies. Carroll (1981) observed correlations of .6 to .85 among different standardized patient pairs in the rating of communication skills for the same clinician. Dawson-Saunders et al. (1987) compared the mean scores of students seen by different standardized patients simulating the same case in 7 cases. Significant differences were observed in 5/7 cases. Differences in student scores were more pronounced when the score depended to a greater extent on patient generated rating. Inequivalency in the student groups seen by different patients in this study may have contributed to the observed differences.

The equivalency of two faculty observers rating the same clinical encounter has been evaluated in four studies (Table 4.4). Observed agreement has been reported to be as low as 44% and as high as 86%. Pearson product moment correlations between observers are in the reported range of .64 to .9.

Assumption#4 Inter-Rater Reliability Between Standardized Patient and Faculty /Non-Faculty Observers

In the four studies which have been reported in this area, correlations between standardized patient and faculty/non-faculty observers tend to be high (Table 4.3: Range r=.52 to .93). Observed agreement was 95% in the one study in which it was reported. Lower correlations are observed for the rating of communication skills, a trend which is evident in the results of the studies of inter-rater agreement among faculty. This observation is likely attributable to the content of the area being evaluated rather than the type of rater which is being used.

TABLE 4.4 STUDIES ESTIMATING THE RELIABILITY OF FACULTY RATING

Author	Year	#raters	#episode per rater	Tot	Pair	Content	Reliability Estimate % agree	K	R
Andrew	1977	15	4-5	?	facul	PE Actions			
						Cardio			.94 (intra class)
						Pulmon			.93
						Neuro			.80
						Eyes			.82
Templeton et al.	1978	?24	?	?	facul	History		.78	
						PE			
						Manage			
						Communic			
Newble	1980	18	5*2	180	facul	PE actions			
						no train	70%		r=.9
						min train	63%		r=.9
						max train	44%		r=.8
Carroll	1981	2	2	4	facul	Communic Rating			r=.64
Mumford	1987	4	1	4	facul	Communic Rating			r=.87
		4	11	44	self	same			r=.75 (interclass)
Van der Vleuten & Lyk	1987	?	?	4510	facul	Hist,PE actions	med=86%		r=.84
						Communic Rating	med=75%		
Cohen et al.	1988	50	71	1775	facul	Data Coll Comm	Significant differences between raters in 5/25 cases		
Swanson & Norcini	(in press)	10-14/yr * 3 yrs		429	facul	PE			.68-.78 (intraclass)
						pt. educ			.5
						procedure skills			.76

Legend: Tot: number of raters * number of episodes/rater

Pair: faculty raters were either compared with themselves or other faculty

Content: the content of items rated: history (hist), physical examination (PE), communication (communic)

% Agree: observed agreement for total score or the items rated

K: Kappa

R: Pearson product moment correlation unless otherwise indicated

A SUMMARY OF THE MEASUREMENT PROPERTIES OF STANDARDIZED PATIENTS

The available evidence to evaluate measurement properties of the standardized patient has been summarized in the following eight points.

1) The inability to detect a standardized patient from a real patient provides weak but consistent evidence in support of the assumption that the standardized patient is a valid representative of a real patient. (i.e. the absence of systematic or random errors in presentation)

2) Although the evidence is limited, clinical behaviour in response to a standardized patient stimulus appears to be no different from that observed with real patients for the evaluation of competence in diagnosis, investigation and empathy. Differences in estimates of competence for data collection (history and physical exam) were found between a standardized patient and real patient stimulus in some cases. These differences appear to be attributable to systematic and/or random errors made in the content of the patient's presentation. Unlike paper problem methods, systematic biases associated with the stimulus format do not appear to be present.

3) There are systematic differences in the potential to measure important components of competence between real and standardized

patient methods of presenting the problem. These differences are confined to two areas: on-going management and technical/surgical skills. The potential to measure the remaining components of competence (see chapter 1); is comparable in both formats. Systematic differences in competency estimates would be anticipated if they include or are confined to competence in ongoing management and technical/surgical skills.

4) There do not appear to be systematic or random errors in the standardized patient's recall of actions taken in the clinical encounter when compared to direct observation. Problems have been noted in one study in the standardized patient's ability to recall actions which did not occur.

5) Although the test-retest reliability of standardized patients is good and equivalent to that of faculty observers, the evidence is too limited to draw firm conclusions.

6) Systematic and/or random errors in rating appear to be present when two standardized patients are presenting and recording actions taken on the same case for both communication and data collection skills.

7) Rater type (i.e. standardized patient vs. faculty/nonfaculty) does not appear to adversely influence inter-rater correlations of data collection or communication scores. Lower correlations among faculty and between faculty and patients are found for communication skills. Correlations of a similar magnitude are found for data collection skills for these two types of rater pairs.

8) Observed and corrected (for unreliability) correlations of standardized patient-based competency scores and clinical performance measures tend to be larger than those observed for paper methods of presenting the problem. Substantial intra-subject variability in competence scores is consistently

noted across cases. Erviti's (1980) analysis of variability in real patient-based chart audit scores suggests that inter-case variability in competence or performance scores is present within subjects when either standardized patient or real patients are used as the measurement stimulus. It is no known whether the magnitude of inter-case variability is affected by the format of problem presentation (i.e. increased or decreased by using real vs. standardized patients).

Variability in the content of the patient's presentation of the problem and in the rating/recording of the resulting clinician behaviour are both potential sources of measurement error in the evaluation of clinical competence. Variability in the content of patient presentation was the explanation offered for differences in scores with real and standardized patient formats in two studies. Although this would suggest that variability in the content of presentation may be an important source of measurement error, no study has provided data on the variability in presentation and it's impact on resulting estimates of competence. Inter-rater differences in the recording of clinical behaviour is a consistent finding in most studies. Although the precision and generalizability of these estimates are limited by small sample sizes, the magnitude of these differences does not appear to vary with rater type. A better understanding of the factors contributing to inter-rater differences would be required for this source of measurement error to be effectively minimized.

The problem with these two sources of measurement error (problem presentation and performance rating) is that they are both confounded with case in the evaluation procedure (i.e. different patients and raters are used for different cases). As a result, they both may contribute to the inter-case variability which has been consistently noted in performance--based measures of competence.

The consequences of inter-case variability is that overall estimates of competence or performance are imprecise if a small number of cases are used in the evaluation procedure. What is not clear is the proportion of

inter-case variability which is due to true case related differences in competence/performance and that which is due to those component sources of measurement error which are confounded with each case in the measurement procedure.

If most of the inter-case variability is the result of true differences in ability then more precise estimates can only be made if the sample size of cases used in the evaluation is increased, the domain from which cases are drawn is narrowed or the number of areas of competence evaluated with each case is reduced. These solutions are unrelated to the choice of measurement stimulus (i.e. real vs. standardized patient) or rater type(faculty vs. standardized patient).

Alternatively, if component sources of measurement error which are confounded with case are contributing to inter-case variability, then a more precise estimate of competence/performance can be obtained by minimizing their contribution. Components of the measurement instrument which are confounded with case include the standardized patient's presentation of the problem, the specification of case-specific performance criteria, the assignment of numerical values to criteria specified and the rating of observed clinical behaviour. The evidence to date would suggest that both standardized patient presentation and raters may contribute to case-related sources of measurement error.

The recent report by Swanson (1988) on reproducibility in standardized patient-based testing provided an estimate of the size of one of these case-confounded sources of measurement error: the effect of raters relative to persons tested and cases. Two data sets (Stillman,1987 & Newble,1988) were used to estimate the proportion of explained variance in test scores which were attributable to raters, cases, students and their respective interactions. In the Stillman (1987) study, the two standardized patients who presented the case were the raters. In the Newble(1988) study two faculty observers per case were used in the estimation of rater effects. In both studies, about half of the explained variance was accounted for by the combination of student*case and rater*case interactions. The independent effect of raters within case

accounted for a mere 2% of the explained variance in data collection skills in both studies. The main effect of raters was larger for communication skills (9%).

In Swanson's (1988) report, cases tended to have a larger main effect(20-40% of explained variance) than persons(4-8% of explained variance) in data collection skills whereas the reverse was true in communication skills. This may be a result of artifacts in the measurement instruments since communication skills are measured by the same instrument across cases whereas data collection skills are measured with case specific checklists. This issue relates to the two additional component sources of case confounded measurement error; criteria specification and numerical score assignment to criteria. Swanson and Norcini, in summarizing sources of measurement error in performance-based measures, concluded that systematic and random errors in criteria specification and numerical score assignment are likely operative when a small number of 'experts' are used in the generation of case-specific criteria and scoring mechanisms. Finally, the last identified case-confounded component of measurement error, standardized patient presentation has not been addressed in any study to date. It's impact on inter-case variability in scores is unknown.

CONCLUSIONS

In conclusion, case-confounded measurement error and true differences in performance ability across cases are likely both contributing to inter-case variability. In either instance, increasing the number of cases included in the estimation of overall competence/performance will improve the precision of the estimate. Efforts to identify and control important case-confounded sources of measurement error is a complementary strategy which could be used to improve the precision of competency estimates. The necessity to investigate these sources of measurement error can be justified on a number of grounds. They include:

Reduction in Evaluation Costs

The operating cost of evaluating competence on each case used in the evaluation has been estimated to be approximately $1,000/case for every 100 students (Klass, 1989). The estimated number of cases needed to provide a reproducible estimate of competence varies by the area of competence being measured with as many as 40-50 cases being required for a stable estimate of diagnosis and as few as 15 cases required for estimating communication skills (Stillman et al., 1986). Even if the number of cases could be reduced by only 5 through better control of case confounded sources of measurement error, it would have an appreciable impact on the cost and feasibility of evaluation.

The Identification and Control of Systematic Sources of Case-Confounded Measurement Error which Could Bias Competency Estimates

If systematic differences exist between patients presenting the same case in either case presentation and/or rating, the resulting estimates of competence for that case may be biased. This applies to situation where there is an interest in drawing inferences about competency to a domain of cases of this type. To avoid introduction of this source of bias, the magnitude of the problem needs to be estimated and methods of control formulated.

The Control of Sources of Measurement Error which would be Confounded with Evaluation Site in Instances where Multiple Sites are Being Used in the Evaluation of Competence

Two components of case-confounded measurement error: the content of standardized patient presentation and raters would be confounded with evaluation site if more than one setting was used for evaluation. Although measurement criteria, scoring, case selection and measurement procedure can be standardized across sites, for practical reasons, different patients and raters are used in each site to present the same clinical problem and rate the resulting performance. Measurement error

attributable to patients and raters could act to both attenuate differences which may exist between centres (through random error) or, more importantly, bias the estimates of competence generated for individuals tested in different centres through systematic errors in rating or problem presentation. These issues are of relevance to credentialing bodies who are considering the feasibility of implementing these methods of evaluation for national licensure and certification purposes. Multi-centre evaluation will be a requirement for test administration and therefore the contribution of these potential sources of measurement error will need to be addressed.

The subsequent chapter describes the research protocol developed to address these two areas of case confounded sources of measurement error; that attributable to systematic or random errors made in the patient's presentation of the problem and that attributable to the rating/recording of clinical behaviour by the standardized patient.

ABSTRACT

CHAPTER 5

THE EVALUATION OF SELECTED MEASUREMENT PROPERTIES OF STANDARDIZED PATIENTS: AN OVERVIEW OF THE THREE STUDIES AND GENERAL METHODS

The standardized patient method is a potentially powerful tool which can be used in the evaluation of clinical competence and performance. The major advantage of this method is the ability to achieve greater experimental control over the clinical situation without substantial loss in fidelity. It has been assumed that variance in competence or performance scores attributable to patients is virtually eliminated by the use of this method. This assumption has never been evaluated.

When standardized patients are used in the evaluation of competence or performance, they are used to present the clinical situation and rate/record the actions taken by the clinician during the encounter. Standardized patients are typically confounded with case in the measurement procedure. Individual standardized patients are used to present only one of the cases which may be included in the evaluation. As a result, measurement error attributable to standardized patient rating or presentation may be contributing to the observed within subject variance in competence score across cases. In addition, when more than one centre are being used in the evaluation process, standardized patients are also confounded with evaluation centre. Systematic differences in the accuracy of problem presentation or rating of behaviour by patients in different centres may bias the resulting estimates of competence.

In order to measure the potential contribution of standardized patients to measurement error, three studies were designed. The first study measured the accuracy with which standardized patients in two universities presented the critical clinical features of the case. A stratified random sample of 537 patient-student encounters were videotaped in two universities who were collaborating in a joint standardized patient-based evaluation of fourth year medical students in 1987. The sample provided equivalent representation of the 2 universities and 15/16 standardized patient cases which were used in the evaluation. After review, 456 of the

videotapes were of adequate technical quality for use in the evaluation of patient accuracy.

The second study was designed to investigate the relationship between the accuracy of patient presentation and competence score as well as factors which may be predictive of patient accuracy. A stratified random sample of 448 patient-student encounters were videotaped at one of the two participating universities in 1988. The encounters sampled provided equivalent representation of 16/18 of the standardized patient cases used in 1988. After review, 383 videotapes were of sufficient technical quality for use in the evaluation of patient accuracy. The method of measuring student competence score was prescribed by medical faculty for each case in a document referred to as the 'case blueprint'. Components of competence measured and score calculations are described.

The third study was designed to investigate the reliability of standardized patient raters. The sample of 456 usable encounters drawn in 1987 was used in this question in the analysis. Standardized patients who had presented each case in the two participating universities were used to estimate three forms of reliability: within rater, between raters from the same university and between raters from different universities. Characteristics of the rating form which might have influenced rater agreement were also evaluated.

CHAPTER 5

THE EVALUATION OF SELECTED MEASUREMENT PROPERTIES OF STANDARDIZED PATIENTS: AN OVERVIEW OF THE THREE STUDIES AND GENERAL METHODS

The standardized patient methodology was first developed by Barrows (1964). It has been used over the past two decades for educational, evaluation and research purposes in the health professions. It has untapped potential for broader application in the evaluation of clinical competence, performance and health care research.

The major advantage of this method for evaluation and research is the ability to achieve greater experimental control over the clinical situation without substantial loss in fidelity. Standardized patient methodology provides the investigator with the opportunity to select the precise clinical situation which may be of interest for research or evaluation purposes. In addition, it is assumed that variation in patient presentation, a common source of measurement error in evaluation studies with real patients, is eliminated with standardized patient methodology. This assumption has never been evaluated.

Substantial inter-case variability in competence scores has been consistently noted in performance-based tests. As a result, overall estimates of competence are imprecise unless either a large number of cases are used in the evaluation or the domain is narrowed. True differences in clinician ability and case-confounded sources of measurement error are likely both contributing to the variability in competency scores across cases. Further investigation and control of case-confounded sources of measurement error will:

1) permit the cost of evaluation to be reduced by reduction in the number of cases required for a precise estimate of competence

2) identify case-confounded systematic sources of measurement error which would bias competency estimates if unbalanced across the number of cases included in the evaluation.

3) identify and describe systematic and random sources of measurement error associated with two components of the test instrument: content of patient presentation and rater effects; both of which would be nested within site in multi-site evaluations.

The following series of studies have been designed to evaluate these two case-confounded sources of measurement error: measurement error attributable to systematic or random errors in the presentation of the patient problem and measurement error attributable to the reliability of recording/rating of actions taken by the clinician in the patient encounter. The overall objective of the three proposed studies is to estimate the contribution of these two sources of measurement error to variation in competence score and identify predictive factors which may subsequently be used to develop strategies for control.

RESEARCH QUESTIONS

STUDY 1 THE CONTENT OF STANDARDIZED PATIENT PRESENTATION

1. When the content of the real patient case is used as the 'gold standard', how accurate is the content of standardized patient presentation?

2. Is there a difference in the accuracy of standardized patient presentation when different patients are trained for the same case by different trainers in two institutions?

STUDY 2 PREDICTORS OF THE ACCURACY OF STANDARDIZED PATIENT PRESENTATION AND THE IMPACT OF ACCURACY ON PERFORMANCE SCORE

1. Are any of the following groups of factors associated with the accuracy of standardized patient presentation:

Group 1: Factors Which Could be Applied in Patient and Case Selection

a) Case Attributes
- *Case Complexity
- *Type of Clinical Features Included (history, physical findings, affect)

b) Patient Attributes
- *Age
- *Gender
- *Previous Acting Experience
- *Previous Simulation Experience
- *Previous Experience with the Health Problem

Group 2: Factors Which Could Be Applied During/at the Completion of Training

a) Patient Attributes
- *Patient Confidence in His/Her Ability to Accurately Present the Problem Post-Training

b) Training Attributes
- *Trainer's Confidence in Patient's Ability to Accurately Present the Problem
- *Training Length
- *Physician Assistance with Training

Group 3: Factors Which Could be Applied During or at the Completion of the Measurement Procedure

a) Procedural Attributes
- *Number of Sessions
- *Time Since Training

b) Encounter Attributes
- *Patient Confidence in the Quality of the Performance
- *Patient Satisfaction
- *Student Performance

2. Is there an association between accuracy of patient presentation and competency score?

   a) For Component Scores

   b) For Overall Competency Score

3a) Are there differences in the percent of items provided spontaneously by different patients presenting the same case to equivalent groups of students?

b) Are there differences in the variance in competency scores between patients presenting the same case between equivalent groups of students?

STUDY 3 THE RELIABILITY OF STANDARDIZED PATIENT RATERS

1. What is the reliability of standardized patient ratings/recordings of clinician's behaviour during the clinical encounter?

Within Rater Estimates

a) What is the test-retest reliability for the same standardized patient rating the same encounter on two separate occasions?

Between Rater Estimates

b) What is the inter-rater reliability for two standardized patients who were trained together for rating the same clinical encounter?

c) What is the inter-rater reliability for two standardized patients trained in two universities for rating the same clinical encounter?

2. Are there systematic differences in the:

a) competency scores derived from standardized patient rating?

b) in the proportion of students passing and failing as a result of standardized patient rating?

3. Are any of the following factors associated with the observed agreement of standardized patient raters?

a) Rater Pair Type

b) Rating Form Factors
*number of items rated
*type of items rated/recorded for the case
*the level of judgement required to rate/record the item
*the ambiguity of the item rated/recorded

THE THEORETICAL CONTEXT FOR THE PROPOSED RESEARCH

In order to place the proposed research into context, the relationship of the questions posed to the theoretical model of competence (Chapter 1) and the components of competency measures (Chapter 3) is described.

THEORETICAL MODEL OF COMPETENCE

As described in Chapter 1, clinical competence is conceptualized as the

ability of the clinician to generate those behaviours which are important determinants of patient outcome in response to the clinical situation being presented. This definition is atypical in that it provides explicit rather than implicit recognition of the relationship between expected clinical behaviour and patient outcome. It includes those behaviours required for a 'safe' performance and those required for 'quality of care', the more conventional labels employed in the literature. These behaviours are generally described in five categories (data collection, diagnosis, management, doctor-patient relationship and professional communication). They are most commonly identified by an expert panel of physicians. Prerequisites of these behaviours are assumed to include relevant knowledge, skill, judgement and attitudes. Performance of these behaviours in the practice setting is assumed to be related to both clinical competence and other attributes of the provider and practice setting. Finally patient outcome is conceptualized as being related to both the adequacy of provider performance as well as to other patient and health care system factors.

Two aspects of this theoretical model are being addressed in the studies proposed: the clinical situation and provider competence. In study 1, the clinical situation is the focus of study. The assumption which is being evaluated is whether the standardized patient provides a valid representation of important elements of a real patient situation. The real patient situation serves as the 'gold standard' against which the validity of standardized patient presentation will be judged. Evidence in support of the validity of the standardized patient will be evaluated by examining the accuracy with which the standardized patient presents important elements of the real patient situation.

In study 2, factors which may be associated with random and/or systematic errors in standardized patient presentation are evaluated. Factors derived from the theoretical model include attributes of the clinical situation selected and the interaction between the presentation and the resulting clinical behaviour. Characteristics of the training process and standardized patient are also being evaluated. In the second study question, the impact of measurement error attributable to the content of

the presentation on the measures of clinical behaviour produced will be quantified. This question addresses the relationship between the two aspects of the model, the clinical situation and provider competence. The aim of both questions is to provide guidelines for enhancing the validity of the clinical situation and the resulting measures of clinical competence.

Study 3 addresses one aspect of the theoretical model; provider competence. The effect of measurement error attributable to raters on the resulting estimates of competence is examined. Both systematic and random errors in the documentation and rating of clinical behaviour within and between raters will be evaluated. It is assumed that the criteria which are used to record/rate competence are important determinants of patient outcome or if not, that the resulting estimates of rater effects would be the same if different criteria were used.

FRAMEWORK FOR COMPONENTS OF COMPETENCY MEASURES

In Chapter 3, the components of competency measures were reviewed. To summarize, two levels of the measurement process were identified: the overall procedure and the individual instruments which are used in the procedure.

Components of the overall procedure included: specification of the domain of practice to which inferences are to be drawn, sample size and method measurement procedure(eg. test site) and the derivation and classification of overall competency scores. Components of the instrument included: the method of presenting the clinical situation (test stimulus), the specification of criteria of a competent performance, the method of assigning numerical scores to criteria and cases and the method of rating/recording clinical behaviour in response to the measurement stimulus.

The validity of estimates of clinical competence rest on the absence of important sources of systematic and random error in each of these components of the measurement process. As was noted earlier, the emphasis

to date has been placed on one of these components, the sample size required for a precise estimate of competence. The three studies proposed address two alternate sources of random and systematic error in measurement; the method of presenting the clinical problem (the test stimulus) and the rating/recording of behaviour exhibited in response to the measurement stimulus.

Study 1 and 2 address sources of measurement error attributable to the measurement stimulus and their predictors. Study 3 focuses on the contribution of rater effects on the measurement of competence using prospectively defined and standardized criteria of competence and related numerical scores.

OVERALL CHARACTERISTICS OF STUDY DESIGN

A multi-site, standardized patient-based evaluation of clinical competence is the data source which will be used to evaluate the research questions. In 1987 and 1988, the University of Manitoba and Southern Illinois University collaborated in the development and implementation of a common performance-based commencement examination of final year medical students. This examination was the first documented effort to train and use standardized patients to present the same clinical problem for evaluation in two different evaluation locations. This project, then, provides a prototype for evaluating the feasibility of multi-centre credentialing examinations. Most importantly, it provides the first opportunity to evaluate the effects of the two sources of measurement error which will be naturally confounded with evaluation site in the future application of this method; content of patient presentation and rater effects. Secondly, this project allows the independent contribution of these sources of measurement error to be evaluated while other case-confounded sources of measurement error (i.e.case specific performance criteria and numerical weighting) are controlled by standardization across evaluation sites.

The research questions will be evaluated using a cross-sectional sample survey design for study 1 and 3. The population will be stratified by

institution and clinical problem and an equivalent sample drawn from each stratum to enhance efficiency in estimation.

A prospective cohort design will be used for study 2. The population from one institution will be stratified by clinical problem and an equivalent sample drawn from each stratum. Data on potential predictors will be collected prior to the evaluation of patient performance(except for those measured during the encounter).

The sampling unit in all three studies is the standardized patient-student clinical encounter. In order to control for the effect of student performance in the evaluation of standardized patients and rater effects across problems, students were sampled so that performance capability would be equivalent across problems.

OVERALL CHARACTERISTICS OF THE METHOD

The source population, sample calculations, sampling methods, clinical evaluation procedure and performance measures are the same in all three studies. They will therefore be described as a group in this section. The specific procedures and instruments used in each study and the results will be described separately in subsequent chapters. The final chapter will summarize and discuss the results of all three studies along with recommendations for future applications and research on this method.

POPULATION

Target population

The general objective of this study is to gain information on the measurement properties of standardized patients which could be generalized to any setting which is training and using standardized patients for evaluation or research purposes. For practical reasons a representative sample of standardized patients cannot be drawn from the target population. Members of the target population have not been enumerated. The representativeness of the source population from which the study

sample will be drawn is therefore unknown.

Source population

The source population for study 1 and 3 is defined as all standardized patients used in the collaborative evaluation of final year medical students in two university settings in 1987(University of Manitoba and Southern Illinois University). Sixteen clinical problems were used in the evaluation in both universities. The 16 problems were identical. Standardized patients were recruited and trained in each institution to portray one of the sixteen clinical problems. A total of 22 patients were trained at Southern Illinois. Two patients were trained for 5 of the 16 problems. At the University of Manitoba, 35 patients were trained for the 16 problems with 2 or more patients being trained for each problem.

The performance of each student in each institution was evaluated on all 16 clinical problems. At the University of Manitoba 95 students were evaluated. At Southern Illinois University, 65 students were evaluated. The resulting number of standardized patient-student encounters from which the sample was drawn was 1040 encounters in Southern Illinois and 1520 encounters in Manitoba (total=2560). Figure 5.1 provides an overview of the process used to generate standardized patient-student encounters in both institutions.

FIGURE 5.1 GENERATION OF STANDARDIZED PATIENT-STUDENT ENCOUNTERS AT SOUTHERN ILLINOIS UNIVERSITY AND UNIVERSITY OF MANITOBA:1987

UNIVERSITY OF MANITOBA	SOUTHERN ILLINOIS	OUTPUT
Joint Definition of Expected Competencies		Competencies to be Tested
Joint Selection of 16 Clinical Problems For the Evaluation		Clinical Problems to be Used in Exam
3 Clinical Problems Developed for Evaluation	13 Clinical Problems Developed for Evaluation	Real Patient Case Selected Draft Case Blueprint Developed
Case Blueprints Reviewed & Revised (i.e. suitability of real case selected, competencies to be tested, performance criteria and scoring):Joint Review		Training Protocol of Essential Case Content Defined Case Specific Checklists + Questionnaires + Scoring Defined
Standardized Patients Recruited + Trained N=35	Standardized Patients Recruited + Trained N=22	Training Tapes Produced & Shared
Evaluation of 95 Final Year Medical Students on 16 Clinical Problems Standardized Pt Complete Case-Specific Checklists	Evaluation of 67 Final Year Medical Students on 16 16 Clinical Problems Standardized Pt Complete Case-Specific Checklists	Student-Patient Encounters: U of M=1520 S.I.U.=1040 Videotaped Sample of Encounters: U of M=240(15/case) S.I.U.=240(15/case)
Scoring of Written Case Responses by Faculty	Scoring of Written Case Responses by Faculty and Assistants	Case-Specific Checklists from Encounter & Written Response Score Used To Produce Student Scores
Analysis of Differences in Student Scores		
Comparison of Scores With Other Local & National Performance Measures	Comparison of Scores With Other Local & National Performance Measures	

For study 2, the source population consisted of all standardized patients used at the University of Manitoba in the comprehensive clinical evaluation of final year medical students in 1988. Eighteen clinical problems were used in the evaluation. Forty-six standardized patients were trained to portray one of the 18 problems. In three problems, more than two standardized patients needed to be used for each encounter(eg.mother and son). Two standardized patients or 2-6 standardized patient pairs were trained for each of the 18 problems. Ninety-eight students were evaluated on each of the 18 problems providing 1764 encounters from which a sample was drawn. Figure 5.2 provides an overview of the process used to generate encounters and predictive data in Study 2.

FIGURE 5.2 GENERATION OF STANDARDIZED PATIENT-STUDENT ENCOUNTERS AT THE UNIVERSITY OF MANITOBA:1988

UNIVERSITY OF MANITOBA	SOUTHERN ILLINOIS	OUTPUT
Joint Definition of Expected Competencies		Competencies to be Tested
Joint Selection of 19 Clinical Problems For the Evaluation		Clinical Problems to be Used in Exam
10 Clinical Problems Developed for Evaluation	4 Clinical Problems Developed for Evaluation 5 From '87 Exam	Real Patient Case Selected Draft Case Blueprint Developed
Case Blueprints Reviewed & Revised (i.e. suitability of real case selected, competencies to be tested, performance criteria and scoring):Joint Review		Training Protocol of Essential Case Content Defined
		Case Specific Checklists + Questionnaires + Scoring Defined
Standardized Patients Recruited + Trained N=35	Standardized Patients Recruited + Trained N=22	Training Tapes Produced & Shared
Predictive Data Collected Patient Questions.		Trainer & Standardized
Evaluation of 92 Final Year Medical Students on 19 Clinical Problems Standardized Pt Complete Case-Specific Checklists	Evaluation of 65 Final Year Medical Students on 19 Clinical Problems Standardized Pt Complete Case-Specific Checklists	Student-Patient Encounters: U of M=1748 S.I.U.=1235 Videotaped Sample of Encounters: U of M=532(28/case)
Scoring of Written Case Responses by Faculty	Scoring of Written Case Responses by Faculty and Assistants	Case-Specific Checklists from Encounter & Written Response Score Used To Produce Student Scores

Study Sample

Sample Size Calculations

Study 1:

The sample size required was estimated in two ways; the first was based on the desired width of the confidence interval for an estimate of the accuracy of patient problem presentation for each case; the second was based on an estimate of the sample size which would be required to detect a difference in the accuracy of patient presentation between the two universities.

No prior estimates of the variation in patient presentation are available for use in the calculations. In theory, the standard deviation has been assumed to be 0. A standard deviation of 5 and 10% were used in both calculations.

For the confidence interval estimates, a 95% confidence interval was specified. The sample size required for a confidence interval width of 4-10% was estimated using standard deviations of 5% and 10%.

For the second estimate, the sample size required to detect a difference of 5% in the average accuracy of the two universities was estimated using a standard deviation of 5% and 10%, a two-tailed test and a Type 1 error of 5% and a Type 2 error of 5%. Sample size estimates for both methods of calculation are displayed in Table 5.1.1.

TABLE 5.1.1 SAMPLE SIZE CALCULATIONS FOR STUDY 1: ESTIMATIONS BASED ON THE DESIRED WIDTH OF THE CONFIDENCE INTERVAL AND A 5% DIFFERENCE IN MEAN ACCURACY SCORE BETWEEN THE TWO UNIVERSITIES

Stand. Dev. Estimate	95% Confidence Interval Width				5% Mean Difference
	4%	6%	8%	10%	
5%	24	11	6	4	26
10%	96	43	24	15	104

The larger sample size was required to provide case specific estimates of average patient accuracy. A sample of 24 encounters per case was felt to be the maximum number which could be practically collected during the evaluation. With this number of encounters, the true value would be 4% above or below the estimated value 95% of the time(with a standard deviation of 10%) and 2% above or below the estimate if the standard deviation were 5%. This would result in a total sample size of 384(16*24) encounters, 12 drawn from each of the 16 cases presented in each university.

Finally, since encounters were being sampled by videotape, it was anticipated that as many as 15% of the encounters may be technically unusable (suboptimal taping facilities were present in both universities). An additional 48 encounters were added to the total number to be sampled in case of this eventuality.

Study 2:

Confidence interval estimates were used in the same manner to calculate the sample size requirements for the second study. A total of 24 encounters were estimated as being required for each of the 18 cases. This was inflated to 28 per case when the 15% potential discard factor was added. Using this method of sample size estimation, a total of 504 (18 cases * 28 encounters/case) encounters would be required for Study 2.

The sample size required to evaluate predictive factors was estimated using Cohen's (1977) sample size tables for multiple regression analysis (see Table 5.1.2). Eighteen predictive factors were to be estimated in relationship to standardized patient accuracy score(the dependent variable). A sample size of 298 would be required to detect an R^2 of 10% for 20 independent variables (273 for 16 independent variables) with a power of 95% and a type 1 error of 5%. Clearly, more than adequate power is provided to detect methodologically important predictive factors for patient accuracy using the former estimate of 504.

TABLE 5.1.2 SAMPLE SIZE CALCULATIONS FOR STUDY 2 USING COHEN'S POWER TABLES FOR MULTIPLE REGRESSION ANALYSIS

Number Independent Variables	Type 1 Error	Power	R^2 to Detect	L	Sample Size
16	.05	95%	.10	28.45	273
20	.05	95%	.10	30.72	298

Legend:

$$L = \frac{R^2_{y.x}}{1-R^2_{y.x}} * (N-\#\text{independent variables}-1)$$

$$N = \frac{L(1-R^2)}{R^2} + \#\text{ independent variables} + 1$$

Study 3:

For study 3, the sample size estimates were based on the number of encounters which would be required to detect a difference of 5% in the mean score of two raters rating each case or one rater rating the same case on two occasions. Specifying a two-tailed test, a Type I error of 5%, a Type II error of 5% and a standard deviation of 5% a sample size of 13 would be required for the within rater estimates and 26 for the between rater estimates for each case. A total sample of 413 would be required on

the basis of 26/case with 13 drawn from each case for each university. Adding a 15% discard factor, a total sample size of 480 would be required(15/case/university).

Sampling Procedure

In order to improve efficiency, a common set of encounters was used for both Study 1 and 3. Different research questions are being addressed in the two studies. However, the use of a common set of encounters provides the added advantage of allowing the link between presentation accuracy and rater reliability to be investigated in secondary analysis. The larger of the two sample size requirements dictated that a sample of 15 encounters needed to be drawn from each of the 16 cases in each university.

Videotaping was the method selected to capture the encounters for analysis. This provided the only means of carrying out the rater reliability studies and provided the opportunity to use the same observer to evaluate the accuracy of each case in both university settings.

In order to balance for the effect of student performance across cases and raters, encounters were sampled by drawing a random sample of students (15/university) and taping the 16 encounters those students had with the 16 cases used in the evaluation.

The taped encounters from the two university settings were grouped by case and reviewed for technical quality. Tapes which were inaudible were discarded. The remaining tapes were numbered. Institutional identifiers were removed. A random sequence of numbers was used to place videotaped encounters in random order. One tape of randomly ordered encounters was then created for each case. These 16 case tapes were then used for study 1 and 3.

The same procedure was used to sample encounters for study 2. Twenty-eight students were randomly selected from the 98 taking the examination at the University of Manitoba. The encounters of those 28 students with the 18 cases were taped. Tapes were reviewed for technical quality and placed in

random order.

CLINICAL EVALUATION PROCEDURE

The Selection of Clinical Problems for the Evaluation

Clinical evaluation committees were formed in both university settings. They were primarily composed of clerkship co-ordinators representing the major clinical departments (medicine, surgery, paediatrics, family medicine, psychiatry). The two university evaluation committees agreed on a common set of competencies which would be expected of their graduating medical students. Clinical problems were selected which would allow the expected competencies to be evaluated. The clinical problems selected were confined to those which would be commonly seen in a primary care setting or they were important because recognition and treatment would make a difference in patient outcome. The two university committees reviewed and came to consensus on a common problem list which would be used in the evaluation. The procedure was carried out for the 1987 and 1988 collaborative clinical evaluations. The clinical problems selected have been listed in Tables 5.2 and 5.3.

TABLE 5.2 CLINICAL PROBLEMS SELECTED FOR THE 1987 CLINICAL EVALUATION AND THE SPECIFIC AREAS OF COMPETENCE EVALUATED WITH EACH CASE

Clinical Problem	Areas of Competence				
	Data Collect	Diagnosis & Diff.	Management Test Interp	Communic Pt.Satis	Knowledge
1. Uncontrolled Hypertension	X	X	X		X
2. Episodic Chest Pain	X	X	X		
3. Lower Back Pain	X	X	X		
4. Sore Throat	X	X	X	X	
5. Hypertension	X	X	X		
6. Acute Abdominal Pain	X	X	X		X
7. COPD & Pneumonia	X	X	X		
8. Febrile Convulsions	X	X	X		
9. Progressive Memory Loss	X		X	X	
10. Sciatica	X	X	X	X	X
11. Headache & Wife Abuse	X	X	X	X	X
12. Infant Gastro	X	X			X
13. Diabetic Polyneuropathy	X	X	X		X
14. Weight Loss & Lymphadenopathy	X	X	X		
15. Anaphylaxis		X	X	X	X
16. Jaundice		X	X		X

TABLE 5.3 CLINICAL PROBLEMS SELECTED FOR THE 1988 CLINICAL EVALUATION AND THE SPECIFIC AREAS OF COMPETENCE EVALUATED WITH EACH CASE

Clinical Problem	Areas of Competence				
	Data Collect	Diagnosis & Diff.	Management Test Interp.	Communic Pat.Satis.	Knowledge
1. COPD & Pneumonia	X	X	X	X	
2. Pre-op Evaluation	X		X	X	
3. Sciatica	X	X	X	X	X
4. Paraplegia	X	X	X	X	
5. Urethritis	X	X	X	X	
6. Accident Prevention	X	X	X	X	X
7. Asthma	X	X	X	X	
8. Endometriosis	X	X	X	X	X
9. Jaundice		X	X	X	
10. Dysphagia	X		X	X	
11. Dizziness	X	X	X	X	
12. Panic Attacks	X	X	X	X	X
13. Abdominal Pain		X	X	X	X
14. Short Stature	X	X	X	X	
15. Hemiparesis & Headache	X	X	X	X	
18. Undiagnosed Hypertension	X		X	X	X
19. Alzheimers	X	X	X	X	
20. Uncontrolled Hypertension	X	X	X	X	X

The Development of the Clinical Problems for Evaluation

Members of the two university evaluation committees were assigned one or two of the problems selected to develop for the evaluation. The faculty member responsible for developing a clinical problem for the evaluation completed the following activities:

1) A specific real patient case was selected to represent the clinical problem.

2) The essential clinical features of the case were abstracted.

3) The competencies which could be evaluated with the case were listed.

4) The data which would be provided to the student prior to their evaluation of the patient was specified(eg. setting, complaint, initial lab data).

5) The actions which would be expected of the student during the clinical encounter were identified(eg. history, management, communication). Recording/rating forms to be used by the standardized patient to document those actions were developed.

6) Actions which may be critical in the evaluation of the patient (if applicable) were specified. Weights if appropriate were assigned to indicate the relative importance of the actions expected.

7) Activities to be completed at the completion of the encounter were specified. These activities were based on the objectives to be evaluated and included such activities as documentation of critical findings, diagnosis & differential, management plans, interpretation of lab data and tests of relevant clinical knowledge.

8) An answer key was developed outlining the acceptable responses for all activities listed.

9) The competencies to be evaluated with the case were linked to the related actions and post-encounter activities.

10) The level of performance required to successfully pass the case was identified.

11) Remedial activities required for students who did not meet the specified level of performance were prescribed.

The cases and materials developed by each faculty member were reviewed and revised by the two university committees. The resulting document was referred to as the case blueprint.

Standardized Patient Recruitment and Training

Once the case blueprint was developed, standardized patients were recruited and trained to portray the specific real patient problem selected for the evaluation. The standardized patient was selected to match the age and essential physical attributes of the case. If hard physical findings were required, a real patient who possessed these findings was recruited and trained for the specific case problem to be presented.

Standardized patients trained at Southern Illinois University were drawn from a pool of standardized patients which had been developed in the institution for teaching and evaluation purposes. Standardized patients trained at the University of Manitoba were drawn from community volunteers. None of the individuals used in the 1987 evaluation in Manitoba had had previous standardized patient experience.

The training process was co-ordinated and carried out by a standardized patient trainer identified in each university setting. Both university trainers had attended the same training workshop in the use of this technique. The trainer at Southern Illinois University had a number of years of experience in standardized patient training. The trainer at the University of Manitoba had no experience prior to the 1987 evaluation.

The essential clinical features listed in the case blueprint were used by both university trainers to prepare the standardized patients for the cases they were to portray. The case blueprint was described in the previous section. It contained all relevant information about the case to be presented as well as the methods of rating and scoring performance. The faculty person responsible for developing the case participated in the training process. Their clinical counterpart in the alternate university carried out the same function. One or two patients were trained to portray each case. When two or more patients were trained for a case, they were trained as a group to enhance comparability of presentation. The training procedure used was that developed by Barrows (1971) (Chapter 4). To improve the comparability of standardized patient presentation between the two universities, videotapes of the training process and/or standardized patient presentations were used when possible.

Standardized patients were oriented to the rating forms they were to use to rate student performance at the completion of the encounter. When two or more patients were trained for a problem, this was carried out in groups. Practice sessions were provided in form completion along with opportunities to discuss areas of ambiguity. Pre-testing of intra-rater and inter-rater reliability in form completion was not carried out.

The Examination Procedure

The evaluation took place over a two week period. Students were randomly assigned to time slots during the evaluation period. Each student completed the evaluation of the 16 clinical cases used in the examination over three consecutive days of testing. When two or more standardized patients were used to portray a clinical problem they were alternated on a daily basis between morning and afternoon testing sessions.

The Scoring of Student Performance

Two types of raters/recorders were used to score student performance. The standardized patient who portrayed the case was used to record/rate the actions taken by the student during the student-patient encounter. They completed rating forms developed by the case developer at the end of the encounter with the student. They usually had 5-10 minutes to complete this activity.

Faculty were responsible for scoring the written responses to post-encounter activities for each case. The faculty member who developed the case (or their clinical counterpart in the alternate university) was usually responsible for scoring. This policy was the most practical solution to the task of student scoring. The use of an independent faculty rater would be a better solution to control potential bias in scoring. The answer key developed as part of the case blueprint was used to score the quality of the written responses.

The Generation of Student Performance Scores

Three types of scores were generated for each student: a clinical case score, an overall competency score and overall specific competency scores. An example of the calculation of each score is provided in Figure 5.3.

FIGURE 5.3 EXAMPLE OF THE GENERATION OF STUDENT SCORES: OVERALL COMPETENCY SCORE, CASE SCORE & OVERALL SPECIFIC COMPETENCY SCORES

Case	Competency	Maximum Score	Student "x" Percent Score
#1			
Renal	Data Collect	11	72.73 (8/11)
Artery	Diagnosis	5	80.00 (4/5)
Stenosis	Management	6	33.33 (2/6)
	Knowledge	3	66.67 (2/3)
#2			
Chest	Data Collect	10	50.0 (5/10)
Pain	Diagnosis	8	75.0 (6/8)
	Management	12	75.0 (9/12)
	Patient Communic	8	75.0 (6/8)
.			
.			
.			
#16			
Jaundice	Diagnosis	9	44.4 (4/9)
	Management	18	66.7 (12/18)
	Knowledge	15	66.7 (10/15)

Student "x" Scores

Case Scores: Case 1= 63.18 [(72.73 + 80.0 + 33.3 + 66.67)/4]
Case 2= 68.75 [50.0 + 75.0 + 75.0 + 75.0)/4]
Case 16=37.25 [44.4 + 66.7 + 66.7)/3]

Overall Competency Score: Sum of (Case 1—Case 16 Case Scores)/16
eg.Overall Competency Score= 56.39 (63.18 + 68.75 + 37.25)/3

Overall Specific Competency Scores:

Data Collection= 61.37 (Case 1(72.73) + Case 2(50.0))/2
Diagnosis= 66.47 (Case 1(80.0) + Case 2(75.0) + Case 3(44.4))/3
Management= 58.32 (Case 1(33.3) + Case 2(75.0) + Case 3(66.7))/3
Knowledge= 66.67 (Case 1(66.67) + Case 3(66.67))/2
Patient Communication= 75.0 (Case 2(75.0))/1

Clinical Case Score:

The maximum score for each of the competencies tested with a case was calculated on the basis of the actions, post-encounter activities linked to each competency and their respective weights if applicable(the denominator). The sum of scores for actions taken by the student and the scores achieved on their written responses were used to calculate the numerator for each of the competency areas

tested. Percent scores were calculated for each of the case specific areas of competence. The clinical case score was calculated by taking the average of the scores achieved for case specific areas of competence. The areas of competence evaluated with each clinical case in 1987 and 1988 are listed in Tables 5.2 and 5.3.

Overall Competence Score:

The overall competence score was calculated by taking the average of all clinical case scores.

Overall Specific Competency Scores:

It can be noted in Tables 5.2 and 5.3 that specific areas of competence were usually evaluated in more than one case. The overall specific competency score (eg.data collection) was calculated by taking the average of all the scores for individual cases in which this specific area of competence was evaluated.

RESULTS

THE RESULTS OF SAMPLING

Table 5.4 displays the number of videotapes obtained for each of the 16 cases used in the 1987 evaluation in the two university settings. Southern Illinois University did not find that it was feasible to tape a random sample of their students due to co-ordination difficulties. Videotapes from Southern Illinois were sampled on a non-systematic basis with an effort to provide a sample of videotaped presentations for each case over the 2 weeks devoted to the examination procedure. This problem may result in a biased estimate of accuracy for patients at Southern Illinois University.

In 1987, it can be noted that no videotapes were obtained for Case 10 from Southern Illinois University. For this reason, Case 10 was discarded from the study population for Study 1. Case 10 was retained for Study 3. Videotapes from the University of Manitoba were sampled according to the procedure outlined. The percent of tapes which had to be discarded for inadequate audio quality were approximately the same in both university settings(SIU=15.06%; U. of M.=

15.1%). The resulting study population available for analysis for Study 1 and was 456 with 23 to 41 videotapes available for each case.

In 1988, there were no tapes collected for Case #5. The problem presented involved a male reproductive examination and the patients were not willing to be videotaped. Difficulties were encountered with Case 7. The standardized patient failed to show up during the second week of the evaluation. As a result, only 11 students were videotaped. This case was omitted from further analysis. A final sample size of 383 videotaped student-patient encounters was used in Study 2. Similar to 1987, 15% of the videotapes were discarded for inadequate technical quality.

TABLE 5.4 SAMPLING RESULTS FOR THE 1987 CLINICAL EVALUATION

CASE	SCHOOL				TOTAL	
	S.I.U.		U. OF M.			
	# taped	#adequate	# taped	#adequate	#taped	#adequate
1	33	23	18	15	51	38
2	14	14	19	17	33	31
3	12	12	20	17	32	29
4	13	12	19	19	32	31
5	7	7	19	19	26	26
6	10	8	18	15	28	23
7	18	16	17	16	35	32
8	12	10	16	15	28	25
9	13	12	18	15	31	27
10	-	-	18	16	18	16
11	23	15	17	17	40	32
12	12	11	18	17	30	28
13	24	20	20	16	44	36
14	12	12	19	18	31	30
15	24	21	22	20	46	41
16	12	10	20	17	32	27
	239	203	298	253	537	456
Percent Not Usable	15.06%		15.10%		15.08%	

Table 5.5 displays the number of videotapes obtained from the 1988 evaluation. Videotapes were sampled according to the procedure outlined for 1987.

TABLE 5.5 SAMPLING RESULTS FOR THE 1988 EVALUATION AT THE UNIVERSITY OF MANITOBA

CASE	# TAPED	# ADEQUATE
1	28	25
2	28	21
3	28	22
4	28	20
5	-	-
6	28	30
7	-	-
8	28	23
10	28	22
11	28	27
12	28	28
13	28	24
14	28	22
15	28	28
16	28	26
18	28	21
19	28	19
20	28	22
Total	448	383

Percent Not Usable: 14.5%

ABSTRACT

CHAPTER 6

STUDY 1: THE CONTENT OF STANDARDIZED PATIENT PRESENTATION

In the measurement of clinical competence, the standardized patient has been one method used to present the clinical situation (the test stimulus). It has been assumed that the standardized patient can provide an accurate reproduction of the important clinical features of a real patient case, thereby eliminating patients as a source of error in the evaluation of competence. This assumption was evaluated in Study 1.

A cross-sectional sample survey of 451 videotaped student-patient encounters from two universities, representing 49 standardized patients and 15 cases was used to evaluate standardized patient accuracy. Presentation accuracy was measured by recording the number of clinical features the standardized patient presented correctly in each encounter. The clinical features to be presented by each standardized patient were based on a real patient case which was identified by medical faculty.

A percent accuracy score was calculated for each standardized patient, case and university. If patients were perfectly standardized, they would have an accuracy score of 100% and a standard deviation of 0. This theoretical optimum was met by 7 of the 49 patients and an accuracy score of 95% or greater was achieved by 26 patients. For 7 patients, the mean accuracy score was below 75%. There was a statistically significant difference in accuracy score between the two universities, between two patients trained by the same trainer and among different cases.

Errors in the presentation of physical findings and patient affect were more common than for features of the patient history. Errors in these former two categories were made on more than 50% of occasions evaluated.

Both random and systematic errors in the presentation of the clinical features of each case contributed to suboptimal accuracy scores. Over all cases, 40% of the errors in presentation were systematic, the remaining

60% were random. Systematic errors are most likely due to problems in training. Systematic errors were more common in the university with the least standardized patient experience. University experience appeared to have no impact on errors made in the presentation of physical findings.

Patients who presented the same case spontaneously provided different amounts of data about their clinical problem. These differences could have been due to inequivalencies in the student groups seen by different patients. Alternatively, they could represent a potential source of bias in the estimation of clinical competence.

The impact of suboptimal accuracy in the presentation of the case on competency score will be evaluated in Study 2.

CHAPTER 6

STUDY 1: THE CONTENT OF STANDARDIZED PATIENT PRESENTATION

THE RESEARCH PROBLEM

In the evaluation of clinical competence and performance the standardized patient is one method which has been used to present the clinical problem. Other methods of presenting the clinical problem were reviewed in Chapter 3. They include: paper or computer simulation methods (eg. Patient Management Problem), oral examiners and real patients.

Systematic omissions of important aspects of the clinical problem have been the major disadvantages of paper and computer methods. Both systematic and random errors in the content of presentation are potential problems with the use of oral examiners. Both methods may lead to bias in the estimates of clinical competence if the aspects omitted are important determinants of the provider's ability to perform. This problem has been documented with paper simulation methods (Page & Fielding,1980; Rethans & Boven,1987).

The use of real patients in the measurement of clinical competence and performance has a number of disadvantages. In order to control for variability in the content of the clinical situation from one subject to the next, patients with chronic and stable findings must be used. This limits the range of problems which can be employed for evaluation purposes and provides no information on the subject's ability to manage problems of a more acute or sensitive nature. Alternatively, different patients with the same presenting problem or clinical diagnosis can be used. In doing so it must be assumed that the variability in competence/performance is a function of differences among providers rather than differences among patients. In addition, since patients are customarily screened for inclusion by virtue of the provider's documentation of the presenting problem or clinical diagnosis, estimates of provider performance are biased by exclusion of those patients who were inaccurately classified. This problem has been discussed by Tugwell (1979) in his critique of methods of evaluating quality of care.

The standardized patient has been proposed as a method of presenting the problem which overcomes many of the disadvantages noted with these alternate formats. It is assumed that systematic omissions in the content of the clinical situation presented noted with paper, computer and oral examiner methods are virtually eliminated. The only areas of competence which cannot be evaluated are long term management and some technical/procedural skills. If the cases selected for evaluation do not require the demonstration of provider ability in these areas, bias in the estimation of competence would not be anticipated.

In contrast to real patients, the use of standardized patients permits competence/performance to be estimated with a range of both chronic and acute problems. It is assumed that the problem of patient variability is eliminated from estimates of provider performance by standardization of the clinical presentation and the use of the same problem for all providers evaluated. Since the important clinical features of the patient's situation are known in advance, prospective case specific performance criteria may be established and the problem of detection bias noted with the use of real patients is eliminated.

If these assumptions are correct, the standardized patient provides a powerful methodological tool for the measurement of clinical competence and performance. The available evidence to support these assumptions was reviewed in Chapter 5.

Evidence offered in support of these assumptions includes:

1) standardized patients cannot be detected from real patients in a practice or evaluation setting (absence of systematic and random errors in presentation) (see Table 1, Chapter 5).

2) there is no difference in the estimates of competence in diagnosis, management and interpersonal skills with standardized patient and real patient problems (absence of systematic errors in competency estimates) (Norman,1982; Sanson & Poole,1980)

3) correlations between standardized patient-based estimates of competence and clinical performance are larger than with paper problem methods despite poorer precision (absence of systematic error in competency estimates) (Swanson,1989).

The evidence which does not support these assumptions includes:

1) the observation that significant differences in estimates of competence in data collection exist between real and standardized patient formats, a problem attributed to systematic and/or random errors in the content of patient presentation (Norman,1982; Nowotony & Grove,1982).

2) significant differences in estimates of case specific competence were found between presumably equivalent groups of students seeing two standardized patients who were trained to present the same clinical problem. This finding was attributed to either systematic differences in the content presented by the two standardized patients and/or differences in their rating of student ability (Dawson-Saunders,1987).

Although variability in the content of standardized patient presentation has been the suspected cause for differences in the estimates of provider competence between presentation formats or provider groups, no study has provided direct quantification of this potential source of measurement error. The research questions which will be evaluated in Study 1 will address this issue. The resulting data will allow two important assumptions about this methodology to be verified or negated.

Assumption #1: Variability within and between patients is eliminated when standardized patients are used to estimate provider competence.

Related Propositions:

a) The content of standardized patient presentation will be the same from one clinician to the next (i.e. no random error in the content of the presentation).

b) The content of standardized patient presentation will be the same for two or more patients trained to present the same clinical problem (i.e. no systematic error in the content of presentation between patients).

c) the content of standardized patient presentation will be the same for two patients trained to present the same problem by different trainers in different institutions (i.e. no systematic error in the content of presentation between patients).

Assumption #2: Standardized patients provide an accurate reproduction of the important clinical features of a real patient problem.

Related Proposition:

a) There will be no systematic or random errors in the content of standardized patient presentation when compared to the content of the real patient case on which it was based.

RESEARCH QUESTIONS

1. When the content of the real patient case is used as the gold standard, how accurate is the content of standardized patient presentation?

2. Is there a difference in the accuracy of standardized patient presentation when different patients are trained for the same case by different trainers in two institutions?

Definition of Terms

Accuracy: the extent to which all important clinical features of the real patient case are presented by the standardized patient. 100% accuracy = absence of systematic and random errors in the content of standardized patient presentation.

DESIGN

A cross-sectional stratified sample survey design was used to evaluate the accuracy of standardized patient presentation in the two study institutions: Southern Illinois University and the University of Manitoba. The same 16 clinical problems were used in the evaluation of students in both institutions. Strata is the term used to define the 2 levels of institution and the 16 levels of case which are represented in the study population. A stratified sampling approach was used to improve on the efficiency of estimation of differences in standardized patient accuracy between the two universities and to provide balanced representation of all standardized patients used in the evaluation since standardized patients were nested within clinical problem. The same number of standardized patient-student encounters was to be drawn from each stratum. In order to balance for the effects of student performance on standardized patient presentation, a random sample of students was drawn at the University of Manitoba and their encounters with each of the 16 clinical problems were used as the study sample. This was not feasible at Southern Illinois University. A convenience sample of encounters was drawn for each of the 16 clinical problems presented.

METHOD

SAMPLING PROCEDURE AND RESULTS

Patient-Student Encounters

Figure 6.1 provides an overview of the sampling procedure.
The student-patient encounters included in the study sample were videotaped during the conduct of the student evaluation in each university. All videotapes were reviewed for technical adequacy at the University of Manitoba. Those with inadequate sound or picture were discarded(approximately 15% from each university). The remaining videotapes were grouped by case. University identifiers were removed and all taped encounters from both universities were placed in random order and retaped to produce one or more randomly ordered tapes for analysis for each of the cases used in the evaluation. For Case 10, no student-patient

encounters were taped at Southern Illinois University. Since only 16 student-patient encounters were available for analysis of patient presentation, Case 10 was eliminated from the data set used in the evaluation of patient accuracy.

FIGURE 6.1 SAMPLING PROCEDURE FOR PATIENT-STUDENT ENCOUNTERS FROM THE 1987 STANDARDIZED PATIENT EVALUATION

S.I.U. N=65	Method	Tot.	Adeq.	Total/Case	Adeq.	Tot.	Method	U of M N=95
Case #1	C	33	20	35	15	18	R	Case #1
Case #2	C	14	14	31	17	19	R	Case #2
Case #3	C	12	12	29	17	20	R	Case #3
Case #4	C	13	12	31	19	19	R	Case #4
Case #5	C	17	7	26	19	19	R	Case #5
Case #6	C	10	8	23	15	18	R	Case #6
Case #7	C	18	16	32	16	17	R	Case #7
Case #8	C	12	10	25	15	16	R	Case #8
Case #9	C	13	12	27	15	18	R	Case #9
Case #10	C	0	0	16	16	18	R	Case #10
Case #11	C	23	15	32	17	17	R	Case #11
Case #12	C	12	11	28	17	18	R	Case #12
Case #13	C	24	20	36	16	20	R	Case #13
Case #14	C	12	12	30	18	19	R	Case #14
Case #15	C	24	21	41	20	22	R	Case #15
Case #16	C	12	10	27	17	20	R	Case #16
Total	C	239	200	453	253	298	R	Total

Legend: C: convenience sample of encounters

R: random sample of students selected and their encounters with the 16 cases was videotaped

Tot.: the total number of encounters videotaped in each university

Adeq.: the number of videotapes which were of adequate technical quality to rate patient accuracy.

Standardized Patients

Standardized patients were selected by the trainers in the two respective institutions. In Southern Illinois, most patients were selected from a pool which had been developed for evaluation and teaching purposes. At the University of Manitoba, standardized patients were recruited from the community.

For most cases, two standardized patients were trained for each case. In Case #9, two couples (mother, son) were trained to present the case. The total number of patients trained at Southern Illinois was 34 patients. Twenty-one of these patients were included in the convenience sample of videotapes provided. As a result, selection bias in the estimation of accuracy scores for Southern Illinois patients cannot be ruled out. In Manitoba, 32 standardized patients were trained (only one patient was trained for Case #6). All 32 patients are included in the study sample. Patient pairs used in Case #9 are treated as single patients in the subsequent analysis.

INSTRUMENT DEVELOPMENT

Accuracy Rating-Instrument Content

For each of the 15 cases, the 'blueprint' developed for each case was used as the basis for the development of accuracy checklists (see Chapter 6). The content of the real patient case served as the 'gold standard' for standardized patient accuracy. The essential clinical features of the real patient case were abstracted by faculty during case development and included in the case blueprint. These essential clinical features defined the content to be included in the accuracy checklists. They included important negative and positive findings. They can be categorized as including items related to the patient's affect, items to be presented on history and items to be presented on physical examination. The encounter checklists and scoring keys for written responses to the case were also reviewed. Clinical data which would need to be provided by the patient to achieve a correct response were also included if not mentioned in the list of essential clinical features.

The number and type of items contained in the checklists developed for each of the 15 cases are displayed in Table 6.1. The number of items per case varies considerably from 7 for Case 7 (an emergency situation of anaphylaxis) to 31 for Case 16 (a complicated problem of a middle-aged woman hospitalized on a number of occasions for undiagnosed jaundice). Most clinical features identified by faculty related to data which were to be provided by the patient on history. Physical findings and affect constituted 3.2% and 5.6% of items respectively.

This disproportionate breakdown of clinical feature items is likely due to two phenomena. For most cases, the data derived from the patient's history are thought to be more important than patient affect and physical findings in the formulation of the correct diagnosis and management plan. Secondly, faculty are more apt to select real patient cases which do not require the patient to present a large number of physical findings. Selection of cases with no physical findings eliminates the need to search for real patients with these hard findings or to train healthy individuals to simulate the physical findings required.

The small number of physical examination and affect items identified limits the generalizability of conclusions which can be drawn about the standardized patient's ability to accurately present important clinical features in these areas.

TABLE 6.1 THE NUMBER OF ITEMS (ESSENTIAL CLINICAL FEATURES) IN EACH ACCURACY CHECKLIST BY CASE AND TYPE

Case	Total # Items	Number By Clinical Feature Type		
		History	Physical	Affect
1-Uncontrolled Hypertension	11	10	0	1
2-Episodic Chest Pain	19	18	1	0
3-Lower Back Pain	17	16	1	0
4-Sore Throat	11	10	1	0
5-Undiagnosed Hypertension	8	8	0	0
6-Acute Abdominal Pain	19	16	1	2
7-COPD & Pneumonia	19	16	1	2
8-Febrile Convulsion	23	22	0	1
9-Progressive Memory Loss	8	7	0	1
11-Headache & Wife Abuse	26	26	0	0
12-Infant Gastroenteritis	17	15	0	2
13-Diabetic Polyneuropathy	16	16	0	0
14-Weight Loss & Lymphadenopathy	18	16	0	2
15-Anaphylaxis	7	3	2	2
16-Jaundice	31	30	0	1
Overall	250	228	8	14
% Breakdown		91.2%	3.2%	5.6%

Accuracy Rating-Instrument Categories

Seven nominal categories were developed to rate each content item included in the 15 case checklists. Three categories were used to characterize items which could not be assessed during the encounter: 1) the student didn't ask for information about the item, 2) the student didn't examine the patient to obtain information about the item or 3) the item could not be assessed because of the technical limitations of the videotape (eg. rater couldn't hear or see the patient's response). Four categories were used to characterize items which could be assessed in the encounter: 1) a correct response was provided spontaneously by the patient 2) a correct response was provided in response to student inquiry or examination 3) an incorrect response was provided spontaneously by the patient or 4) an incorrect response was provided by the patient in response to student inquiry or examination.

The categories, as constructed, provided the essential information necessary to calculate the accuracy of patient presentation by the dichotomous rating of whether the response to an item was correct of incorrect. It provided additional information on the conditions of the response; made spontaneously or to inquiry/examination. The data were used in secondary analysis of patient presentation to address two concerns: 1) are certain patients more apt to provide more of the essential clinical features of the problem to a student spontaneously thereby potentially inflating the resulting estimates of competence in data collection abilities and 2) are correct/incorrect responses more apt to occur under one response condition than the other (i.e. data were provided spontaneously or to student inquiry)?

The 15 case checklists which were used in the evaluation are found in Appendix 1.

RECRUITMENT AND TRAINING OF ACCURACY RATERS

Three graduate students in social work were trained to rate the accuracy of patient presentation. One rater was assigned to all of the encountered sampled for one case to eliminate the introduction of inter-rater sources

of measurement error in accuracy rating. The research associate for the project trained the rater to use the accuracy checklist. Training consisted of a review of the meaning of each checklist item and a practice session where both the research associate and rater trainee rated a sample of 2-3 encounters. Intra-rater reliability was then pre-tested by first having the assigned rater rate the first 10 encounters of each case tape. These were returned to the research associate and the rater was then asked to re-rate the same 10 encounters within one week. The interval time between the first and second rating may not have been long enough to be confident that recall did not positively bias observed agreement. Raters were available for a limited time which necessitated the use of a shorter interval between first and second rating. The observed agreement between the first and second rating was calculated. If the agreement in rating items was less than 90%, the rater was retrained and the pretest repeated until an observed agreement of at least 90% was achieved. The resulting intra-rater agreement achieved for each of the 15 cases is displayed in Table 6.2. Two categories of observed agreement are provided: agreement based on the congruence in rating for all 7 categories in the scale and agreement based on congruence between the rating of an item as correct or incorrect. It can be noted that observed agreement for the dichotomous rating of correct or incorrect was in the range of 96% to 100%. Observed agreement based on a 7 category scale was in the range of 90 to 100%. The majority of errors in rating were made in the conditions in which the response was provided (i.e. spontaneously or to inquiry).

TABLE 6.2 TEST-RETEST AGREEMENT OF PATIENT ACCURACY RATERS

CASE	N	OBSERVED AGREEMENT FOR RATINGS OF Correct vs. Incorrect (2 category)	Overall
#1	110	100%	90%
#2	190	99.5%	90%
#3	170	99.3%	90.7%
#4	110	99%	95.5%
#5	80	96.3%	91.3%
#6	190	100%	91.1%
#7	190	99%	95.36%
#8	230	100%	93.9%
#9	80	100%	90%
#11	260	100%	100%
#12	170	100%	95.3%
#13	160	100%	98.7%
#14	180	100%	94.7%
#15	70	95.7%	91.4%
#16	310	99.8%	95.8%

Legend: N: number of items to be rated * 10 pairs of encounters rated twice

Overall: Observed agreement for all scale categories, correct vs.incorrect (2 categories) response conditions (2 categories) and not evaluated (3 categories)

Correct vs. Incorrect: observed agreement for accuracy rating

PROCEDURE FOR RATING ACCURACY OF PATIENT PRESENTATION

Once a rater had achieved an intra-rater observed agreement of 90% on a 7 category scale, the rater was instructed to complete the rating of the remaining encounters on the case tape. The second rating of the first 10 encounters was used in the analysis of patient accuracy. Any queries about the rating of specific encounters or items were noted on the rating form. When it was unclear whether the patient had provided a correct or incorrect response, the patient was given the benefit of the doubt and the response was recorded as being correct.

ANALYSIS

Data Entry

Data were entered and verified directly from the accuracy checklists to computer to avoid errors in coding.

The Calculation of Accuracy Scores and Characterization of Error Type

Accuracy Score

An accuracy score was calculated for each standardized patient presentation using the following formula:

$$\text{Percent Accuracy Score} = \frac{\text{number of items correct}}{\text{(total number of items - number of items not asked/examined/evaluated)}} * 100$$

The score corrects the denominator for the number of potentially correct responses the patient was able to provide contingent on the actions taken by the student and the number of items which could be technically evaluated. The numerator represents the number of clinical features of the real patient case which were presented correctly by the standardized patient when the opportunity to do so was provided.

Characterization of Errors in Presentation By Type

The percent accuracy score falls below 100% when the standardized patient fails on one or more encounters to present an important clinical feature

of the problem correctly. These errors may be characterized as being systematic or random. This typology provides some insight into their likely origin and potential impact on competency score estimates.

Systematic errors in presentation are most likely due to errors in training whereas random errors may be due to a variety of other factors (eg. patient characteristics, student performance). Whereas random errors would act to reduce the precision of competency estimates and comparisons, systematic errors have the potential to bias comparisons if systematic errors are associated with the determinants being studied.

Errors in presentation are classified for each of the clinical features listed in the accuracy checklist for each case and standardized patient as being: absent, systematic or random according to the following definitions:

Absent: No errors were made in the presentation of the clinical feature in any of the opportunities in which it could be evaluated.

Systematic: An incorrect response for a clinical feature was provided in all encounters evaluated.

Random: An incorrect response for a clinical feature was provided in one or more of the encounters evaluated, but not in all encounters.

Descriptive Analysis and Hypothesis Testing

Missing Data

In order to examine for the presence of response bias in the evaluation of patient accuracy, descriptive statistics will be employed to summarize clinical features which could not be evaluated by case, type of clinical feature and institution. Clinical features are grouped and reported in three categories: patient affect, history items and physical examination items. They are the important findings in the real patient case which were identified by faculty and used in the checklist to score patient accuracy.

As was noted in Chapter 5, the proportion of tapes deemed to be technically inadequate was comparable across cases and institutions. The

introduction of response bias in accuracy comparisons among cases and between institutions through this route is unlikely.

Patient Accuracy

Descriptive statistics will be used to summarize the accuracy of presentation by standardized patient, case and institution. Frequency analysis will be employed to examine the number and type of errors made in presentation by patient, case and institution. The percent of items provided spontaneously and to inquiry will also be summarized by patient, case and institution.

Hypothesis Testing

It has been assumed that two or more standardized patients can be trained by the same trainer to provide an accurate presentation of a real patient case. This assumption will be evaluated by calculating the difference in accuracy scores for the patients trained to present the same case at the University of Manitoba and Southern Illinois. The null hypothesis that will be tested is that there is no difference in average accuracy score between patients who are trained together by the same trainer.
An independent t-test will be used to test the hypothesis on a case by case basis. Because the number of encounters is not balanced across cases, multiple regression analysis will be used to test the same assumption across cases. Case and standardized patient (nested within case) are the defined independent variables and accuracy score is the defined dependent variable.

It has also been assumed that trainers in different institutions can train different patients to provide an accurate presentation of the same real patient case. This assumption will be evaluated by examining the difference in accuracy scores for patients trained for the same case at the University of Manitoba and Southern Illinois. The null hypothesis which will be tested is that there is no difference in accuracy score between patients trained by different trainers for the same case.
An independent t-test will be used to test the hypothesis for each case and across all cases.

Multiple regression analysis will be used to estimate the relative contribution of university, patient and case to variation in accuracy score. Since patients are nested within case, a nested model will be employed to estimate the proportion of variance attributable to each factor.

The same approach will be used to evaluate differences in the percentage of items provided spontaneously (as opposed to in response to inquiry) between universities and among cases and patients.

Secondary Analysis

Secondary analyses will be carried out on the data with the aim of identifying additional factors which may have an influence on accuracy score. These factors include the relationship of accuracy score to the number of clinical features which must be presented with each case (range 7-31) and the time the presentation occurred in the course of the evaluation (week 1 to week 4). A one-way analysis of variance model will be used to examine the effect of time and patient accuracy. Regression analysis will be used to examine the effect of the number of items. Multiple regression analysis will be used to estimate the proportion of variance in accuracy attributable to these factors in addition to university, case and patients. The same approach will be used to evaluate the relationship between these factors and the percentage of items provided spontaneously by the patient during the student encounter.

RESULTS

MISSING DATA

Table 6.3 displays, for each case within each university, the percentage of times items which could not be evaluated by case and university. The percentages are divided into those where the encounter could not be evaluated for technical reasons and those which could not be evaluated because the student did not ask about or examine the clinical feature to be presented.

TABLE 6.3 FREQUENCY (%) WITH WHICH ACCURACY ITEMS COULD NOT BE EVALUATED BY UNIVERSITY, CASE AND REASON

CASE	# ITEMS	UNIVERSITY					
		Southern Illinois			University of Manitoba		
		Total No. of Items to be Evaluated	Freq. (%) Not Evaluated		Total No. of Items to be Evaluated	Freq. (%) Not Evaluated	
			(a)	(b)		(a)	(b)
1	11	220	0	14.1	165	0	21.2
2	19	266	.8	22.6	323	.6	27.2
3	17	187	5.3	25.7	306	1.0	21.2
4	11	132	0	22.0	209	2.4	36.4
5	8	56	0	35.7	144	0	36.1
6	19	152	0	22.4	285	1.8	30.5
7	19	304	3.6	16.1	304	6.3	23.0
8	23	207	0	40.6	345	0	39.7
9	8	96	0	14.6	120	2.5	37.5
11	26	390	.3	66.0	442	0	58.1
12	17	187	7.0	26.7	289	1.7	27.0
13	16	304	.7	25.0	256	0	39.5
14	18	168	0	19.1	252	0	21.4
15	7	154	4.0	15.0	140	3.6	14.3
16	31	310	2.3	19.7	527	0	22.8
Total	280	3133	1.69	27.71	4107	1.14	31.29

Legend: Total Items to be Evaluated: (the number of items) * (the number of student-patient encounters)

(a): the percentage of items which could not be evaluated because of the technical quality of the tape

(b): the percentage of items which could not be evaluated because the student did not ask or examine the patient

At the University of Manitoba, items could not be evaluated on approximately 32% of the 4107 opportunities to do so in contrast to 29.4% of the 3133 opportunities at Southern Illinois University. For all cases at both universities, the major reason an item could not be evaluated was because the student did not ask or examine the patient (U of M=31% S.I.U=28%). Technical reasons accounted for a 1.1% of the 32.4% of times items could not be evaluated at the University of Manitoba and 1.7% at

Southern Illinois. For Case 8 and Case 11 at both universities, items could not be evaluated 40% to 66% of the time. With the exception of one item on Case 8 (mother did not have pica during pregnancy) all items were evaluated on at least 10 occasions. Student performance was the major reason why items could not be evaluated in these two cases. The number of clinical features(items) available for presentation and the time constraints placed on the encounter (20 minutes) are likely responsible for the greater percentage of items not evaluated in these two cases.

Using a chi-square test, the difference in the proportion of times items could not be evaluated between the two universities is statistically significant (see Table 6.4). Differences are small (3%) and the analysis may be biased by violations in the assumption of independence among items. Differences in student performance between the two universities are responsible for this modest difference in the proportion of times items could not be evaluated. The potential bias created by this difference on patient accuracy score between the two universities is likely negligible. If average accuracy score at Southern Illinois were significantly less than that at Manitoba, differences in the challenge provided to the patients in the two universities could not be ruled out as a possible explanation.

TABLE 6.4 CHI-SQUARE ANALYSIS: ASSOCIATION BETWEEN % MISSING AND UNIVERSITY

University	Times Evaluated		
	Not Evaluated	Evaluated	Total
Southern Illinois	921 (29.4%)	2212 (70.6%)	3133
Univer. of Manitoba	1332 (32.4%)	2775 (67.6%)	4107
Total	2253	4987	7240

$X^2(1\ df) = 7.64 \quad p<.01$

Table 6.5 provides, for each case, a breakdown of the percentage of times that items could not be evaluated in the 3 clinical feature categories. History items constituted the greatest proportion of items evaluated and the category where a greater percentage of items could not be evaluated (32% of the opportunities available). Physical examination items could not be evaluated in 18% of opportunities available and affect items in only 10%. Again student performance was the major reason why history and physical examination items could not be evaluated. Affect items were relatively independent of student performance. In the evaluation of physical examination and affect items, the rater was more dependent on camera angle and picture quality (i.e. they needed to see the patient and their response to examination). This would explain why a larger percentage of these items could not be evaluated for technical reasons (physical exam=5.9% and affect=9.1%).

TABLE 6.5 FREQUENCY (%) WITH WHICH ACCURACY ITEMS OF VARIOUS TYPES COULD NOT BE EVALUATED

		ITEM TYPE								
		History			Physical			Affect		
CASE	# ITEMS	Total Items to be Eval.	Freq. (%) Not Eval.		Total Items to be Eval.	Freq. (%) Not Eval.		Total Items to be Eval.	Freq. (%) Not Eval.	
			(a)	(b)		(a)	(b)		(a)	(b)
1	11	350	0	18.9	35	0	0	--	--	--
2	19	558	.4	26.2	31	6.5	6.5	--	--	--
3	17	464	2.8	16.2	29	0	62.1	--	--	--
4	11	310	.3	32.9	31	12.9	9.7	--	--	--
5	8	200	0	36.0	--	--	--	--	--	--
6	19	345	0	33.9	46	10.9	8.7	46	0	0
7	19	512	1.8	22.3	32	12.5	12.5	64	50.0	0
8	23	528	0	41.9	--	--	--	24	0	0
9	8	216	1.4	27.3	--	--	--	--	--	--
11	26	832	.2	61.8	--	--	--	--	--	--
12	17	420	.5	30.5	--	--	--	56	28.6	0
13	16	560	.4	31.6	--	--	--	--	--	--
14	18	480	0	17.9	--	--	--	60	0	0
15	7	168	4.8	24.4	84	0	0	84	3.6	2.4
16	31	810	8.6	22.4	--	--	--	27	0	0
Total	280	6591	.73	31.08	253	5.93	12.25	396	9.09	.51

Legend: #items: number of essential clinical features identified from the real patient case

Total items to be evaluated: (number of items) * (number of student-patient interactions)

(a): % of times items could not be evaluated because of the technical quality of the tape

(b): the percent of times items could not be evaluated because they were not asked or examined by the student

There were significant differences in the proportion of times items could not be evaluated among the 3 clinical feature categories (see Table 6.5). Violations in the assumption of independence among items may have biased this estimate. These differences would be important if the probability of a correct response was greater or less for those items

which could not be evaluated on history. If this were the case, accuracy scores may be under or over estimated for cases with a large proportion of history items and a greater percentage of missed evaluation opportunities (eg. Case 8 and Case 11). Similarly accuracy scores may be under or over estimated for the standardized patient's ability to provide history items relative to those on physical examination or affect. The patients, cameramen and students were not aware that accuracy of presentation was being evaluated during the course of the examination. It is therefore unlikely that systematic differences in the probability of a correct response were introduced through these routes.

TABLE 6.6 CHI-SQUARE ANALYSIS: ASSOCIATION BETWEEN % OF ENCOUNTERS WHERE ITEMS COULD NOT BE RATED BY ITEM CATEGORY

Item Type	Times Evaluated		
	Not Evaluated	Evaluated	Total
History	2169 (32.9%)	4422 (67.1%)	6591
Physical	46 (18.2%)	207 (81.8)	253
Affect	38 (9.6%)	358 (90.4)	396
Total	2253	4987	7240

$X^2(2\ df) = 114.9 \quad p<.001$

PATIENT ACCURACY

The theoretical optimum for standardized patient presentation is an accuracy score of 100% and a standard deviation of 0. When this situation prevails the standardized patient has correctly presented all important clinical features identified in the real patient case, when given the opportunity, and has done so on repeated occasions across all student-patient encounters evaluated. A score of less then 100% would be achieved under two possible conditions:

1. The standardized patient provides an incorrect response to one or more items on all occasions in which the problem is presented. In this situation, a score of less than 100% is achieved but the standard deviation is 0. The patient in this situation provides a standardized performance (the absence of random errors in presentation) meeting the conditions of assumption #1 (variability in patient presentation is eliminated) but makes consistent or systematic errors in the presentation of the problem thereby failing to meet assumption #2 (accurate reproduction of a real patient case).
2. The standardized patient provides an incorrect response on one or more items on some but not all of the occasions in which the problem is presented. In this situation, a score of less than 100% is achieved and the standard deviation is greater than 0. Since more than one item is contributing to accuracy score, a combination of both systematic and random errors or pure random error in clinical feature presentation could be contributing to the observed variability in score. In either situation, assumption #2 would not be met. To determine the extent to which assumption #1 is met, a breakdown of item errors into those which are random versus those that are systematic is required.

Table 6.7 displays the mean percent accuracy score, standard deviation and 95% confidence interval of the mean for all standardized patients by case and university. The theoretical optimum for standardized patient presentation was met by 7 of the 49 standardized patients, 4 of those patients were from Southern Illinois and 3 were from Manitoba. There was no case in which the theoretical optimum was achieved by all patient presenters. Accuracy scores of 95% or greater were achieved by 26/49 standardized patients (13/20 from S.I.U. and 13/29 from U of M) and for all patients in 5 of 15 cases. Accuracy scores of 90% or greater were achieved by 31/49 patients(S.I.U.=14/20; U of M=17/29) and 6 of 15 cases. For 7/49 standardized patients, the mean accuracy score was below 75%, with 2 cases (Case 4 and Case 13) accounting for 6 of these scores.

TABLE 6.7 % ACCURACY SCORES BY UNIVERSITY, PATIENT AND CASE

CASE	N	UNIVERSITY	PT	N	% ACCURACY	S.D.
#1	35	S.I.U.	#1	20	89.3	1.8
		U of M	#1	5	83.1	4.9
		U of M	#2	10	78.0	10.6
		Overall		**35**	**85.2**	**7.8**
#2	31	S.I.U.	#1	14	95.9	4.0
		U of M	#1	10	99.4	1.9
		U of M	#2	7	89.8	6.3
		Overall		**31**	**95.6**	**5.4**
#3	29	S.I.U.	#1	4	82.3	22.2
		S.I.U.	#2	7	98.4	.2
		U of M	#1	12	90.3	8.0
		U of M	#2	6	97.3	4.4
		Overall		**29**	**92.6**	**10.7**
#4	31	S.I.U.	#1	12	74.6	9.6
		U of M	#1	3	70.6	24.1
		U of M	#2	16	69.3	18.4
		Overall		**31**	**71.5**	**15.8**
#5	25	S.I.U.	#1	7	100.0	0.0
		U of M	#1	16	88.2	10.0
		U of M	#2	2	100.0	0.0
		Overall		**25**	**92.5**	**9.8**
#6	23	S.I.U.	#1	3	100.0	0.0
		S.I.U.	#2	5	98.5	3.4
		U of M	#1	15	98.0	3.4
		Overall		**23**	**98.4**	**3.1**
#7	32	S.I.U.	#1	16	74.2	5.6
		U of	#1	5	91.0	7.5
		U of M	#2	11	79.9	9.6
		Overall		**32**	**78.8**	**9.3**
#8	24	S.I.U.	#1	9	100.0	0.0
		U of M	#1		98.8	2.8
		U of M	#2	10	100.0	0.0
		Overall		**24**	**99.7**	**1.3**
#9	27	S.I.U.	#1	12	93.4	8.8
		U of M	#1	5	83.5	10.6
		U of M	#2	10	94 6	9.2
		Overall		**27**	**92.0**	**9.9**

CASE	N	UNIVERSITY	PT	N	% ACCURACY	S.D.	
#11	32	S.I.U.	#1	8	100.0	0.0	
		S.I.U.	#2	7	97.0	5.2	
		U of M	#1	8	98.2	3.4	
		U of M	#2	9	100.0	0.0	
		Overall		**32**	**98.9**	**3.1**	
#12	28	S.I.U.	#1	11	97.2	5.4	
		U of M	#1	5	86.9	4.5	
		U of M	#2	12	84.1	5.1	
		Overall		**28**	**89.7**	**8.0**	
#13	35	S.I.U.	#1	11	95.0	11.1	
		S.I.U.	#2	8	74.0	15.4	
		U of M	#1	13	69.5	9.7	
		U of M	#2	3	73.4	9.5	
		Overall		**35**	**79.3**	**15.7**	
#14	30	S.I.U.	#1	12	89.2	8.2	
		U of M	#1	5	89.8	5.5	
		U of M	#2	13	88.0	4.2	
		Overall		**30**	**88.6**	**6.2**	
#15	42	S.I.U.	#1	11	97.0	6.7	
		S.I.U.	#2	11	98.5	5.0	
		U of M	#1	9	95.2	10.1	
		U of M	#2	11	98.5	5.0	
		Overall		**42**	**97.4**	**6.7**	
#16	27	S.I.U.	#1	10	97.3	3.4	
		U of M	#1	15	99.7	1.1	
		U of M	#2	2	97.7	3.2	
		Overall		**27**	**98.7**	**2.5**	
Overall Cases				**451**	**90.2**	**12.2**	**89.1, 91.3**

Legend: N: the number of taped student-patient encounters overall and by patient

PT: the patient presenting the problem (different patients presented each case)

In all instances where the average standardized patient accuracy score fell below 100%, the standard deviation was greater than 0. This suggests that pure random error or a combination of systematic and random errors are contributing to score variance.

Categorization of Item Errors as Systematic or Random

A breakdown of all items in which one or more errors was made by one or more patients was carried out. Table 6.8 provides a breakdown of the percentage of errors by case, university and error type.

TABLE 6.8 THE CATEGORIZATION OF ERRORS IN ACCURACY ITEMS AS SYSTEMATIC OR RANDOM: BY CASE AND UNIVERSITY

Case	University	Total # Items	Total # Items with Error (%)	Breakdown of Item Errors %(n) Systematic	%(n) Random
1	S.I.U.	11	1 (9.1)	100.0(1)	
	U of M	22	6(27.3)	50.0(3)	50.0(3)
	Total	33	7(21.2)	57.1(4)	42.9(3)
2	S.I.U.	19	3(15.8)		00.0(3)
	U of M	38	4(10.5)	50.0(2)	50.0(2)
	Total	57	7(12.3)	28.6(2)	71.4(5)
3	S.I.U.	33	3 (9.1)		100.0(3)
	U of M	34	6(17.7)		100.0(6)
	Total	67	9(13.4)		100.0(9)
4	S.I.U.	11	4(36.4)		100.0(4)
	U of M	22	10(45.5)	30.0(3)	70.0(7)
	Total	33	14(42.4)	27.2(3)	63.6(11)
5	S.I.U.	8	0		
	U of M	16	1 (6.3)		100.0(1)
	Total	24	1 (4.2)		100.0(1)
6	S.I.U.	38	1 (2.6)		100.0(1)
	U of M	19	2(10.5)		100.0(2)
	Total	57	3 (5.3)		100.0(3)
7	S.I.U.	19	7(36.8)	57.1(4)	42.96(3)
	U of M	38	11(29.0)	36.4(4)	63.6(7)
	Total	57	18(31.6)	44.5(8)	55.5(10)
8	S.I.U.	23	0		
	U of M	46	1(2.2)		100.0(1)
	Total	69	1(1.5)		100.0(1)
9	S.I.U.	8	2(25.0)		100.0(2)
	U of M	16	2(6.3)	50.0(1)	50.0(1)
	Total	24	4(16.7)	25.0(1)	75.0(3)

Case	University	Total # Items	Total # Items with Error (%)	Breakdown of Item Errors %(n) Systematic	%(n) Random
11	S.I.U.	52	1(1.9)	100.0(1)	
	U of M	52	1(1.9)		100.0(1)
	Total	104	2(1.9)	50.0(1)	50.0(1)
12	S.I.U.	17	2(11.8)	50.0(1)	50.0(1)
	U of M	34	6(17.6)	66.7(4)	33.3(2)
	Total	51	8(15.7)	62.5(5)	37.5(3)
13	S.I.U.	32	13(40.6)		100.0(13)
	U of M	31	8(25.8)	75.0(6)	25.0(2)
	Total	63	21(33.3)	28.6(6)	71.4(15)
14	S.I.U.	18	4(22.2)		100.0(4)
	U of M	36	6(16.7)	33.3(2)	66.7(4)
	Total	54	10(18.5)	20.0(2)	80.0(8)
15	S.I.U.	14	2(14.3)		100.0(2)
	U of M	14	4(28.6)		100.0(4)
	Total	28	6(21.4)		100.0(6)
16	S.I.U.	31	3 (6.5)	33.3(1)	66.7(2)
	U of M	62	2 (3.2)		100.0(2)
	Total	93	5 (5.4)	20.0(1)	80.0(4)
Overall		812	116(12.3)	32.8(38)	67.2(78)

Legend: Total # Items: the number of items evaluated in each item category times the number of patients presenting those items

Total # Items with Errors: the number of items where an error was present times the number of patients making an error on that item (# patients=1-2/item)

% Systematic: An incorrect response for a clinical feature was provided in all encounters evaluated

% Random: An incorrect response for a clinical feature was provided in one or more of the encounters evaluated, but not in all encounters

Over all cases, 116 item errors were made in the 812 opportunities to do so. Of these 116 errors, 33% met the criteria of a systematic error and 67% met the criteria of a random error. This overall split varied among cases. In cases where very few errors were made, stable estimates of error type are not possible. Ignoring the cases with less than a 10% error rate, 5 of the remaining 10 cases had a larger proportion of random errors

varying from 70-100% (Cases #3,9,13,14,15), four were about equally divided between the two types of errors (Cases # 1,2,4,7,) and in Case #12, a greater proportion of errors were systematic. In the 3 cases with the greatest proportion of errors (Case #4,7,13), random and systematic errors were comparably distributed in one case and were mostly random in the remaining two cases.

Systematic errors are fairly clearly attributed to a problem in training. Over all cases, 33% of the errors were likely due to inadequacies in the training process. This varied across problems with as many as 62% of errors in Case #12 attributable to training problems and as few as 0% in Case #15. This suggests that the challenge of training patients may vary with the problem to be presented. In order to better understand other factors which may contribute to error type, item errors were separated by patient, university and item type.

Categorization of Errors by Patient

Table 6.9 provides a summary of the types of errors made by patients in the two universities. As was noted previously, 7 patients made no errors. An additional 6 patients made systematic presentation errors only. In this situation although the content of the presentation was not 100% accurate, there was no variability in the presentation from student to student. These patients may have contributed to a bias in competency score estimate but had minimal potential to contribute to a lack of precision in case-specific competency scores. Both systematic and random errors in presentation contributed to accuracy scores of less than 100% in the remaining 36 patients. These patients have the potential to contribute to both a bias in estimating case-specific competency scores and a lack of precision in the estimation.

TABLE 6.9 CATEGORIZATION OF STANDARDIZED PATIENTS BY TYPE OF ERRORS MADE IN PRESENTATION

Error Type	University		Total
	S.I.U.	U of M	
No Errors	4	3	7
Random Errors Only	9	10	19
Systematic Errors Only	1	5	6
Mixed-Random + Systematic	6	11	17
Number of Patients	20	29	49

The stability of these estimates of error type by patient is compromised by the small number of events per patient. It is assumed that classification errors are equivalently distributed between systematic and random error types.

Categorization of Errors by University

Table 6.10 displays errors by frequency and type for each university. Errors were made on more items by patients at the University of Manitoba (22% of items) than by patients from Southern Illinois (17% of items). The majority of items were presented correctly by patients from both universities in all encounters evaluated (SIU = 83% of items; U of M = 78% of items). For items where one or more errors were made, the frequency of errors was about the same for both Universities (SIU = 14%; U of M = 15%). For errors made in both universities, there is a significant difference in the proportion of systematic errors made by patients in Southern Illinois in comparison to patients at the University of Manitoba ($p<.05$). Forty percent of all errors are systematic at the University of Manitoba in contrast to 22% at Southern Illinois. This observation is likely due to differences in the standardized patient training experience. The University of Manitoba had no standardized patient training experience prior to 1987. None of the patients trained at Manitoba had previous experience as a standardized patient. In contrast, Southern Illinois University has been using standardized patients for teaching and

evaluation purposes for over 10 years and had a well established pool of experienced patients and trainers. A university with no prior experience with standardized patients may initially experience more systematic errors in standardized patient presentation. Systematic differences in the accuracy with which critical case findings are presented may have implications for the comparison of student scores across universities.

TABLE 6.10 CATEGORIZATION OF ERRORS IN THE PRESENTATION OF ACCURACY OF ITEMS BY UNIVERSITY

	UNIVERSITY	
Statistic	**Southern Illinois**	**University of Manitoba**
Total Number of Items Evaluated	250	250
Number Items Where No Errors Present	208 (83%)	196 (78%)
Number Items Where One or More Errors Present	42 (17%)	54 (22%)
Total Number of Opportunities to Make an Error for Items with Error Present (items * # encounters where item could be rated)	332	480
Number Errors Made	46 (14%)	70 (15%)
Error Breakdown		
A. Systematic	10 (22%)	28 (40%)
B. Random	36 (78%)	42 (60%)

Categorization of Errors by Clinical Feature (Item) Type

Table 6.11 displays the breakdown of item errors by clinical feature type. There are striking differences in the percentage of errors among the 3 types of clinical features. Although 75% of history items were presented correctly on all occasions evaluated, only 38% of physical exam items and 43% of affect items were presented correctly on all occasions evaluated. Unfortunately the number of opportunities to sample performance in these

latter categories was limited. The generalizability of these findings is therefore uncertain. Nevertheless it suggests that accuracy of patient presentation in these areas may be far from adequate.

The breakdown of errors by type shows a 70:30 split of random: systematic errors for the presentation of the medical history. A slightly greater proportion of systematic errors is noted with affect items suggesting that inadequacies in training may have been a greater factor in these areas. Errors in the presentation of physical findings were exclusively random.

TABLE 6.11 CATEGORIZATION OF ERRORS IN THE PRESENTATION OF ACCURACY ITEMS BY CLINICAL FEATURE TYPE

Statistic	Clinical Feature Category		
	History	Physical Findings	Affect
Total Number of Items Evaluated	228	8	14
Number of Items Where No Errors Present	170 (75%)	3 (38%)	6 (43%)
Number of Items Where One or More Errors Present	58 (25%)	5 (62%)	8 (57%)
Total Number of Opportunities to Make an Error for Items with Error present (# items * # encounters where item could be rated)	1216	106	211
Number Errors Made	91 (8%)	9 (9%)	16 (8%)
Error Breakdown			
A. Systematic	31 (34%)	0 (0%)	7 (44%)
B. Random	63 (66%)	9 (100%)	9 (56%)

Since differences in the errors attributable to training were noted between the two universities, differences in error type by university and clinical feature type were assessed. The results are displayed in Table

6.12. It is clear that error rates for physical examination and affect items are high for both universities. The greater contribution of training problems to error rates noted for the University of Manitoba persists in 2 item categories (history and affect). The greater contribution of training problems to affect items appears to be entirely due to the problems experienced at the University of Manitoba. In contrast, random errors accounted for all errors made in the presentation of physical examination items in both universities.

TABLE 6.12 CATEGORIZATION OF ERRORS IN THE PRESENTATION OF ACCURACY ITEMS BY UNIVERSITY & CLINICAL FEATURE TYPE

	UNIVERSITY					
	Southern Illinois			University of Manitoba		
	History	Physical	Affect	History	Physical	Affect
Total Number of Items Evaluated	228	8	14	228	8	14
Number of Items Where No Error Present	198 (87%)	4 (50%)	6 (43%)	187 (82%)	3 (37%)	6 (43%)
Number of Items Where One or More Errors Present	30 (13%)	4 (50%)	8 (57%)	41 (18%)	5 (63%)	8 (57%)
Total Number of Opportunities to Make an Error for Items with Error Present (# items * # encounters where item could be rated)	578	50	95	638	56	116
Number of Errors Made	37 (6%)	3 (6%)	6 (6%)	54 (9%)	6 (11%)	10 (9%)
Error Breakdown						
A. Systematic	7 (18.9%)	0 (0)	2 (33%)	25 (46%)	0 (0%)	5 (50%)
B. Random	30 (81.1%)	3 (100%)	4 (66%)	29 (54%)	6 (100%)	5 (50%)

Summary

Both systematic and random errors contribute to variance in accuracy score across most cases. The proportion of systematic errors (errors which are attributable to training problems) varies among cases and between universities. As might be expected, the university with the least standardized patient experience had a greater proportion of errors attributable to training inadequacies. This was true for the training of history and affect items but there was no apparent benefit of greater university experience for the training of physical exam items. Substantial differences in error rates were noted across the 3 clinical feature categories. There were errors in the presentation of more than one half of the physical examination and affect items in contrast to one-quarter of the history items. Limited sampling opportunities make the generalizability of these findings uncertain.

Random errors make a substantial contribution to the proportion of errors in certain cases and for physical examination items. The factors which may contribute to random errors are numerous. They may include attributes of the patient, encounter, student performance and case. Clearly, if patient accuracy is to be improved, a better understanding of factors which contribute to random errors in performance needs to be gained as well as of those factors in the training process which are important in minimizing systematic and random errors in performance.

THE EVALUATION OF DIFFERENCES IN ACCURACY SCORE BETWEEN UNIVERSITIES AND PATIENTS

Since the two assumptions about standardized patients were met by only 7/49 standardized patient presenters, the proposition which must be evaluated is whether systematic differences existed between the accuracy of patient presentation between the two universities. Such a difference, if present, could act to confound a comparison of measures of clinical competence between universities if patient accuracy was associated with competency score.

Table 6.13 provides a breakdown of accuracy score by university and case.

In 12 of the 15 cases evaluated the average accuracy score for patients from Southern Illinois University was higher than the score for patients at the University of Manitoba. In 5 cases, the estimated difference was statistically significant. Over all cases, a difference of 2.47% in accuracy score was observed (S.I.U.=91.57% ; U of M=89.10%) between the two universities which was statistically significant (p=.03). Although most differences in accuracy score were small, there was a systematic trend for patients from Southern Illinois to be more accurate than those from the University of Manitoba. This may be a function of a difference in the experience of the two universities in standardized patient training. Southern Illinois had 10 years of experience and the University of Manitoba had none prior to 1987. Alternatively some of these differences might be explained by a difference in the patient used to present the problem rather than inherent differences in the trainer or training process. If this were the case, patient accuracy might be improved by patient selection rather than changes in the training process or trainer experience.

The second proposition which was therefore evaluated was whether two patients trained simultaneously by the same trainer would differ in the accuracy with which they presented the problem. The bulk of the data to answer this question is provided by the University of Manitoba where the two patients who presented 15/16 problems were evaluated. Four comparisons of patient differences are available from Southern Illinois University. Table 6.14 provides data on the differences between patients trained by the same trainer by case.

TABLE 6.13 DIFFERENCES IN ACCURACY SCORE BY UNIVERSITY AND CASE

CASE	N	% ACCURACY (S.D.)	UNIVERSITY				P
				S.I.U.		U of M	
			n	% Accuracy (s.d.)	n	% Accuracy (s.d.)	
1	35	85.3 (7.8)	20	89.4 (1.8)	15	79.7 (9.3)	.001[u] *
2	31	95.6 (5.4)	14	95.9 (4.0)	17	95.5 (6.4)	.839
3	29	92.6 (10.73)	12	92.5 (14.3)	17	92.7 (7.82)	.964[u]
4	31	71.5 (15.75)	12	74.6 (9.6)	17	69.5 (18.6)	.322[u]
5	25	92.5 (9.79)	7	100.0 (0)	18	89.5 (10.1)	.000[u] *
6	23	98.4 (3.14)	8	99.0 (2.7)	15	98.0 (3.4)	.482
7	32	78.8 (9.3)	16	74.2 (5.61)	16	83.3 (10.2)	.005[u] *
8	24	99.7 (1.2)	9	100.0 (0)	15	99.6 (1.6)	.451
9	27	92.0 (9.8)	12	93.4 (8.8)	15	90.9 (19.8)	.52
11	32	98.9 (3.1)	15	99.3 (2.87)	17	98.5 (3.3)	.518
12	28	89.7 (8.0)	11	97.2 (5.4)	17	84.9 (5.00)	.000 *
13	36	79.3 (15.7)	20	86.5 (16.2)	16	70.2 (9.5)	.000[u] *
14	29	88.6 (6.2)	11	88.8 (8.5)	18	88.4 (4.5)	.898[u]
15	42	97.4 (6.7)	22	97.7 (5.9)	20	97.0 (7.68)	.739
16	27	98.7 (2.5)	10	97.3 (3.4)	17	99.5 (1.42)	.084[u]
All	451	90.2 (12.2)	199	91.6 (11.3)	252	89.1 (12.8)	.03 *

Legend: (s.d.): standard deviation

P: the probability of observing such a difference or a bigger difference in means by chance alone under the null hypothesis of no difference. Estimated using an independent t-test

u: t-test calculated on the basis of unequal variances

*: significant difference present after Bonferroni's correction for multiple comparisons

TABLE 6.14 DIFFERENCES IN % ACCURACY SCORE FOR PATIENTS TRAINED TOGETHER IN THE SAME UNIVERSITY BY CASE

CASE	N	UNIVERSITY	% ACCURACY SCORE BY PATIENT				P
			Patient #1		Patient #2		
			n	% Accuracy(s.d.)	n	% Accuracy(s.d.)	
1	15	U of M	5	83.1 (4.9)	10	78.0 (10.64)	.33
2	17	U of M	10	99.4 (1.9)	7	89.8 (6.3)	.006[u]*
3	18	U of M	11	90.3 (7.9)	6	97.3 (4.4)	.06
	11	S.I.U.	4	82.4 (22.3)	7	98.4 (4.2)	.247[u]
4	19	U of M	3	70.6 (24.1)	16	69.3 (18.4)	.911
5	18	U of M	16	88.2 (10.0)	2	100.0 (0)	.000[u]*
6	8	S.I.U.	3	100.0 (0)	5	98.5 (3.4)	.374[u]
7	16	U of M	5	90.9 (7.5)	11	79.9 (9.6)	.04
8	15	U of M	5	98.7 (2.8)	10	100.0 (0)	.165
9	15	U of M	5	83.5 (10.6)	10	94.6 (9.2)	.056
11	17	U of M	8	98.1 (3.4)	9	100.0 (0)	.125
	15	S.I.U.	8	100.0 (0)	7	97.0 (5.2)	.12
12	17	U of M	5	86.9 (4.5)	12	84.1 (5.2)	.308
13	16	U of M	13	69.5 (9.7)	3	73.4 (9.5,	.535
14	18	U of M	5	89.8 (5.5)	13	87.9 (4.2)	.451
15	20	U of M	9	95.2(10.1)	11	98.5 (5.0)	.398[u]
	22	S.I.U.	11	97.0 (6.7)	11	98.5 (5.0)	.557
16	17	U of M	15	99.7 (1.0)	2	99.7 (3.2)	.06

Legend: (s.d.): standard deviation

P: the probability of observing such a difference in means by chance under the null hypothesis of no difference (estimated using an independent t-test)

u: t-test calculated on the basis of unequal variances

*: significant difference present after correction for multiple comparisons

Differences were in the range of 1.5% for Case #6 to 20.9% for Case #13, both values being generated by pairs from Southern Illinois. Three of the 19 possible comparisons were statistically significant. Since smaller sample sizes limited the power of this latter analysis, the contribution of patients as opposed to university training site was assessed by comparing the average magnitude of the differences between patients trained at different universities to those trained at the same university. The results are displayed in Table 6.20. The average absolute difference between the 38 possible pairs of patients trained for the same case at different universities was 5.5% in contrast to an average difference in accuracy score of 6.2% for patients trained by the same trainer at the same university. This difference is not significant.

The data from this analysis do not support the assumptions which have been made about the standardized patient. Systematic differences are present between patients trained with the same protocol by different trainers in different universities. Secondly, it cannot be assumed that two patients simultaneously trained by the same trainer will provide an equivalently accurate performance.

The methodological implications for these observed differences are at the moment uncertain. The relationship between patient accuracy and the resulting estimates of clinical competence score will need to be evaluated to determine the size of deviation from 100% accuracy score which are of methodological importance. This question will be addressed in Study 2.

EVALUATING POTENTIAL DETERMINANTS OF ACCURACY SCORE-SECONDARY ANALYSIS

If accuracy score is to be improved, the factors which contribute to a less than optimal score will need to be understood. The data gathered in Study 1 provides the opportunity to evaluate the effect of differences among patients, cases and between universities on accuracy score. The impact of three additional factors on accuracy score was also explored: the number of items that the standardized patient was required to present with each case, the conditions in which the response was

provided (spontaneously or in response to student inquiry) and when, during the course of the 4 week evaluation period, patient accuracy was sampled. It was hypothesized that the complexity of the case would increase with the number of items the patient would have to present and that an inverse relationship would exist between accuracy score and the number of items to be presented. In relation to the conditions of response, it was hypothesized that patients who were less confident about their presentation would be less apt to voluntarily provide data to the student and that a positive relationship would therefore exist between the percent of items provided spontaneously and accuracy score. Finally it was hypothesized that practice would improve the quality of standardized patient performance and that there would be a trend for accuracy score to increase across the 4 week evaluation period.

Univariate analysis was initially used to evaluate the contribution of these factors to accuracy score. Table 6.15 provides a breakdown of accuracy score by university, case, number of items, evaluation week and the percent of items provided spontaneously. Accuracy scores by patient are displayed in Table 6.7. The probabilities, calculated by regression analysis, are also provided.

TABLE 6.15 UNIVARIATE ANALYSIS OF POTENTIAL DETERMINANTS OF STANDARDIZED PATIENT ACCURACY SCORE

Determinant	Categories	N	% Accuracy (s.d.)	P	β (95% C.I.)
University	S.I.U	199	91.57 (11.26)	.03	
	U of M	252	89.10 (12.82)		
Case	# 1	35	85.16 (7.77)	.0001	
	# 2	31	95.63 (5.35)		
	# 3	29	92.63 (10.73)		
	# 4	31	71.47 (15.75)		
	# 5	25	92.47 (9.79)		
	# 6	23	98.39 (3.14)		
	# 7	32	78.76 (9.34)		
	# 8	24	99.74 (1.28)		
	# 9	27	92.01 (9.83)		
	#11	32	98.88 (3.08)		
	#12	28	89.74 (7.95)		
	#13	36	79.28 (15.74)		
	#14	29	88.58 (6.20)		
	#15	42	97.39 (6.71)		
	#16	27	98.69 (2.54)		
Patient	49 pts.		see Table 6.7	.0001	
Evaluation Week	Week 1	34	88.68 (12.56)	.48	
	Week 2	85	89.88 (12.58)		
	Week 3	71	90.26 (12.78)		
	Week 4	62	87.06 (13.42)		
Number of Items in case	< 10	94	94.53 (8.85)	.0001	.41 (.17,.57)
	10-15	66	78.73 (13.91)		
	16-20	208	88.32 (11.82)		
	21-25	24	99.74 (1.28)		
	> 25	59	98.79 (2.82)		
% Score Spontaneous	< 10	91	90.95 (12.77)	.0001	.10 (.01,.14)
	11-20	97	84.28 (14.42)		
	21-30	82	89.08 (11.81)		
	31-40	53	89.35 (11.83)		
	41-50	43	93.95 (7.41)		
	51-60	17	93.48 (7.78)		
	61-70	9	95.04 (8.74)		
	71-80	8	93.40 (9.84)		
	81-90	2	91.67 (11.79)		
	91-100	40	97.26 (6.85)		

In order to examine trends, the number of items and percent provided spontaneously are categorized and the respective accuracy scores are provided for each category. Both were treated as continuous variables in the estimation of their relationship to accuracy score with the estimated regression coefficients and standard errors for these analyses provided. The evaluation of the impact of evaluation week was conducted on Manitoba data only. These data were not available for Southern Illinois patient-student encounters.

Patient, case and university are all factors associated with accuracy score in the univariate analysis. Evaluation week appears to have no relationship to accuracy score. Average accuracy score was more or less equivalent across the 4 weeks of evaluation. There was a significant positive relationship between the number of items to be presented for a case and accuracy score. This relationship was significant ($p=.0001$) but was the reverse of that hypothesized. Since number of items was correlated with case, it is not possible to obtain an unbiased estimate of the relationship of case complexity to accuracy score in this data set. If an inverse relationship does exist, it may be due to the selection of more capable patients for difficult cases rather than number of items per se. Finally, the percent of items provided spontaneously by the patient was positively related to accuracy score ($p=.0001$). With each 1% increase in the percent of items provided spontaneously, the patient accuracy score is estimated to increase by approximately 1/10 of a percentage point (95% C.I.: .01-.14). This factor however accounted for less than 1% of score variance.

Multiple regression analysis was employed to estimate the independent contribution of these factors to accuracy score. The results are displayed in Table 6.16.

TABLE 6.16 THE EVALUATION OF FACTORS ASSOCIATED WITH PATIENT ACCURACY: THE RESULTS OF MULTIPLE REGRESSION ANALYSIS

Factor	D.F.	R^2	F	P
All Factors	50	38.16%	13.70	.0001
University	1	0.76%	8.24	.004
Case	14	23.8%	18.42	.0001
Patient (Univ * Case)	33	13.5%	4.45	.0001
Number Items	1	0.48%	5.18	.02
% Spontaneous	1	.07%	.85	.36

Notes:

1. The corrected total degrees of freedom was 450.
2. Partial correlation coefficients were calculated for each factor using the formula outlined by Kleinbaum and Kupper(1978).
3. Dummy variables were defined for the three class variables (case, university and patient).
4. Patient was defined as being nested within university and case.
5. All factors were initially treated as being fixed. Since the interest was in evaluating potential predictors of accuracy, only main effects were assessed.

All factors combined explained 38.6% of the variance in patient accuracy score. Case and patient nested within university and case were the two factors which explained the greatest proportion of variance. University and number of items were significantly associated with accuracy score but explained a small proportion of the variance (.76% and .48% respectively). The same rank-ordering of factors was produced when case and patient were treated as random factors in the analysis. When other factors were included in the model, the condition of response (percent of times data was provided spontaneously) was not associated with accuracy score.

THE PERCENT OF ITEMS PROVIDED SPONTANEOUSLY

The Percent of Items Provided Spontaneously by University, Patient and Case

One of the unique aspects of using standardized patients to present the clinical problem is that they are informed about those aspects of the problem which are of clinical importance in diagnosis and management. There is the potential therefore for the data provided by the standardized patient to be manipulated by deliberate efforts to make the problem 'easier' or 'harder' for the clinician examinee. In an effort to help a struggling examinee, the patient may spontaneously provide more clinically relevant information about his/her problem or alternately restrict access to information requested to an examinee whose demeanour the patient finds offensive.

It has been assumed that effective training will eliminate this potential source of bias in the generation of competency scores. Typically, in the training process, the conditions in which clinical data are to be provided are specified. (i.e. spontaneously or in response to specific inquiry). The conditions in which data are provided by the patient should be legitimately considered as part of the accuracy of patient presentation. However, in the current study, the protocol used to train patients at the two universities did not specify what data were to be provided spontaneously and which data were to be provided only in response to specific types of inquiry. Because this was the case, the difference in the average percent of items provided spontaneously by different patients and the two universities was examined. If it is assumed that the student groups seen by different patients and in the two universities were equivalent, then patients, if accurate, would be expected to provide the same amount of data spontaneously.

In order to evaluate whether differences existed in the percentage of items provided spontaneously by patients presenting the same case, descriptive statistics were generated for the conditions in which the clinical data were provided. Table 6.17 provides a breakdown of the percent of responses provided spontaneously by case and patient. It can be

noted that the percent of items provided spontaneously varies from 0% for Case #11 (a problem of a depressed woman, abused by her husband who presents with headache) to 98.3% for Case #15 (a patient presenting with anaphylaxis with a hospital nurse requesting direction for clinical management). These differences are appropriate for the types of cases and situations developed.

TABLE 6.17 THE AVERAGE PERCENT OF ITEMS PROVIDED SPONTANEOUSLY BY CASE, UNIVERSITY AND PATIENT

CASE	N	UNIVERSITY	PT	N	% SPONTANEOUS	S.D.	P	R^2
1	35	S.I.U.	#1	20	22.4	9.1	.39	.05
		U of M	#1	5	23.6	4.1		
		U of M	#2	10	28.1	14.6		
2	31	S.I.U.	#1	14	42.1	7.5	.0001	.55
		U of M	#1	10	38.8	7.2		
		U of M	#2	7	21.0	9.9		
3	29	S.I.U.	#1	4	21.4	26.0	.64	.06
		S.I.U.	#2	7	22.0	20.5		
		U of M	#1	12	31.0	17.4		
		U of M	#2	6	22.2	12.6		
4	31	S.I.U.	#1	12	13.0	7.3	.44	.06
		U of M	#1	3	19.8	7.7		
		U of M	#2	16	12.5	10.3		
5	25	S.I.U.	#1	7	11.7	11.2	.96	.00
		U of M	#1	16	10.3	14.3		
		U of M	#2	2	12.5	17.7		
6	23	S.I.U.	#1	3	25.7	2.6	.69	.04
		U of M	#1	5	30.2	6.9		
		U of M	#2	15	30.0	8.9		
7	32	S.I.U.	#1	16	26.1	8.7	.0001	.65
		U of M	#1	5	14.4	2.5		
		U of M	#2	11	4.3	7.5		
8	24	S.I.U.	#1	9	16.0	11.0	.22	.14
		U of M	#1	5	13.8	5.5		
		U of M	#2	10	21.8	8.5		

CASE	N	UNIVERSITY	PT	N	% SPONTANEOUS	S.D.	P	R^2
9	27	S.I.U.	#1	12	61.7	17.2	.25	.11
		U of M	#1	5	43.7	22.7		
		U of M	#2	10	54.8	21.8		
11	32	S.I.U.	#1	8	0.0	0.0	-	-
		S.I.U.	#2	7	0.0	0.0		
		U of M	#1	8	0.0	0.0		
		U of M	#2	9	0.0	0.0		
12	28	S.I.U.	#1	11	21.4	7.1	.25	.11
		U of M	#1	5	21.5	5.2		
		U of M	#2	12	16.9	7.4		
13	35	S.I.U.	#1	11	26.6	13.2	.18	.14
		S.I.U.	#2	8	17.0	21.7		
		U of M	#1	13	22.4	8.9		
		U of M	#2	3	37.0	3.4		
14	30	S.I.U.	#1	12	29.5	9.4	.003	.34
		U of M	#1	5	39.4	19.0		
		U of M	#2	13	48.9	13.2		
15	42	S.I.U.	#1	11	100.0	0.0	.63	.04
		S.I.U.	#2	11	97.0	6.7		
		U of M	#1	9	97.8	6.7		
		U of M	#2	11	98.2	6.0		
16	27	S.I.U.	#1	10	45.8	15.9	.02	.27
		U of M	#1	15	49.5	18.2		
		U of M	#2	2	11.4	3.2		

Legend: N: the number of taped patient-student encounters by case and by patient

PT: the patient presenting the problem

% Spontaneous: the average percent of items provided to the student spontaneously by the patient (both correct and incorrect responses)

S.D.: standard deviation

P: the probability of observing differences this big or bigger among the percent of items provided spontaneously by different patients by chance (estimated using one-way ANOVA)

R^2: the proportion of variance in the percent of items provided spontaneously attributable to differences among patients.

A one-way ANOVA was used to evaluate whether there were differences in the percentage of items provided spontaneously by the 3-4 patients presenting each case. In this analysis it must be assumed that the student groups seen by each of the patients were equivalent. Significant differences were noted among patients in three cases; case #2,7 and 14. In these cases, differences among patients accounted for 34-65% of the variance in the percent of items provided spontaneously.

The next question is whether these differences are a function of differences between universities or differences among different patients. Table 6.18 provides a breakdown of the percent of items provided spontaneously by university and case.

A t-test was used to evaluate differences in scores between universities for each case. Significant differences were noted between universities on the same 3 cases: case #2, #7 and #14 with Southern Illinois patients providing more data spontaneously than University of Manitoba patients on 2 of 3 cases. This would suggest that the differences noted previously may be a function of differences between trainers, student groups or patients evaluated.

TABLE 6.18 DIFFERENCES IN % OF ITEMS PROVIDED SPONTANEOUSLY BY UNIVERSITY AND CASE

CASE	N	% SPONTANEOUS (S.D.)	UNIVERSITY				P
			S.I.U.		U of M		
			n	% Spontan.(s.d.)	n	% Spontan.(s.d.)	
1	35	24.21 (10.54)	20	22.4 (9.1)	15	26.6 (12.1)	.252
2	31	36.27 (11.48)	14	42.1 (7.5)	17	31.5 (12.1)	.008 *
3	29	25.69 (18.17)	12	20.6 (20.7)	17	29.3 (15.8)	.214
4	31	13.39 (9.01)	12	13.0 (7.3)	19	13.7 (10.1)	.846
5	25	10.87 (13.12)	7	11.7 (11.2)	18	10.6 (14.1)	.854
6	23	29.44 (7.86)	8	28.5 (5.9)	15	30.0 (8.9)	.688
7	32	16.78 (12.48)	16	26.1 (8.7)	16	7.5 (7.9)	.000 *
8	24	17.95 (9.33)	9	16.0 (11.0)	15	19.1 (8.4)	.436
9	27	55.80 (20.32)	12	61.7 (17.2)	15	51.1 (21.9)	.18
11	32	0.0 (0)	15	0.0 (0)	17	0.0 (0)	-
12	28	19.50 (7.08)	11	21.4 (7.1)	17	18.3 (7.0)	.251
13	36	23.80 (14.14)	20	22.7 (17.0)	16	25.1 (9.9)	.596[U]
14	29	39.97 (15.44)	11	29.7 (9.8)	18	42.6 (15.1)	.003 *
15	42	98.25 (5.47)	22	98.5 (4.9)	20	98.0 (6.2)	.778
16	27	45.32 (19.13)	10	45.8 (15.9)	17	45.1 (21.37)	.924
All	451	32.08 (27.47)	199	33.6 (29.0)	252	30.9 (27.2)	.293

Table 6.19 provides a breakdown of the differences in the percent of items provided spontaneously between patients simultaneously trained by the same trainer in the same university. Differences between patients were significant in 4 of the 15 cases evaluated (Case #2,7,13 and 16). In Case #14, a difference of 10% was noted between patients, a difference which was slightly smaller than the difference in scores between universities for this case. Since the power of the comparison for differences between

patients is less than that for the differences between universities, a comparison of the number of significant differences does not provide meaningful information on the relative contribution of university vs. patients on differences in the percent of items provided spontaneously.

TABLE 6.19 DIFFERENCES IN % OF ITEMS PROVIDED SPONTANEOUSLY FOR PATIENTS TRAINED TOGETHER IN THE SAME UNIVERSITY BY CASE

CASE	N	UNIVERSITY	% PROVIDED SPONTANEOUSLY BY PATIENT				P
			Patient #1		Patient #2		
			n	% Spontan.(s.d.)	n	% Spontan.(s.d.)	
1	15	U of M	5	23.6 (4.16)	10	28.1 (14.6)	.385[U]
2	17	U of M	10	38.8 (7.2)	7	21.0 (9.9)	.000 *
3	18	U of M	12	31.0 (17.4)	6	22.2 (12.6)	.280
	11	S.I.U.	4	21.4 (26.0)	7	22.0 (20.5)	.966
4	19	U of M	3	19.8 (7.7)	16	12.5 (10.3)	.26
5	18	U of M	16	10.3 (14.3)	2	12.5 (17.7)	.843
6	8	S.I.U.	3	25.7 (2.6)	5	30.2 (6.9)	.329
7	16	U of M	5	14.4 (2.5)	11	4.3 (7.5)	.001[U]*
8	15	U of M	5	13.8 (5.5)	10	21.9 (8.5)	.08
9	15	U of M	5	43.7 (22.7)	10	54.8 (21.8)	.376
11	17	U of M	8	0.0 (0)	9	0.0 (0)	-
	15	S.I.U.	8	0.0 (0)	7	0.0 (0)	-
12	17	U of M	5	21.5 (5.2)	12	16.9 (7.4)	.231
13	16	U of M	13	22.4 (8.9)	3	36.9 (3.4)	.016 *
14	18	U of M	5	39.4 (19.0)	13	48.9 (13.2)	.242
15	20	U of M	9	98.0 (6.7)	11	98.2 (6.0)	.889
	22	S.I.U.	11	97.0 (6.7)	11	98.5 (5.0)	.557
16	17	U of M	15	49.5 (18.2)	2	11.4 (3.2)	.000[U]*

In order to gain an understanding of the relative contribution of these two factors, numerical differences in the scores obtained for all possible comparisons for patients trained in different universities were compared to differences in scores obtained for patients trained in the same university. The results are displayed in Table 6.20.

The average difference in scores for patients trained in the same university is 7.93% in contrast to an average difference of 6.66% for patients trained in different universities for the same case. This difference is small and is neither methodologically important nor statistically significant. What it suggests however is that differences among different patients presenting the case are present and that these differences are a function of the patient presenter rather than the trainer or evaluation site. It is possible that the observed differences between patients are a function of inequivalencies in the student groups evaluated. Since the Manitoba comparison was based on a random sample of students who were randomly assigned to a patient, this possibility seems a less likely explanation of the differences noted. The importance of these differences can only be determined by an evaluation of their impact on student score, a question which will be addressed in Study 2.

TABLE 6.20 COMPARISON OF THE SIZE OF THE DIFFERENCE IN PERCENT ACCURACY SCORE AND PERCENT ITEMS PROVIDED SPONTANEOUSLY BETWEEN PATIENTS TRAINED IN DIFFERENT UNIVERSITIES AND PATIENTS TRAINED IN THE SAME UNIVERSITY BY CASE AND OVERALL

CASE	PATIENT PAIRS	TRAINED IN DIFFERENT UNIVERSITY		TRAINED IN THE SAME UNIVERSITY	
		% Accuracy %	% Spontan.	% Accuracy	% Spontan.
1	1	6.2	1.2	5.1	4.5
	2	11.3	5.7		
2	1	3.6	3.3	9.6	17.8
	2	6.1	21.1		
3	1	8.0	.8	16.0	.6
	2	14.9	9.7	7.0	8.9
	3	8.1	.2		
	4	1.1	8.1		
4	1	4.0	6.9	1.4	7.4
	2	5.4	.5		
5	1	0	1.4	11.8	2.2
	2	11.8	.8		
6	1	2.0	4.5	1.5	.3
	2	.4	4.2		
7	1	16.7	11.6	11.1	10.1
	2	5.7	21.8		
8	1	1.3	2.2	1.3	8.1
	2	0	5.8		
9	1	9.9	18.1	11.1	11.1
	2	1.2	7.0		
11	1	1.9	0	3.0	0
	2	0	0	1.9	0
	3	1.2	0		
	4	3.0	0		
12	1	10.4	0.4	2.8	4.6
	2	13.2	4.5		
13	1	25.5	4.2	20.9	14.5
	2	21.6	10.4	4.0	9.7
	3	4.6	20.0		
	4	.6	5.5		

CASE	PATIENT PAIRS	TRAINED IN DIFFERENT UNIVERSITY		TRAINED IN THE SAME UNIVERSITY	
		% Accuracy	% Spontan.	% Accuracy	% Spontan.
14	1	.6	9.9	1.9	9.5
	2	1.3	19.4		
15	1	1.7	2.2	1.5	3.0
	2	1.5	1.8	3.2	.4
	3	3.2	.8		
	4	0	1.2		
16	1	2.4	3.8	2.0	38.2
	2	.4	34.4		
Overall					
average difference		5.5	6.7	6.2	8.0
standard deviation		6.3	8.0	5.7	9.0
number of comparisons		38	38	19	19
probability difference between different & same university due to chance (indep. t test)		.7	.6		

Legend: Different University Pair 1 = Patient 1 in Manitoba with patient 1 in S.I.U.
Pair 2 - Patient 2 in Manitoba with patient 1 in S.I.U.

Same University Pair 1 = Patient 1 in Manitoba with patient 2 in Manitoba

Footnotes:

1. Differences in scores for patients trained in different universities were generated by subtracting the score for the U of M patient from the score of the S.I.U. patient who presented the same case. Three patients presented most cases(one from S.I.U. and two from U of M) allowing two possible comparisons of inter-university differences. Four comparisons were possible when 4 patients presented the case (i.e. 2 from S.I.U. and 2 from U of M).

2. Differences in scores for patients trained in the same university were generated by subtracting the score of the first U of M or S.I.U. patient from the score of the second U of M or S.I.U. patient(eg. U of M Pt#1-U of M Pt#2). Most intra-university comparisons were generated by differences in the two patients trained at U of M (i.e. 15/19).

3. Overall average differences were calculated ignoring the signs.

Factors Associated with the Percent of Items Provided Spontaneously

In order to gain an understanding of factors which may be associated with the percent of items which are provided spontaneously, the effect of university, case, patient, number of items and evaluation week on the percent of items provided spontaneously was estimated. As was indicated earlier, legitimate differences would be expected among the percent of items provided spontaneously in different cases. With equivalent groups of students, no differences would be expected among different patients presenting the same case or between universities for the same case. It is hypothesized that patients would be less apt to provide clinical data spontaneously as the number of items or case complexity increased. Similarly it is hypothesized that the effect of practice over the 4 week evaluation period would act to increase the number of items provided spontaneously as patients became more confident in their presentation.

The percent of items provided spontaneously for each factor being evaluated is provided in Table 6.21 along with the results of univariate regression analysis. Patient scores are displayed in Table 6.17. Number of items has been treated as a continuous variable in the analysis but has been broken down categorically to illustrate trends in mean score. As was also the case for accuracy score, patients and cases in the univariate analysis accounted for the largest proportion of variance in score. Neither university nor evaluation week were significantly associated with the percent of items provided spontaneously.

An inverse relationship between the number of items to be presented with a case and score was found ($p=.0001$). It is estimated that the percent of items provided spontaneously decreases by approximately 1 1/2 % with each additional item to be presented with a case (95% C.I. -1.17, -1.89). The correlation between these two factors was -.36. This observation is compatible with the hypothesis that case complexity acts to reduce the percent of items are provided spontaneously. However, since number of items is correlated with case, an unbiased estimate of this relationship is not possible.

TABLE 6.21 UNIVARIATE ANALYSIS OF POTENTIAL DETERMINANTS OF THE PERCENT OF ITEMS PROVIDED SPONTANEOUSLY BY STANDARDIZED PATIENTS

Determinant	Categories	N	% Spontan. (s.d.)	P	β(95% C.I.)
University	S.I.U.	199	33.6 (29.0)	.29	
	U of M	252	30.9 (27.1)		
Case	# 1	35	24.2 (10.5)	.0001	
	# 2	31	36.3 (11.5)		
	# 3	29	25.7 (18.2)		
	# 4	31	13.4 (9.01)		
	# 5	25	10.9 (13.1)		
	# 6	23	29.4 (7.9)		
	# 7	32	16.8 (12.5)		
	# 8	24	18.0 (9.3)		
	# 9	27	55.8 (20.3)		
	#11	32	0.0 (0)		
	#12	28	19.5 (7.1)		
	#13	36	23.8 (14.1)		
	#14	29	40.0 (15.4)		
	#15	42	98.3 (5.5)		
	#16	27	45.3 (19.1)		
Patient	49 pts.	see Table 6.17		.0001	
Evaluation Week	Week 1	34	30.3 (19.5)	.44	
	Week 2	85	34.4 (28.3)		
	Week 3	71	30.0 (24.5)		
	Week 4	62	27.1 (27.3)		
Number of Items	< 10	94	62.8 (38.5)	.0001	-1.53
	10-15	66	19.1 (11.2)		(-1.17,-1.89)
	16-20	208	27.1 (15.1)		
	21-25	24	18.0 (9.3)		
	> 25	59	20.7 (26.1)		

The independent contribution of each of these factors was assessed using multiple regression analysis. The regression model was defined in the same manner as that used to evaluate factors associated with patient accuracy. All factors were initially treated as being fixed and analysis was restricted to the evaluation of main effects. The results are displayed in Table 6.22.

TABLE 6.22 THE EVALUATION OF FACTORS ASSOCIATED WITH THE PERCENT OF TIMES PATIENTS PROVIDED DATA SPONTANEOUSLY: THE RESULTS OF MULTIPLE REGRESSION ANALYSIS

Factor	D.F.	R^2	F	P
All Factors	49	50.3%	44.67	.0001
University	1	00%	.58	.45
Case	14	46.4%	85.90	.0001
Patient	33	3.9%	3.03	.0001
Number Items	1	00%	.62	.43

Notes:

1. The corrected total degrees of freedom was 450.
2. Partial correlation coefficients were calculated for each factor using the formula outlined by Kleinbaum and Kupper (1978).

The multiple regression analysis provides the same rank ordering of factors. Number of items was not an important predictor of the percent of items provided spontaneously when case and patient are taken into account. This is likely because item number was acting as a surrogate for case in the univariate analysis. The university in which the standardized patient was trained is not associated with the outcome variable in either the univariate or multivariate analysis.

DISCUSSION & CONCLUSIONS

The primary objective of Study 1 was to evaluate the two main assumptions which have been made about the content of standardized patient presentation. On the basis of the evidence provided in this study we can draw the following conclusions about these assumptions.

Assumption #1: Variability in patient presentation is eliminated when standardized patients are used to estimate provider competence.

Proposition 1

The first proposition assumes that the standardized patient will provide the same presentation from one subject to the next. When scores are corrected for differences in the actions taken by different students, seven patients met this assumption. Variability in the content of patient presentation was minimal in an additional 6 patients where the same error was made in the content of presentation in all opportunities evaluated. Variability in the content of presentation existed in the remaining 36 patients where either random error or a combination of random and systematic error contributed to variability in presentation. From the data collected, it can be concluded that variability in the important content of the patient's presentation can be eliminated as a potential source of measurement error but was not eliminated in the majority of standardized patients evaluated.

Proposition 2

The second proposition assumes that patients trained together by the same trainer will present the same clinical problem. Nineteen pairs of patients trained by the same trainer were evaluated (4 from SIU and 15 from U of M). Differences in average accuracy score were present in all patient pairs evaluated. These differences were attributable to both differences in the items in which errors were made and the frequency with which content was erroneously presented. The average difference in accuracy score for patients trained together by the same trainer was 6.16%. The difference was less than 5% in 11/19 pairs and greater than 20% in one pair.

Differences were also present in the percent of items provided spontaneously between patients trained by the same trainer. The average difference was 7.93% with a range of 0 to 38.18%. Part of these differences may be accounted for by inequivalencies in the students seen by different patients.

From the data collected, it cannot be assumed that patients trained together by the same trainer will present the same clinical problem or provide the same data about the problem under equivalent student performance conditions.

Proposition 3

The third proposition assumes that patients trained for the same case using a common protocol by different trainers in different settings will present the same clinical problem. This assumption was evaluated in 15 cases and in 38 possible comparisons between patients trained for the same case in different universities.

Over all cases, there was a systematic difference in the accuracy scores achieved by patients trained in different university settings with Southern Illinois patients being more accurate (Mean accuracy S.I.U.=-91.57%; U of M =89.10%). However, there were no differences in average accuracy in 4/38 possible comparisons, differences of less than 1% in an additional 8 comparisons and of less than 5% in 22/38 comparisons. The

average difference in accuracy score for patients trained in different universities was 5.54%. Differences in accuracy score are attributable to differences in the number and type of items in which errors were made by the respective patients as well as their frequency.

Differences in the percentage of items provided spontaneously were also noted between patients trained in different settings, the average difference being 6.66% with the range from 0 to 34.43%.

In conclusion, it is possible to use a common protocol and different trainers and settings to train two standardized patients to present the same clinical problem. However, in the majority of instances evaluated, this assumption was not met either for the content of the problem presented or in the conditions in which data were provided. The magnitude of the differences in many instances was small and may not be of importance methodologically in the comparison of student performance between evaluation settings.

Assumption #2: Standardized patients provide an accurate reproduction of the important clinical features of a real patient problem

Proposition 1

It was assumed that the standardized patients would not make errors (either random or systematic) in their presentation of the clinically important content of the real patient case. This assumption was met by seven of the 49 patients evaluated. In the remaining 42 patients this assumption was not met. The average accuracy of patient presentation varied among patients and across cases. In addition systematic differences were present in the average accuracy of patient presentation between universities and among different types of clinical items. Although sampled on a limited basis, standardized patients made errors in their presentation of more than half of the physical examination and affect items in contrast to one-quarter of the history items. The frequency with which they made errors on these type of items was also greater (30%-42% of the time in contrast to 12% of the time for history items).

In conclusion, standardized patients can be trained to provide an accurate presentation of the content which medical faculty feel is of clinical importance in a real patient case. This assumption, however, was not met by the majority of standardized patients evaluated. The systematic differences in patient accuracy between universities suggest that a university with more experience with the technique may have better patient accuracy (either because of more experienced patients and/or more effective selection and training). Differences in the proportion and frequency of errors made with different types of items raise some concern about the use of standardized patient to present clinically important findings in the area of patient affect and physical examination.

Secondary Analysis: Factors Associated with Accuracy Score and the Percent of Items Provided Spontaneously

The relationship between patient accuracy and five factors (university, case, patient, number of items presented with a case and the percent of items provided spontaneously by the patient) was evaluated. Four of these factors were significantly associated with accuracy score (university, case, patient and number of items). Of these four factors, case and patient explained the greatest proportion of variance. In order to construct useful guidelines for future patient selection and training, the attributes of the case and patient which contributed to variance in accuracy score need to be identified. The number of clinically essential items the patient was required to present with a case was the only case attribute evaluated in Study 1. Although this attribute was significantly associated with accuracy score, the relationship was in the opposite direction to that hypothesized. It is unclear whether the relationship between number of items as an index of case complexity and accuracy score was biased by the selective use of 'better patients' for the more complex cases. This relationship will be re-evaluated in Study 2. Additional attributes of the case which may be important predictors of accuracy were suggested in the analysis of item errors. The proportion of physical examination and affect items included in the essential clinical features of the case may be an important predictor of accuracy and will be evaluated in Study 2.

No data on patient attributes were collected in Study 1. Factors which may be important in patient selection and training were identified in Chapter 5. Those factors which are amenable to economical measurement will be evaluated in Study 2. They include standardized patient age, acting/simulation experience, familiarity with the health problem, confidence in their presentation ability as well as the characteristics of their training.

Two factors were associated with the extent to which patients provided data spontaneously to the clinician during the patient encounter: the case being presented and the actual patient presenter. Case differences were expected and appropriate for the real patient cases trained. Differences among patient presenters presenting the same case could have been due to inequivalencies in the student groups seen. However, they could also be the result of true differences in the conditions under which relevant clinical data was provided by different patients. This latter possibility requires further study since it could bias the estimation of case-specific competency scores. This question will be addressed in Study 2. Despite differences in standardized patient trainers and presentation site, there were no systematic differences in the percent of data provided spontaneously by patients in the two universities.

Summary

In most instances it cannot be assumed that the use of standardized patients will provide an entirely accurate reproduction of a real patient case or eliminate variation, attributable to patient presentation, from estimates of clinical competence. As such, standardized patients <u>may</u> contribute to bias in the estimation of clinical competence and case-confounded sources of random error. This would only be true if the accuracy of standardized patient presentation was associated with competency score estimates. If it is associated with competency score, the strength and shape of the relationship will determine how large the deviation from the theoretical optimum could be before it would be of methodological concern in research and evaluation.

From the literature, it could be hypothesized that standardized patient accuracy would have the strongest association with estimates of competence in data collection. This component of clinical competence is usually scored on the basis of the clinician's ability to identify and/or collect important data on history and physical examination. It was this component which was influenced by patient format in the studies of Norman et.al. (1982) and Nowotony & Grove (1982), a finding attributed to errors in patient presentation.

The impact of patient accuracy on estimates of competence in diagnosis and management would likely be less direct. In the Norman et.al. (1982) study, no association was found between patient format and diagnosis and management (where errors in presentation content were thought to be associated with format type). In Barrows et.al. (1978) study of the clinical reasoning process of the physician, it was noted that physicians on average collect only 60% of the clinically important data on history and physical examination prior to formulating a diagnosis and management plan. The 60% of data gathered varied from physician to physician. This observation would suggest that the formulation of an acceptable diagnosis and management plan may not depend on the standardized patient's ability to correctly present <u>all</u> clinically important elements of the real patient case. Rather, it could be hypothesized that some minimum threshold level of accuracy is required in the presentation of clinically important items. If this were the case, a curvilinear relationship would be expected between patient accuracy and competence estimates in diagnosis and management. A relationship would exist only below a certain accuracy threshold. Bergman & Beck's (1986) study of pediatric residents provides some data which would refute this hypothesized relationship. They noted that the clinical appearance of the patient influenced the management plan selected and the clinician's confidence in the most probable diagnosis. In their study, more aggressive steps in management were taken with an infant who appeared 'sick' than an infant with the same history who did not. The findings of this study suggest that errors which may be made in the presentation of certain types of clinical items rather than some threshold percentage of presentation accuracy may have an effect on estimates of

competence in diagnosis and management. Clearly, both types of hypothetical relationships would need to be evaluated in order to understand the relationship of patient accuracy to estimates of clinical competence in these areas.

The data from the literature do suggest that a relationship would exist between the accuracy of patient presentation and estimates of clinical competence. Furthermore, they suggest that the strength and shape of the relationship may differ depending on the components of competence measured and the contribution those components make to overall case and competency score. The nature of this relationship will be evaluated in Study 2 (Chapter 8).

If patient accuracy is associated with estimates of clinical competence, then the continued use of this method will depend on the ability to identify and manipulate factors which will improve the accuracy of presentation. The case and patient were identified in multiple regression analysis as the two factors accounting for the greatest proportion of variance in patient accuracy. The specific attributes of the case and patient which contribute to presentation accuracy need to be identified. These attributes can be considered in three operational groups.

The first group of attributes are those which could be used to select patients and cases who/which are more likely to be accurate. This group includes such patient attributes as age, gender, acting and simulation experience, health-problem related experience and case attributes such as number and type of clinical items. Identification of attributes in this group would provide the most efficient means of improving patient accuracy since no resources are spent in training patients or cases which have a low probability of providing an accurate presentation.

The second group of attributes are those which could be used to improve standardized patient training or exclude patients during or at the completion of training who are less likely to be accurate. They include training characteristics such as length and the trainer's and patient's

perception of the patient's ability to accurately present the important clinical elements of the problem. The identification and manipulation of attributes in this group have greater cost implications for the use of the method since training resources must be spent before the likelihood of presentation accuracy can be adequately predicted.

The third group of attributes provides the least efficient means of controlling for the accuracy of patient presentation. This group of attributes are those which would be operational during the actual course of standardized patient use in evaluation or research. They include attributes of the clinician-patient encounter and the standardized patient's perception of the quality of their performance. The identification of attributes associated with accuracy of presentation in this group would practically mean that some cases/encounters would have to be eliminated from the estimates of competence after the evaluation had been completed. If this were the case, a safety factor of a certain additional number of cases would have to be built into the evaluation procedure to supplement those which are discarded in order to permit a precise estimate of competence to be obtained.

In order to identify methods which could be used to improve patient accuracy, attributes of the case and patient in each of these three groups will be evaluated in Study 2.

ABSTRACT

CHAPTER 7

STUDY 2: PREDICTORS OF THE ACCURACY OF STANDARDIZED PATIENT PRESENTATION AND THE IMPACT OF STANDARDIZED PATIENT ACCURACY ON COMPETENCY SCORE

In Study 1, the accuracy of the standardized patient's presentation of the clinical situation was evaluated. Of the 49 patients evaluated, 42 made errors in the presentation of important clinical features of the problem. In 3 of the 15 cases, patients presenting the same case provided significantly different amounts of data spontaneously.

In Study 2, three groups of factors which may predict the accuracy of patient presentation were evaluated: attributes of the patient and case, the training procedure and the measurement process. The impact of the accuracy of case presentation and the amount of clinical data provided by the patient on competence score were also evaluated.

Predictors of patient accuracy were evaluated using a random sample of 383 videotaped patient-student encounters drawn from the 1988 evaluation of 98 fourth year medical students at the University of Manitoba. The sample included 16 of the 18 cases and 40 of the 44 standardized patients used in the evaluation. The relationship between patient accuracy and the percentage of data provided spontaneously by the patient was evaluated using the 636 encounters sampled at the University of Manitoba in 1987 and 1988.

Patient attributes which were associated with presentation accuracy included the patient's reported understanding of the problem being presented, previous simulation and acting experience, and experience with the health problem being presented. The type of clinical feature to be presented was the only attribute of the case which was associated with presentation accuracy. Physical findings and patient affect were presented less accurately than patient history. Patients who had 2 training sessions and 3 hours of training had the highest accuracy scores. Physician assistance at the training session improved the accuracy of presentation for physical findings and patient affect but had no impact on the accuracy

of the history. The accuracy of the presentation of physical findings and affect was adversely influenced by the number of encounters the patient had presented in a day. Weeks since training was negatively associated with the accuracy of presentation of patient affect.

Patient accuracy was inversely related to student scores in data collection, interpersonal skills and management but not to scores for overall competence or diagnosis. The strength and direction of the relationship varied among different cases and competence scores. It was concluded that some cases may be more sensitive than others to the content of the patient presentation and that an alternate method for measuring patient accuracy would be recommended.

Differences in the percent of data provided spontaneously by different patients presenting the same case were observed in both 1987 and 1988. This may influence competence score. However, there was no evidence to suggest that conditions which might have led the patient to provide more data during the encounter influenced student rating.

CHAPTER 7

STUDY 2: PREDICTORS OF THE ACCURACY OF STANDARDIZED PATIENT PRESENTATION AND THE IMPACT OF PATIENT ACCURACY ON COMPETENCY SCORE

THE RESEARCH PROBLEM

In study 1, two assumptions about the content of standardized patient presentation were evaluated: the assumption that variability in patient presentation is eliminated by standardization and the assumption that standardized patients can accurately reproduce the important clinical features of a real patient case. The performance of 14% of the 49 patients evaluated met these two assumptions. Since these two assumptions were not met by the majority of patients evaluated, the impact of patient inaccuracy on the object of measurement, clinical competence, needs to be explored.

The Relationship of Standardized Patient Accuracy and Competency Score

As reviewed in Chapter 4 and 5, the content of standardized patient presentation represents a potential source of measurement error which is confounded with case. It is confounded with case by virtue of the fact that different standardized patients are used to present each clinical situation used in the measurement procedure. Variability in the accuracy of standardized patient presentation within and across cases was noted in Study 1. If accuracy of presentation is associated with clinical competence score, then the observed phenomenon of a large case effect in competency score may be partially attributable to this potential source of measurement error (i.e. variation in the presentation of the test stimulus-the standardized patient).

There are a number of reasons why this potential source of measurement error should be pursued. These were reviewed in Chapter 4 and will be reiterated for reader convenience.

1) If variation in standardized patient presentation is contributing to variation in competency scores across cases, reduction or

elimination of this source of variance would reduce the number of cases required to obtain a reasonably precise estimate of clinician competence for the target domain of clinical problems to which inferences are to be drawn. At present, the number of cases which would be required to derive precise estimates of competence is considered, by some, to be too large to make this method a viable alternative for large scale evaluation/ research (Norcini,1988).

2) In some instances, failure to meet an established level of performance with a case results in planned remediation for the clinical content area tested. If accuracy of presentation is associated with clinical competence score, systematic and/or random errors in the classification of acceptable and unacceptable performances would be possible. The costs of remediation for both the examinee and evaluating body are usually substantial. For this reason, even a small proportion of errors in the classification of performance is important.

3) In instances where more than one setting is being used to measure clinical competence, different patients are used to present the same clinical problem. In this situation, standardized patients are confounded with evaluation setting as well as with case. If some attribute of the evaluation setting itself is the object of measurement (eg. American vs. Canadian students) and patient accuracy is associated with competence score, then systematic differences in patient accuracy may bias the comparison. Alternatively, random error in patient presentation which is non-differentially distributed may obscure the detection of true differences which may be present. Finally, if multiple settings are being used for large scale evaluation rather than for research purposes, candidates who undertake the evaluation may have a different probability of success depending on the accuracy with which clinical problem is presented in their evaluation setting.

4) In Chapter 2, the literature related to clinical competence and its theoretical relationship to practice performance, patient outcome and the clinical situation were reviewed. Numerous gaps in our understanding of these relationships were identified. Minimization of 'noise' in the measurement process is necessary if we are to better understand these theoretically complex relationships.

In Study 2, the relationship of patient accuracy and competency score will be evaluated. In the discussion of the results in the previous chapter, it was noted that the content of patient presentation had been identified as a factor influencing estimates of competence in data collection in two studies and diagnosis and management actions in one. Since the components contributing to overall competency score vary across cases, the relationship of patient accuracy to component scores, case scores and overall scores will be evaluated.

In the last chapter, two plausible relationships between patient accuracy and competency score were identified. The first was curvilinear, an association may be found below a certain minimum accuracy threshold. The second type of relationship which will be considered is a linear relationship between overall accuracy of presentation, as well as the accuracy of presentation of certain clinical features (eg. affect items), with clinical competence scores.

The Relationship of the Conditions of Response to Competency Score

In Study 1, differences in the percentage of items provided spontaneously were noted among patients presenting the same case. As reviewed in the previous chapter, there is the unique potential for standardized patients to manipulate the amount of clinically relevant data provided to the student during the patient encounter which would not occur if the real patient had presented. Therefore it is possible that estimates of student performance will be biased. Bias created by this source is difficult to identify since the standardized patient is customarily used to evaluate

the doctor-patient relationship and patient communication and to record actions taken on history and physical examination. Conditions which may lead the patient to provide more or less data spontaneously (eg. such as the struggling student) may also act to bias the patient's evaluation of the student in a similar direction. If the conditions in which the patient provides clinical data and evaluates actions differ from one student to another, no relationship between the percent of items provided spontaneously and competency score may be found. In Study 2, a theoretical model of the conditions of response and competency rating will therefore be evaluated. This model, in the absence of a second independent evaluator, may help detect this type of bias.

Predictors of Standardized Patient Accuracy

The identification of factors which are predictive of the accuracy of patient presentation will be important if a relationship is found between clinical competence scores and the accuracy of problem presentation. As was noted in Chapter 4, there is no empirical basis for identifying patients who have a higher probability of providing a standardized and accurate presentation of a real patient's problem. Since accuracy surveillance during the measurement procedure is costly and time-consuming, the continued viability of standardized patient use for research and evaluation purposes will rest on the ability to identify patients who have a higher probability of accurately presenting the clinical problem.

From an operational perspective, these factors can be considered in three groups: those which could be applied in patient and case selection, those which could be applied during the training process and those which could be economically applied during the measurement procedure. As was pointed out in the previous chapter, the ability to predict patient accuracy using the first group of factors is the most desirable since it avoids spending resources to train patients who have a low probability of being able to provide an accurate presentation. From the cost perspective, factors in the second group are the next most desirable and those in the third, the least desirable method of excluding patients and cases with a low

probability of presentation accuracy.

Two factors, the case being presented and the patient presenter, were associated with accuracy score in Study 1. Important attributes of the case, the standardized patient and the training process have been identified in all three groups by authors experienced in the standardized patient method (see Chapter 4-Figure 4.1).

Two attributes of the case selected will be evaluated: the number of critical features and the type of features which must be presented by the standardized patient. The three major types of critical features which will be evaluated are history, physical findings and affect. Critics of the standardized patient procedure contend that the standardized patients are unable to accurately present the physical findings and affect of the real patient case. It is also hypothesized that standardized patients may have difficulty recalling a large number of critical features. If this were true, the accuracy of patient presentation would be inversely related to the number of critical features to be presented.

Seven attributes of the standardized patient will be evaluated: age, gender, previous acting and simulation experience, familiarity with the health problem presented, confidence in their ability after training and their ratings of the quality of their presentation during the evaluation. The first five factors have been identified by Barrows (1988) and Stillman (1986) as important determinants of the quality of standardized patient performance. The patients' perception of their ability is included because it is easy to measure and, if associated with accuracy, would provide a feasible method of excluding patients who would be inaccurate in case presentation. The rating of performance quality during the evaluation is included because it may provide a less costly alternative to videotape surveillance for identifying cases which were inaccurately presented.

Four training attributes and one attribute of the evaluation procedure will be evaluated. Guidelines related to the length of training, number of sessions and physician assistance have been suggested by Barrows (1988) and Stillman (1986) (see Figure 4.3). The relationship of these training

characteristics to patient accuracy will therefore be evaluated. Barrows (1988) comments that patients who may have difficulty presenting the case can usually be identified by the trainer during the training process. Since this provides a cheap way of excluding patients who may be less accurate, the trainers perceptions of the patient's ability and accuracy score will be evaluated. The effects of fatigue on patient accuracy have been suspected but never evaluated. It is hypothesized that patients may become less accurate after consecutively presenting the case on numerous occasions. The number of times the patient can accurately present the case during a test day is an important issue to resolve if large scale testing of subjects is required. The effect of the number of consecutive sessions on patient accuracy will therefore be evaluated.

It has been hypothesized that the accuracy of the standardized patient's presentation may be compromised by his/her reaction to clinicians who are socially or clinically inept in the patient encounter. The relationship of the patient's ratings of the clinician's interpersonal and data collection skills to accuracy of presentation will therefore be evaluated.

RESEARCH QUESTIONS

1. Are any of the following factors associated with the accuracy of standardized patient presentation?

   Group 1: Factors Which Could Be Applied in Patient and Case Selection

   a) Case Attributes
      - *Case Complexity
      - *Type of Clinical Features Included (history, physical findings, affect)

   b) Patient Attributes
      - *Age
      - *Gender
      - *Previous Acting Experience
      - *Previous Simulation Experience
      - *Previous Experience with the Health Problem

Group 2: Factors Which Could Be Applied During/at Completion of Training

a) Patient Attributes *Patient's Confidence in Their Ability to Accurately Present the Problem Post-Training

b) Training Attributes *Trainer's Confidence in Patient's Ability to Accurately Present the Problem Post-Training
*Training Length
*Physician Assistance with Training

Group 3: Factors Which Could be Applied During/At the Completion of the Measurement Procedure

a) Procedural Attributes *Number of Sessions
*Time Since Training

b) Encounter Attributes *Patient Confidence in the Quality of Performance
*Student Interpersonal Skills
*Student Data Collection Skills

2. Is there an association between the accuracy of patient presentation and competency score?
   a) For Component Scores (eg. data collection)
   b) For Overall Competency Score

3a) Are there differences in the percent of items provided spontaneously by different patients presenting the same case to equivalent groups of students?

b) Are there differences in the variance of competency scores between patients presenting the same case to equivalent groups of students?

DESIGN

A prospective cohort design was used to evaluate the relationship between potential predictors of patient accuracy and standardized patient accuracy score. The study cohort consisted of the 44 standardized patients used in the 1988 comprehensive clinical evaluation of 98 final year medical students at the University of Manitoba. Data related to factors in Groups 1 and 2 were prospectively gathered; those in Group 3 were collected at the time of the evaluation procedure. A random sample of patient-student encounters was drawn to evaluate standardized patient accuracy. An equal sample of encounters was drawn for each of the 18 cases used in the 1988 evaluation. To control for the effect of student performance on patient accuracy, all encounters for a random sample of the students participating in the 1988 evaluation were sampled across cases. Although a balanced experimental design would have provided a superior means of evaluating hypothesized predictors, it was not feasible to conduct during the 1988 evaluation. The major limitations of the design are the inefficiencies in predictor estimation produced by imbalances across categories of potential determinants, restriction in the range of determinants such as age and number of case items, and a limited number of patients presenting each case.

A cross-sectional sample survey design will be used to evaluate the relationship between standardized patient accuracy, the conditions of response and competency score. The study sample was drawn from the patient-student encounters generated at the University of Manitoba during the 1987 and 1988 evaluations. In this setting it was feasible to randomly sample students and record their encounters with all cases used in the evaluation in 1987 and 1988 thus controlling for the effects of student performance on patient accuracy and assuring equivalency of student groups across cases and patients. The study sample included the 15 cases used in the 1987 evaluation and 18 cases in the 1988 evaluation. Four of the cases used in 1987 were also used in 1988. Both 1987 and 1988 data were used in order to increase the sample size available for estimating the relationship between patient accuracy and competence score.

METHOD

SAMPLING PROCEDURE AND RESULTS

The study population and sample size estimates are described in Chapter 5. The procedure used to sample patient-student encounters for the 1987 evaluation at the University of Manitoba was described in Study 1 and Chapter 5. Twenty students were randomly sampled and their encounters with each of the 15 cases used in the 1987 evaluation were videotaped. A total of 280 student-patient encounters were videotaped, 18-22/case of which 253 were found to be technically adequate (15-20/case).

The procedure used in 1987 was used again to sample student-patient encounters in the 1988 evaluation. Twenty-eight students were randomly sampled from the 98 participating in the 1988 evaluation and their encounters with all but one of the 18 cases used in the evaluation were taped. The exception was case 5: it involved an examination of the male reproductive system and the standardized patients were unwilling to have this aspect of the encounter videotaped. Case 5 was therefore dropped from the study sample. In Case 7, the standardized patients did not participate after the first week of the evaluation. This case was dropped from the evaluation procedure. A total of 17 standardized patient cases were used in the evaluation procedure with student-patient encounters sampled in 16 of the 17 cases. The number of patients and accuracy scores calculated for each case are displayed in Table 7.1. One accuracy score was calculated for each of the two pairs of patients used in Case 19. One accuracy score was calculated for the each of the 5 possible pairs of 5 patients used in Case 11.

Values for patient-related predictors of accuracy in Group 1 and 2 were used for 34 of the 40 patients. In Case 19, values of the predictor variables were used for the one member of each pair who was responsible for presenting most of the accuracy items evaluated. This resulted in values for 38 patients.

TABLE 7.1 THE NUMBER OF STANDARDIZED PATIENTS AND AVERAGE SCORES FOR PRESENTATION ACCURACY BY CASE FOR THE 1988 EVALUATION

CASE	NUMBER PATIENTS	NUMBER ENCOUNTERS EVALUATED	NUMBER AVERAGE ACCURACY SCORES (1 per patient or per patient pair)
1 COPD & Pneumonia	2	25	2
2 Pre-Op Evaluation	2	21	2
3 Back Pain	2	22	2
4 Paralysis	2	20	2
5 Urethritis$_1$	2	-	-
6 Infant Fall	2	30	-
7 Asthma$_2$	2	-	-
8 Endometriosis	2	23	2
10 Jaundice	2	22	2
11 Dysphagia$_3$	4 (5)	27	5
12 Dizziness	2	28	2
13 Panic Attacks	2	24	2
14 Changed Bowel Habits	2	22	2
15 Short Stature	6	28	6
16 Hemiplegic Migraine	2	26	2
18 Hypertension	2	21	2
19 Alzheimers$_4$ Disease	4	19	2

CASE	NUMBER PATIENTS	NUMBER ENCOUNTERS EVALUATED	NUMBER AVERAGE ACCURACY SCORES (1 per patient or per patient pair)
20 Renal Artery Stenosis	2	22	2
Total	44	383	39

Notes:

1. No encounters were videotaped
2. Patients did not participate for the second week
3. Five patients were used in 5 different pair combinations. One patient, who participated in a few encounters, was also used in Case 14. A score was produced for each pair.
4. One accuracy score was produced for each pair of patients

In Case 11, the number of encounters for each of the 5 possible pairs of patients was small (n=2-7). In addition, two values of each of the patient-related predictors were available for each pair and the patients in each pair were not mutually exclusive. For these reasons, the four patients in Case 11 were not used in the analysis of patient-related predictors for Group 1 and 2. These four patients were used in the analysis of Group 3 Factors.

Of the 28 students videotaped with each case, 19 to 28 tapes/case were found to be of sufficient technical adequacy to evaluate patient accuracy. The precision of the performance estimates for some cases will likely be compromised by this reduction in sample size contingent on the variation in accuracy score with each case. The total sample size available for each case from the 1987 and 1988 evaluation samples is displayed in Table 7.2.1 & 7.2.2.

The total number of student-patient encounters available for analysis of questions #2 and #3 in Study 2 is 636; 253 from 1987 and 383 from 1988. The 383 encounters generated in 1988 will be used to evaluate question #1: potential predictors of patient accuracy, a sample size which was estimated to be more than adequate to detect 10% of explained variance with a power of 95% and a Type 1 error of 5% (see Table 5.1.2-Chapter 5).

TABLE 7.2.1 THE NUMBER AND BREAKDOWN OF ESSENTIAL CLINICAL FEATURES (ITEMS) BY CASE AND CLINICAL TYPE FOR THE 1988 EVALUATION

CASE	NUMBER ITEMS	ITEM BREAKDOWN			# ENCOUNTERS
		History	Physical	Affect	'88
*1 COPD & Pneumonia	19	16	1	2	25
2 Pre-Op Evaluation	30	29	0	1	21
3 Back Pain	15	12	3	0	22
4 Paralysis	26	12	13	1	20
5 Urethritis	-	-	-	-	-
6 Infant Fall	13	12	0	1	30
7 Asthma	-	-	-	-	-
8 Endometriosis	19	19	0	0	23
*10 Jaundice	31	30	0	1	22
11 Dysphagia	22	19	0	3	27
12 Dizziness	26	26	0	0	28
13 Panic Attacks	22	21	0	1	24
14 Changed Bowel Habits	21	21	0	0	22
15 Short Stature	13	13	0	0	28
16 Hemiplegic Migraine	24	23	1	0	26
*18 Hypertension	8	8	0	0	21
19 Alzheimers	30	19	9	2	19
*20 Renal Artery Stenosis	11	10	0	1	22
Total	530	290	27	13	383
%		87.9	8.21	4.01	

Legend: *: Cases used in both years

TABLE 7.2.2 THE NUMBER AND BREAKDOWN OF ESSENTIAL CLINICAL FEATURES (ITEMS) BY CASE AND CLINICAL TYPE FOR THE 1987 EVALUATION

CASE ENCOUNTERS	NUMBER	ITEM BREAKDOWN			#
	Items	History	Physical	Affect	'87
*1 Uncontrolled Hypertension	11	10	0	1	16
2 Episodic Chest Pain	19	18	1	0	17
3 Lower Back Pain	17	16	1	0	17
4 Sore Throat	11	10	1	0	19
*5 Hypertension	8	8	0	0	18
6 Acute Abdominal Pain	19	16	1	2	15
*7 COPD & Pneumonia	19	16	1	2	16
8 Febrile Convulsion	23	22	0	1	15
9 Progressive Memory Loss	8	7	0	1	15
11 Headache & Wife Abuse	26	26	0	0	17
12 Infant Gastroenteritis	17	15	0	2	17
13 Diabetic Polyneuropathy	16	16	0	0	16
14 Weight Loss & Lymphadenopathy	18	16	0	2	18
15 Anaphylaxis	7	3	2	2	20
*16Jaundice	31	30	0	1	17
Total	250	228	8	14	253
%		91.2%	3.2%	5.6%	

Legend: *: Cases used in both years

VARIABLE DEFINITIONS AND MEASUREMENT

Predictor Variables

Sixteen potential determinants of patient accuracy have been identified in three groups. The definition of each variable is provided along with the source and method of measurement. The questionnaires developed to collect data on the variables identified are found in Appendix 2 and Appendix 3.

Group 1: Case Attributes

Case Complexity: is defined as the number of items the standardized patient has been trained to present for the case. The items represent clinically important features of the real patient case which have been identified by the faculty member responsible for case development (see Chapter 5). The number of items for cases used in the 1988 evaluation are found in Table 7.1. The range is 8 to 31 and it was treated as a continuous variable in the analysis.

Type of Clinical Features: The clinical features or items to be presented with each case have been broken down into three categories: items on patient history, physical findings and patient affect. The number of items for each case in each of these categories and overall cases is displayed in Table 7.2.1. Item type will be treated as an independent variable in the prediction of overall accuracy score. It will also be used as three dependent variables in the analysis of the relationship of predictive factors with accuracy on history, physical findings, and affect.

Group 1: Patient Attributes

Data on patient attributes were collected for each standardized patient at the completion of training using the self-administered questionnaire found in Appendix 2.

Age: is defined as the standardized patient's reported age at the time of training. Age was treated as a continuous variable in the analysis.

Gender: is defined as male or female. Gender was treated as a dichotomous nominal variable in the analysis.

Previous Acting Experience: is defined as previous experience and/or training in role-playing or acting. Acting experience was treated as a nominal, dichotomous variable in the analysis.

Previous Simulation Experience: is defined as having had previous simulation experience either with the case being presented in 1988 or with other cases. Since most of the patients in the study cohort would not have had extensive previous simulation experience, the effect of the amount of past experience on accuracy was not evaluated. Simulation experience was treated as a dichotomous variable (no experience, experience) in the analysis.

Previous Experience with The Health Problem: is defined as the proximity of the standardized patient's own personal life experience with the problem he/she is simulating. Three variables were created from the 4 questions used to probe this area. The first variable is defined as having had **personal experience** with the health problem being simulated. The response to 2 questions was used to create a binary scale: 0-never had this problem or symptoms similar to this problem, 1-have had symptoms similar to this problem or have had this problem. The second variable is defined as thee patient's **perceived understanding of the problem being simulated.** The standardized patient's response to a 3 category ordinal scale('not very well' to 'very well') was used. The third variable was defined as **vicarious knowledge of the health problem** he/she was simulating. The dichotomous response to the 1 question, knows someone with the problem being simulated, was used to measure this variable.

Group 2: Patient Attributes

Patient Confidence in His/Her Ability to Accurately Present the Problem Post-Training: is defined as the patient's self-rated confidence in his/her ability, at the completion of training, to accurately present the problem he/she was simulating. Confidence in his/her ability to present history, physical examination and affect items when applicable was rated on three, 5-point Likert scales using the self-administered questionnaire in Appendix 2. Scale ratings were combined as a percent confidence rating and treated as a continuous variable in the analysis.

Group 2: Training Attributes

Training Length: Two indices of training time were measured: number of sessions and total time spent in training. The standardized patient trainer recorded data for both measures on the Trainer questionnaire (Appendix 3) for each of the 49 standardized patients trained. Number of training sessions was treated as a nominal variable in the analysis and total training time as a continuous variable (most patients had one or two sessions).

Physician Assistance: is defined as the number of training sessions attended by the resource physician for the clinical problem being trained. Data were recorded by the trainer for each standardized patient using the Trainer questionnaire (Appendix 3); the information was treated as a nominal variable in the analysis.

Trainer Confidence in the Standardized Patient's Ability to Accurately Present the Clinical Problem: is defined as the trainer's rating of the standardized patient's ability to present the clinical problem. The trainer rated her predictions of the patient's ability to provide both a consistent and accurate presentation of the history, physical exam findings and affect at the completion of training using six 5 point Likert scales (Trainer questionnaire-Appendix 3.) Scale ratings were combined in the form of a percent confidence rating and treated as a continuous variable in the analysis.

Group 3: Procedural Attributes

Number of Sessions: In order to determine if patient fatigue may have adversely influenced the accuracy of presentation, data were collected on the number of times the patient had presented their problem prior to the encounter sampled. These data were collected from the from the timetable for patients, cases and students for each patient-student encounter sampled. Session number was treated as a continuous variable in the analysis.

Time Since Training: In order to determine if patient presentation accuracy improved with practice or alternatively deteriorated with the length of the interval since training had occurred, the week in which the student-patient encounter was sampled was recorded from the student,case and patient timetable. Weeks since training was treated as a nominal variable in the analysis (patients were used for two weeks after training).

Group 3: Encounter Attributes

Patient Confidence in the Accuracy of Presentation for the Encounter Sampled: In order to determine if patients were able to identify encounters in which there may have been a problem in the accuracy of their presentation, they were asked to rate their performance for each of the encounters sampled on a five point Likert scale. The ratings were completed at the end of each sampled encounter using the self-administered questionnaire in Appendix 2. Encounter ratings were treated as a nominal variable in the analysis.

Student Interpersonal Skills: is defined as the rating of the student-patient relationship provided by the patient at the completion of each student-patient encounter. It was hypothesized that the patient's ability to accurately present the clinical problem may be influenced by the standardized patient's like or dislike of the student. The sum of the patient's ratings of the doctor-patient relationship (20 Likert scales) was used as an index of patient satisfaction. It was converted to a percent score and treated as a continuous variable in the analysis (see Appendix 4).

Student Data Collection Skills: is defined as the case-specific data

collection score achieved by the student. It was rated by the patients. The method of score calculation is detailed in Chapter 5. It was hypothesized that student performance in data collection may influence the ability of the patient to accurately present the problem. Competence in data collection was treated as a continuous variable in the analysis.

Patient Accuracy

Instrument Development

The same procedure, as described in Study 1, was used to develop case-specific accuracy checklists. For the 4 cases which were used in both the 1987 and 1988 evaluations, the same accuracy checklists were used in both years. The scale used to evaluate the response to each item was the same as that used in Study 1. Two nominal categories were devoted to the correctness of response, two to the conditions of response and three to the reasons why items could not be evaluated, when applicable. The twelve new accuracy checklists developed for the 1988 cases are found in Appendix 5.

Recruitment and Training of Accuracy Raters

Two graduate students in nursing and one in social work were used to evaluate patient accuracy. Raters were trained and pre-tested in the same manner as carried out for Study 1. The test-retest reliability of raters for the first 10 encounters of each case rated is displayed in Table 7.3. In 8 of the 16 cases, rater reliability was 100%. In the remaining cases, observed agreement between the first and second rating never fell below 98.9%.

TABLE 7.3 TEST-RETEST RELIABILITY OF RATERS FOR ACCURACY AND THE CONDITIONS OF PATIENT RESPONSE (SPONTANEOUS VS. TO INQUIRY)

Case	% Observed Agreement For Accuracy	% Observed Agreement For Response Conditions
#1	100.0%	99.2%
#2	99.7%	98.4%
#3	100.0%	99.4%
#4	99.6%	98.1%
#6	100.0%	100.0%
#8	100.0%	98.4%
#10	99.8%	99.4%
#11	99.8%	98.2%
#12	100.0%	100.0%
#13	99.6%	98.5%
#14	98.9%	98.9%
#15	99.7%	99.7%
#16	100.0%	100.0%
#18	100.0%	100.0%
#19	100.0%	99.6%
#20	99.5%	98.6%

Accuracy Rating Procedure

One rater was used to evaluate all encounters sampled for a patient case. Taped student-patient encounters were placed in random order and evaluated consecutively. Any queries about the rating of an item or encounter were noted on the rating form. When the correctness of the patient's response was ambiguous, the patient was given the benefit of the doubt and the response was recorded as being correct.

Calculation of Accuracy Score

The method of calculating accuracy score has been outlined in Study 1. The same formula will be used in Study 2. As was noted in the previous chapter the percent accuracy score is calculated after correcting the denominator for the number of opportunities the patient had to provide a correct response. The opportunities to provide a correct response depended on student performance and the number of items which could be technically evaluated in the taped encounter.

The Percent of Items Provided Spontaneously-the Response Conditions

Instrument Development

The accuracy checklists developed for each case were used to record the conditions of response for each item evaluated. The condition of response was recorded on a dichotomous scale as either being provided spontaneously by the patient during the encounter or in response to inquiry or examination by the student.

Rater Recruitment and Training

The same raters used to evaluate the accuracy of standardized patient presentation were used to record the conditions of response for each item. The test-retest reliability for scoring the first 10 encounters of each case is displayed in Table 7.3. In 4 cases, rater agreement was 100%. In the remaining cases, agreement never fell below 98.2%.

Rating Procedure

The conditions of response were rated at the same time as response accuracy using the procedure outlined previously for accuracy rating. When in doubt about the conditions of the response, the response was recorded as having been provided spontaneously.

Calculation of Percent of Items Provided Spontaneously

The percent of items provided spontaneously was calculated using the following formula:

$$\text{Percent Spontaneous} = \frac{\text{Number of items provided spontaneously}}{\text{(Total Number Items - [Number Items not Asked/Examined/Evaluated])}} * 100$$

The denominator is corrected for the number of items the rater was able to evaluate contingent on the student's performance and the technical quality of the videotape. The method of calculating the conditions of response is the same as that used in Study 1.

Competency Score

Instrument Development

The overall method of selecting and developing cases for the evaluation in 1987 and 1988 has been described in Chapter 5. The case blueprint was the term employed to describe the content of each clinical problem used in the evaluation. It included: the identification of the essential clinical features of the real patient case selected (used in patient training and the development of accuracy checklists), the components of competence to be evaluated, the instruments to be used in evaluating the components specified, the numerical value of each of the items evaluated and method of scoring each of the components, the minimum cut-off score for an acceptable level of competence with the case, and remedial action to be taken for students failing to meet the minimum cut-off level. The components of competence evaluated with each case have been outlined in Tables 7.4.1 & 7.4.2 for both the 1987 and 1988 cases. The number of items used to measure each component with each case is indicated along with the maximum score points assigned.

TABLE 7.4.1 THE BREAKDOWN OF CASE-SPECIFIC COMPETENCY SCORES FOR THE 1987 CASES BY NUMBER OF ITEMS AND POINTS AWARDED FOR EACH COMPONENT MEASURED

CASE	SCORE COMPONENTS						
	Overall	Data Collect	Diagnosis	Test Sel & Interp	Manage	Commun & Pt Sat	Knowledge
1 Hypertension							
# items	23	10	4	6	-	-	3
Points	25	11	5	6	-	-	3
2 Chest Pain							
# items	34	10	5	4	15	-	-
Points	36	10	6	4	16	-	-
3 Back Pain							
# items	21	10	3	5	3	-	-
Points	75	10	11	36	18	-	-
4 Sore Throat							
# items	53	21	10	13	1	8	
Points	118	21	20	13	8	56	-
5 Hypertension							
# items	34	15	-	-	14	-	5
Points	49	26	-	-	18	-	5
6 Abdominal Pain							
# items	24	10	7	1	1	-	5
Points	33	10	7	2	4	-	10
7 COPD & Pneumonia							
# items	44	20	11	2	11	-	-
Points	43	20	8	3	12	-	-
8 Convulsions							
# items	44	24	14	6	-	N.U.	-
Points	70	24	28	18	-	N.U.	-
9 Memory Loss							
# items	24	5	4	-	11	4	
Points	31	6	8	-	13	4	-
10 Sceatica							
# items	69	31	4	3	11	8	12
Points	71	29	5	3	14	8	12
11 Headache							
# items	78	27	15		11	8	17
Points	79	28	15	-	11	8	17

CASE	SCORE COMPONENTS						
	Overall	Data Collect	Diagnosis	Test Sel & Interp	Manage	Commun & Pt Sat	Knowledge
12 Infant Gastro							
# items	43	22	13				8
Points	100	34	34	-	-	-	32
13 Polyneuropathy							
# items	28	11	12	4			1
Points	35	11	15	8	-	-	1
14 Weight Loss							
# items	38	21	12	5	-	-	-
Points	36	21	15	7	-	-	-
15 Anaphylaxis							
# items	52	22	6		20	-	4
Points	65	31	6	-	20	-	8
16 Jaundice							
# items	27	-	6	12	-	N.U.	9
Points	42	-	9	18	-	N.U.	15
Overall							
# Items(%)	636	259(40.7)	126(19.8)	61(9.6)	98(15.4)	28(4.4)	64(10.1)
Points(%)	908	292(32.2)	192(21.1)	118(13.0)	134(14.8)	76(8.4)	103(11.3)
Item:Point Ratio	1:1.4	1:1.1	1:1.5	1.1.9	1:1.4	1:2.7	1.1.6

Legend. N.U.: measured but blueprint excluded component from competency score calculations
Data Collect: history and physical examination
Test Sel & Interp: the selection and interpretation of lab tests and radiographic procedures
Commun & Pt Sat: communication and patient satisfaction

TABLE 7.4.2 THE BREAKDOWN OF CASE-SPECIFIC COMPETENCY SCORES FOR THE 1988 CASES BY NUMBER OF ITEMS AND POINTS AWARDED FOR EACH COMPONENT MEASURED

CASE	SCORE COMPONENTS						
	Overall	Data Collect	Diagnosis	Test Sel & Interp	Manage	Commun & Pt Sat	Knowledge
1 COPD & Pneumonia							
# Items	68	20	11	2	11	24	-
Points	115	20	8	3	12	72	-
2 Pre-op Assess							
# Items	55	16	-	12	3	24	-
Points	148	22	-	36	18	72	-
3 Sciatica							
# Items	73	31	4	3	11	24	12
Points	135	29	5	3	14	72	12
4 Paraplegia							
# Items	46	13	9	-	-	24	6
Points	126	35	19	-	-	72	6
5 Urethritis							
# Items	61	18	3	5	11	24	-
Points	150	27	14	14	23	72	-
6 Accident							
# Items	68	-	-	-	20	24	24
Points	128	-	-	-	28	3	28
7 Asthma							
# Items	77	25	14	10	4	24	-
Points	132	25	14	13	8	72	-
8 Endometriosis							
# Items	59	8	7	6	8	24	6
Points	115	9	11	6	10	72	7
10 Jaundice							
# Items	51	-	6	12	-	24	9
Points	114	-	9	18	-	72	15
11 Dysphagia							
# Items	72	36	4	-	8	24	-
Points	144	32	16	-	24	72	-
12 Dizziness							
# Items	79	12	25	8	10	24	-
Points	147	12	38	14	11	72	-

CASE	SCORE COMPONENTS						
	Overall	Data Collect	Diagnosis	Test Sel & Interp	Manage	Commun & Pt Sat	Knowledge
13 Panic Attacks							
# items	76	17	14	-	7	24	14
Points	124	17	14	-	7	72	14
14 Abdominal Pain							
# items	49	10	1	9	-	24	5
Points	115	10	10	13	-	72	10
15 Short Stature							
#items	71	16	31	-	-	24	-
Points	156	30	54	-	-	72	-
16 Migraine							
# items	76	20	16	15	1	24	-
Points	127	20	18	15	2	72	-
18 Hypertension							
# items	58	15	-	-	14	24	5
Points	121	26	-	-	18	72	5
19 Alzheimers							
# items	70	18	19	3	6	24	-
Points	130	24	19	5	10	72	-
20 Hypertension							
# items	47	10	4	6	-	24	3
Points	97	11	5	6	-	72	3
Overall							
#items(%)	1156	285(24.7)	168(14.5)	91(7.9)	114(9.9)	432(37 4)	84(7.3)
Points(%)	2324	349(15.0)	254(10.9)	146(6.3)	185(8.0)	1296(55.8)	100(4.3)
Item:Points Ratio	1:2.0	1:1.2	1:1.5	1:1.6	1:1.6	1:3.0	1:1.2

Legend: N.U.: measured but blueprint excluded component from competency score calculations
Data Collect: history and physical examination
Test Sel & Interp: the selection and interpretation of lab tests and radiographic procedures
Commun & Pt Sat: communication and patient satisfaction

It can be noted that the sampling base of items used to measure the respective component and case competency scores varies. In 1987, the largest proportion of items was devoted to data collection and diagnosis; in 1988 to data collection and communication. Points awarded show a similar distribution. Clearly, the precision with which each of the component and case scores is measured will vary according to the sample of items used in the evaluation. The ratio of items to points is provided. It indicates that relatively more points per item are assigned to communication, diagnosis and test selection, fewer to data collection. Even when all components are weighted equally when calculating case scores and overall scores, the precision of the estimates with the largest item sample base and lowest item:point ratio would be better.

The instruments used to evaluate each component of each case were case-specific in all instances except for the measurement of the doctor-patient relationship where one common evaluation form was used across all cases. The case-specific instruments employed are available from the author on request. The instrument used to evaluate the doctor-patient relationship was developed by Schnabl(1988) and is found in Appendix 4.

Raters and Rating Procedure

The standardized patient who presented the case was used to rate the "doctor-patient" relationship as well as actions taken on history and physical examination during the encounter. Ratings were completed at the end of each patient-student encounter. The standardized patients were trained, by the standardized patient trainer, to complete the instruments related to their case . Training consisted of orientation to the form as well as a number of practice sessions. Patient reliability in rating was not pre-tested. The reliability of standardized patient raters in 1987 is evaluated in Study 3, Chapter 8.

The faculty person responsible for developing the case was used to rate the written responses provided at the completion of each student-patient encounter. These included the questionnaires which had been developed to evaluate hypothesis generation and diagnosis, management, working knowledge, test-ordering and interpretation and data synthesis. One

faculty member was designated to rate all cases. Scoring keys, which were developed prospectively, were used to rate open-ended questions. The test-retest reliability of the faculty rater was not pre-tested. The contribution of faculty raters to score variance is confounded with case. The magnitude of the contribution of this potential source of measurement error to score variance across cases is unknown.

Competency Score Calculation

The items and values used in calculating the component scores for each case were case-specific. The method of calculating the three scores of interest in Study 2 are outlined in Chapter 5. It can be noted that each component of competence was given equal weight in case-specific score calculations despite differences in the sample of items used in its estimation. Similarly, each case was given equal weight in the overall score estimates despite differences in the test length with each case and the number of items contributing to case-specific scores. This method of score calculation provides equivalent representation for each case and component tested. It does not reflect the underlying precision of measurement of each component and case. This method of score calculation may reduce precision in competency score and thereby attenuate the relationship between patient accuracy and competency score.

The raw data scores for each item evaluated were entered into the computer and all data entry verified by repeat entry. Scores were then calculated, by computer according to the rules specified in the case blueprint. Misclassification of students through computation errors should therefore be minimal.

ANALYSIS

Descriptive Analysis

Response Bias

In order to characterize potential sources of response bias, the frequency of missing data was tabulated for each of the variables under study. Evidence in support of the presence of response bias was considered to exist if there:

*were significant differences between the accuracy scores for standardized patients who had complete data on predictor variables vs. those who had incomplete or absent data

*was a significant relationship between the overall percent of items which could not be evaluated and accuracy score, percent spontaneous score and/or overall student competency score

*were significant differences in the frequency with which items could not be evaluated within categories of each predictor variable or among different cases.

Appropriate statistical tests were used to evaluate each of these potential sources of response bias with the null hypothesis being that no differences/relationship exists.

Descriptive Summary of the Variables Under Study

Univariate and bivariate descriptive statistics were employed to describe the characteristics of the variables under study. The data summary will include:

*Predictor Variable Data: the percent frequency of response or mean response for each variable

*Patient Accuracy: mean accuracy score by case and patient for the 1988 evaluation along with a case by case comparison for the

four cases common to both the 1987 and 1988 evaluations

*Response Conditions: the mean percentage of items provided spontaneously by case and patient

Hypothesis Testing-Bivariate and Multivariate Analysis

Predictors of Patient Accuracy

In the evaluation of predictors of patient accuracy, the patient-student encounter was used as the unit of analysis. One observation of accuracy score is provided for each student-patient encounter. Multiple observations of accuracy score are provided fo each of the patients included in the study. In order to evaluate the contribution of each of the predictor variables to accuracy score, accuracy score variance will be partitioned into that attributable to repeated measures on the same patient, that attributable to differences among patients in the same predictor category and that attributable to differences between categories of the predictor variable. Variance attributable to repeated measures on the same patient and between patients in the same category will be treated as error and the proportion of variance explained by each predictor will be calculated using the general formula provided by Kleinbaum and Kupper(1978). The only exception to this rule will be in the analysis of item type. An accuracy score for each category of item type was calculated for each encounter. The proportion of variance explained by item type will be calculated by partitioning score variance into that due to patient-student encounter and that due to item type.

The overall proportion of variance explained by each of the three groups of predictive factors will be evaluated separately using multiple regression analysis. As indicated in Chapter 5, the explanation of 10% of the variance by Group 1 factors, 20% by Group 2 factors and 30% by Group 3 factors will be considered of practical importance. The independent contribution of each factor in the 3 groups to the proportion of variance explained will be estimated and reported using the partial correlation coefficient (Kleinbaum & Kupper,1978). The correlation among factors within each group will be initially evaluated to determine if problems with multicollinearity may be encountered. A Pearson product moment

correlation coefficient will be used to examine the association between continuous variables. The phi coefficient will be used to examine the association between categorical variables. Secondary analysis will be carried out on the relationship of each group of factors to accuracy score for the three types of items; history, physical findings, and affect.

The Relationship of Patient Accuracy to Competency Score

The shape of the relationship between patient accuracy score and competency scores will initially be examined by plotting the residuals of linear regression analysis. On the basis of inspection of the residuals, a variety of curvilinear relationships may be explored. If the additional variance explained by a curvilinear relationship is significantly greater than that explained by a linear model, a curvilinear model will be used to estimate the relationship between patient accuracy and competency score.

A repeated measures multiple regression model will be used to evaluate the relationship of patient accuracy with competency score over all cases. The repeated measure will be students, the independent variable will be patient accuracy and the dependent variable will be competency score. A linear or polynomial regression model will be used to evaluate the relationship in individual cases. Bonferroni's correction for multiple comparisons will be used to adjust the Type I error.

The Conditions of Patient Response

Differences in the percent of items provided spontaneously between the two patients presenting each case will be evaluated by an independent t-test. An unbiased estimate of the relationship between the percent of items provided spontaneously and student score cannot be obtained because the patient was used to rate data collection and patient communication scores. If the conditions which led the patient to provide more data spontaneously also influence the rating of student actions, then no relationship between the conditions of response and student score will be found.

Standardized patients are at particular risk to provide more clinically relevant data to students who are having difficulty with the case. If the

evaluation is being conducted to detect students who are unsafe or ineffective in the clinical practice domain, then biases in this direction are more serious (students who are unsafe are classified as competent). If standardized patients were providing more data to these type of students and rating them in a more lenient fashion, then the variance in scores for patients who provide more data spontaneously would be smaller than for patients who provide less data spontaneously. In order to evaluate this possibility, differences in the variance of student scores for interpersonal skills and data collection for patients presenting the same case will be evaluated.

Differences in the variance of student scores between two patients presenting the same case will be evaluated by using the folded form of the F statistic (SAS,1985). It uses a two-tailed F test to examine the null hypothesis that the two variances are equal. In cases where more than two patients were involved in case presentation, a one-way ANOVA model will be used to estimate the respective variance components. The assumption of equal variances will be evaluated by using Hartley's test (Winer,1972). A potential bias will be considered to be present if the student score variance is significantly smaller in patients who provided more data spontaneously.

RESULTS

RESPONSE BIAS

Missing Values For Patient Related Predictors

Five patients had missing values for one or more of the patient-related predictor variables. In addition, patients who presented Case # 11 were excluded from the evaluation of patient related predictors for the reasons indicated in the section on sampling procedure. They have been added to the 5 patients with missing values. When the accuracy scores for these 9 patients are compared with the 29 who had no missing values, accuracy score was significantly larger for patients with missing values (96% versus 93%) (see Table A7.1 in Appendix 7). With the exclusion of the 4 patients for Case #11, there is no significant difference in accuracy score between the two groups. It was concluded, therefore, that there were no systematic differences between the accuracy scores for patients with and without missing predictor data. The potential for response bias in the evaluation of patient-related predictors through this route is unlikely.

Missing Values For Accuracy Items For Each Case

For each case, 8-31 accuracy items were to be evaluated in each of the patient-student encounters sampled. The number of items for each case is found in Table 7.2.1. Over all cases and patient-student encounters, there were 7368 occasions in which accuracy items could be rated. The percentage of times that items could not be evaluated with each case and patient are displayed in Table A7.2 in Appendix 7. Over all cases, items could not be rated on 34% of occasions. There were significant differences in the percentage of times items could not be evaluated among the 16 cases. In all cases, the most frequent reason that an item could not be evaluated was because the student did not inquire about the item. Response bias attributable to differences in the data available to estimate the accuracy of patient presentation for individual cases may be present.

Missing Values For Accuracy Items For Each Category of Predictor Variable

The number of opportunities to rate accuracy items for each of the predictor variables is displayed in Appendix 7 Table A7.3. The number of opportunities is the product of the number of encounters rated in each category and the number of items which could be rated for each encounter. The percentage of times items could not be evaluated for each category of a predictor variable is also found in Table A7.3.

For item type, there were differences in the proportion of times that items could not be rated. More items could not be rated for history (33%) and physical examination items (48%) than for affect (6%). Missing values for history items are a result of student performance (the student did not ask about the information contained in the item). For physical examination items, technical quality of the videotape and student performance each explained about half of the missing values. A biased representation for the estimated accuracy of patient performance for items from these two categories can not be ruled out. Since the reason an item could not be rated probably is independent of patient accuracy, a biased estimate of these two categories seems unlikely.

There was a greater proportion of missing values generated for patients in the 50-59 years of age group than for other age groups. This may bias the estimation of accuracy for this group. Small sample sizes in this group also may contribute to poor precision in the estimation of patient accuracy.

With respect to the patient's understanding of the health problem being presented, only one patient indicated that they did not understand the problem well. Small sample sizes and a greater proportion of missing values in this category may contribute to bias and a lack of precision in the estimate of patient accuracy. The same problem is present for the patient's rating of his/her performance during the examination. Only 2 patients rated their performance as poor and the largest proportion of missing values is present in this category (65%). In the interpretation of results, no inferences will be drawn about patient accuracy in these two categories for the reasons outlined.

There is a slightly greater proportion of missing values for patients who had 3 training sessions (46% versus 31% and 16% for 1 and 2 sessions)

and/or 4 hours of training (46% versus 24%-35% for patients having 1-3 hours of training). Estimates of patient accuracy for the longest training length may be biased by a greater proportion of missing values.

The Relationship of Missing Values for Accuracy Items and Accuracy and Competence Score

The percentage of times accuracy items could not be evaluated was associated with some of the predictor variables and with case. The relationship between percentage of items missing and the two main dependent variables used in the subsequent analyses was evaluated in order to determine if the percentage of items missing could act as a confounder in the analysis. There was no significant linear relationship between the percentage of accuracy items missing and overall patient accuracy score and overall student competence score. Since there is no association with the two main dependent variables in the analyses, it is unlikely that the percentage of items missing could act as a confounder in the estimated relationships of predictors and patient accuracy and patient accuracy and competence score.

PREDICTORS OF THE ACCURACY OF STANDARDIZED PATIENT PRESENTATION

Descriptive Statistics of the Three Groups of Potential Predictor Variables

In Table 7.5, the number of patients (and cases where applicable) in each category of the potential predictor variables is provided. For Group 3 variables, multiple observations of student performance and patient rating of the quality of performance were collected, one for each encounter evaluated. The number of patients contributing to each category of a predictor variable is indicated. The maximum number per category is 38 patients.

The number of observations of accuracy score for each category of the predictive variables is provided in the second column of the table. It represents the number of patient-student encounters in which accuracy was calculated. Overall, accuracy scores were calculated for 374 patient--student encounters. The number of patient-student encounters where values

for the predictor variable were missing is indicated.

The average number of accuracy items rated for each case was 19.7 with a range of 8 to 31. The number of items the patient must present with each case is used as an index of case complexity. Number of items is categorized to help see the distribution of encounters sampled for each category within the range. It can be seen that there is a reasonably equivalent distribution of cases within the four categories of the range. Number of items is moderately correlated with two of the other variables in Group 1: health problem experience(phi=.55) and vicarious knowledge of the health problem (phi=.50). Those without health problem experience or vicarious knowledge of the health problem were more likely to present cases with larger numbers of accuracy items.

An accuracy score for each type of item was calculated, when applicable, for each encounter. History items account for 88% of all accuracy items evaluated while physical exam and affect account for 8% and 4% respectively. This distribution is similar to the breakdown of accuracy items evaluated in Study 1 and reflects the same rationale in case selection discussed previously (see Chapter 6). The distribution of item type by case was displayed earlier in Table 7.2.1. Physical exam items are positively associated with the number of items to be presented with each case.

Approximately two-thirds of the standardized patients were adults under 50 years of age. In Case 15, 6 five year old boys and their mothers presented the problem. Since all accuracy items assessed were provided by the mother, the children are not included in the analysis of predictive factors. The age of the standardized patient was associated with the three variables related to the health problem. Older patients were more apt to have had the health problem or symptoms similar to those presented (phi=.56). Younger patients were more apt to have had vicarious knowledge of the health problem (phi=.45).

Two-thirds of the standardized patients were female. Gender was associated with age, with more females being in the younger age groups (phi=.62). Females were more likely than males to rate that they understood the problem very well.

Most of the standardized patients had experience in either role-playing (66%) and/or acting (26%). Of this latter group, 23% had acting training. Younger patients were more likely to have had acting experience (phi=.39). Those with acting experience were more apt to indicate that they had a good understanding of the problem they were presenting.

About half of the standardized patients had prior experience simulating health problems. The length of experience was short for all patients with all having had experience with only one other case which was presented on a maximum of two other occasions. Those who had not simulated a problem before were more apt to have had experience with the health problem and indicate that they had a better understanding of the problem they were presenting (phi=.41).

Almost half of the patients had either had the health problem they were presenting(16%) and/or had had symptoms similar to the health problem they were presenting(42%). Experience with the health problem was positively associated with the patient's reported understanding of the problem they were presenting (phi=.34). About half of the patients knew someone who had the problem they were presenting and again this subgroup had a better understanding of the case they were presenting (phi=.39).

Patients were more confident than the trainer about their ability to accurately present the clinical problem. The average patient rating was 91% with only 5 patients rating their ability in the interval of 51%-75%. The average trainer rating was 81% with 11 patients being given ratings in the 51%-75% interval. The correlation between the patient's and trainer's ratings was Pearson's r=.24. Patient's confidence in their ability showed a modest negative association with training attributes (r=-.25 to-.27 for number of sessions, hours trained and MD assistance). Trainer confidence was not associated with any of the training variables except for a modest association with the number of sessions with a physician resource (r=.23).

Most patients had two training sessions averaging 2.5 hours in total. For most patients (81%), a physician resource was present at least one session. Variables characterizing the patient's training were all strongly

correlated with each other. The number of sessions correlated at r=.89 with the number of hours spent in training and r=.61 with the number of sessions attended by a physician resource. It may not be possible to obtain separate estimates of the association of each of the training variables with patient accuracy.

Number of sessions already presented that day by the patient prior to the videotaped sample of performance has been categorized to permit visual inspection of the breakdown of the encounters evaluated. Most patients were taped in all three times of the day. The number of encounters done by the patient that day was not associated with any of the other variables to be evaluated in Group 3.

Slightly more encounters were sampled in the second week after training (59%) however most patients were represented in both weeks. Weeks since training was not associated with any other variable in Group 3.

During the student evaluation, patients rated their confidence in the quality of their own performance at the end of each encounter. The average rating was 3.9%. It can be noted from the categorical breakdown provided that only two patients rated their performance as poor. The majority rated their performance as good or very good. In view of the small range of values, patient confidence rating will be treated as a nominal variable in the analysis. Patient confidence rating was not associated with any of the other variables in Group 3.

The variation of the two student performance variables is small. The average score for interpersonal skills was 72.8% with a standard deviation of 6.2%. There were no students who received a score of less than 41%. Only 16 of the patients provided interpersonal skills ratings in the 41%-60% or 81%-100% intervals. Each patient rated at least one student in the 61%-80% interval. The average score for data collection was 67.2% with a standard deviation of 6.2%. All student scores were in the interval of 41%-80%. The number of patients contributing scores to these two intervals was about equivalent. The two student scores were correlated at r=.37. Student scores were not correlated with any of the other variables in Group 3.

TABLE 7.5 DESCRIPTIVE STATISTICS OF POTENTIAL PREDICTOR VARIABLES: PERCENT FREQUENCY BY CATEGORY AND MEANS FOR EACH PREDICTOR VARIABLE

Predictor Variables		Number	Number of Times Accuracy (% Frequency) Evaluated (patient*student)	
Group 1				
A) Case Attributes				
Number of Items/Case				
8-15		5 cases	123	(32.9)
16-20		2 cases	48	(12.8)
21-25		3 cases	121	(32.4)
26-31		4 cases	82	(21.9)
[mean=19.7, s.d.=6.31]				
Item Type				
History		290 items	373	(55.3)
Physical Findings		27 items	111	(16.4)
Affect		13 items	191	(28.3)
B) Patient Attributes				
Age				
20-29 years		8 patients	81	(23.7)
30-39 years		10 patients	92	(26.9)
40-49 years		3 patients	38	(11.1)
50-59 years		1 patient	10	(2.9)
60-69 years		7 patients	82	(24.0)
>70 years		3 patients	39	(11.4)
Missing		6 patients	32	
[mean=44.1, s.d.=17.9]				
Gender				
Male		11 patients	121	(35.4)
Female		21 patients	221	(64.6)
Missing		6 patients	32	
Previous Experience				
Acting	Yes	21 patients	221	(64.6)
	No	11 patients	121	(35.4)
	Missing	6 patients	32	

Predictor Variables		Number	Number of Times Accuracy (% Frequency) Evaluated (patient*student)
Hours	1	7 patients	49 (14.3)
	2	7 patients	82 (24.0)
	3	13 patients	154 (45.0)
	4	5 patients	57 (16.7)
	Missing	6 patients	32
	[mean=2.43, s.d.=.84]		
MD Assistance (# sessions)	0	6 patients	65 (19.0)
	1	10 patients	91 (26.6)
	2	13 patients	155 (45.3)
	3	3 patients	31 (9.1)
	Missing	6 patients	32
Group 3			
A) Procedural Attributes			
*Number of Sessions Done that Day	1-3	31 patients	119 (31.8)
	4-6	35 patients	99 (26.5)
	7-10	36 patients	156 (41.7)
	[mean= 5.46; s.d.=2.9]		
*Time Since Training	1 week	37 patients	150 (40.1)
	2 weeks	34 patients	224 (59.9)
B) Encounter Attributes			
*Patient Confidence in Performance	1-2	2 patients	2 (.8)
	3-4	30 patients	212 (83.8)
	5	13 patients	39 (15.4)
	Missing	33 patients	121
	[mean=3.85; s.d.=.17]		
*Student Performance Interpersonal Skills	0-20	0 patients	0
	21-40	0 patients	0
	41-60	16 patients	16 (4.5)
	61-80	38 patients	327 (91.0)
	81-100	16 patients	16 (4.5)
	Missing	15 patients	43 students
	[mean=72.8, s.d.=6.15]		

Predictor Variables		Number	Number of Times Accuracy (% Frequency) Evaluated (patient*student)
Simulation	Yes	16 patients	173 (50.6)
	No	12 patients	169 (49.4)
	Missing	6 patients	32
Has Health Problem	Yes	14 patients	178 (52.0)
	No	18 patients	164 (48.0)
	Missing	6 patients	32
Vicarious Knowledge Of Problem	Yes	15 patients	152 (45.5)
	No	16 patients	182 (54.5)
	Missing	7 patients	40
Understands Patient Problem	Well	17 patients	187 (55.2)
	Fair	13 patients	138 (40.7)
	Not	1 patient	14 (4.1)
	Missing	7 patients	35
Group 2			
A) Patient Attributes			
Patient Confidence			
51-75%		5 patients	47 (14.1)
76-100%		25 patients	286 (85.9)
Missing		8 patients	41
[mean=91.3, s.d.=9.99]			
B) Training Attributes			
Trainer Confidence			
51-75%		11 patients	130 (40.8)
76-100%		16 patients	189 (59.2)
Missing		11 patients	55
[mean=80.62, s.d.=11.1]			
Training Length			
# Sessions	1	9 patients	71 (20.8)
	2	18 patients	214 (62.6)
	3	5 patients	57 (16.6)
Missing		6 patients	32

Predictor Variables	Number		Number of Times Accuracy (% Frequency) Evaluated (patient*student)
*Student Performance Data Collection	0-20	0 patients	0
	21-40	0 patients	0
	41-60	38 patients	54 (16.3)
	61-80	33 patients	277 (83.7)
	81-100	0 patients	0
	Missing	27 patients	13 students
	[mean=67.17, s.d.=6.14]		

Legend:
*: 38 patients were used in the analysis. In group 3, multiple values for each patient are present, one for each encounter evaluated. The number of patients (of the 38) represented in each category of the prediction variable are indicated. If perfectly balanced, all 38 patients would be represented in each category of a predictor variable.

Descriptive Statistics of the Dependent Variable: Accuracy Score

Table 7.6 displays the accuracy scores for the 16 cases evaluated in 1988. The overall accuracy score for standardized patients presenting in 1988 was 93.4% with a standard deviation of 8.7%. Significant differences were present among the 16 cases (p=.0001). The lowest score was 80.6% for Case 20 and the highest score was 100% for Case 18. In five cases the accuracy score was below 90% (Case #4,#12,#14,#20 & #15). Of these cases, accuracy items could not be rated on approximately 50% of occasions for 2 cases (Case #4 & #12) but could be rated on more than 80% of occasions for the lowest scoring case, Case #20.

Accuracy score was above 95% for 8 of the 16 cases (Case #3, #6, #8, #10, #11, #16, #18 & #19). In two of these cases, accuracy items could not be rated on 40-50% of occasions; in the one case with the highest score accuracy items could be rated on approximately 80% of all occasions. After using Bonferroni's correction for multiple comparisons, there were significant differences in accuracy score between patients presenting the same case in 2 of the 16 cases (Case #1, a 10% difference and Case #4, a 9% difference).

The variation in standardized patient accuracy score in the 1988 evaluation is small. This may limit the ability to identify important predictor of patient accuracy in the current study.

TABLE 7.6 MEAN % ACCURACY SCORE BY CASE, PATIENT AND OVER ALL CASES: 1988

Cases With 2 Pts.	Number Encount.	Number of Items	Case Accuracy	Accuracy by Patient: Patient #1 N	Patient #1 Mean	Patient #1 S.D.	Patient #2 N	Patient #2 Mean	Patient #2 S.D	P-value (t-test or f-test)
1	25	19	93.4%	10	99.4	(2.0)	15	89.4	(9.7)	.001
2	21	30	94.1%	16	93.0	(3.8)	5	97.8	(3.3)	.02
3	22	15	98.5%	14	99.5	(2.1)	8	96.8	(4.4)	.14
4	20	26	89.9%	9	85.0	(4.4)	11	93.9	(6.6)	.003
6	30	13	96.8%	15	99.4	(2.3)	15	94.2	(10.6)	.08
8	23	19	98.6%	9	99.1	(2.8)	14	98.3	(3.4)	.57
10	22	31	99.0%	13	100.0	(0)	9	97.6	(2.9)	.006
12	22	26	88.3%	12	87.7	(6.3)	10	89.1	(5.9)	.60
13	24	22	92.8%	11	93.5	(5.1)	13	92.2	(4.6)	.53
14	22	21	83.9%	8	84.1	(3.5)	13	83.8	(8.2)	.89
16	26	24	96.9%	14	98.7	(2.6)	12	94.8	(7.2)	.10
18	21	8	100.0%	5	100.0	(0)	16	100.0	(0)	-
19	19	30	97.2%	9	98.1	(5.6)	10	96.4	(9.1)	.63
20	22	11	80.6%	8	81.3	(10.9)	14	80.2	(6.1)	.77

Cases With > 2 Patients	Number Encount.	Number of Items	Case Accuracy		Pt#1	Pt#2	Pt#3	Pt#4	Pt#5	Pt#6	P-value
11	27	22	99.0%	N	9	7	6	2	3		.52
				Mean	98.8	100.0	97.2	100.0	100.0		
				S.D.	3.7	0	4.4	0	0		
15	28	13	84.7%	N	3	2	3	6	2	12	.17
				Mean	77.8	79.2	88.4	93.7	70.1	84.3	
				S.D.	19.2	5.9	11.1	7.0	10.8	11.8	

Total	373	330	93.4 (8.7)	.0001

Legend: Number Encount.: the number of patient-student encounters used to calculate accuracy score for each case

Number Items: the number of items used to rate accuracy with each encounter

Comparison of Standardized Patient Accuracy for University of Manitoba Patients For the 1987 and 1988 Cohorts

The overall accuracy score for standardized patients from the University of Manitoba in 1987 was 89.1% with a standard deviation of 12.8%. In 1988, the overall accuracy score was 93.4% with a standard deviation of 8.7%, an improvement of 4.3%. In order to clarify the source of the difference between the two years, the scores for patients participating in both 1987 and 1988 and those participating in only one of those years were calculated. The results are displayed in Table 7.7. An independent t-test was used to evaluate whether the observed differences between patients scores in 1987 and 1988 for the two groups of patients could be attributable to chance. In addition the scores between the 4 cases which were used in both 1987 and 1988 were compared.

Thirteen patients were used in 1987 and 1988. Of these 13 patients, only 3 presented the same case in both years. The mean score for these patients in 1987 was 94.6% and in 1988 it was 92.5%. This difference was not significant.

Twenty-two patients were used only in 1987 and 31 patients were used only in 1988. The scores for the 1987 patients was 85.7% and for the 1988 patients was 93.8%. This difference is statistically significant ($p<.006$).

Those responsible for selecting and training patients in 1988 were not aware of the results of patient accuracy in 1987. If the observed differences in accuracy score were due to better training in 1988, one would expect that the scores for patients used in both years would be lower in 1987. This was not the case. The observed pattern of scores is more compatible with the hypothesis that the trainer (who was used in both years) was better at selecting patients who were more likely to be accurate in 1988.

Of the four cases used in both 1987 and 1988, there was a significant difference in accuracy score between the two years for 2 cases (Case #2, a 10% difference and Case #3, a 10% difference). In both cases, accuracy

scores were higher in 1988. For the remaining two cases, mean accuracy score was 80% for one case and 99% for the other. These findings suggest that both the patient and the case selected may be important factors in patient accuracy.

TABLE 7.7 COMPARISONS OF ACCURACY SCORES FOR STANDARDIZED PATIENTS IN 1987 AND 1988 AT THE UNIVERSITY OF MANITOBA

	Accuracy Score for Patients Used in One Year Only		Accuracy Score for Patients Used in Both Years	
By Year	N	Mean (s.d.)	N	Mean (s.d.)
1987	22	85.7% (13.6%)	13	94.6% (9.2%)
1988	31	93.8% (8.5%)*	13	92.5% (9.2%)
By Cases in Common for 1987 & 1988	1987		1988	
	N	Mean (s.d.)	N	Mean (s.d.)
Case 1	15	79.7% (9.3%)	22	80.6% (8.0%)
Case 2	18	89.5% (10.1%)	21	100.0% (0)*
Case 3	16	83.3% (10.2%)	25	93.4% (9.0%)*
Case 4	17	99.5% (1.4%)	22	99.0% (2.2%)

Legend: *: significant difference ($p<.006$) after correction for multiple comparisons between 1987 and 1988 using an independent t-test

N: number of patients or number of patient-student encounters

Notes: Of the 13 patients used in 1987 & 1988, only 3 presented the same case in both years.

Group 1 Factors: The Relationship of Factors Which Could be Applied in Patient & Case Selection to Accuracy Score

The Relationship of Group 1 Factors to Overall Accuracy in Case Presentation

Table 7.8 displays the mean accuracy score for 9 potential predictive factors in Group 1. Case complexity and age were treated as continuous variables in the analysis but are categorized to facilitate inspection of the distribution of accuracy scores. The proportion of variance explained (R^2) by each factor for the bivariate analysis is reported in the fourth column of the table. The proportion of variance explained by each variable when all factors are included in the regression model is reported in the fifth column of the table.

Overall, Group 1 factors explained only 11.8% of the variance in standardized patient accuracy score. In the bivariate analysis, the patient's reported understanding of the problem he/she was presenting, the type of clinical item being presented, simulation and health problem experience, and age were the five factors which explained the largest proportion of the variance.

Age appeared to have a non-linear relationship to patient accuracy. However, smaller sample sizes in the middle of the range studied limit the precision of this estimate. Patients in the 40-69 years of age interval were less accurate than younger patients. Patients in the over 70 years of age group had the highest accuracy score. When all factors are taken into consideration, age explained only 1% of the variance. Within the range of ages studied, age does not appear to be of practical importance in patient selection.

Previous simulation experience explained 2.8% of the variation in accuracy score. Patients with experience scored approximately 1% higher than those without experience. The same trend was present for health problem experience. Those reporting health problem experience scored about 2% higher

than those who did not. This factor explained 1.4% of variance in accuracy score.

The patients' understanding of the problem they were presenting explained the largest proportion of variance (5.2% in the model with all factors). Those who understood the problem well scored 5% higher than those who had only a fair to poor understanding. As was noted previously, the patient's reported understanding of the problem is correlated with age (better in the younger age groups), acting and simulation experience and with knowledge of the health problem being presented.

Consistent with the findings in the evaluation of patient accuracy in 1987, physical findings are presented least accurately (79.4%) followed by patient affect (89.5%). Item type explained 4.5% of the variance in accuracy score in the bivariate analysis. Since there were three values for accuracy for each encounter, one for each category of clinical feature type, it could not be included in the analysis of all factors combined.

Patient gender and the number of items to be presented with the case appeared to have little association with accuracy score; these two factors explained less than 1% of the variance.

The Relationship of Group 1 factors to Accuracy in the Presentation of the History, Physical Findings and Patient Affect

Whereas the patients' reported understanding of the problem had been the one factor which explained the greatest proportion of variance in overall accuracy, its effect is limited to the patients' ability to accurately present the patient history and affect (explaining 6% of the variance in accuracy score for history and 3% for affect). There is a negative relationship between this factor and percent accuracy on physical findings. Patients who understood the problem well had a mean score of 95% on history, 90% on patient affect and 73% on physical findings. Those having a fair understanding of the problem had a score of 90% on history, 78% on physical findings and 83% on affect.

Simulation and health problem experience had little association with the accuracy of the patient history. The major impact of these factors is on the accuracy of presentation of the physical findings and patient affect. Simulation experience is associated with better accuracy on both physical findings (83% vs. 75% for no experience) and affect (94% vs. 82% for no experience). It explained 1% of the variance for physical findings and 8% for patient affect.

Experience with the health problem was associated with higher scores on patient affect (90% vs. 82% for those with no experience) but lower scores on physical findings (75% vs. 83% for those with no experience). This factor explained 9% of the variance on physical findings and 2% for patient affect. Few patients who had health problem experience had previous simulation experience. The absence of previous simulation experience is the most likely explanation for patients with health experience doing more poorly on physical findings.

Acting and/or role-playing experience was not associated with accuracy for patient history or affect; it was associated with the accuracy of physical finding presentation (96% for those with experience vs. 67% for those without). It explained 2% of the variance in the presentation of physical findings.

Although gender was not associated with overall accuracy score or with accuracy on patient history, female patients provided a less accurate presentation of the patient affect (85% vs. 100% for males); gender explained 10% of the variance in score.

Those in the youngest age group (20-29) provided the most inaccurate presentation of physical findings (53% compared with 80% for 30-39 yrs., 90% for 60-69 yrs. and 96% for >70 yrs.). Age explained 2% of the variance in accuracy of presentation of physical findings. However, this observation was limited to one case with 21-25 items. It is not possible therefore to draw any conclusions about the independent effects of age, and case complexity in relationship to the accuracy of physical finding presentation. Although it has been suspected that older patients would do

less well overall and in particular in the presentation of physical findings, there was no evidence that this was the case with the patients included in this study. Those over the age of 60 had the highest scores of any age group on physical findings and had equivalent scores on history and affect.

TABLE 7.8 GROUP 1 PREDICTIVE FACTORS: FACTORS WHICH COULD BE APPLIED IN PATIENT AND CASE SELECTION - PERCENT ACCURACY SCORE AND PROPORTION OF VARIANCE EXPLAINED BY EACH FACTOR

Predictive Factor		Percent Accuracy Score			Proportion of Variance Explained	
		N	Mean	S.D.	Bivariate	All factors Included
Case Complexity						
# items:	8-15	123	92.0%	11.0%	.8%	.8%
	16-20	48	95.9%	7.3%		
	21-25	120	94.7%	7.2%		
	26-31	82	92.2%	7.1%		
Item Type						
History		374	93.5%	8.9%	4.5%	
Physical Findings		374	79.4%	37.5%		
Affect		374	89.5%	30.7%		
Patient Age						
20-29		81	94.9%	6.8%	3.5%	1.2%
30-39		92	94.3%	6.8%	-	
40-49		37	90.9%	9.5%		
50-59		10	89.1%	5.9%		
60-69		82	90.7%	9.7%		
>70		39	95.5%	6.9%		
Patient Gender						
Male		120	93.2%	8.3%	.00%	00%
Female		221	93.2%	8.8%		
Previous Experience						
Acting	Yes	221	94.0%	8.5%	3.2%	00%
	No	120	91.8%	8.8%		
Simulation	Yes	172	93.6%	8.6%	2.5%	2.8%
	No	169	92.8%	8.7%		
Has Health	Yes	178	94.1%	8.4%	2.4%	1.4%

Predictive Factor		Percent Accuracy Score			Proportion of Variance Explained	
		N	Mean	S.D.	Bivariate	All factors Included
Case Complexity						
Problem	No	163	92.2%	8.8%		
Vicarious	Yes	152	94.4%	9.2%	.8%	00%
Knowledge	No	181	92.0%	8.1%		
Understands	Well	187	95.1%	7.3%	13%	5.2%
Patient	Fair	137	90.0%	9.6%		
Problem	Not	14	91.1%	13.3%		
Overall		374	93.4%	8.7%		11.8%

Group 2 Factors: The Relationship of Factors Which Could be Applied During or at the Completion of Training to Accuracy Score

The Relationship of Group 2 Factors to Overall Accuracy in Case Presentation

Table 7.9 displays the mean accuracy score and standard deviation for case presentation for each category of the five predictor variables. The R^2 for both the bivariate and multiple regression analysis are reported in the fourth in fifth columns of the table respectively. Overall, Group 2 factors explained 10.1% of the variance in the accuracy of patient presentation. Attributes of the training process were the only factors which explained a substantial proportion of the variance. Patient confidence was positively associated with accuracy score but explained only 1% of the variance. Trainer confidence was negatively associated with presentation accuracy with higher scores by the trainer being associated with slightly poorer accuracy scores. This factor explained 1.8% of the variance in the bivariate analysis and 3.4% of variance in multiple regression analysis.

Of the training attributes, the number of sessions and sessions attended by an MD are the two factors most strongly associated with patient accuracy in the bivariate analysis. Since training hours and number of

sessions were strongly correlated, only number of sessions was included in the multiple regression model to avoid problems of multicollinearity. The independent variance contribution of these two factors was 3.3% for number of training sessions and 2.5% for number of sessions attended by a physician resource. In inspecting the means for these two attributes, it is apparent that the most frequent training category is associated with the best patient accuracy. Those patients who had received two sessions had the best accuracy score. Those with more training sessions were less accurate. The explanation of this observation is likely one of selection bias: the most common training practice for standardized patients is two sessions with one or two of those sessions being attended by a physician resource. More sessions are added if the patient(s) appears to be having some difficulty with the case presentation and/or if the case is one which is particularly difficult to present. Therefore, an unbiased estimate of the effect of three or more sessions is not possible. The relationship of training attributes to case performance is more clearly understood when the relationship of these factors to accuracy on history, physical findings and affect is evaluated.

The Relationship of Group 2 Factors to Accuracy in Presentation of the History, Physical Findings and Patient Affect

The patient's confidence in his/her ability to accurately present the problem was modestly and positively related to the presentation of the history (explaining 1.4% of score variance). Patients who rated their ability in the 76-100% interval had an accuracy score of 93% in contrast to 90% for those rating in the 51-75% interval. Patient confidence rating was not associated with accuracy in the presentation of the physical findings or patient affect.

The trainer's confidence in the patient's performance was positively associated with accuracy in history presentation (explaining 5.6% of score variance) but was negatively associated with accuracy in presentation of affect (explaining 3.4% of score variance). There was no association between trainer rating and the presentation of physical findings. The

converse was true for patient affect: the mean score was 80% for patients with ratings over 75% and 94% for patients with lower ratings.

Number of training sessions explained 2% of the variance on history and affect and 23% of the variance in accuracy score for physical findings. Accuracy in the presentation of the history and patient affect were better with two training sessions than one (History: 89% for one session and 95% for two sessions; Affect: 81% for one session and 95% for two). Patients who had three training sessions were slightly less accurate in the presentation of the history (92%) than those with two sessions but were better than those with one session. These findings suggest that one session is likely not a sufficient amount of training time. All patients who presented physical findings had two or three sessions. Those with two training sessions had an accuracy score of 93% while those with three sessions had an accuracy score of 49%. This trend is likely explained by the problem of patient or case selection bias discussed previously.

The evaluation of number of hours spent in training shows a similar trend. Accuracy in history and affect presentation improve linearly up to three hours of training. Patients presenting physical findings all had at least three hours of training. Those with three hours of training were more accurate than those with four, the four hour group being the same as those receiving three training sessions. The estimate for the four hour group is biased for reasons outlined in the previous paragraph.

The assistance of a physician resource with the training session has a positive effect on the accuracy of patient presentation for physical findings and patient affect. This factor explained 17% of the variance in the accuracy score for physical findings and 1% for patient affect. Those with no sessions attended scored 77% on accuracy of affect; with one session attended, the score was 97% and with three, the score was 100%. A physician resource was present at one or more sessions for all patients who were trained for physical findings. For patients with physician attendance at one session, mean accuracy in physical finding presentation was 90%, for two sessions it was 73% and for three sessions 100%. There

was no relationship between physician assistance and accuracy of presentation on the history.

TABLE 7.9 GROUP 2 PREDICTIVE FACTORS: FACTORS WHICH COULD BE APPLIED DURING OR AT THE COMPLETION OF TRAINING - PERCENT ACCURACY SCORE AND THE PROPORTION OF VARIANCE EXPLAINED BY EACH FACTOR

Predictive Factor		Percent Accuracy Score			Proportion of Variance Explained	
		N	Mean	S.D.	Bivariate	All Factors Included
Patient Confidence in Ability						
51-75%		47	90.5%	9.9%	1.2%	.9%
76-100%		285	93.5%	8.4%		
Trainer Confidence in Patient Ability						
51-75%		130	93.3%	8.4%	1.8%	3.4%
76-100%		188	94.0%	8.0%		
Training Length						
Number of Sessions:	1	70	88.3%	10.8%	12.6%	3.3%
	2	214	95.2%	7.4%		
	3	57	91.7%	7.4%		
Training Hours	1	49	90.2%	11.6%	5.2%	
	2	81	90.7%	9.0%		
	3	154	96.0%	6.8%		
	4	57	91.7%	7.4%		
Physician Assistance						
Number of Sessions:	0	65	93.1%	10.2%	16.1%	2.5%
	1	90	89.5%	9.5%		
	2	155	97.0%	6.0%		
	3	31	87.3%	5.7%		
Overall		374	93.4%	8.7%		10.1%

Group 3 Factors: The Relationship Between Factors Which Could Be Applied During Or At The Completion Of The Measurement Procedure

The Relationship of Group 3 Factors to Overall Accuracy in Case Presentation

Table 7.10 provides the breakdown of mean accuracy scores for the 5 factors evaluated in Group 3. The proportion of variance explained by these factors in the bivariate analysis is found in the fourth column of the table. The proportion of variance explained when all factors are included in the regressions model is found in the fifth column of the table.

Overall, Group 3 factors explained 7.4% of the variance in accuracy score. Patient rating of the quality of their performance during the student evaluation was the only factor which was associated with accuracy score in the multiple regression model. There was an inverse relationship between patient rating and accuracy of case presentation. There are too few observations in the poor category to draw any conclusions. However, those who rated their performance as fair were more accurate then those who rated their performance as good. This observation is interesting, possibly suggesting that patients who are more critical of their performance do better. This factor is of little practical importance as a method of identifying encounters where the accuracy of presentation was sub-optimal.

The Relationship of Group 3 Factors to Accuracy in the Presentation of the History, Physical Findings and Affect

Patient rating of performance explained 8.4% of the variance in accuracy score for the patient history and 19% of the variance in score for patient affect. Higher ratings in both instances were associated with lower scores. There was no relationship between patient rating and the accuracy of presentation of physical findings.

The number of sessions performed by the patient earlier in the day had no relationship to accuracy in the presentation of the history but was negatively associated with the accuracy in the presentation of the physical findings and affect. Mean accuracy in physical findings was 88% for the first three sessions, 76% for the next three sessions and 73% for seven-ten sessions. This factor explained 2.1% of the variance in accuracy score for physical findings. Mean accuracy in the presentation of patient affect was lowest at the beginning of the test day; 87% for patients who were evaluated during the first, second or third session of the day. Affect score was highest for patients evaluated during their fourth to sixth session of the day, deteriorating slightly thereafter. Number of sessions explained 3.4% of the variance in the presentation of patient affect. This observed relationship may reflect the difficulty some patients may have in assuming the role at the beginning of the test day. Fatigue may explain the slight deterioration in performance towards the end of the day. The implications of these findings for the organization of the test procedure will be discussed in the conclusions of this chapter.

Time since training was not associated with accuracy in the presentation of physical findings or patient history. It did explain 3.7% of the variance in the presentation of patient affect. Patients were more accurate in their presentation of patient affect in the first week after training (93%) than in the second week (86%).

Student performance on data collection and interpersonal skills had no relationship to the accuracy of patient presentation. The standardized patient was responsible for rating both of these aspects of performance. This may have biased the estimation of this relationship.

TABLE 7.10 GROUP 3 PREDICTIVE FACTORS: FACTORS WHICH COULD BE APPLIED DURING OR AT THE COMPLETION OF THE MEASUREMENT PROCEDURE - PERCENT ACCURACY SCORE AND THE PROPORTION OF VARIANCE EXPLAINED BY EACH FACTOR

Predictive Factor	Percent Accuracy Score			Proportion of Variance Explained	
	N	Mean	S.D.	Bivariate	All Factors Included
Number of Sessions Done That Day					
1-3	119	94.12%	8.8%	.2%	00%
4-6	98	92.9%	8.2%		
7-10	156	93.1%	9.0%		
Time Since Training					
1 week	150	93.7%	8.3%	.7%	00%
2 weeks	223	93.2%	9.0%		
Patient Rating of Performance					
1-2 (poor)	2	100%	0%	5.0%	7.4%
3-4	211	94.3%	7.8%		
5 (good)	39	88.8%	11.9%		
Student Performance Interpersonal Skills					
41-60	16	93.8%	9.7%	.3%	00%
61-80	326	93.4%	8.7%		
81-100	16	92.2%	10.4%		
Student Performance Data Collection Skills					
41-60	54	95.4%	8.0%	2.1%	00%
61-80	276	92.8%	8.9%		
81-100	0	-	-		
Overall	374	93.4%	8.7%		7.4%

PATIENT ACCURACY AND ITS RELATIONSHIP TO COMPETENCY SCORE

Patient accuracy scores and student scores from the 1987 and 1988 University of Manitoba evaluation samples were used in the analysis of the relationship of presentation accuracy and student competency score. In the literature, the accuracy of patient presentation has been identified as a factor which may influence component competency scores. The component scores which have been identified are data collection, interpersonal skills (or patient communication), diagnosis and management (see Chapter 4). Overall competency score may be influenced by patient accuracy if these components of competence contribute in an appreciable way to the total pool of items evaluated. The breakdown of the components of competence measured with each of the cases used in 1987 and 1988 is provided in Table 7.4.1 and 7.4.2. Since the components of competence which contribute to overall competency score may vary among different evaluation procedures, the relationship between patient accuracy and competency score will be evaluated separately for each of the four component scores identified as well as for overall competence score.

Descriptive Statistics of the Relationship Between Patient Accuracy and Competency Score

Table 7.11 displays mean student competence scores for seven categories of patient accuracy score. Inspection of the mean scores associated with each category of patient accuracy does not suggest the presence of any obvious relationship between the accuracy of patient presentation and competency score. The overall score for students who saw patients with accuracy scores of less than 70% is 65%. Scores were of similar magnitude (67%) for students who saw patients who were accurate 100% of the time. The standard deviation of all scores except diagnosis are small; thus it is difficult with this data set to detect the presence of a possible relationship.

TABLE 7.11 THE RELATIONSHIP BETWEEN THE ACCURACY OF PATIENT PRESENTATION AND COMPETENCE SCORE: OVERALL AND COMPONENT COMPETENCE SCORES FOR CATEGORIES OF PATIENT ACCURACY FOR 1987 AND 1988

Accuracy Score of Patient Presentation	N	Student Competence Score				
		Overall	Data Coll.	Diag.	Manage	Interpers
< 70%	34	65.4%	74.2%	72.8%	71.0%	72.4%
71-75%	25	66.6%	64.6%	73.4%	69.7%	74.2%
76-80%	38	72.9%	73.8%	76.7%	57.1%	77.6%
81-85%	63	71.2%	75.2%	72.7%	58.4%	78.7%
86-90%	76	68.1%	69.9%	70.7%	51.8%	73.5%
91-95%	83	69.0%	71.4%	72.9%	53.9%	74.7%
96-100%	306	66.5%	65.45	71.2%	56.4%	72.2%
Overall Mean		68.2%	67.2%	68.8%	55.6%	72.8%
Standard Dev.		3.7%	6.1%	10.4%	7.7%	6.2%

Legend: Overall: student competency score for all cases (see Chapter 5 for calculation)

Data Coll: student competency score for data collection over all cases

Diag: student competency score for diagnosis over all cases

Manage: student competency score for management over all cases

Interpers: student competency score for patient communication and doctor-patient relationship over all cases

N: the number of patient-student encounters

The Evaluation of The Relationship Between Patient Accuracy and Student Competency Score

It was hypothesized that there may be two possible forms of the relationship between patient accuracy and competency score: a positive linear relationship and a curvilinear relationship. In the latter, it was hypothesized that patient accuracy would not influence student score unless accuracy fell below some minimum threshold level. In order to evaluate these two potential forms of the relationship, the linear relationship and residuals were plotted. Inspection of the plotted relationship did not suggest the presence of a threshold level of effect. Inspection of the residuals suggested that the assumption of

homogeneity of error variance was not met. Variance increased with increasing values of accuracy score. There was no evidence of a curvilinear relationship in the plotted residuals.

The linear relationship between accuracy score and student competency score was therefore examined. The results of repeated measures linear regression analysis are displayed in Table 7.12. Even after correction for multiple comparisons, a significant relationship was found between accuracy score and competence scores for interpersonal skills and management. The direction of each relationship examined is the reverse of that expected except for diagnosis where no relationship exists. With increasing values for accuracy score, competency score was diminished. In order to stabilize the variance of competency score in the estimation of this relationship, a log transformation of competency score was carried out. In stabilizing residual variance, the negative relationship between competence in data collection and accuracy score was significant ($p=.01$). The conclusions for the remaining relationships were unchanged.

The estimated influence of patient accuracy on student competence score is relatively small and may not be of practical importance. Student competence score decreases by 1/10th to 3/10th of a percentage point for every 1% increase in patient accuracy.

TABLE 7.12 THE LINEAR RELATIONSHIP BETWEEN ACCURACY OF PATIENT PRESENTATION AND COMPETENCY SCORE USING REPEATED MEASURES MULTIPLE REGRESSION ANALYSIS OVER ALL CASES ON THE 1987 AND 1988 PATIENT STUDENT COHORTS

Dependent Variable	Estimated Beta For Accuracy Score	99% C.I. For Beta	P Value [Ho:B=0]
Overall Competency	-0.06	-0.15, .03	.10
Data Collection	-0.15	-0.36, .06	.07
Diagnosis	-0.00	-0.14, .14	.95
Interpersonal Skills	-0.12	-0.19, -.05	.0001
Management	-0.30	-0.50, -.10	.0001

This paradoxical situation is difficult to explain for the observed relationship between patient accuracy and competency in management. Data

collection and interpersonal skills were rated by the standardized patient whereas competence in management was rated by medical faculty. The relationship between accuracy and scores in data collection and interpersonal skills could be explained if patients who were less accurate were also apt to be more lenient raters.

This hypothesis can be evaluated by use of the data gathered in Study 3. Standardized patients participating in the 1987 evaluation at the University of Manitoba were asked to re-rate videotaped encounters of a random sample of student presentations drawn from the joint evaluation conducted in both universities. The ratings of these encounters can be used to calculate data collection score and interpersonal skills. By examining patients who had accuracy scores at the two extremes of the range, those with presentation accuracy below 80% and those with accuracy scores equal to or above 95%, the hypothesis of systematic difference in encounter rating can be evaluated. Table 7.13 displays the results of this analysis. Patients whose accuracy of presentation was less than 80% did score the same students approximately 3% higher than patients whose accuracy was equal to or above 95%. This difference, when evaluated with an independent t test is not statistically significant ($p=.07$). The trend however is compatible with the stated hypothesis. For this reason, evaluation of the relationship between patient accuracy and competence in data collection and interpersonal skills should be re-evaluated using an independent rater for student performance.

TABLE 7.13 DIFFERENCES IN RATING OF STUDENT DATA COLLECTION SKILLS AND INTERPERSONAL SKILLS FOR PATIENTS WITH DIFFERENT LEVELS OF PRESENTATION ACCURACY IN THE 1987 EVALUATION COHORT AT THE UNIVERSITY OF MANITOBA

Accuracy Score	Number of Patients	Number of Encounters Rated	Mean Score for Data Collection and Interpersonal Skills (s.d.)
<80%	6	207	69.62% (20.8)
>94%	12	167	66.14% (16.6)

The Relationship Between Accuracy Score and Individual Case Scores

The accuracy of patient presentation may be more important in some cases than in others. Since the mix of cases will vary from one evaluation to the next, it is important to examine the presence of a potential relationship between patient accuracy and competence score for individual cases. In Appendix 7, Tables A7.4 to A7.8 provide a breakdown of the relationship of patient accuracy with the five competency scores evaluated for cases used in 1987 and 1988. Cases which were common to both years were combined if there was no significant difference in patient accuracy score. Each table provides the mean accuracy score for each patient presenting the case, the mean student score for students seen by each patient, the estimated beta derived from linear regression and the p-value associated with the test of whether Beta is equal to zero.

Overall Competency Score

After using Bonferroni's correction for multiple comparisons, there was no case in which a relationship was found between patient accuracy score and overall competence score (see Appendix 7, Table A7.4). Of the 29 cases evaluated, the relationship was, if anything, the reverse of that expected in 20 cases (i.e. competency score diminished with increasing accuracy score). This was true for 10 of the 14 cases used in 1988, the 2 cases which were combined for 1987 and 1988 and 8 of the 13 cases in 1987. The estimated beta was zero in 3 cases. In the remaining 6 cases, competency score increased with increasing values of patient accuracy (1987: Case #3, #11, #14 & #15; 1988: Case 13 & Case 16). In these 6 cases the estimated beta was in the range of b=.04 to 1.15. For the two 1988 cases, differences in mean student score were 0 and 2%. Mean accuracy score for patients presenting both of these cases was high (92%-98%) with differences between the 2 patients presenting the same case being 4% and 1% respectively.

In 1987, there was a greater range in both student score and patient accuracy; however, small sample sizes compromise the power of most comparisons. In the 4 cases in 1987 with a positive slope for the

relationship of patient accuracy and student score, differences of 2%-7% in patient accuracy and 1%-8% in mean student score were observed. The largest estimated beta was for Case 11 where mean accuracy for patients was 98% and 100% and the corresponding scores for students were 57% and 63%. In this case, errors were made on one item by the standardized patient. The patient denied she were otherwise healthy, providing more information on other problems besides the problem being presented. This may have resulted in students choosing an alternate approach to data collection, diagnosis and management (3/5 components measured with this problem). This possibility is supported by the estimated slope for diagnosis (beta=2.1) and management (beta=1.03). For the case with the largest difference in student score (Case #15, 8%), average patient accuracy was 95% and 99%. Errors were made on 3 of the 7 items by the patient with the lowest score, errors which would likely have a consequence for data collection and management. The estimated slope for data collection is b=.46. There is a 17% difference in student score for the two patients who presented this case. For management, the slope is b=.23 with a 5% difference in scores received by students who saw the two patients.

In the 20 cases where the reverse of the expected relationship was present, the estimated slope of the relationship was in the range of -.02 to -.95. The largest differences for both patient and student scores were for Case #7 in 1987 where mean patient accuracy differed by 11% (84% and 95%) and student score differed by 9% (71% and 80%). Unfortunately, the hypothesis of systematic differences between raters with different accuracy levels cannot be evaluated in this case.

Competence Scores in Data Collection

Table A7.5 in Appendix 7 displays the results of the evaluation of patient accuracy and scores in data collection with each case. Data collection was measured in 28 cases. Of these 28 cases, the relationship was significant in only one case after Bonferroni's correction for multiple comparisons (Case #7). In this case, the estimated beta was -.82, the reverse of expected. As indicated earlier, it is possible that patients with lower

accuracy scores were also more lenient raters which may explain this paradoxical relationship.

The relationship was in the direction expected in 8 of the 12 cases evaluated in 1987 (where the lowest accuracy scores were observed) and 2 of the 14 cases used in 1988. In the remaining 18 cases, there was no relationship found in 1 and the relationship was the reverse of expected in 17 (i.e. the slope was negative). For the 9 cases where a positive slope was found, accuracy was in the range of 69% to 100%. The largest difference in student score is for Case #15 (an 18% difference) associated with a 3% difference in accuracy score. This case was discussed earlier in relationship to overall score. The largest point estimate of the slope of the relationship is for Case #2 (beta=1.0). In this case competence in data collection could have increased by as much as 2.3% for every percent increase in accuracy or been diminished by 2/10 of a percentage point for every 1% increase in accuracy (the 95% confidence interval for the beta). The accuracy in presentation for this case was 90% for one patient and 99% for the other. The corresponding student scores were 77% and 88% respectively.

Competence Scores in Diagnosis

Table A7.6, in Appendix 7 displ 's the results of the analysis of the relationship between accuracy score and competence score in diagnosis. In the measurement of diagnosis, medical faculty were responsible for scoring the written write-up of the case. Standardized patient rating would not be expected to have an influence on scores in diagnosis. No significant relationship between diagnosis and accuracy score was found in any of the 18 cases in which diagnosis was evaluated. A positive slope was present in 14 of the 28 cases (6 cases in 1987 and 8 cases in 1988). For these cases, the largest point estimate of the slope was 2.1 % (1987, Case #11) with a 95% confidence interval of -2.1 to 6.3%. This case was discussed earlier in relationship to overall competency score. Smaller sample sizes limit the precision of the estimates in 1987.

For the remaining 14 cases, no relationship was found in 3 and a negative relationship was found in 11. The largest point estimate of a negative relationship was for Case #8 in 1987 (beta=-4.6). The 95% confidence interval for this estimate was -11.9 to 2.8. The only possible explanation for this paradox is that patients who were less accurate could have provided data to the student which could have been helpful in diagnosis. This kind of problem would not necessarily be detected by accuracy rating. Accuracy rating focused on the correctness of data provided by the patient on certain critical features. It did not assess all information which was provided by the patient, some of which may have been helpful in arriving at the correct diagnosis.

Competence Scores in Management

The relationship of competence scores in patient management and patient accuracy is displayed in Table A7.7 in Appendix 7. Competence in management was evaluated in 24 cases. After using Bonferroni's correction for multiple comparisons, no case showed a significant relationship between accuracy and competency score. For most cases, management was scored by medical faculty so patient rating should have no influence on the observed relationship.

In 11 of the 24 cases evaluated, the estimated slope of the relationship was positive (3 cases in 1987, 1 case used in both years and 7 cases in 1988). The largest point estimate of the beta is for Case # 11 (B=1.03) with a 95% confidence interval of -2.9% to 4.9%. There was a difference of 17% in student scores for the two patients presenting this case accompanied by a 2% difference in patient accuracy. The problems with this case were discussed previously. Insufficient sample sizes limited the power of most of these comparisons.

There was no relationship between patient accuracy and management in 3 cases and the point estimate of the slope was negative in the remaining 10.

Competence in Interpersonal Skills

Interpersonal skills were measured in 18 cases, 2 in 1987 and the remainder in 1988 and in two of the cases used in both years. Interpersonal skills were rated by the patient which, as discussed previously, may bias the estimation of the relationship. The results of the evaluation of the relationship between accuracy score and competence in interpersonal skills with each case are displayed in Table A7.8 in Appendix 7. None of the relationships evaluated were statistically significant. In 13 of these cases, the relationship was the reverse of expected with the point estimate of the slope being negative. In 4 cases the estimate of the slope was positive (all cases being presented in 1988). In 1 case the point estimate of the slope was 0.

For those cases where a positive relationship existed between the accuracy of patient presentation and competency score, the point estimates were small ranging from .01 to .32. The 95% confidence interval around these estimates was also small suggesting that a relationship does not exist.

For cases where a negative relationship was observed between accuracy and student score, Case #8 had the largest point estimate. The 95% confidence interval on this estimate was -4.2% to 2.1%. A relationship of larger magnitude in the same direction was noted between patient accuracy and diagnosis for Case #8. Since patients rated interpersonal skills, there is a possibility for the reasons discussed previously, that this may have biased the estimated relationship.

The implications of these results will be discussed in the final section of this chapter.

THE CONDITIONS OF RESPONSE: THE PERCENT OF ITEMS PROVIDED SPONTANEOUSLY AND IT'S EFFECT ON STUDENT SCORE VARIANCE

Descriptive Statistics of the Percent of Items Provided Spontaneously

Items which are to be provided spontaneously by the patient during the patient encounter should be specified in the training protocol for each case. Errors in the type of data provided spontaneously should be considered as problems in the accuracy of presentation. In 1987 and 1988, the protocol used to train standardized patients did not specify which data were to be provided spontaneously and which data were to be provided only in response to specific inquiry. For this reason, the accuracy of the patient's response in the provision of clinical data under the specified conditions cannot be evaluated. Differences in the percent of items provided spontaneously by patients presenting the same case can be evaluated. If it is assumed that the student groups seen by different patients are equivalent, then no difference should be observed in the percent of items provided spontaneously by 2 or more patients presenting the same case. Since students were randomly assigned to patients and encounters were randomly sampled, this assumption is likely valid.

In Table 7.14, the percent of items provided spontaneously by patients in 1988 is provided. The corresponding data for patients in 1987 are provided in Chapter 6, Table 6.16. Differences in the percent of data provided spontaneously by two patients presenting the same case were evaluated by an independent t test. When more than two patients presented the case, a one-way ANOVA was used. The probability of observing this big a difference between and among patients due to chance alone is provided in the right hand column of the table.

Overall, 21% of patient data were provided spontaneously in 1988 in contrast to 32% in 1987. This difference could be explained either by differences among the patients used in both years or differences in the cases for 1987 and 1988. The percent of data provided spontaneously in 1988 varied from 0% to 65% for different patients and cases. In 1987 it varied from 0% to 100% among different cases and patients. In both years

there was a difference in the percent of data provided spontaneously in different cases. This finding is to be expected.

For one of the cases in 1988, there was a significant difference in the percent of items provided spontaneously by different patients (after using Bonferroni's correction for multiple comparisons). This was in Case #2 where one patient provided 29% of the patient data spontaneously in contrast to 18% for the other patient. In the remaining cases, all observed differences were less than 10%. In 1987, there were significant differences among patients presenting the same case in 3 cases. The difference between patients in these cases was in the range of 9%-17%. The impact of differences in patient presentation of the same case on variance in student score will be reviewed in the next section.

TABLE 7.14 PERCENT OF ITEMS PROVIDED SPONTANEOUSLY BY PATIENT AND CASE IN 1988

Cases Presented By 2 Patients	Number of Encounters	Number of Items/Encounter	Percent Spontaneous Patient #1			Patient #2			P-Value
			N	Mean	S.D.	N	Mean	S.D.	
1	25	19	10	28.7	(9.2)	15	17.8	(4.1)	.004
2	21	30	16	32.3	(12.0)	5	26.8	(11.3)	.38
3	22	15	14	4.1	(6.2)	8	6.0	(5.8)	.47
4	20	26	9	14.9	(6.6)	11	13.5	(5.1)	.58
6	30	13	15	65.8	(13.4)	15	60.9	(16.8)	.34
8	23	19	9	11.0	(7.2)	14	12.8	(8.0)	.60
10	22	31	13	31.8	(14.6)	9	26.0	(13.8)	.10
12	28	26	12	0.0	(0)	10	0.0	(0)	-
13	24	22	11	25.3	(4.7)	13	21.6	(6.4)	.12
14	22	21	8	21.2	(10.0)	13	11.2	(9.6)	.03
16	26	24	14	7.8	(7.2)	12	2.3	(3.4)	.47
18	21	8	5	0.0	(0)	16	4.9	(8.9)	.24
19	19	30	9	27.6	(11.6)	10	20.5	(7.6)	.13
20	22	11	8	32.5	(13.4)	14	30.7	(18.7)	.81

Cases With > 2 Patients				Pt#1	Pt#2	Pt#3	Pt#4	Pt#5	Pt#6	P-Value
11	27	22	N	9	7	6	2	3		.03
			Mean	24.1	33.2	37.6	32.4	25.7		
			S.D.	5.2	9.6	8.2	4.2	11.0		
15	28	13	N	3	2	3	6	2	12	.07
			Mean	13.9	12.5	24.5	13.2	18.1	1.9	
			S.D.	17.3	17.7	13.2	17.6	9.8	4.3	
Total	373	330				20.77 (19.1)				.0001

Legend: P-value: A t-test was used to evaluate differences between cases with 2 patients and a one-way ANOVA to evaluate differences in cases with >2 patients. Differences among cases was evaluated using one-way ANOVA.

The Impact of Differences in the Percent of Items provided Spontaneously on Variance in Student Score

Standardized patients who present the case also act as raters of interpersonal skills and recorders of actions taken in data collection. An unbiased estimate of the effect of the percent of items provided spontaneously on student score in these two areas is therefore not possible. It is hypothesized that circumstances which lead the patient to provide more data spontaneously will also influence the manner in which they score the same encounter. If patients provide more data to students who are having difficulty and score those students in a more lenient manner, then score variance for patients providing more data spontaneously would be expected to be smaller. This hypothesis was tested for each of the cases presented in 1987 and 1988. Cases used in both 1987 and 1988 were combined if there was no difference in the accuracy of patient presentation

After correction for multiple comparisons, no significant differences in variance were found. The number of cases in which the results were consistent with the hypothesis are summarized in Table 7.15. For data collection 7/11 cases were consistent with the hypothesis in 1987 and 8/15

in 1988. For interpersonal skills, 2/2 were consistent with the hypothesis in 1987 and 8/15 in 1988. These data do not support the presence of a relationship between the percent of data provided spontaneously and variance in student score. An independent rater would need to be used to rate interpersonal skills and data collection skills to determine if the amount of clinical data provided by the patient influences competence score.

TABLE 7.15 A SUMMARY OF THE RELATIONSHIP BETWEEN THE PERCENT OF PATIENT DATA PROVIDED SPONTANEOUSLY AND VARIANCE IN STUDENT SCORE

Year	Cases Which Are Comparable With The Hypothesis		Cases Which Are Incompatible With The Hypothesis	
	Data Coll	Interpers	Data Coll	Interpers
1987	7/11	2/2	4/11	0/2
1988	8/15	8/15	7/15	7/15

Legend: Data Coll: for data collection score

Interpers: for interpersonal skills score

Notes:

1. The hypothesis: Patients who provide more data spontaneously may do so in response to students who are having difficulty. If they also rated these students in a more lenient fashion, variance in score would be less for patients who provided more data spontaneously.

2. Cases where the score was not calculated were excluded from the denominator. Cases where both patients provided no data spontaneously were also excluded from the denominator

3. In cases where there was more than two patients, the hypothesis was examined using the patient with the highest and lowest percent of data provided spontaneously.

CONCLUSIONS AND DISCUSSION

Predictors of Patient Accuracy

Three groups of predictive factors were evaluated: those which could be applied in patient and case selection, those which could be applied during or after standardized patient training and those which could be applied during the measurement procedure. Of these three groups of factors, those which could be applied in patient and case selection explained the largest proportion of variance in accuracy score (11.8%) followed by those which could be used during the training process (10.1%). Selected variables in all 3 groups were of importance in predicting accuracy of presentation on the patient history, physical findings, and affect.

Case Selection

Accuracy score does not seem to be adversely affected by the number of items the standardized patient is required to present with each case. This conclusion is limited to cases where no more than 31 clinical features are to be presented. The major limitation in this analysis is that cases are confounded with number of items and patients are nested within case. An unbiased estimate of the effect of number of items on accuracy score would require the same patients and cases to be studied under conditions which varied the number and type of items to be presented.

Lower accuracy scores would be expected in cases where physical findings and patient affect are part of the clinical features to be presented. This conclusion is limited by the small number of items sampled in these two categories. However, the same trend was noted in the evaluation of patient accuracy for both 1987 and 1988. Only 4 cases and 13 patients were common to both years. There are patient, training and procedural attributes which are associated with the accuracy of presentation of these two types of clinical features. Attention to these factors may improve accuracy scores for cases which would require the presentation of physical findings and affect.

Patient Selection

There is no defined protocol, reported in the literature, for the recruitment and selection of standardized patients. A standard protocol was not used in the recruitment and selection of patients in this study. The selection factors which may have been operative in the definition of the subset of individuals included in this study are therefore unknown. The comparison of patient accuracy in 1987 and 1988 suggests that the trainer was better in 1988 than 1987 at selecting patients who were more likely to be accurate. Although this suggests that the trainer may have applied different criteria in 1988 and 1987 in standardized patient selection, the nature of these criteria are unknown. This will be an important area for future study. It should be noted that the evaluation of individual factors associated with patient accuracy in this study is limited to patients who have already been screened for inclusion.

Three patient characteristics were associated with overall accuracy score: the patient's reported understanding of the health problem to be presented, previous simulation experience, and experience with the health problem or symptoms being presented. These three factors accounted for 9.4% of the variance in overall accuracy score. Younger age groups, those with previous acting or health problem experience and female patients reported a better understanding of the health problem they were presenting. Patients who reported that they had a good understanding of the problem had higher accuracy scores for the presentation of the history and patient affect but not for physical findings. Barrows(1987) has identified this factor as being one of the most important in successful standardized patient training. The patient's understanding of the problem being presented is highlighted and reinforced during the training procedure proposed by Barrows (see Chapter 4).

Patients who had the health problem being presented had better accuracy in the presentation of the history and patient affect. Knowing someone with the health problem being presented, on the other hand, was not associated with presentation accuracy. This finding is not that surprising. First

hand experience with the problem being presented probably facilitates an understanding of the problem and its consequences for patient affect. It may also help the patient recall important clinical features.

Barrows(1987) has indicated that previous simulation experience is helpful in reducing the amount of training time required for a standardized patient. This factor also acts as a selection variable. Those with experience represent a subset of patients who have been selected by themselves and the trainer for additional standardized patient roles. Previous experience as a standardized patient was associated with accuracy in presentation of the physical findings and patient affect. Previous simulation experience does not appear to have any added benefit for accuracy in the presentation of the history.

Acting or role-playing experience was associated with accuracy in the presentation of the physical signs but had no direct benefit for accuracy in the presentation of the patient history or affect. Those with previous experience are likely a self-selected group of individuals who may be more adept at simulating the presence of physical signs they normally do not possess.

It has been anecdotally noted that older patients are more difficult to train and that they may not be able to provide a consistently accurate presentation of the clinical problem. Older patients in this study (i.e. over 60 years) were clearly equivalent or superior to those in the younger age groups in the accuracy of their performance. The age of the patient had no relationship of practical importance to accuracy score in this study.

In summary, patients with a good understanding of the problem who have had previous simulation and health problem experience are more apt to provide an accurate presentation of the clinical situation. For cases that involve the presentation of physical signs, those with acting and/or role-playing experience in addition to previous simulation experience will be more apt to provide an accurate presentation of the physical signs.

Patient Training

At the completion of training, neither the patient nor the trainer were able to predict the accuracy with which the patient would subsequently present the case. In the presentation of the physical findings and patient affect, the trainer, in fact was more apt to score patients, who subsequently had poorer accuracy, higher. This problem may have been because there was no opportunity during the training process to actually view and critique the performance of the standardized patient trained in a 'dry-run' session with an independent clinician examiner.

The number of sessions used to train the standardized patient and the number of sessions attended by a physician resource are the two training attributes which showed the strongest association with accuracy score. Two training sessions and three hours of training appear to be the minimal number required for optimal accuracy. In this study, patients who received more training hours or more sessions were less accurate. The modal number of sessions and hours was two and three respectively. Those receiving more sessions probably represent patients who were having difficulty in presentation or were presenting more difficult cases. This would suggest that cases or patients requiring more than the usual two sessions are not apt to be accurate in their presentation of the problem and might be best eliminated.

Physician assistance at the training session provides no apparent benefit in training the patient to accurately present the history. Physician assistance is important in improving the accuracy of presentation of physical findings and patient affect.

The Measurement Procedure

Although it has been hypothesized that student performance may reduce the accuracy of patient presentation, there was no evidence that this was so in this study. The limitation in the evaluation of this relationship was that the same patients were used to rate both of these two aspects of

student performance. An unbiased estimate of this relationship would require these two aspects of performance to be rated by an independent evaluator.

Patients were unable to identify those encounters in which they had inaccurately presented the case. In fact, those who rate their performance as good were the least accurate. Patients ratings of the quality of their performance could not be used to identify encounters in which accuracy of presentation was less than adequate.

In this study, the number of sessions done in a day by patients was 10. Accuracy in the presentation of the history was not adversely affected by this requirement. Accuracy in the presentation of physical findings and affect were sensitive to the number of student encounters the patient had already presented earlier that day. Accuracy of physical finding presentation deteriorated linearly from 88% during the first three encounters to 73% for the last three encounters. Patient affect was most accurately presented by the 4-6th session in the day, the lowest value being at the beginning of the day. The patient may initially find it difficult to assume the patient role. Barrows(1987) has identified this as a potential problem and has advised the use of a 'warm-up' session by the trainer prior to the standardized patient's use in the evaluation.

Patients were evaluated one and two weeks after training. Within this time span, weeks since training was not associated with accuracy in the presentation of the history or physical findings. Accuracy in the presentation of the patient affect was lower in the second week after training. This would suggest that patients required to present a specific affect may benefit by a weekly review of this component of their presentation.

THE ACCURACY OF PATIENT PRESENTATION AND ITS RELATIONSHIP TO COMPETENCY SCORE

The relationship of the accuracy of patient presentation and student competence score was evaluated in 18-29 cases and 632 student-patient

encounters. Five competence scores were evaluated: overall score and scores in data collection, diagnosis, management and interpersonal skills. Inspection of the residuals suggested that the assumption of linearity was not violated however residual variance increased with increasing values of accuracy score. After using a log transformation of student competency score to stabilize residual variance, a significant negative association was found between patient accuracy score and three competence scores: data collection, management and interpersonal skills.

One possible explanation of this paradoxical phenomenon is that patients who were less accurate were also more lenient raters. This explanation would apply to the two scores where standardized patients were used as raters: data collection and interpersonal skills. This hypothesized explanation was evaluated by examining the scores provided by patients with accuracy scores less than 80% and those with accuracy scores greater than or equal to 95% using their ratings of videotaped encounters of the same students. Although the trend in the resulting mean scores supported this hypothesis, the difference between patients with different accuracy levels was not significant. When evaluated on a case by case basis, a negative relationship between patient accuracy and competence score was noted more frequently for scores where standardized patients were used as raters. A negative relationship was found in 64% (18/28) of cases in data collection and 72% (13/18) of cases for interpersonal skills in contrast to 42% (10/24) of cases for management and 39% (11/28) of cases for diagnosis. This trend is compatible with the hypothesis that less accurate patients are more apt to be lenient raters. In order to evaluate this possible problem, it is recommended that an independent rater be used to record student performance during the patient encounter and that accuracy and its relationship to student score be re-evaluated under these conditions.

The second possible explanation is that patients who were less accurate in the presentation of the critical features of the case were also more apt to provide the students with information which would improve their resulting competence score. This explanation would apply to scores in diagnosis, management and data collection but would not likely be relevant

to scores on interpersonal skills. This possible explanation of the results cannot be evaluated with this data set. Errors in the presentation of the specified clinical features of the case were the only aspects of patient presentation measured. Although standardized patients are not usually sophisticated about the management of their case, they are usually familiar with the diagnosis of the problem they are presenting and the type of actions which they can expect during data collection. A patient who has been ineffectively trained could provide the student with some of this data during the encounter which in turn could result in improved scores. A different approach to the evaluation of patient accuracy would have to be used to evaluate this potential problem. Anecdotally, this kind of problem has occurred when both real and standardized patients are used for evaluation purposes.

In order to determine if patient accuracy may be more critical in certain cases, the relationship between accuracy and the 5 competence scores was evaluated on a case by case basis. The limitation in this analysis is that there is inadequate power in most instances to detect relationships which may be of practical importance. In addition, for the two scores where patients are used as raters (data collection and interpersonal skills), rater reliability has been noted in Study 3 to be poor. Random error, attributable to raters, may attenuate the estimated effect of patient accuracy on competence score.

After correction for multiple comparisons, a significant relationship between patient accuracy and competence score was found in only one case (Case #7 for data collection). The point estimate of the slope of this relationship was negative (-.82). Point estimates of the slope of the relationship for individual cases were as large as -4.6% for cases where a negative slope was found and 2.1% for cases where a positive slope was found.

A minimum score for acceptable patient accuracy was not evident in this study. For example, in Case #11, mean accuracy scores for the two patients presenting this case were 98% and 100%. Errors made in the presentation of one item in the case appear to have had consequences for both diagnosis

and case management. This suggests that the method of calculating patient accuracy in this study may not be sensitive to the relative importance of different items for specific competence scores. A method which weights each item by its relative importance to data collection, diagnosis and management may be a more appropriate method of calculating presentation accuracy.

The case-specific analysis suggests that some cases may be more sensitive than others to the accuracy of patient presentation. This sensitivity likely relates to the aspects of competence being measured (eg. diagnosis, data collection) and the relative importance of data provided by the patient to decisions about clinical actions in these areas. The sensitivity of overall competency score to patient accuracy will obviously depend on the components of competence measured. In future applications of this technique, special attention should be directed to the accuracy of patient presentation for clinical data which is important for diagnostic, and management decisions.

The proportionate contribution of patient accuracy to competence score was not estimated in this study. Evaluation of the two potential sources of bias which could have contributed to a negative relationship would be recommended prior to this analysis. The association of patient accuracy with competence score should be re-evaluated using an independent rater for data collection and interpersonal skills score. The method of measuring accuracy should take into consideration additional information beyond that in the training protocol which is provided by the patient.

THE PERCENT OF ITEMS PROVIDED SPONTANEOUSLY BY THE PATIENT AND IT'S RELATIONSHIP TO STUDENT SCORE VARIANCE

In future applications of the standardized patient technique, the items to be provided spontaneously during the encounter by the patient should be identified in the training protocol. Failure to comply with the stated protocol should be considered as a problem in the accuracy of patient presentation. The conditions of response were not specified in the 1987 and 1988 training protocols for standardized patients used in this

study. Differences in the percent of items provided spontaneously by patients presenting the same case were therefore examined. It was assumed that random sampling of encounters and random allocation of students to patients would result in equivalent groups of students for each patient. Differences in the percent of items provided spontaneously between patients presenting the same case would therefore be attributable to patients. Patients differed in the percent of items they provided spontaneously for 3 of the 15 cases presented in 1987 and 1 of the 16 cases presented in 1988. An unbiased estimate of the effect of these differences on competency score was not possible with this data set. An independent rater of data collection and interpersonal skills would be required for an unbiased estimate of this relationship.

It was hypothesized that patients who provided more data spontaneously would have smaller score variances for their respective students. There was no significant difference in student score variance for patients who provided different amounts of clinical data. In this data set, it does not appear that patients who provided more data spontaneously would also have scored the student more leniently and thereby reduced score variance.

ABSTRACT

CHAPTER 8

THE RELIABILITY OF STANDARDIZED PATIENTS AS RECORDERS/RATERS

In the evaluation of clinical competence, standardized patients are used to present the clinical problem and rate and record actions taken by the clinician. Ratings by standardized patients have been compared to ratings by faculty and research personnel. Correlations from .52 to .93 are reported. When two or more standardized patients have been trained to rate the same case, systematic differences in the scores assigned have been noted. No study to date has compared standardized patients who have been trained by different trainers in different settings.

This study examined the reliability of standardized patients who had been trained to present and rate the same case. Three comparisons were made: the agreement between patients from two universities, the agreement between patients from the same university and the agreement between two ratings carried out by the same patient on two occasions. These three comparisons provide data on the extent to which rater reliability is influenced by different trainers, different patients and/or inconsistencies within the same patient. Secondly, it provided a means of examining whether systematic differences in rating existed in different universities. Attributes of the rating form which may influence agreement were also examined in order to provide guidelines for the construction of rating forms. Finally, the data were used to estimate the extent to which patient raters, who are typically nested within cases, contributed to measurement error in the evaluation of clinical competence.

A cross-sectional stratified survey design was used to sample patient studer encounters in the two universities. Strata were defined by university (n=2) and case (n=16) and an equivalent sample was drawn from each of the 32 strata. A total of 456 videotape encounters were of sufficient technical quality for use in the study. Five to twenty-nine items were rated for each case for a total of 252 items. Fifteen standardized patients from Southern Illinois and twenty-nine patients from Manitoba participated as raters. Standardized patients rated the case they

had presented 3 months previously in the joint evaluation conducted between the two universities.

Three indices of agreement were used to summarize rater reliability: observed agreement and kappa for individual items and the intra-class correlation coefficient for total score. The average observed agreement between raters was 81%. The average kappa was .45. Kappa values reflected slight to poor agreement for 36% of items, fair to good agreement for 23% and excellent to perfect agreement for 40%. Raters contributed significantly to competence score in this study, accounting for 22% of score variance.

The average agreement for items was significantly better for the within rater comparison (kappa=.52) than for the two between rater comparisons (kappa=.40 and .40). This was not true for the overall scores. The intra-class correlation coefficient for the within rater comparison was .37 in contrast to .41 and .42 for the two between rater comparisons. Agreement between raters trained in different universities was as good as that between raters trained in the same university.

The content being rated was the only rating form factor which was significantly associated with observed agreement. Item ambiguity, number of items rated and judgement level were not associated with observed agreement.

Systematic differences existed between raters who were trained in different universities. Raters at Southern Illinois scored on average 7% lower than raters from Manitoba. This difference in score was associated with a trend for Southern Illinois raters to fail more students than raters from Manitoba.

Patients who were more reliable in the repeated ratings of the same encounter were also patients who were more apt to provide an accurate presentation of their case. There was a significant linear association between the intra-class correlation coefficient calculated for each standardized patient rater and his/her mean accuracy score in problem presentation (p=.0003).

CHAPTER 8

STUDY 3: THE RELIABILITY OF STANDARDIZED PATIENTS AS RECORDERS/RATERS

THE RESEARCH PROBLEM

Standardized patients have been used to record or rate actions and behaviors carried out by the clinician during the patient encounter. These include the rating of communication and doctor-patient relationship skills and the recording of actions taken on history and physical examination.

In the rating of communication skills the standardized patient acts as a proxy for the rating which would be carried out by a real patient. The validity of the standardized patient rating using the real patient as a 'gold standard' has not been reported. In the recording of actions taken by the clinician on history and physical examination, the standardized patient serves as a proxy for a faculty or 'expert' rater. The relationship between standardized patient rating with that of faculty and other raters who are directly observing the encounter has been evaluated. The results of these studies have been summarized in Chapter 4, Table 4.3. Measures of association and agreement have been variable. Correlations range from .52 to .93 and observed agreement from 85%-100%. Poorer agreement tends to be present in the rating of communication skills when faculty are used as the gold standard. This finding is also true when ratings of different faculty are compared suggesting that it is likely the content of this area of measurement which is problematic rather than the type of rater being used.

Newble (1980) noted in his study of faculty raters that the observed agreement for faculty rating the same encounter was poor. Dawson-Saunders (1987) and Schnabl (1989) provide the only data for agreement among standardized patient raters. Both authors noted that there were systematic differences in the scores of students rated by two standardized patients presenting and rating the same case. Dawson-Saunders (1987) concluded that systematic differences likely existed between the two patients presenting the same case either in the content of their presentation or in their rating of students. Poor agreement or systematic differences among

standardized patient raters has a number of implications for the measurement of clinical competence. These will be reviewed subsequently.

POTENTIAL CONTRIBUTION OF RATERS TO MEASUREMENT ERROR

Potential Contribution to Variance Across Cases

Estimates of clinical competence are usually based on the clinician's performance with a number of patient problems which are sampled from an identified clinical domain. As was noted earlier, the number of cases required to achieve a stable estimate of clinical competence varies from 30-50 for diagnosis and management and 10-20 for communication and data collection. Variation in the clinician's true ability across cases has been imputed as the likely cause of this observed variation. If this is the case, reduction in the sample size of cases required could only be achieved by narrowing the content of the domain from which cases are drawn or measuring only those components of competence where fewer cases are required to achieve a stable estimate (eg. communication & data collection).

An alternate or supplementary explanation of the observed variation in performance across cases is the contribution of sources of measurement error which are confounded with case. Included in this group are the content of standardized patient presentation and standardized patient raters. The content of standardized patient presentation was the subject of study in Chapter 6 and 7.

Standardized patient raters may theoretically contribute to variation in scores across cases by inflating the main effect of cases within subjects or contributing to the residual error term. Measurement error attributable to raters could result from any of the following sources:

1) Poor intra-rater reliability in rating the same case: poor agreement between a standardized patient's repeated ratings of the same clinical encounter

2) Poor inter-rater reliability in rating the same case: poor agreement

between two standardized patients rating the same clinical encounter either as a function of:

a) differences between individuals who have simultaneously undergone the same training
b) or differences in individuals who have undergone training for the same rating form but by different trainers in different evaluation sites

3) Poor inter-rater reliability between raters presenting and rating different cases: there is poor agreement between standardized patients rating the same student on two or more cases. This source of error is less amenable to empirical investigation since standardized patients are used to rate only the one case they present and the rating forms used for cases are case-specific. As a result, agreement in rating can not be assessed under the usual evaluation conditions. Differences in rating of the same components of competence can be assessed across cases but, if present, may be due to other factors such as differences in the rating form or student ability with different cases. A potential problem in this area would be expected, however, if poor agreement is demonstrated between two patients presenting the same case. It could be hypothesized that agreement would be no better for two standardized patients presenting different cases.

Swanson and Norcini (in press) estimated the relative contribution of raters relative to subjects and cases in competency score and found a small proportion of variance attributable to the independent effect of raters. In this study, it was assumed that the student groups being rated by the two patients presenting the same case were equivalent. Although no study has attempted to replicate these findings, the results of this study would suggest that raters, despite systematic differences in score, contribute a negligible amount to measurement error in overall competence scores for long tests.

Potential Contribution to Bias in the Estimation of Competence

Biased estimates of clinical competence are a potential problem when there

are systematic differences among raters. The problem of bias in estimation varies in importance depending on the type of scoring method being used, the purpose of the evaluation and the number of sites in which measurement is being conducted.

Potential bias created by systematic differences in the competence scores tabulated for different raters is not a concern when:

1) one evaluation site is being used and all individuals measured are equivalently affected by biases in estimation

2) an overall estimate of competence for a sample of cases is being calculated and 'lenient' and 'tough' raters are balanced across examinees (either deliberately or through random assignment to patients within cases to examinees)

Bias created by systematic differences in raters is a concern when:

1) pass/fail decisions, or other methods of classification, are made on the basis of the performance of certain actions or the achievement of a pre-specified score value with a specific case

2) raters are systematically altering their rating of different subsets of students/clinicians (eg. rating is systematically lower or higher for certain age groups, for a specific gender)

3) raters are confounded with the variable being studied and systematic differences in the scores tabulated for those raters are present

4) a multi-site evaluation is being conducted with raters nested in site. Systematic differences in the rating among sites would influence the proportion passing/failing with both normative and criterion based scoring methods

The majority of studies reported do not provide data on the presence/absence of systematic differences between raters presenting the

same case. A Pearson product moment correlation coefficient is often used as a summary measure of rater reliability. This can be misleading. Large correlations can be obtained in the presence of substantial differences in ratings.

No reported studies have examined the reliability of standardized patients trained in different sites. Rater related bias in the estimation of clinical competence requires further exploration if multiple sites are being considered for evaluation or if criterion-based scoring methods are being used to categorize individuals into various levels of competence.

The proposed study has been designed to estimate three of the four potential sources of measurement error attributable to raters in the recording and rating of behaviour in the patient-clinician encounter: intra-rater reliability and the two conditions of inter-rater reliability for standardized patients presenting the same case (i.e. trained together, trained in different universities). Equivalency of student groups rated and rating conditions will be assured by videotaping and using the same patient-student encounters for all raters. Systematic differences in competency score and in the proportion of students passing and failing within and between raters will be evaluated to identify potential bias. This will provide the first estimate of bias for standardized patient raters who are confounded with evaluation site.

Finally, factors which may influence the reliability of raters will be explored to identify areas where selection, training or measurement conditions could be improved.

RESEARCH QUESTIONS

1. What is the reliability for standardized patients rating/recording behaviour during the clinical encounter?

Within Rater Estimates

a) What is the test-retest reliability for the same standardized patient rating the same encounter on two separate occasions?

Between Rater Estimates

b) What is the inter-rater reliability for two standardized patients who were trained together for rating the same clinical encounter?

c) What is the inter-rater reliability for two standardized patients trained in two universities for rating the same clinical encounter?

2. Are there systematic differences in the:

a) component competency scores derived from standardized patient rating?

b) the proportion of students passing and failing as a result of standardized patient rating, for raters trained in different university sites?

3. Are any of the following factors associated with the observed agreement of standardized patient raters?

a) Rater Pair Type

b) Rating Form Factors

*number of items rated/recorded for the case

*type of items rated/recorded(history, physical exam, communication)

*the level of judgement required to rate/record the item

*the ambiguity of the item rated/recorded

RESEARCH DESIGN

A cross-sectional stratified survey design was used to sample standardized patient-student encounters from the two university settings. Stratum was the term used to define the 2 levels of university and 16 levels of case in the study population. The same number of encounters was to be sampled from each stratum to improve the efficiency in the estimation of inter-rater differences. Sampled encounters were recorded on videotape and

the same sample was used for all three estimates of rater reliability. Patient raters were 'nested' within case in all three estimates of reliability.

Inter-Rater Estimates—Raters Trained in Different Universities

For each of the 16 cases presented, two raters from University of Manitoba and one rater from Southern Illinois University rated the encounters sampled for their respective cases. This provided two pairs of inter-rater comparisons for each case; rater #1 from Manitoba and rater #1 from Southern illinois and rater #2 from Manitoba with rater #1 from Southern Illinois.

Inter-Rater Estimates—Raters Trained in the Same University

In 15 of the 16 cases at the University of Manitoba, two standardized patients presented and rated each case in the clinical evaluation. Inter-rater estimates of reliability for patients trained in the same university were produced by pairing the ratings of the two standardized patients who rated each case. One pair of inter-rater comparisons was generated for each of the 15 cases.

Intra-Rater Estimates

A test-retest design was used to estimate intra-rater reliability. Within rater estimates of reliability were made using Manitoba raters and Manitoba generated clinical encounters only. The rating provided by the patient during the conduct of the clinical examination was used as the first rating, the second rating was that carried out after a review of the videotape of the same encounter. Two within rater comparisons for each case were produced, one for each of the standardized patients used in the original examination.

METHOD

SAMPLING PROCEDURE & RESULTS

The Student-Patient Encounters to Be Rated

A description of the sample size estimates and study population is

provided in Chapter 5. A total of 537 videotaped patient-student encounters were sampled from the 1987 clinical evaluation according to the sampling procedure outlined in Chapter 5. Fifteen percent of the tapes were discarded as being technically inadequate resulting in a study population of 456 encounters, 16-41 available for the analysis with each individual clinical problem (see Table 8.1).

TABLE 8.1 SAMPLE SIZES BY ITEM, CASE AND RATER PAIR TYPE

CASE	ITEMS	RATER PAIR TYPES				
	Number Items Rated/Encounter	Between University No. Encounters Rated		Within University No. Encounters Rated	Within Rater No. Encounters Rated	
		Pr#1	Pr#2	Pr#1	Pr#1	Pr#2
1	12	35	35	35	5	9
2	24	31	31	31	10	7
3	9	30	-	-	12	-
4	20	31	31	31	3	16
5	8	26	-	-	2	-
6	10	23	-	-	15	-
7	20	33	-	-	5	-
8	8	25	25	25	5	10
9(s)	12	-	-	27	10	5
9(m)	9	-	-	27	10	5
10	29	16	-	-	11	-
11	27	32	32	32	8	9
12	5	28	28	28	4	12
13	11	36	36	36	13	3
14	21	30	30	30	5	13
15	19	42	42	42	9	11
16	8	27	27	27	15	2

Legend: Pr#: refers to the rater pair of which a maximum of two were present for the between university and within rater comparisons and one for the within university comparison

Notes: For Case 9, two standardized patients were used (i.e. mother, son), each with their own rating form

For the within rater estimates, the actions recorded by the standardized patient during the 1987 clinical evaluation at the University of Manitoba for each of the encounters sampled were used as the first recording. These results were retrieved from rating form data entered and verified after

the completion of the 1987 evaluation. The second recording of actions for the same encounter was derived from the results of videotape observation. The number of encounters which were used in the estimate of within rater reliability for each case are displayed in Table 8.1.

The Standardized Patient Raters

The standardized patients who presented and rated the 16 clinical problems at each respective university in the original evaluation procedure were asked to participate in the rater reliability study. At Manitoba, the 33 patients who presented the 16 cases (two patients were trained for all cases except case #6 and #9; one patient was trained for case #6 and 4 patients for case #9) were asked to participate. At Southern Illinois, the one patient who presented most of the evaluation encounters for each of the 16 cases was asked to participate (total=17).

The two patients from Southern Illinois who presented Case #9 were unable to complete the videotape review. As a result, 15 standardized patient raters from S.I.U.,one for each of the 15/16 cases used in the evaluation were used in the inter-rater estimates of reliability for patients trained in different universities. Four of the 33 patients presenting the 16 clinical problems at the University of Manitoba were unable to participate in the study. Two had died shortly after the original evaluation, one had moved out of province and one could not find the time to participate. Inter-rater estimates of reliability for patients trained in the same university could therefore be made on 11 of the 16 problems presented. The number of rater pairs contributing to the estimate of between rater reliability for each case and training condition are displayed in Table 8.1.

The Number of Items Used to Rate/Record Actions with Each Case

The number of items rated/recorded by the standardized patient with each case were prospectively defined by the faculty member responsible for case development. The number of items contributing to overall case estimates of observer agreement and systematic score differences are displayed in Table 8.1. These range from 5 to 29. The majority of items required the standardized patient to record actions taken by the student as a proxy for

a faculty observer. Action recording was done on a dichotomous scale of present or absent. Rating of communication ability was required in 6 cases and was usually done on a 7 category Likert scale. In these situations the standardized patient was acting as a proxy for the real patient. In this study, communication rating was completed on only two cases, Case 8 and 16. In the remaining 4 cases, action recording and communication rating was required. Only the action recording was completed by patient raters participating in this study.

ENCOUNTER RATING/RECORDING

Rating Instrument

Encounter rating forms were developed for each case by the faculty responsible for developing the evaluation in the two universities. The development process is described in Chapter 5. Individual rating forms were prospectively developed for each case. The content of the rating form was case-specific and was dependent on the objectives to be measured with each case. The only exception was in the rating of communication and doctor-patient relationship where one common form was used across all cases. Most rating forms required the presence or absence of an action to be recorded on a dichotomous scale. In some instances, certain conditions (eg. examination technique) had to be satisfied before an action was recorded as being present. Competence in communication was rated in all but one instance on a 7 category Likert scale. Rating forms used for each case are found in Appendix 6. The same rating forms which were used in the 1987 evaluation were used to rate the sample of encounters videotaped for the rater reliability studies. The only difference in rating forms was the inclusion of an additional category for each action. It was used when the rater was unable to record/rate the action in question due to technical limitations of the videotape.

Rating Procedure

Standardized patients had been oriented to the rating forms which they were to complete approximately one to two weeks before the actual examination. The same patients were paid to participate in the subsequent rating of videotaped encounters which were sampled from the actual

examination in both universities. On average, the rating of the sample of videotaped encounters occurred 3-4 months after the actual examination. This time interval should minimize the risk of recall bias for the within rater estimates of reliability.

Two identical tapes were produced for each case (one for each university). Each case tape included the encounters sampled from both universities. The order in which encounters was presented on each tape was randomly determined. Randomly ordered encounters were numbered consecutively.

The procedure for rating the sampled videotapes was standardized across all cases rated and both university settings. The protocol for rating the videotapes was provided to the standardized patient trainer in each university who used it to provide instructions to the standardized patient raters. A copy was given to each standardized patient rater.

The protocol requested the standardized patient to review each encounter as consecutively presented on the tape produced for each case. The standardized patient was instructed to watch the encounter. At the completion of the encounter, a prompt on the videotape asked them to record/rate the actions taken by the student using the rating form developed for that case. They were then asked to review the next encounter and complete the same procedure.

This procedure for rating the encounter mimics the procedure used during the actual evaluation. The only differences between the actual evaluation and the videotaped rating is that the standardized patient did not have to present the case as well as remember the actions taken by the student. The introduction of this artificiality would likely act to facilitate more accurate recall of the actions taken by the student. Since the rating conditions were the same for all standardized patient raters, this aspect of the study procedure would not bias the estimate of between rater estimates of reliability. It may bias the within rater estimates of reliability since the first rating was done in the evaluation setting where the patient would have been presenting the case as well as recording

actions taken at the completion of the encounter.

Limitations in The Use of Videotaped Encounters

The advantage of using videotaped encounters is that it allows both the student and the rating conditions to be standardized. The resulting estimates of rater reliability could therefore be more confidently attributed to differences between raters rather than students or rating conditions. This advantage holds for the between rater estimates of reliability. For within rater estimates of reliability, the student encounter is standardized but the rating conditions are different. The resulting estimates of intra-rater reliability may therefore partly be a function of differences in the rating conditions rather than the reliability of the rater per se.

The disadvantage of videotape is that some actions may be difficult to evaluate from a videotape recording (eg. liver palpation). Secondly, rating at the completion of videotape review does not mimic the exact conditions for rating in the evaluation setting because the patient is not required to present the case. Unfortunately, there is no means of standardizing the conditions and subject of measurement under the usual live evaluation conditions since only one patient can present the case. It is anticipated that the use videotape rating may bias the estimate of rater reliability in a positive direction (i.e. improve reliability) since recall is not hampered by the simultaneous requirement of case presentation.

Data Produced by the Videotape Rating Procedure

Ratings for each encounter were produced by the standardized patient who presented each case at Southern Illinois and the two standardized patients who presented each case at the University of Manitoba. As a result, three data sets are produced, one for each of the rater reliability estimates. An example of the construction of each of the three comparisons is provided in Table 8.2.

TABLE 8.2 AN EXAMPLE OF THE CONSTRUCTION OF THE THREE RATER COMPARISONS USING CASE #1

	CASE #1									
	Actual Evaluation Rating						Videotape Rating			
	Manitoba (A)		Manitoba (B)		S.I.U. (C)		Manitoba (D)		Manitoba (E)	
	Rater 1 n=5		Rater 2 n=9		Rater 3 n=35		Rater 1 n=35		Rater 2 n=35	
Action Taken:	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Item #1	5	0	9	0	14	14	13	15	19	11
Item #2	0	5	9	0	11	22	12	21	6	28
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
Item #12	5	0	9	0	9	26	34	0	34	1
Total Score	90%		81%		44%		81%		76%	

Notes: Inter-Rater Comparison--Different Universities: The agreement between Rater 3(C) and Rater 1(D) as well as Rater 3(C) and Rater 1(E) was calculated for the 35 encounters; one score for each item and one total score.

Inter-Rater Comparison--Same University: The agreement between Rater 1(D) and Rater 2(E) was calculated for the 35 encounters for individual items and total score.

Intra-Rater Comparison: The agreement for the 1st rating by Rater 1(A) and the 2nd rating for Rater 1(D) was calculated for the five students rated in the actual evaluation for each item and overall encounter score. The same procedure was used to calculated agreement for Rater 2(B) and Rater 2(E) for the nine students rated in the actual evaluation.

1) Within Rater Estimates of Reliability:

Within rater estimates of reliability are calculated by using the rating provided by each Manitoba standardized patient during the actual evaluation and the second rating produced after videotape review. A pair of ratings is therefore produced for each encounter. A total of 244 encounters were evaluated on two occasions. Twenty-nine patients from Manitoba produced paired data which could be used in the estimate. A total of 252 rating items contributed to the overall estimate with a range of 5-29 per case (see Table 8.1).

2) Between Rater Estimates of Reliability for Raters Trained Together:

Between rater estimates of reliability for common training conditions are estimated by pairing the ratings produced for each encounter by the two standardized patients presenting the same case at the University of Manitoba. There were two raters from Manitoba for 11 of the 16 cases. In case #9, 4 raters participated resulting in two pairs of ratings, one for the two standardized patients who played the mother and one for the two standardized patients who played the son. A total of 12 pairs of ratings were used in estimating the reliability of patients trained together. The number of encounters used in estimating reliability was 371 for the 12 rater pairs. The number of items was 176 with 5-29 per case (see Table 8.1).

3) Between Rater Estimates of Reliability for Raters Trained in Different Universities:

Between rater estimates for different training conditions are estimated by pairing the ratings produced for each encounter by the patient from Southern Illinois University with the patient from Manitoba for each case. Since with most cases, two patients from Manitoba rated each encounter, two comparisons of inter-university differences in rating are possible with 12 of the 16 cases. A total of 445 student-patient encounters were rated by the 39 standardized patients from the two universities; 15 patients from Southern Illinois and 24 patients from Manitoba. The number of items contributing to the estimate overall cases was 231 with 5-29 for each case (see Table 8.1).

4) Item Ratings:

A total of 252 items were rated across the 16 cases. For each item, three to five estimates of observed agreement were obtained, one to two for each type of rater pair. Each estimate of observed agreement was based on the ratings of 5-40 patient encounters, the number of encounters depending on the rater pair type and case (see Table 8.3 for the breakdown of sample size by case and rater pair type). The total number of observations of observed agreement for all items was 990.

GENERATION OF STUDENT SCORE AND PASS/FAIL CLASSIFICATION

The rating form completed by each rater for each encounter was used to calculate a student encounter score. The case blueprint described in Chapter 5 specified the manner in which scores from the patient rating form were to be calculated. In cases where the patient was recording actions taken on history, physical exam, management or teaching, the percent of actions taken to those listed was calculated by simple addition. With the exception of one case, each action listed had a weight of one. In cases where the patient was rating communication ability on a Likert scale, scale points were converted to a numerical continuous scale and the numerical rating achieved on each scale was added to produce an overall score.

The case blueprint in 10 of 16 cases defined the minimum number of actions which must be taken by the student to pass the case. This minimum passing level was applied to the scores produced by each patient for all encounters rated. The proportion of students passing and failing was then calculated for each rater.

In the remaining 6 cases, no pass/fail criterion was specified for the patient-rated component of the case. An overall passing score of 60% was arbitrarily specified in our university. This was applied to all encounters rated by raters from the two universities in order to examine trends in the proportion of students passing and failing as a function of evaluation site.

PREDICTOR VARIABLE DEFINITION AND MEASUREMENT

Rater Pair Type: was defined in one of three mutually exclusive categories:

1) Between University Pair Type: consisted of two standardized patients who presented the same case but were trained and used in different university settings.

2) Within University Pair Type: consisted of two standardized patients who presented and rated the same case and were trained and used in the same university setting.

3) Within Rater Pair Type: consisted of the same standardized patient rating the same encounter on two separate occasions.

Each rater pair, for the three defined types, generated one observation of observed agreement for each of the items rated for the case. The number of items varied by case (see Table 8.1). Each observation was based on the ratings of 5-40 student encounters, the number dependent on the case and pair type (see Table 8.3 for breakdown).

TABLE 8.3 THE NUMBER OF OBSERVATIONS OF OBSERVED AGREEMENT FOR ITEMS BY CASE AND RATER PAIR TYPE

CASE	ITEMS RATED	NUMBER OBSERVATIONS/ITEM BY RATER PAIR TYPE			TOTAL OBSERVATIONS/ CASE (items*pairs)
		Bet-Univ	With-Univ	With-Rat	
1	12	2	1	2	60
2	24	2	1	2	120
3	9	1	0	1	18
4	20	2	1	2	100
5	8	1	0	1	16
6	10	1	0	1	20
7	20	1	0	1	40
8	8	2	1	2	40
9(s)	12	0	1	2	36
9(m)	9	0	1	2	27
10	29	1	0	1	58
11	27	2	1	2	135
12	5	2	1	2	25
13	11	2	1	2	55
14	21	2	1	2	105
15	19	2	1	2	95
16	8	2	1	2	40
Total	252				990

Legend: Bet-Univ: between university; With-Univ: within university; With-Rat: within rater

Rating Form Factors:

Item Type: was defined as the content of the item to be rated in one of seven mutually exclusive categories: History, Physical Examination, Diagnosis, Management, Teaching, Communication and Auxiliary items. The 252 items rated across the 16 cases used in the evaluation were classified by the investigator. The reliability of classification was estimated by reclassifying a 15% random sample of 252 of the items. The agreement on the first and second classification was 98.7%. The resulting classification of item type by case is displayed in Table 8.4.1.

TABLE 8.4.1. THE BREAKDOWN OF ITEMS BY TYPE AND JUDGEMENT LEVEL FOR THE 16 CLINICAL PROBLEMS

CASE #	NO. ITEMS BY TYPE							NO. ITEMS BY JUDGEMENT LEVEL				TOTAL
	His	Phy	Dx	Mg	Tea	Com	Aux	Act Y/N	Crit	A+Q	Qual	
1	-	8	-	-	4	-	-	-	11	1	-	12
2	7	3	-	13	1	-	-	23	1	-	-	24
3	9	-	-	-	-	-	-	8	1	-	-	9
4	14	6	-	-	-	-	-	18	2	-	-	20
5	-	-	-	8	-	-	-	8	-	-	-	8
6	6	3	-	-	-	-	1	8	2	-	-	10
7	12	8	-	-	-	-	-	14	6	-	-	20
8	-	-	-	-	2	6	-	-	-	2	6	8
9(s)	-	-	-	6	4	2	-	11	-	1	-	12
9(m)	1	3	-	-	-	5	-	3	2	1	3	9
10	15	14	-	-	-	-	-	24	5	-	-	29
11	18	9	-	-	-	-	-	26	1	-	-	27
12	-	-	-	-	-	5	-	1	-	-	4	5
13	-	10	-	-	-	-	1	11	-	-	-	11
14	12	9	-	-	-	-	-	21	-	-	-	21
15	1	3	1	13	-	1	-	15	3	1	-	19
16	-	-	-	-	-	8	-	-	-	2	6	8
Total	95	76	1	40	11	27	2	191	34	8	19	252
%	38	30	.3	16		4 11	.8	76	14	3	8	

Legend:
- His: history items
- Phys: physical exam items
- Dx: diagnosis items
- Mg: management items
- Tea: teaching items
- Com: communication items
- Aux: auxillary items

- Act Y/N: records action present or absent
- Crit: records action present according to criteria
- A & Q: records presence and quality of action taken
- Qual: rates quality of action(s) taken

Judgement Type: was defined as the type of judgement required of the standardized patient in recording/rating an item in one of four mutually exclusive categories:

1) Action Taken-Yes/No
2) Action Taken-Yes requires the judgement that explicit criteria were met
3) Action Taken-Yes requires that the behaviour occurred and the quality of behaviour was perceived to be acceptable(no criteria specified)
4) Rated Quality of Behaviour on a Likert scale. The 252 items rated were classified by the investigator. The reliability of classification was estimated using the same procedure as for item type. The agreement was 96.4%. The resulting classification of items by judgement type is displayed in Table 8.4.1.

Item Ambiguity: was defined as the number of actions which were included in the rating of one item. Ideally, one action should be specified for each item to minimize ambiguity in rating (Des Raj,1972). When two or more actions are specified (eg. examined range of motion for knees and hips) in an item, the rater must determine whether the performance of one or both actions is required to qualify as a Yes response. The number of actions included in each item was classified by the investigator in one of three mutually exclusive categories: one action, two or more explicitly stated actions, multiple actions implied but not explicitly stated in an item (eg. examines extremities). The reliability of classification was estimated by reclassifying a 15% random sample of the 252 items classified. The agreement between the first and second classification was 95.2%. The resulting classification of items by case is displayed in Table 8.4.2.

TABLE 8.4.2 THE BREAKDOWN OF PROBLEMS BY ITEM AMBIGUITY FOR THE 16 CLINICAL PROBLEMS

CASE	ITEM AMBIGUITY			TOTAL
	1 action/item	multiple explicit	multiple implicit	
1	2	10	0	12
2	19	2	3	24
3	8	0	1	9
4	15	2	3	20
5	8	0	0	8
6	2	6	2	10
7	9	4	7	20
8	0	2	6	8
9(s)	3	2	7	12
9(m)	0	4	5	9
10	11	12	6	29
11	10	7	10	27
12	1	0	4	5
13	0	2	9	11
14	10	3	8	21
15	14	3	2	19
16	0	2	6	8
Total	112	61		79
252				
%	44.4	24.2		31.3

Item Number: was defined as the number of items the standardized patient was required to rate for each encounter. The number of items recorded/rated for each case was prescribed by the faculty person responsible for case development (see Chapter 5). The number of items ranges from 5 to 29 and will be treated as both an ordinal and continuous variable in the analysis.

ANALYSIS

Missing Data and Response Bias

Four standardized patients from Manitoba and two patients from Illinois were unable to participate in the study. In order to determine whether response bias may have been introduced by the exclusion of these patients from the study, the original rating scores produced, during the

comprehensive evaluation, by participating patients will be compared to scores produced by patients who were unable to participate. This comparison will be limited to the 3 patients from the University of Manitoba because original evaluation scores are only available for Manitoba students. Failure to find significant differences in scores between participants and non-participants will not necessarily rule out the presence of response bias but would suggest that the non-participating patients were not different in the way in which they rated students in the original evaluation.

The use of videotapes rather than live encounters as the format for presenting the clinical situation introduces an additional factor which may influence rater agreement. The standardized patient had the option, when reviewing the videotape, to indicate that the videotape was inadequate to make a decision on an item. As a result, a biased representation of the agreement between raters by predictor variables could result if certain items were more apt to be classified as 'unratable' by different raters. In order to estimate the presence of this type of response bias, the percentage of times items were not rated by each rater pair type and each rating form factor will be evaluated by one-way ANOVA. If an association is found between the percentage of times items were not rated, the relationship with the observed agreement will be evaluated. The percent of items missing will be identified as a confounder if associated with rater pair type or rating form factors and observed agreement.

Rater Reliability

The primary interest in the examination of rater effects is to determine the extent to which two raters agree with themselves or each other in the rating/recording of the same encounter. Three methods of summarizing agreement will be used: observed agreement and kappa for categorical data (items) and an intra-class correlation coefficient for continuous data (overall encounter score).

1. Observed Agreement

Observed agreement will be calculated for each item, case and rater type in the following manner:

a) Observed Agreement to an Item: the percentage of times two ratings of the same item are in agreement for encounters evaluated.

b) Observed Agreement for the Case: the average agreement for items in each case.

c) Observed Agreement for Rater Pair Type: the average agreement for all cases for the three type of rater pairs.

2. Kappa and Standardized Kappa

The kappa statistic will be calculated for each item as well as the average kappa for each case and rater pair type. Kappa provides a summary measure of the observed agreement correcting for the expected agreement which could occur by chance alone(Fleiss,1981; Cohen,1960). Since the maximum value of kappa cannot be obtained in situations where there is high or low prevalence of the event and differences in the marginal proportions, standardized kappa will also be estimated. - Standardized kappa adjusts the estimate of kappa for the maximum possible kappa which could have been obtained given the marginal prevalence of events reported(Cohen,1960).

3. Intra-Class Correlation Coefficient

The intra-class correlation coefficient will be calculated for each rater pair type using the overall student score generated by the ratings of each encounter. The intra-class correlation coefficient provides a summary estimate of the magnitude of measurement error relative to the differences in the estimate of true ability among students (Winer,1962). The correlation coefficient assumes a value in the interval from <0 to 1 with 1 indicative of the theoretically optimal situation of no measurement error (i.e. variance in rating is attributable to students rather than raters).

Finally, a Pearson product moment correlation coefficient will be calculated for each case and rater pair type. Although it is not an index of agreement (Bland & Altman,1986), it provides data which would permit comparison to other reported rater studies reviewed in Chapter 4.

Systematic Differences Between and Within Raters For Overall Score and the Proportion of Students Passing and Failing

The primary interest in examining for systematic differences between raters is to estimate the potential for bias when raters are confounded with evaluation site or more than one rater is being used to present a case. Large scale evaluation of competence using standardized patients is being considered by licensing bodies in both the United States and Canada (Melnick,1989; Bérard,1989). In order to conduct such an evaluation, multiple evaluation sites will be required with local recruitment of standardized patients and raters. A fundamental prerequisite of such an evaluation is that performance scores and pass/fail status are not biased by the site in which the evaluation is conducted.

To evaluate whether systematic differences exist between raters and sites, differences in overall scores generated by raters in different and the same evaluation site will be calculated. The null hypothesis being tested is that no systematic differences exist between raters in the same or different sites. This hypothesis will be evaluated for each case and rater pair using a paired t-test.

In order to explore the impact that observed differences between raters might have on pass/fail status, differences in the proportion of students failing or passing for each rater pair evaluated in each case will be calculated. The null hypothesis of no difference will be tested using McNemar's test for paired categorical data.

Overall trends in the number of students passing and failing the clinical evaluation will also be evaluated. An arbitrary cut-off of 60% will be used as the minimum passing score. The proportion of student encounters which failed to meet this criterion will be calculated by rater and university site.

The Evaluation of Predictive Factors

Two groups of potential predictive factors will be evaluated, those related to the type of rater pair and those related to attributes of the rating form (eg. item type, item ambiguity) The dependent variable in this analysis is the observed agreement produced for each item evaluated by each rater pair. There are 5 possible pairs of raters, two for the between rater pair type, one for the within university pair type and two for the within rater pair type. One observation of agreement per rater pair type for each item will be produced by taking the average of the two pairs when present. This will result in 3 observations of agreement for each item (one for each rater pair type).

Bivariate relationships between potential predictive variables and observed agreement will be initially evaluated using a one-way ANOVA model for nominal level factors and linear regression for number of items. A mixed ANOVA model will be used to estimate the independent contribution of each factor to observed agreement. The random factor is the number of items and the fixed factors are item type, item ambiguity, judgement level, and rater pair type. Since the design is unbalanced, multiple regression analysis will be used to estimate the variance components. The varying precision of each estimate of agreement will be taken into account by carrying out a weighted regression analysis with the weight being defined as the sample size used to produce each observation of the dependent variable.

RESULTS

MISSING DATA AND RESPONSE BIAS

The Standardized Patients

Table 8.5 displays the mean student scores calculated for the standardized patient ratings in the original Manitoba evaluation. Students were essentially randomized to standardized patient so differences in the ratings between patients can be attributed to rater differences rather than differences in the population of students evaluated. Since small sample sizes limited the precision of the estimate, the observed differences in scores rather than those of statistical significance will be discussed. It can be noted that the average difference between participating and non-participating standardized patients was smaller than the average difference between participating patients for the same case (6.91% vs.11.23%). Although this would suggest that there were no major differences between non-participants and participants in terms of the ratings provided in the original evaluation, it is notable that all non-participants had higher mean student scores than their participating counterpart for the same case. This might indicate that standardized patients who did not participate were more lenient raters. If participation was limited to less lenient raters, rater reliability estimates would probably be biased in a positive direction.

TABLE 8.5 STUDENT SCORES CALCULATED FROM 1987 EXAMINATION RATINGS FROM PARTICIPATING AND NON-PARTICIPATING STANDARDIZED PATIENTS

Cases Where One Standardized Patient Did Not Participate

Case	Non-participant S.P. Student Score		Participant S.P. Student Score		Difference
	N	Mean (s.d.)	N	Mean (s.d.)	
#3	6	92.59% (11.48)	12	87.96% (11.07)	4.63
#5	17	73.53% (9.7)	2	68.75% (8.84)	4.78
#7	11	86.36% (10.27)	5	73.00% (12.04)	13.36*
#10	5	58.57% (13.74)	11	53.90% (18.45)	4.85
Average Difference					6.91

Cases Where Both Standardized Patients Participated

Case	1st S.P. Participant Student Score		2nd S.P. Participant Student Score		Difference
	N	Mean (s.d.)	N	Mean (s.d.)	
#1	5	66.00% (37.82)	9	86.67% (11.18)	20.67
#2	10	88.00% (12.29)	7	64.29% (35.52)	23.71
#4	3	75.00% (10.0)	16	60.94% (12.55)	14.06
#8	5	76.00% (6.75)	16	77.25% (6.40)	1.25
#9(M)	5	60.00% (35.66)	10	70.00% (22.25)	10.00
#9(S)	5	31.67% (18.07)	10	21.67% (11.92)	10.00
#11	8	62.98% (11.98)	9	73.93% (12.13)	10.95
#12	4	100.00% (0.0)	12	90.00% (28.92)	10.00
#13	13	73.43% (25.30)	3	72.73% (9.09)	0.70
#14	5	67.62% (7.82)	13	63.37% (20.52)	4.25
#15	9	38.60% (16.01)	11	56.94% (8.43)	18.34*
#16	15	69.17% (9.85)	2	80.00% (14.14)	10.83
Average Difference					11.23

Legend: N: the number of students in the study sample

Mean: the average student score calculated on the basis of the standardized patient ratings in the 1987 examination

*: the difference is statistically significant using an independent t-test after using Bonferroni's correction of the Type I error.

S.P.: standardized patient

Potential Predictive Factors

Table 8.6 provides a breakdown of the percent of times that items could not be rated for each of the potential predictor variables. There are 986 observations in the data set, one corresponding to each observation of observed agreement. The percent missing is calculated on the basis of the number of times the item could not be evaluated in the 5-42 student encounters rated by each rater pair type for each case.

Differences among the categories of each potential predictor were evaluated using a one-way ANOVA design and regression analysis with dummy variables. The null hypothesis being evaluated is that there are no differences in the percent missing among categories of each predictor. It can be noted that the null hypothesis was rejected for each of the variables considered.

TABLE 8.6 THE PERCENT OF TIMES ITEMS COULD NOT BE EVALUATED WITHIN CATEGORIES OF EACH OF THE POTENTIAL PREDICTOR VARIABLES TO BE EVALUATED

Predictor Variable	N	Average % Missing	S.D.	F	R^2	P-Value
Rater Pair Type						
Between Universities	384	7.48%	11.71	15.49	3.1	.000
Within University	176	4.84%	9.30			
Within Rater	424	3.43%	9.52			
Item Type						
History	343	3.33%	6.11	16.59	9.2	.000
Physical Exam	299	9.80%	13.71	*17.63	6.8	.0001
Diagnosis	5	15.24%	18.79			
Management	162	2.08%	6.00			
Teaching	47	3.35%	8.26			
Communication	121	3.89%	12.85			
Auxiliary	7	9.61%	7.28			
Judgement Level						
Action Y/N	745	4.11%	8.45	33.03	9.2	.000
Criteria	114	14.01%	15.15			
Action+Quality	36	2.2%	5.6			
Quality	89	4.95%	14.76			
Item Ambiguity						
One Action/Item	438	3.83%	7.52	9.08	1.8	.0001
Multiple explicit	227	7.53%	14.13			
Multiple implicit	321	6.21%	13.13			
Item Number						
0-5	25	8.96%	19.64	13.83	6.6	.0001
6-10	161	4.55%	10.62	+14.34	1.4	.0002
11-15	151	9.08%	12.55			
16-20	235	8.51%	15.41			
21-25	225	1.79%	4.83			
>25	189	3.45%	6.19			
Overall	986	5.45%	11.37			

Legend: N: the number of observations of observed agreement per category

Average % Missing: for each observation of observed agreement for each item and rater pair type, the percentage of times the item could not be evaluated was calculated. The average of these values for all items in the respective category was calculated and is tabled above.

F, R2 & P: calculated on the basis of a one-way ANOVA design using linear regression analysis with dummy variables

+: the F, R^2 and P-value when item number is treated as a continuous variable

*: the F, R^2 and P-value when diagnosis and auxiliary items are excluded from the analysis

In order to determine whether the rateability of an item from videotape would be a confounder in the analysis of potential predictors of agreement, the relationship between percent missing and observed agreement was evaluated using linear regression analysis. Table 8.7 provides the estimated betas, F statistic, R^2 and P-value by case and overall for this relationship.

TABLE 8.7 RELATIONSHIP BETWEEN % MISSING AND OBSERVED AGREEMENT BY CASE

CASE	N	% MISSING (S.D.)	R^2	B	F	P-value
#1	60	16.42 (15.66)	6.63	-0.48	4.12	.047
#2	120	.91 (3.67)	0.19	-0.19	0.23	.633
#3	18	12.78 (16.25)	1.89	-0.10	0.31	.587
#4	100	8.03 (13.57)	0.31	0.08	0.31	.58
#5	16	.72 (1.55)	0.00	-0.16	0.00	.951
#6	20	9.13 (6.13)	27.77	-1.25	6.92	.017*
#7	40	11.67 (15.64)	1.85	-0.07	0.07	.792
#8	40	5.8 (15.5	23.30	-0.63	11.54	.001*
#9(s)	36	.31 (1.04)	3.26	-2.51	1.14	.292
#9(m)	28	1.1 (3.22)	0.35	-0.32	0.09	.771
#10	56	3.33 (8.11)	0.44	0.19	0.24	.626
#11	133	3.50 (5.22)	11.86	-0.91	17.63	.0001*
#12	25	8.96 (19.65)	3.28	0.08	0.78	.386
#13	55	6.72 (7.40)	0.65	-0.20	0.35	.558
#14	105	2.79 (5.74)	8.37	-1.01	9.41	.003*
#15	93	7.25 (16.95)	5.72	-0.21	5.52	.021*
#16	40	.91 (2.13)	4.88	1.82	1.95	.171
Overall	984	5.45 (11.37)	3.12	-0.32	31.63	.0001*

Legend: N: the number of observations of observed agreement by case and overall

F, R2, B & P-value: estimates produced from bivariate linear regression

*: statistically significant after correction for multiple comparisons

There is a significant association between the percent of times that items could not be rated and observed agreement. Over all cases, an inverse relationship exists. As the percent of times items could not be evaluated increases the observed agreement decreases. Observed agreement is estimated to diminish by approximately a third of a percent with every 1% increase in the percent of times that items could not be evaluated. This

varies by case from a 1.82% increase in observed agreement to a 2.51% decrease.

This relationship likely reflects the difficulty in rating some items by videotape. Observed agreement would be compromised if some items were difficult to rate through videotape observation and yet raters attempted to provide a rating even when they had insufficient data from the videotaped encounter to do so. This explanation is supported by the nature of items where the percent missing was higher: physical examination and items which required a number of implicit or explicit criteria to be met. Both categories of items would require adequate visualization and sound to determine if an action had been taken. Since the rateability of items by videotape is a confounder in the relationship of potential predictors and observed agreement, it will be adjusted for in the analysis.

There were more occasions where items could not be evaluated for the between university pair. Southern Illinois patient raters indicated that items could not be rated more frequently. As a result, the between university rater pair estimate may be biased. Observed agreement may be positively biased by the tendency for Illinois raters to conservatively rate only those items where the presence or absence of an action on videotape was less ambiguous. If this were the case the between university pair would demonstrate better agreement than the within university or within rater pair types.

With respect to item type, there was a greater percentage of times when physical examination, diagnosis and auxiliary items could not be evaluated. The number of observations of observed agreement for auxiliary and diagnosis items is small. When this problem is coupled with the difficulty in rating these items, the ability to draw inferences about observed agreement with these item types is sufficiently compromised to warrant their exclusion from the analysis of predictive factors. Physical examination items were probably more difficult to evaluate by videotape review due to inadequate visualization of the patient and examiner. All physical examination items were rated on at least 10 occasions.

Items which required the standardized patient to employ specific criteria in making a decision about the presence or absence of an action were more frequently not rated than those requiring other categories of judgement. This may be the result of the difficulty in observing all required criteria on videotape. Alternatively, the standardized patient may experience greater difficulty in recalling whether all criteria were met. It is not possible to determine which of these two possibilities was operative in this situation. The data on item ambiguity would suggest that items with multiple explicit or implicit criteria are more frequently found to be unratable. This would suggest that part of the problem may be attributed to difficulty in recall. This hypothesis would require further evaluation. Those items which were rated in the criteria and multiple action categories may provide biased representation of better agreement than would actually exist.

There were fewer occasions when an item could not be rated for rating forms with 21 or more items. This observation seems to be counter-intuitive. One would expect that recall problems alone would produce a greater frequency of missing ratings with longer response forms. This phenomenon is likely explained by the fact that the longer forms contained items which no student was able to perform within the given time constraints. Typically, these items related to psycho-social evaluation which few students had time to pursue given the priority placed on the evaluation of the physical problem. Response forms with 21 or more items therefore probably represent an artificial extension of the scale for number of items. Most raters, operatively speaking, rated forms of 5 to 20 items in length. The frequency of missing ratings for these categories is more or less equal.

RATER RELIABILITY

Observed Agreement

Observed agreement was calculated for each item on the basis of the videotaped encounters rated by each standardized patient rater pair type. The number of encounters contributing to each observation of observed agreement is displayed in Table 8.1. The number of observations

of observed agreement by item and case is displayed in Table 8.3. The average agreement across items was calculated for each case and rater pair type. When one of the two raters in a pair did not rate an item for an encounter, the encounter(s) was(were) excluded from the calculation of agreement on the item. The average observed agreement for items rated in a case represents only those items and encounters where both raters provided a rating.

For the 2 cases where a 7 category Likert scale was used to rate communication abilities (Case 8 and Case 16), the scale was collapsed into a 3 category scale and the observed agreement for the 3 category scale was used in the calculations. The reason for doing so was that the agreement for the 7 category scale rarely rose above 40%. There was no major improvement with reduction to a 5 category scale, therefore a 3 category scale was used. The results are displayed in Table 8.8.

TABLE 8.8 THE AVERAGE OBSERVED AGREEMENT FOR ITEMS RATED: BY CASE AND RATER PAIR TYPE

Case	Number Items Rated	Overall n	Overall %(s.d.)	RATER PAIR TYPE Between-U % (s.d.)	Within-U % (s.d.)	Within-R % (s.d.)
1	12	60	70.03 (29.3)	66.93 (23.7)	85.94 (9.9)	65.17 (37.8)
2	24	120	86.23 (15.8)	85.24 (13.5)	87.75 (12.6)	86.46 (19.2)
3	9	18	79.43 (12.2)	78.47 (11.2)		80.39 (13.8)
4	20	100	75.69 (19.6)	76.70 (16.2)	71.82 (19.4)	76.63 (22.9)
5	8	16	89.83 (14.6)	85.90 (10.4)		93.75 (17.7)
6	10	20	79.87 (15.4)	79.02 (17.5)		80.72 (13.9)
7	20	40	75.47 (23.8)	68.94 (24.7)		82.00 (21.4)
8	8	40	66.43 (20.2)	49.46 (9.3)	67.00 (13.8)	83.13 (16.6)
9(s)	12	36	85.44 (14.4)		84.66 (12.4)	85.83 (15.5)
9(m)	9	27	80.67 (17.6)		81.98 (17.5)	80.00 (18.1)
10	29	56	77.43 (23.6)	90.61 (8.2)		63.27 (26.6)
11	27	133	84.17 (13.8)	83.20 (14.7)	84.29 (9.6)	85.10 (14.9)
12	5	25	91.68 (8.6)	86.21 (8.1)	89.60 (8.6)	98.18 (3.8)
13	11	55	80.97 (17.9)	82.55 (12.2)	81.50 (14.0)	79.14 (24.0)
14	21	105	81.74 (20.0)	77.17 (20.3)	78.74 (21.4)	87.82 (17.6)
15	19	93	91.85 (9.2)	91.89 (6.7)	90.84 (10.7)	92.30 (10.5)
16	8	40	77.19 (17.6)	80.62 (11.7)	74.77 (15.1)	74.97 (23.2)
Total	252	986	81.30 (17.7)	80.10 (17.7)	82.03 (21.5)	82.2 (15.6)

Legend: Number Items Rated: the number of items rated by case for each videotaped encounter

Between-U: rater pair produced by comparing the ratings of the standardized patient from Southern Illinois and with the standardized patient(s) from Manitoba for the same videotaped encounter

Within-U: rater pair produced by comparing the ratings of the two standardized patients who were trained together in Manitoba for the same encounter

Within-R: rater pair produced by using the two ratings produced by the same standardized patient for the same encounter; the first from the original examination and the second from subsequent videotape review

N: the number of observations of observed agreement produced by the product of number of items rated (8-29) and number of rater pair observations (2-5). Each observation of agreement is based on the review of 16-42 videotapes (see Table 8.1) for Between-U and Within-U and 5-20 encounters for Within-R

The average observed agreement for items rated was 81.3% varying by case from 66.43% for Case #8 to 91.68% for Case #12. In 8 cases, the average agreement was less than 80%. The standard deviation for each case and rater pair type is large suggesting that substantial variation in observed agreement exists among the items rated for each case.

It can be noted that the magnitude of the average agreement for different types of rater pairs for all cases combined is similar, differing by only 2 percentage points. This similarity in average agreement among rater pair types is present in 9 of the 16 cases (2, 3, 4, 6, 9, 11, 13, 15, 16). This finding is surprising since it would be expected that raters would agree with themselves much more frequently than with another rater. This observation may be the result of differences in the rating conditions for the within rater estimates which would tend to lower the observed agreement. Alternatively, agreement may be more dependent on the item rather than the type of rater pair. If observed agreement was high for all possible pairs of raters, then no difference among rater types would be observed. This explanation seems plausible for case #15, where agreement was over 90% for all rater pair types. It is a less likely explanation for similarity in agreement among rater pair types in the remaining cases.

In 7 cases there are differences in the average agreement for different pair types of 10% or more. In Case #1, agreement between the two patients trained in the same university is substantially better than for patients trained in different universities or for the same patient rating twice. Inspection of the observed agreements calculated for each item suggests that much of the difference in the between university pair was attributable to 1 of the 8 physical examination items (observed agreement from 30%-50%) and 2 of the 4 teaching items (observed agreement from 25-28%). For the remaining items standardized patients trained together agreed to the same extent as those trained in different settings. The small sample sizes used to calculate the within rater item agreement contributed to instability in the observed agreement estimate.

In Case #10, raters trained in different settings agreed to a greater extent than the same rater rating twice. This remarkable observation is

attributable to 8 of the 29 items rated. For these 8 items there was a tendency for the rater during the original exam rating to indicate that either all or most students had taken the action indicated. In the second videotaped rating, many more students were rated as having not taken the action. This may be a function of the differences in the evaluation conditions.

In 5 cases (5, 7, 8, 12, 14), the within rater agreement is better than the two types of between rater agreement (between-u and within-u). The trend in these cases is in the direction expected with best agreement being within the same rater followed by raters trained together and raters trained in different universities.

Kappa and Standardized Kappa

The calculation of observed agreement fails to take into consideration that raters may be expected to agree on some occasions by chance alone. The kappa statistic (Cohen, 1960; Fleiss, 1981) provides a means of correcting the observed agreement for that which would occur by chance. Kappa is sensitive to the marginal proportions of events and the maximum value of 1 (perfect agreement) cannot be obtained when the marginal distributions produced by the two raters are different and the prevalences of an event are very high or very low. Standardized kappa corrects the calculated kappa statistic for the maximum possible agreement (Kappa max) which could occur given the marginal proportions. Standardized kappa represents the ratio of kappa to the maximum kappa attainable given the marginal proportions. Both kappa and standardized kappa were calculated for each of the 252 items rated by case and rater pair type. A summary of the average kappa and standardized kappa for each case and rater pair type is displayed in Table 8.9.

TABLE 8.9 AVERAGE KAPPA AND STANDARDIZED KAPPA BY CASE AND RATER PAIR TYPE

CASE	RATER PAIR TYPE							
	Overall		Between Universities		Within Universities		Within Rater	
	K	StK	K	StK	K	StK	K	StK
1	.36	.37	.29	.29	.31	.35	.46	.46
2	.64	.66	.58	.63	.63	.67	.70	.70
3	.25	.25	.14	.14	--	--	.37	.37
4	.32	.35	.28	.33	.23	.28	.40	.40
5	.74	.78	.60	.67	--	--	.88	.88
6	.45	.48	.50	.55	--	--	.40	.42
7	.43	.50	.36	.46	--	--	.50	.55
8*	.24	.27	.07	.11	.04	.10	.49	.50
9	.54	.54	--	--	.49	.49	.42	.43
10	.34	.38	.56	.63	--	--	.09	.12
11	.48	.50	.44	.47	.40	.45	.55	.55
12	.44	.44	.22	.24	.02	.02	.86	.86
13	.44	.49	.44	.48	.40	.60	.45	.46
14	.43	.43	.33	.48	.34	.36	.57	.57
15	.66	.67	.59	.61	.69	.69	.71	.71
16*	.16	.19	.10	.10	.02	.33	.28	.29
Total	.45	.47	.40	.44	.40	.41	.52	.53

Legend: *: cases where communication was rated using a 7 category Likert scale collapsed for this analysis into a 3 category scale. The remaining cases were rated on a dichotomous scale.

K: kappa

StK: standardized kappa

Landis and Koch (1977) have provided a schema for interpreting the meaning of the range of values of kappa for observer agreement. Using their schema, the average kappa value for rater agreement across all cases and for the three rater pair types would be considered to be good (values of .41 to .60). The average agreement was excellent (values of .61-.80) in three cases (2, 5, 15) and good in seven cases (6, 7, 9, 11, 12, 13 & 14). For the remaining cases, the average agreement was fair (values of .21 to .40) in five cases and slight in one case(values of 0 to .20).

Agreement was fair to slight in both cases where a collapsed 3 category scale was used to rate communication. This is consistent with the literature where poorer observer agreement is noted in rating communication skills (see Chapter 5). This problem is usually considered to be attributable to the content of measurement; the rating of sentiments. Rater agreement is also harder to achieve on a multi-category scale: communication skills being the only content area rated on a scale with more than two categories. Within rater agreement for both of these cases is superior to the two types of between rater agreement. In Case 8, the within rater agreement would be considered to be good while the between rater agreement is slight. In Case 16, the within rater agreement is fair where as the between rater agreement is slight. This would suggest that better reliability in rating communication skills might be achieved by selecting and training more comparable raters along with improvements in the measurement scale itself.

Once observed agreement has been corrected for chance, it can be noted that a different pattern emerges with respect to the magnitude of agreement by rater pair type. Better agreement within the same rater is evident over all cases and in 14 of the 16 cases evaluated. This difference is statistically significant ($F_{2,978}=11.50$ $p=.0001$) but accounts for only 2.3% of the variance in kappa values. No differences are present in the average agreement achieved by the two between rater pair types. Raters trained together in the same institution do no better than those trained in different institutions for the same case. This observation is true whether kappa or standardized kappa is used.

In most instances, standardized kappa values tended to be higher than kappa values although the difference was negligible in most cases and rater pair types. One of the difficulties which was encountered in using kappa as a summary statistic for agreement was in instances where one or both raters found the prevalence of an event to be either 0 or 100%. When one rater recorded prevalences of 0 or 100%, the kappa value was always 0 even if the disagreement was limited to one student encounter. The observed agreement in these instances may be as high as 98%. When both raters recorded prevalences of 0 or 100%, the kappa value was, of course, 1. This problem occurred in 32% of the of the 961 observations of item agreement. In order to determine how this phenomenon might influence the conclusions about the overall magnitude of agreement and the differences in agreement by rater pair type, a second analysis was conducted. In this analysis, items where one or both raters recorded prevalences of 0 or 100% were removed. The results are reported using the categorization schema of Landis and Koch (1977) and are displayed in Table 8.10.

TABLE 8.10 THE PERCENT FREQUENCY OF KAPPA VALUES BY QUALITY OF AGREEMENT CATEGORY AND RATER PAIR TYPE FOR ALL ITEMS AND FOR ITEMS WITH PREVALENCES OF GREATER THAN 0% OR LESS THAN 100%

AGREEMENT CATEGORY	RATER PAIR TYPE							
	Overall		Between Universities		Within Universities		Within Rater	
	All	Subcat	All	Subcat	All	Subcat	All	Subcat
Perfect or Excellent (.61 to 1)	40%	41%	41%	45%	43%	46%	50%	53%
Good or Fair (.21 to .60)	23%	35%	27%	33%	26%	33%	16%	31%
Slight or Poor (0 to .20)	36%	25%	32%	23%	35%	20%	33%	16%

Legend: All: All 986 item observations of agreement calculated included in the analysis

Subcat: Items where one or both raters recorded prevalences of 0% or 100% are removed resulting in the inclusion of 674 item observations in the analysis

In Table 8.10, the six categories proposed by Landis and Koch(1977) have been collapsed into three categories in order to allow the trends in agreement by rater pair type to be more easily visualized. In the six category schema, the removal of items with recorded prevalences of 0% or 100% reduced the frequency of agreement for items in the two extreme categories of perfect and poor. Approximately one third rather than one quarter or one fifth of items now fall into the middle category of good or fair.

The observed differences among rater pair types are also evident when kappa is treated as a categorical rather than a continuous variable as done in the earlier analysis. The removal of items with prevalences of 0% and 100% changes the distribution of items among the categories. It does not change the overall trend noted in the previous analysis. Raters tend to agree better with themselves than with other raters. The agreement of raters trained together is no better or worse than raters trained in the two university sites.

The frequency of agreement in the perfect and excellent categories for the within rater pair type likely represents an underestimate because, unlike the between rater estimates, the conditions of measurement between the first and second rating were different. It will be recalled that the first rating was carried out during the student examination, when the standardized patient was also required to present the case, while the second was carried out on the basis of videotape review. The between rater estimates were based on the review of the same videotapes under the same measurement conditions. As a result, differences among rater pair types may be underestimated.

Despite differences in the frequency of perfect or excellent agreement by rater pair type, the frequency of agreement in the slight and poor category for all rater pair types is far from optimal. Overall, one quarter to one third of observations of item agreement fell into this category (depending on whether all or a the defined subcategory of items was used). Factors which may have contributed to item agreement, other

than rater pair type, will be reported in a subsequent section. It could be argued that the agreement on any specific item is not important if similar scores are produced by the same or different raters for the same encounter. This is a reasonable conclusion if performance on the items themselves does not contribute to decisions made about the examinee. In instances where overall score alone is being used, the intra-class correlation coefficient would provide a better index of the agreement between raters for the same encounter. The next section describes the results for rater agreement when the agreement in total encounter score is evaluated.

The Intra-Class Correlation Coefficient For Total Encounter Score

A score for each encounter was produced for each rater. The encounter score was produced by summing the number of actions which were recorded as being taken and dividing by the number of actions which could be rated during videotape review. Scores were converted into percents. An intra-class correlation coefficient was calculated for each rater pair, by case, using the formula described by Winer (1971) and a random effects model for estimating each variance component. The results of this analysis are displayed in Table 8.11. Since a Pearson product moment correlation has been used to summarize rater reliability for total score in a number of studies, this statistic was calculated and is included in the table for comparison purposes. The results of item agreement from Tables 8.8 and 8.9 have also been included to permit comparison between average item agreement and agreement for total score.

TABLE 8.11 SUMMARY OF RATER RELIABILITY FOR ITEM AND OVERALL SCORE AGREEMENT BY CASE AND RATER PAIR TYPE

Case	Average Agreement By Item		Agreement for Total Score	
	% Agreement	Kappa	Intra-Class R	Pearson's R
Case 1				
Between U	66.93	.29	.21	.48
Within U	85.94	.31	.77	.59
Within R	65.17	.46	0	0
Case 2				
Between U	85.24	.58	.52	.82
Within U	87.75	.63	.77	.88
Within R	86.46	.70	.65	.73
Case 3				
Between U	78.47	.14	.12	.17
Within U				
Within R	80.39	.37	.37	.70
Case 4				
Between U	76.70	.28	.32	.43
Within U	71.82	.23	.17	.24
Within R	76.63	.40	.06	0
Case 5				
Between U	85.90	.60	.60	.60
Within U				
Within R	93.75	.88	.80	1
Case 6				
Between U	79.82	.50	.45	.47
Within U				
Within R	80.72	.40	.64	.64
Case 7				
Between U	68.94	.36	.27	.52
Within U				
Within R	82.00	.50	0	.42
Case 8				
Between U	49.46	.07	.27	.44
Within U	67.00	.04	.02	.03
Within R	83.13	.49	.62	.62
Case 9				
Between U				
Within U	81.98	.49	.63	.80
Within R	80.00	.42	.38	.51

In Table 8.11, intra-class correlation coefficients with a negative value were given a value of zero. This applies to within rater estimates for Case 1, 7 and 12. The average item agreement and intra-class correlation coefficient for total score are similar in 7 cases. In the remaining cases, the intra-class correlation coefficient was higher for within and between university rater pairs for Case 1,8, and 10. In these cases, raters may had difficulty agreeing on specific items but overall score assigned to each student assessed was more similar. In Case 1,4,7,12,13 and 15, the within rater estimates of agreement for total score were less than for average item agreement. In these cases, the variance for students was particularly small (most students received a score of 0 or 100 depending on the case). This phenomenon, coupled with the smaller sample size used in the estimations, resulted in lower estimates of agreement for overall score.

Although average item agreement, corrected for chance, shows better agreement for within rater pairs, this difference does not persist when agreement on overall score is evaluated. The average intra-class correlation coefficient for the within rater pair type is the same or slightly worse then between rater pairs. Inspection of the individual pair values for within rater estimates indicates that some raters were remarkably reliable (Kappa and intra-class $R > .8$) while others were extremely poor (Kappa and intra-class $R < .2$). Agreement for the two types of between rater pairs is similar for overall score. This suggests that poor item agreement may primarily be a function of differences between raters rather than trainers or evaluation site.

The Pearson product moment correlation coefficient is larger in magnitude than the intra-class correlation for total score. The values for the Pearson's R tend to be smaller than those reported in the studies reviewed in Chapter 4 (Table 4.3). The estimates reported by other authors were smallest for communication skills (r=.52-.77). The comparable cases in this study are Case 8 and 16 where communication skills were rated. Estimates for these cases varied from r=.03-54. The remaining reported measures of association are for history and physical exam skills ranging from r=.7-.93. In this study, history and physical exam skills

were the only items rated in Cases 3, 4, 6, 7, 10, 11, 13 and 14. The values of Pearson's R for these cases ranged from r=.17-.8. There may be a number of reasons for these differences. Most of these studies compared standardized patient rating with faculty or research assistant rating. There is no comparable study of rater reliability for standardized patient raters. The sample sizes used for reported estimates in other studies were small. The confidence interval around these estimates would likely be large and may include the values of the Pearson's R observed in this study. The student groups studied by other authors may have been more heterogeneous, raters may have been better trained or cases/items easier to rate.

The average value for kappa and the intra-class correlation coefficient over all cases and within specific cases is below conventionally accepted values for rater reliability of .8 or .9. Using the more lenient classification schema of Landis and Koch(1977), the qualitative value of agreement by case and rater pair type is summarized in Table 8.12.

TABLE 8.12 QUALITATIVE INTERPRETATION OF THE AVERAGE AGREEMENT FOR ITEMS AND OVERALL SCORE BY RATER PAIR TYPE FOR THE CASES EVALUATED USING THE CLASSIFICATION OF LANDIS AND KOCH

	Qualitative Category of Agreement					
Rater Pair Type	Slight or Poor (0 to .21)		Fair or Good (.21 to .60)		Excellent or Perfect (.61 to 1)	
	K	ICR	K	ICR	K	ICR
Between University	3/15	1/15	12/15	12/15	0/15	2/15
Within University	3/11	4/11	6/11	3/11	2/11	4/11
Within Rater	1/16	5/16	11/16	6/16	4/16	5/16

Legend: Numerator: is the number of cases which have values belonging to the category

Denominator: total number of cases evaluated for the rater pair type

K: kappa

ICR: intra-class correlation coefficient for all rater pair types, agreement was better in more cases for overall score (intra-class correlation coefficient) than for items (kappa). The maximum number of cases in the excellent to perfect category was for the within rater pair type (cases=5), the least for the between university pair type (cases=2).

Agreement was far from desirable in most cases. In most cases differences between students accounts for only 20-60% of the variance in scores, the remainder being attributable to differences between raters and the interaction of students and raters. These findings would suggest that raters may be making a considerable contribution to case effects and measurement error. The finding is at odds with the reports of Swanson and Norcini (in press). In their reports of two studies, raters within cases contributed a little more than 1% of the total variance. In order to compare Swanson and Norcini's findings with those arising from this study, a random effects repeated measures ANOVA was carried out to derive estimates of the variance attributable to students, cases and raters within case. The results of both studies are displayed in Table 8.13.

TABLE 8.12 THE CONTRIBUTION OF RATERS WITHIN CASES TO VARIANCE IN STUDENT SCORE: A COMPARISON OF THE RESULTS FROM THIS STUDY WITH THOSE OF SWANSON AND NORCINI

Effects	Variance Component		
	Swanson & Norcini Study 1	Swanson & Norcini Study 2	This Study
Students	13.84	24.26	20.45
Cases	76.37	134.53	86.69
Raters:Case	5.73	4.77	81.85
Error (Students*Cases + Students*Patients:Cases + Error)	185.64	140.88	226.83

Swanson and Norcini (in press) provide two estimates of the contribution of raters to score variance. In the first study, standardized patients were used to both present the case and rate the student. Ten cases were studied in a similar population of students. In the second study, two members of the medical faculty rated the case. Fifteen cases were used in the evaluation: faculty raters being nested within case. Neither study provides the same conditions for rating as used in this study.

The contribution of raters within cases in this study is quite a bit larger than that found by Swanson and Norcini (in press). This may be a result of a more homogeneous group of students in this study (the variance attributable to students being somewhat smaller). Swanson and Norcini's estimates were derived from two raters per case whereas the estimates for this study were based on three raters per case. Separate analyses were carried out for each pair of raters within case to determine if this would alter the results. The relative size of the variance components was unchanged for different combinations of raters. Swanson and Norcini's estimates were based on the rating of data collection. In this study, two cases required the rating of communication skills, an area where rater agreement is known to be poor. These cases were removed from the analysis with virtually no effect on the size of the variance components.

In summary, raters contributed in an appreciable way to score variance in this study. These findings are different from those of Swanson and Norcini (in press) who found the rater effect to be negligible. These differences may be partially attributable to differences in the rating conditions in the two studies. In Swanson and Norcini's study, rater scores arising from the actual examination were used. Patients were used to present and rate the case in study 1. Faculty rated the case by direct observation in study 2. Scores from videotape review were used in this study. Since under videotape conditions, the patient is not being asked to present the case as well as recall student actions, one would anticipate that agreement between raters would be better than that observed in Swanson and Norcini's study 1. For this reason, this explanation seems a less likely candidate for the observed differences. A more likely explanation is that the quality of the patient raters or rating forms differed in the two studies. Rating form factors which may have contributed to poorer agreement will be evaluated in a subsequent section of the results. It is conceivable that faculty raters are more reliable (study 2), particularly when rating by direct observation. This hypothesis would require further evaluation.

SYSTEMATIC DIFFERENCES IN ENCOUNTER SCORES AND THE PROPORTION PASSING AND FAILING

Differences in Encounter Scores

In order to evaluate potential bias in the evaluation of competence by evaluation site, the average difference in student encounter scores between the raters from different university settings was examined. The null hypothesis being tested was that the average difference would be zero.

The average difference between raters from the same site was also evaluated. This provides a reference group for the interpretation of differences observed between raters from different sites. It, in addition, provides an estimate of the presence of systematic bias between two standardized patients who are presenting and rating the same case. Previous studies in this area have been limited by either small sample sizes or the possibility that differences in the groups evaluated explained observed differences in score. The results of both analyses are displayed in Table 8.14.

Table 8.14 provides the average score for the student-patient encounters evaluated. The same encounters were rated by all standardized patient raters for a given case. The differences in scores generated are exclusively due to differences among the raters. The average difference in score for rating the same encounter for patients trained in different sites is also displayed.

Twenty-five rater pairs were available for evaluating differences between raters trained in two university sites. In 18 of the 25 pairs, the average score for the Southern Illinois rater was lower than the University of Manitoba rater. These pairs are identified by the negative value for the average difference in score (note: the difference in score was calculated by subtracting the score for the U of M rater from the SIU rater). The magnitude of these differences varies considerably being as low as 2% for one rater pair in Case 4 to a high of 37% for Case 1. In eleven of these comparisons, the null hypothesis that there was no difference between

raters from different sites was rejected.

In order to examine overall trends, the difference in score over all cases was calculated using two summary measures: the average score over all cases rated and the average difference in score for each encounter rated. In order to produce these two summary measures, the average score for the two Manitoba raters for each common encounter rated was first calculated in the 12 cases where there were two Manitoba raters. The average score for the 472 encounters evaluated by the 15 raters from SIU was 62% in contrast to 67% for Manitoba raters.

The average difference in rating for each encounter was 7%, with the SIU raters scoring lower than the U of M raters. Both summary measures were statistically significant ($p<.0001$) using an independent and paired t-test respectively. The impact that this difference in score had on the proportion of students passing and failing will be reported subsequently.

Although SIU raters tended to produce scores which were systematically lower than Manitoba raters, the difference in scores between raters from the same and different sites was of similar magnitude. The difference between raters from the same site was smaller than the difference for the two between site pairs in only two cases (Case 1 & Case 13). The null hypothesis of no systematic difference between raters was rejected in 44% (11/25) of between site pairs and 50% of same site pairs (6/12). These observations suggest that the presence of systematic differences between raters is a function of both differences in evaluation site and individual raters.

TABLE 8.14 DIFFERENCES IN ENCOUNTER SCORE FOR RATERS FROM DIFFERENT AND THE SAME EVALUATION SITE

CASE	N	MEAN ENCOUNTER SCORE			MEAN DIFFERENCES IN ENCOUNTER SCORE		
		Site 1	Site 2		Between Site 1 & 2		Within Site 2
		R^1	R^1	R^2	Pr^1	Pr^2	Pr^1
1	35	44.0%	81.0%	75.5%	-36.7%*	-30.5%*	5.16%
2	31	50.0%	52.2%	45.4%	- 2.2%	4.5%*	6.7%*
3	30	88.7%	74.2%		14.5%*		
4	31	45.1%	61.2%	46.8%	-16.2%*	- 1.7%	14.4%*
5	26	62.4%	65.5%		- 3.2%		
6	23	68.8%	65.7%		3.2%		
7	33	54.6%	73.8%		-19.2%*		
8	25	65.3%	72.3%	79.8%	- 7.2%*	-14.5%*	7.3%*
9(s)	27		23.3%	35.2%			11.9%*
9(m)	27		80.6%	69.0%			11.6%*
10	16	61.9%	65.2%		- 3.3%		
11	32	60.1%	58.2%	66.6%	1.9%	- 6.5%*	8.4%*
12	28	84.6%	98.0%	90.2%	-12.9%	- 5.5%	7.3%
13	36	55.1%	64.0%	61.0%	- 8.9%*	- 6.0%	3.0%
14	30	53.0%	62.2%	73.7%	- 9.3%*	-20.7%*	11.4%
15	42	56.4%	49.6%	54.2%	6.8%	2.2%	4.6%
16	27	75.8%	70.5%	87.9%	5.3%	-12.1%*	17.4%
Overall		62.3%	67.2%		-6.74		

Legend: Site: Site 1=Southern Illinois University and Site 2=University of Manitoba

R: rater

Pr: refers to the rater pair.

*: statistically significant after using Bonferronni's correction for multiple comparisons

Systematic Differences in the Proportion of Students Passing and Failing by Evaluation Site and Rater

In the previous section, an average difference of 5-7% was noted in scores generated by raters in the two universities. A difference of this magnitude may not be of importance for multi-site credentialing examinations if it has no impact on the proportion of students passing and failing. Similarly systematic differences between raters from the same site may not be of importance if they have no impact on pass/fail status. The proportion of students who would have been passed or failed by

virtue of rater scores was therefore examined over all encounters evaluated. Pass/fail status was possible to calculate in 10 of the 16 cases where a minimum score was specified for the percent of actions which had to be taken in the patient encounter to pass. Trends in pass/fail status over all 16 cases were examined by applying an arbitrary passing criterion score of 60% to all encounters evaluated. The results are displayed in Table 8.15.

TABLE 8.15 THE PROPORTION(%) OF STUDENTS PASSING AND FAILING BY CASE, RATER AND EVALUATION SITE

CASE	N	STATUS	EVALUATION SITE			
			S.I.U.	U of N		
			R^1	R^1		R^2
1	35	Pass	9%	83%		71%
		Fail	91%	17%		29%
2	31	Pass	65%	84%		68%
		Fail	35%	16%		32%
3	30	Pass	93%	77%		
		Fail	7%	73%		
4	31	Pass	10%	32%		10%
		Fail	90%	68%		90%
5	26	Pass	89%	92%		
		Fail	11%	8%		
9	27	Pass		93%		67%
		Fail		7%		33%
10	16	Pass	0%	6%		
		Fail	100%	94%		
11	32	Pass	0%	0%		16%
		Fail	100%	100%		84%
12	36	Pass	25%	42%		39%
		Fail	75%	58%		61%
13	30	Pass	0%	3%		13%
		Fail	100%	97%		87%
14	42	Pass	26%	2%		19%
		Fail	74%	98%		81%
Overall	456	Pass	50%		67%	
		Fail	50%		33%	

Legend: N: the number of patient-student encounters evaluated

The most striking observation about Table 8.15 is the number of student failures. Pass/fail status for each case was determined by prospectively defined criteria established by the faculty in the two universities. The proportion of student failures could be attributed to unrealistic performance standards, substandard student populations or differences in the proportion of students failing as a function of the rating method (i.e. by videotape review). The specification of performance standards and the appropriateness of passing and failing students on the basis of their performance on one case are both issues worthy of exploration. However, these policy decisions were not under the control of the investigator and are outside of the scope of this thesis.

The potential contribution of the rating method to the proportion of students passing and failing was evaluated using the examination scores which were available for Manitoba students. The subset of videotaped encounters which were drawn from the Manitoba population were used to evaluate whether differences in overall mean rating and the proportion failing existed by virtue of rating method. The overall average score produced by videotape review was 64.8% in contrast to 72% for ratings carried out during the actual examination. Using an independent t-test, this difference was statistically significant ($p<.0001$). The proportion of students passing and failing as a function of rating method is displayed in Table 8.16.

TABLE 8.16 THE PROPORTION(%) OF STUDENTS FAILING BY VIDEOTAPE REVIEW AND BY RATINGS CARRIED OUT DURING THE ACTUAL EVALUATION: MANITOBA ENCOUNTERS ONLY

CASE	N	VIDEOTAPE REVIEW			ACTUAL EXAMINATION
		SIU^{R1}	UofM^{R1}	UofM ^{R2}	
1	14	100%	29%	64%	7%
2	17	41%	12%	24%	18%
3	18	11%	28%		0%
4	19	84%	68%	84%	63%
5	19	16%	5%		0%
9	15		7%	40%	40%
10	16	100%	94%		94%
11	17	100%	94%	100%	77%
12	16	31%	31%	38%	13%
13	18	100%	100%	100%	100%
14	20	55%	75%	70%	70%

In most cases, the proportion of students failing was lower in the actual examination than by videotape review. The proportion of student failures calculated on the basis of videotape review is therefore likely an overestimate of the proportion who would have been failed on the actual examination. There is no 'gold standard' against which the estimates of the proportion failing can be compared. The extent to which each of the estimates may be a biased representation of the true proportion of failures is therefore unknown. It is possible that standardized patients raters were more apt to score an action as being present in the actual evaluation if they did not recall whether it occurred. Recall problems may have been reduced in videotape ratings since standardized patients were not simultaneously required to both present and rate the case. This hypothesis is supported by Neiman's (1988) study on recall accuracy. Actions which were not taken were more likely to be inaccurately reported.

In Table 8.15 it is apparent that there are substantial differences in the proportion of students who would have failed on the basis of scores calculated for each rater. This is most dramatically seen in Case 1 where the SIU rater's scores resulted in the failure of 32 students in contrast to 6 and 10 students failed by the two Manitoba raters. In examining the

average difference in score displayed in Table 8.14, it appears that differences of 4-5% in overall score between raters has consequence for differences in the proportion of failures. In this data set, average differences of this size would have resulted in a difference, between raters, of 4-5 student failures. This figure serves as only a crude guideline for the size of the difference which may be of importance in the comparison of systematic differences in score between raters and evaluation sites.

It is apparent that there were substantial differences in pass/fail classification by raters for the same case. Over all cases, 50% of students failed to meet the criterion pass level of 60% when rated by S.I.U. raters in contrast to 33% when rated by Manitoba raters. The results of this study indicate that raters may bias the estimate of student competence for an individual case. In addition, raters in different evaluation sites may produce systematically different scores for the students evaluated. In order to evaluate this potential problem, differences in the classification of students was evaluated for raters from the two evaluation sites. The results are displayed in Table 8.17.

Agreement in the classification of pass/fail status varied by case from a low of 31% for Case 1 to 100% for Case 11. The distribution of the untied pairs is provided in the right of the table. The null hypothesis being tested is that disagreements in classification are randomly distributed between raters from the two evaluation sites. If this were the case, the proportion of SIU failures/Manitoba passes would be equivalent to the proportion of Manitoba failures/SIU passes. This hypothesis was tested with McNemar's test for paired proportions. After employing Bonferroni's correction for multiple comparisons, statistically significant differences in classification were present in three pairs and two cases. In Case 1, the SIU rater systematically failed more students than the two Manitoba raters. In Case 15, one Manitoba rater systematically failed more students than the SIU rater.

TABLE 8.17 DIFFERENCES IN THE CLASSIFICATION OF PASS/FAIL STATUS BY RATERS FROM DIFFERENT EVALUATION SITES

CASE	PR.	PERCENT AGREEMENT ON PASS/FAIL STATUS	DISTRIBUTION OF THE UNTIED PAIRS	
			Number of Times SIU failed examinee and Manitoba passed	Number of Times Manitoba failed examinee and SIU passed
1	1	31%	26	0*
	2	45%	22	0*
2	1	81%	6	0
	2	77%	4	3
3	1	77%	1	6
4	1	71%	8	1
	2	87%	2	2
5	1	89%	2	10
	1	94%	1	0
11	1	100%	-	-
	2	84%	5	0
13	1	61%	10	4
	2	75%	7	2
14	1	97%	1	0
	2	87%	4	0
15	1	76%	0	10*
	2	83%	2	5

Legend: PR: rater pair (i.e. SIU rater#1 & UofM rater#1 or #2)

*: statistically significant using McNemar's test for paired proportions and Bonferroni's correction for multiple comparisons.

The power of most of these comparisons was compromised by the small number of events. There is an obvious trend, however, for SIU raters to fail more students than Manitoba raters. In 8/10 cases, the SIU rater failed more students then the Manitoba rater. This suggests that the systematic differences noted in encounter scores between the two evaluation sites is also associated with systematic differences in the classification of pass/fail status. The implication of these results for research and evaluation will be discussed in the concluding section of this chapter.

FACTORS INFLUENCING OBSERVER AGREEMENT

In the previous sections it was noted that agreement (corrected for chance) was slight to poor for at least one third of the items rated in the 16 cases included in the study sample. Agreement for the remaining two-thirds of items evaluated was equivalently distributed between fair to perfect. A similar range of agreement was noted when total encounter score was employed as the unit of analysis.

In order to improve agreement between raters, factors which are associated with better or worse agreement for items rated need to be identified. In this study two groups of factors were evaluated: the type of rater pair and rating form factors. In order to develop future guidelines, the contribution that these factors made to variation in item agreement, after missing data has been taken into account, will be estimated.

The relationship of each factor with observed agreement was initially evaluated. Of the five factors evaluated, three were significantly associated with observed item agreement; item type, judgement level and number of items. The mean agreement for each level of the five factors evaluated is summarized in Table 8.18 along with the F statistic R^2 and P-values for the bivariate analysis.

TABLE 8.18 THE RELATIONSHIP BETWEEN RATER PAIR TYPE AND RESPONSE FORM FACTORS AND OBSERVED AGREEMENT: BIVARIATE ANALYSIS

Potential Predictor	N	Mean Agreement (S.D.)	F	R^2	P-value
Rater Pair Type					
Between University	227	80.30 (16.98)	1.33	.4%	.266
Within University	174	82.16 (15.64)			
Within Rater	246	81.16 (19.63)			
Response Form Factors					
A. Judgement Level					
Action Yes/No	491	82.47 (16.58)	6.06	2.7%	.0005
Criteria for Yes	80	77.57 (21.78)			
Action + Quality Rated	22	70.07 (20.89)			
Quality Rated	54	78.68 (17.36)			
B. Item Ambiguity					
One Action/Item	290	82.56 (16.40)	1.85	.5%	.157
Multiple explicit/Item	149	78.00 (21.21)			
Multiple implicit/Item	208	81.36 (16.43)			
C. Item Type					
History	239	80.27 (16.28)	13.57	7.8%	.0001
Physical Exam	200	78.92 (19.86)			
Management	105	90.60 (11.2)			
Teaching	29	77.87 (22.83)			
Communication	74	77.69 (17.06)			
D. Number of Items					
0-5	15	91.33 (8.35)	5.68	.8%	.02
6-10	118	78.18 (16.81)			
11-15	90	78.54 (20.30)			
16-20	153	81.10 (18.49)			
21-25	135	84.04 (15.22)			
>25	136	81.40 (18.08)			

There was no significant differences among the three types of rater pairs for observed agreement on items. The agreement between standardized patients who are trained in the same university site was no better than between patients trained in different sites for the same case. This finding is consistent with the results of the rater reliability studies discussed previously. The within rater estimates of agreement were similarly no better than the between rater estimates when observed agreement was used as the dependent variable. This finding is different than the results found when kappa was used as the index of agreement. It will be recalled that in this analysis, within rater agreement was significantly better than between the two estimates of between rater agreement. The explanation for this difference is likely attributable to different levels of chance agreement which were operating in the within and between rater pairs; chance agreement being smaller in the within rater estimates.

Two response form factors contributed to the explanation of observed agreement: judgement level and item type. No association was found between item ambiguity and agreement. A modest relationship exists between the number of items rated and agreement accounting for .8% of the variance in the dependent variable. Item number was treated as a continuous variable in the analysis. For descriptive purposes, six categories were created and their respective means provided. Inspection of these means suggests that the relationship may be non-linear. The residuals of linear regression analysis were therefore evaluated. There was no evidence of any kind of non-linear trend. The proportion of variance explained by item number was not improved by using a curvilinear model. One of the problems with this variable is that it is strongly associated with case. The lower end of the range (0-5 items) is exclusively contributed by Case 12 where agreement was notably good. With this data set, it is not possible to separate out the effect of number of items from case at the extreme ends of the scale.

With respect to judgement level, the best agreement was achieved when the standardized patient was asked to make a yes/no decision about the presence of an action. Agreement was worse when the standardized patient was asked to evaluate whether specific criteria were present before rating the action as occurring or when both the quality of an action and

compliance with criteria were being rated. This finding may be explained by difficulty in recalling actions specified by the criteria at the end of an encounter with the student. Somewhat better agreement was demonstrated when raters were asked to rate the quality of an action.

In the six categories of item type, the best agreement was demonstrated in the recording of actions taken on management (90.60%) followed by questions which were asked on history during the patient encounter (80.27%). Poorer agreement was demonstrated in the remaining three categories of teaching, physical examination and communication items. There were significant associations between all response form factors including judgement level and item type. For example, criteria were often specified in the recording of physical examination items and criteria and quality were often characteristic of the judgments required for teaching items. In order to distinguish whether it is the content of the item and/or the type of judgement required which most influences agreement, all factors were included in a mixed repeated measures multiple regression model. The results are displayed in Table 8.19.

Since percent missing was previously identified as a confounder in the estimation of predictive factors and agreement, it was included in the model. Partial F statistics were calculated for each potential predictive factor using the formula outlined by Winer(1971) for mixed models. The Type I error was corrected for multiple comparisons using Bonferroni's method from the conventional .05 to .008 (Kleinbaum & Kupper 1978).

With all factors included in the model, 22% of the variance in observed agreement is explained. When all other factors are taken into consideration, item type is the only factor which is significantly associated with item agreement. The same results are obtained for a model which assumes that all factors are fixed and for one which assumes that all factors are random.

Differences in the reliability of raters for different types of items is consistent with the limited literature available in this area. In these studies, rater reliability tends to be the best for history items followed by physical examination and communication (see Chapter 4). No previous

reports of rater reliability could be found for teaching and management items. From this analysis, it would appear that it is primarily the content of the item rather than the type of judgement required which is contributing to the level of agreement. The implications of these findings and the limitations in the study design, on which they are based, will be discussed in the final section of this chapter.

TABLE 8.19 REPEATED MEASURES MULTIPLE REGRESSION ANALYSIS OF ALL PREDICTIVE FACTORS OF ITEM AGREEMENT

Source of Variation	DF	Mean Square	F	R^2	P-Value
All Factors	67	12115.26	2.37	22%	.0001
Error	579	5107.98			
Component Sources			**Partial F**		**P-Value**
Percent Missing	1	2365.68	.463		>.008
Number of Items	1	2401.81	.47		>.008
Rater Pairtype	2	6511.41	1.09		>.008
Item Ambiguity	2	9403.42	1.57		>.008
Judgement Level	3	12390.41	2.07		>.008
Item Type	4	34655.46	5.78		<.008

DISCUSSION AND CONCLUSIONS

RATER RELIABILITY

Sixteen cases and three different types of rater pairs were assessed in this study. The average agreement for items was 81% and the average kappa for items was .45. Rater reliability for total score using the intra-class correlation coefficient ranged from .37 to .42 for different rater types. In most cases, the reliability of raters for total score and the average rating of individual items would be considered to be fair to good using the classification schema of Landis and Koch(1977). In no case did the reliability of rating for the three pair types exceed .8, the conventional minimum for acceptable rater reliability. Contrary to the results of Swanson and Norcini (in press), raters within cases did contribute to the variance in scores arising from the patient encounter in this study. In Swanson and Norcini's studies, raters contributed a little over 1% to total score variance in contrast to a contribution of approximately 20% by raters within cases in this study. The reason for the difference in these two estimates is unknown. It is hypothesized that it may be due to a difference in the quality of raters, training procedure or rating forms used in the two studies. At least in this study, it is concluded that raters contributed in a significant way to measurement error in the evaluation of competence. Improvements in the training procedure for standardized patient raters would be recommended. In this study, raters were provided with an orientation to the rating form. No practice sessions or pre-testing was carried out. This would be recommended in future.

Generally rater reliability for total score was slightly better than the average for individual items. There was considerable variation in the agreement for individual items for all rater pair types. In some evaluation procedures, pass/fail status has been determined for a case (or overall) on the basis of a student's performance on a pre-specified set of critical items. For example, in this study, in Case 15, failure to administer a correct dose of epinephrine to a patient with an anaphylactic reaction resulted in a failure for that case. Although there may be good philosophical grounds for this type of evaluation policy, it would not be recommended until rater reliability for items is improved.

This study evaluated two groups of factors which may influence rater agreement: attributes of the rating form and the type of rater pair. Of the four attributes of the rating form evaluated; item type, judgement level, item ambiguity and number of items; only item type was significantly associated with agreement. The agreement trends in item type are similar to those reported in the literature with rater reliability being better for history items and worse for physical examination and communication items. Agreement for the rating of teaching and management items by standardized patients has not been reported in the English literature. Agreement was particularly poor for teaching items and very good for management items.

There are a number of possible reasons for these differences in agreement for different content areas. It may be more difficult to recall some of the actions taken by the student than others. For example, it may be easier to recall questions asked than examinations done particularly for the medically unsophisticated standardized patient who has no physical findings to present (real or programmed). Alternatively, the kind of judgement the standardized patient must make may be different for different content areas. Judgement level was evaluated in this study and although significant in the bivariate analysis, it did not significantly contribute to the explanation of variance once item type had been taken into account. The limitations in this analysis is that these factors were strongly correlated. For example, most teaching items asked the standardized patient to determine if certain criteria had been met and rate the quality of an action for a Yes response. The poorest agreement was demonstrated in this category of judgement. For practical purposes, it will be important to distinguish whether it is truly the content of the item or the type of judgement the standardized patient is being asked to make which is contributing to poor agreement. Judgement level is amenable to manipulation in the construction of items where as item content is not. Further study of this issue is therefore recommended using a balanced design which evaluates the four levels of judgement for each category of item type.

There were similar limitations in the evaluation of the number of items on

the rating form and item agreement. At the extreme ends of the scale, case was confounded with item number making it impossible to distinguish the effect of case from the number of items on the form. It was also noted that with longer forms (over 20 items) that students rarely carried out actions listed at the end of the form (usually psycho-social). This operatively limits the extent to which agreement can be evaluated for forms which exceed 20 items. Given these limitations, no conclusion can be confidently reached about the effect of the number of items on the response form and rater agreement.

The failure to find the expected trend in item ambiguity was surprising. Agreement was higher for items which had the optimal one action/item. However, a similar level of agreement was found for items which had multiple implicit actions in an item. It may be that most raters perceived these items in the simplest sense as one action whereas worse agreement arose when the various actions which were included in an item were listed (i.e. the multiple explicit category). Alternatively response bias may have inflated the estimate of agreement since significantly more items were not rated in these two categories.

Three types of rater pairs were evaluated: raters who were trained in two university settings to complete the same form (the between university pair type), raters trained together in the same university (the within university pair type) and the same rater rating on two occasions (the within rater pair type). The objective in evaluating the reliability of these three different types of rater pairs was to sort out the extent to which poorer agreement was attributable to different training sites, different patients or inconsistency within the same rater.

No difference was found among these three types of rater pairs for two of the three indices used to summarize agreement: average observed agreement for items and the intra-class correlation coefficient for total score. It was only when average item agreement was corrected for chance that the within rater pair type demonstrated significantly better reliability than the two between rater pairs. This difference was most notable in the rating of communication skills where raters tended to agree with themselves better than with other raters. The acknowledged limitation of

this study is that within rater agreement may be underestimated since the first rating occurred under different circumstances than the second. Despite this limitation, there was substantial variation among different raters in the estimates of within rater reliability. For example the reliability of two raters was above .8 for the intra-class correlation coefficient and kappa. In contrast, six raters had an intra-class correlation coefficient of zero or less and a kappa of .3 or less.

Differences between the reliability of individuals as raters has been noted by Newble(1980) for medical faculty. Newble concluded that there are some individuals who are more reliable raters. Training seemed to have no impact on reliable and unreliable raters. The attributes of those raters who are more reliable were not identified. This study provides no information on the attributes of patients who were more reliable raters from those who were not. The very practical question which can be answered is whether there is a relationship between the reliability of a standardized patient rater and the accuracy with which they present their case. The data from Study 1, Chapter 6 were used to evaluate whether more accurate patients were also more reliable raters. The results are displayed in Table 8.20.

Standardized patients who were more accurate in their presentation of the case tended to be more reliable raters when reliability was calculated for total score (the intra-class correlation coefficient). The Pearson product moment correlation between percent accuracy and the intra-class correlation coefficient was r=.66. The F statistic derived from linear regression was 18.41(p=.0003). The association between average item agreement and patient accuracy was less pronounced. The correlation for observed agreement was r=.41 and for kappa was r=.26. Using linear regression, only observed agreement was significantly associated with patient accuracy (p=.04). It is evident however that patients who had the worst accuracy scores also had the lowest values for item agreement.

TABLE 8.20 THE RELATIONSHIP OF THE STANDARDIZED PATIENT'S ABILITY TO ACCURATELY PRESENT THE PROBLEM AND THEIR RELIABILITY AS A RATER

Patient Accuracy	N	Indices of Reliability		
		Agreement	Kappa	Intra-Class R
Less than 70%	2	75.80	.35	0
70% to 79%	3	72.16	.45	.14
80% to 89%	7	86.34	.67	.31
90% to 99%	11	83.24	.54	.45
100%	3	87.83	.62	.68

Legend: N: the number of patients who had accuracy scores in the respective category

Agreement: average observed agreement for items rated

Kappa: average kappa for items rated

Intra-Class R: the intra-class correlation coefficient for total score produced from rating each encounter

Patient Accuracy: the accuracy with which the standardized patient presented the critical features of the case

The data from this study provide useful guidelines for those responsible for the setting up standardized patient based evaluations. First, rater agreement does not appear to be influenced by differences in trainer or training site. It should be noted that the same method of training raters was used in the two centres involved in this study. This guideline may not hold if different types of training practices are being carried out in different evaluation sites.

Second, patients who are more accurate in their presentation of the clinical problem will also tend to be more reliable raters. Verification of these findings would be recommended so that in future, pre-testing the patient for one of these two desirable properties would suffice.

Third, rater reliability appears to be an important contribution to

measurement error particularly in the rating of teaching, physical examination and teaching items. Pre-testing and training of raters would be strongly recommended for evaluations which include a significant proportion of these items in the evaluation procedure.

SYSTEMATIC BIAS

Systematic bias in the scores provided by different standardized patient raters who are presenting the same case is a problem if:

1) the examinees are being passed or failed on the basis of the performance on one case
2) examinees are not randomly allocated to patients within case or
3) when more than one evaluation site is being used and raters are nested within site.

In this study, both conditions 1 and 3 were operative. This allowed systematic bias between patients and site in encounter score and pass/fail status to be evaluated.

Systematic differences were present in encounter scores were produced by raters from the two universities. The average difference for each encounter rated was 7% lower for Southern Illinois University raters than University of Manitoba raters. The overall score produced for all encounters by S.I.U. raters was 62% in contrast to 67% by Manitoba raters. Both of these differences were statistically significant.

This difference in score resulted in a trend for more students to be failed by S.I.U. raters than by Manitoba raters. It is roughly estimated that a difference in score of 4-5%, in this study, would have resulted in the failure of an additional 4-5 students. Systematic differences of a similar magnitude were noted between the two standardized patient raters from the same university.

Systematic differences between standardized patient raters have been noted by other authors. Dawson-Saunders(1987) reported the presence of systematic differences in student scores between patients trained for the

same case, the source of this data being Southern Illinois University. In a recent report by Schnabl et.al.(1989), systematic differences in student score for different patients presenting the same case were noted in 10 of 16 cases. In Schnabl's study, the impact of these differences was evaluated by correlating the adjusted scores with the unadjusted scores for all students. Correlations of .94 to .98 were found. This would suggest that systematic differences between patients has little impact on the rank ordering of students when students are randomly allocated to patient within case. Systematic differences between patients presenting the same case in the same evaluation site would be problematic if conditions 1 and 2 stated previously were operative.

There are no previous reports of differences in the scores of standardized patients trained for the same case in different evaluation sites. One can only speculate on why raters from Southern Illinois University systematically rated students lower than raters from the University of Manitoba. One possible explanation relates to the relative inexperience of standardized patient raters from Manitoba. It is hypothesized that when a standardized patient rater is inexperienced that he/she may have difficulties in recalling actions taken by the student for their case and when in doubt are more likely to rate the action as having being done. A related hypothesis is that the inexperienced patient rater is more apt to feel as if his/her rating will have some catastrophic impact on the student's future career and is therefore more likely to be lenient in their rating practice. Future studies using centres with different levels of experience would be recommended to provide more data on potential site differences and the impact of different levels of standardized patient experience.

The results of this study raise concerns about the use of multiple evaluation sites for standardized patient-based evaluations of competence. This is particularly true when multiple sites are being used for licensure or certification examinations. Strategies which may be employed to control for systematic differences in scoring by site will be discussed in the final chapter of this thesis.

ABSTRACT

CHAPTER 9

THE USE OF THE STANDARDIZED PATIENT IN THE EVALUATION OF CLINICAL COMPETENCE: CONCLUSIONS

On the basis of the results of Study 1,2, and 3, the following conclusions are drawn about the accuracy of patient presentation and the reliability of standardized patients raters.

1. It is possible for standardized patients to provide an accurate reproduction of the important features of a real patient case. In the majority of patients, however, the theoretical optimum of 100% accuracy was not met.

2. The features of the real patient case which are least accurately presented are the physical findings and affect.

3. Standardized patients trained in a university with no previous experience with this method may be less accurate.

4. Factors which contribute to better patient accuracy include: previous experience as a standardized patient, previous acting experience, personal experience with the health problem or symptoms, a reported understanding of the health problem by the patient, the number of training sessions, the number of training sessions assisted by a physician resource, the number of encounters presented in a day and the number of weeks since training.

5. The accuracy of standardized patient presentation is associated with competence scores in data collection, interpersonal skills and management. The direction and magnitude of this effect appears to vary among different cases.

6. There are differences in the amount of data patients provide spontaneously in equivalent groups of students. This is a potential

source of bias in the estimation of competence score.

7. The reliability of standardized patient raters in this study was below the conventionally accepted minimum for research and evaluation studies of .8 or .9.

8. The reliability of standardized patient raters is not influenced by training location.

9. Standardized patient raters in different universities may bias competence score by virtue of systematic differences in rating.

10. Standardized patients who are more accurate in their presentation of the problem are also more apt to be reliable raters.

11. Better reliability is demonstrated for the rating of patient management items.

On the basis of this research, guidelines for the application of this method are recommended. Future areas of research of the standardized patient method and the evaluation of clinical competence are identified.

The major contributions of this thesis are in:

1. The evaluation of the basic assumptions of the standardized patient method,

2. The identification of empirically-derived factors which could be used to improve patient accuracy,

3. The quantification of potential sources of bias associated with the use of the standardized patient method in multiple evaluation centres and

4. The quantification of the contribution of standardized patient raters to variance in competence score.

CHAPTER 9

THE USE OF THE STANDARDIZED PATIENT IN THE EVALUATION OF CLINICAL COMPETENCE: CONCLUSIONS

AN OVERVIEW OF THE THESIS

Clinical competence was defined in Chapter 1 as the ability to select and carry out appropriate actions in response to a clinical situation. Appropriate actions are those which are important determinants of health outcome for the type of clinical situation presented. These actions have been conventionally defined in five groups: data collection, diagnosis, management, doctor-patient relationship, and professional communication.

The rationale for measuring clinical competence was reviewed in Chapter 1. It is assumed that clinical competence is one of the major determinants of the quality of day to day practice performance. The measurement of clinical competence is therefore used as a means of identifying individuals who may be unsafe or ineffective in practice.

The evidence to support assumptions about the relationship of competence to performance and health outcome was reviewed in Chapter 2. There was no reported study which assessed the relationship of these three factors in the same study population. Other factors which influence practice performance were identified. They include attributes of the clinical situation, the practice setting, the provider and method of remuneration. The relationship of these factors to clinical competence and the relative contribution that each makes to practice performance has not been studied. Further research in this area is recommended. In order to do so, an effective method of measuring clinical competence must be identified.

The methods of measuring clinical competence were reviewed in Chapter 3. The standardized patient was identified as one option which can be used to present the clinical problem and rate clinical actions occurring during the patient encounter. Research related to the reliability and validity of the standardized patient format was reviewed in Chapter 4. Support for the validity of this technique is provided by evidence that the standardized

patient cannot be detected from a real patient in practice and that there is no difference in competence score when measured with a real or standardized patient.

The assumption that the standardized patient is in fact standardized (i.e. the content of the problem presented does not vary from one subject to the next) had not been evaluated until now. This assumption was evaluated in Study 1. Violations of this assumption were observed. The impact these violations had on competence score were therefore evaluated in Study 2. Factors which may be associated with presentation accuracy were also evaluated in Study 2.

In a small number of studies reported by other authors, systematic and random errors in the rating of clinical actions taken during the encounter by standardized patients have been reported. Small sample sizes limit the precision of most of these estimates. Systematic differences in the competence score for patients presenting the same case have been observed in other studies. There is no reported study which has evaluated the reliability of standardized patient raters who have been trained and used in different evaluation centres.Because standardized patients are nested within evaluation site, this issue will need to be evaluated if multiple centres are used in the measurement of clinical competence. These issues were addressed in Study 3. The reliability of standardized patients was evaluated using three types of comparisons: within rater agreement, between rater agreement for raters trained in the same centre and between rater agreement for raters trained in different centres. Systematic differences in encounter rating between raters trained in different centres were also evaluated. Finally factors which may contribute to rater agreement were assessed in order to establish future guidelines for training.

The major conclusions resulting from these three studies will be reviewed in the next two sections. Guidelines for application will be suggested and future areas for research identified.

CONCLUSIONS AND JUSTIFICATION

Conclusion # 1

It is possible for standardized patients to provide an accurate reproduction of the important features of a real patient case. In the majority of patients, however, the theoretical optimum of 100% accuracy was not met.

Justification

In the 839 student-patient encounters and 89 patients in which the accuracy of patient presentation was evaluated in 1987 and 1988, 7 patients in 1987 and 6 patients in 1988 met the theoretical optimum of 100% accuracy (standard deviation =0). The average accuracy with which standardized patients presented the critical features of the clinical problem was 90.2% in 1987 and 93.4% in 1988. The average accuracy of patient presentation was below 75% for 7 patients in 1987 and 1 patient in 1988.

Conclusion #2

The features of the real patient case which are least accurately presented are the physical findings and affect.

Justification

In 1987, errors in the presentation of more than half of the physical finding and affect items were observed in contrast to one quarter of the history items. In 1988, the average accuracy of physical finding and affect presentation was 79% and 90% respectively. In 1988, the average accuracy for presentation of the history was 94%. The generalizability of these findings is limited by the relatively small number of items evaluated in these two categories (physical findings=35 items, affect=27 items).

Conclusion #3

Standardized patients trained in a university with no previous experience with this method may produce patients who are less accurate.

Justification

In 1987, patients trained at the university with past experience with this method had an average accuracy score of 92% in contrast to an average accuracy score of 89% for patients trained in the university with no previous experience. Significant differences in accuracy in favour of the university with past experience were noted in 5 of the 15 cases evaluated. In only 2 of the 15 cases was accuracy score higher for the patients trained at the university with no previous experience. Training problems contributed to errors in presentation in 49% of instances for the university with no experience. Training problems accounted for 26% of errors in the university with experience. After one year of experience, the average accuracy of patient presentation for the novice university was 93%. The observed improvement in the accuracy of patient presentation appears to be attributable to better patient selection and training. The generalizability of this conclusion is limited by the fact that only two universities were studied.

Conclusion #4

Factors which contribute to better patient accuracy include: previous experience as a standardized patient, previous acting experience, personal experience with the health problem or symptoms, a reported understanding of the health problem by the patient, the number of training sessions, the number of training sessions assisted by a physician resource, the number of encounters presented in a day, and the number of weeks since training.

Justification

Patients with previous experience as a standardized patient had better accuracy scores for the presentation of physical findings (83%) and affect (94%) than patients with no previous experience (75% for physical findings and 82% for affect). Patients with previous acting or role-playing experience were more accurate in the presentation of the physical findings (96%) than those with no experience (67%). Those with personal experience with the health problem were particularly better in the presentation of the patient affect (90% vs. 82%). A reported understanding of the health problem was associated with a more accurate presentation of the history (95% vs. 90% for those with a fair understanding) and affect (90% vs 83% for those with a fair understanding).

Patients with two training sessions had higher scores for patient history (95%) than those with one session(89%). Accuracy in the presentation of the affect was 95% for patients with 2 sessions and 81% for patients with 1 session. Accuracy in the presentation of the patient affect was 77% for patients who were trained with no physician assistance. Patients who had physician assistance at one session had a score of 97% and at 2 sessions 100%. Accuracy of physical finding presentation was 73% for patients having 2 sessions with physician assistance and 100% for patients with 3 sessions of physician assistance.

An average accuracy score of 88% was observed for the presentation of physical findings at the beginning of the test day in contrast to 73% at the end of the day (7-10 encounters). The accuracy of patient affect was better in the middle of the test day (92%) than at the beginning (87%) or end of the day (89%). Accuracy in the presentation of the patient affect was worse two weeks after training (86%) than after one (93%).

The limitation in this analysis is that patient attributes were correlated. The relative independent contribution of each of the patient and case-related factors may be biased. Selection factors may have acted to bias the estimates of training length.

Conclusion #5

The accuracy of standardized patient presentation is associated with competence scores in data collection, interpersonal skills and management. The direction and magnitude of this effect appears to vary among different cases.

Justification

A significant negative relationship between patient accuracy and competence score was found for data collection, interpersonal skills and management. This may be explained by two phenomena: the less accurate patient is also a more lenient rater and/or the less accurate patient provides direction about what data should be collected about his/her problem as well as it's diagnosis and management. Alternate methods of rating accuracy and student competence would have to be used to evaluate these hypotheses.

The direction of the relationship was positive in 34% of cases and scores evaluated. The estimated beta for these cases varied from .06% to 2.1%. In 57% of cases and scores evaluated, the relationship between presentation accuracy and competence scores was negative. For these cases, the estimated beta was in the range of -.03% to -4.6%. Small sample sizes limited the precision of estimates for individual cases.

Conclusion #6

There are differences in the amount of data patients provide spontaneously in equivalent groups of students. This is a potential source of bias in the estimation of competence score.

Justification

Significant differences in the percent of data provided spontaneously by patients presenting the same case were noted in 3/15 cases in 1987 and 1/16 cases evaluated in 1988. An independent rater of the student encounter would be required in order to derive an unbiased estimate of the impact of these differences on student score.

GUIDELINES FOR FUTURE APPLICATION

For Single or Multiple Site Standardized Patient-Based Evaluation of Competence

1. In the selection of cases, accuracy of patient presentation does not appear to be adversely influenced by the need to present as many as 30 clinical features. No information is available on presentation accuracy for cases which require more than 30 clinical features to be presented.

2. In the recruitment of individuals for the standardized patient role, those with previous experience in acting and/or role-playing and those with personal experience with health problem or symptoms to be presented should be given priority.

3. Prior to using a standardized patient in the evaluation of competence, they should have at least one supervised opportunity to present their problem. This provides an opportunity for both the patient and trainer to identify problems which the patient is experiencing in the presentation of the problem. Only patients who have had previous experience in standardized patient presentation should be used in the evaluation of competence.

4. Three hours of training divided into two sessions is recommended for training standardized patients for the role they are to present. If more sessions are required because of difficulties encountered by the patient, an alternate patient should be considered or the case discarded as being too difficult.

5. Physician assistance in training should be sought for cases where the presentation of physical findings or a distinct patient affect (eg. pain, anxiety) is required.

6. For cases which only require the presentation of the history, the patient may be scheduled to present at least 10 encounters in a day without adverse effects on presentation accuracy.

7. For cases which require a distinct patient affect to be presented, a 'warm-up' of che patient for their role is recommended at the beginning of each test day.

8. For cases where physical findings or patient affect are required, pre-testing the patient for accuracy in presentation is recommended. The number of encounters the patient is required to present in a day when physical findings are part of the case should likely be reduced from 10. Data from this study do not suggest a maximum number.

9. The protocol used to train standardized patients should identify the key clinical features of the clinical situation to be presented. These clinical features should include all data which may have an impact on the criteria being used to evaluate competence score. The protocol should also specify which clinical features are to be provided spontaneously and which clinical features are to be provided only in response to certain forms of inquiry.

10. Direct observation or videotape surveillance are the only current means of monitoring the accuracy of patient presentation. To be confident that the estimate of standardized patient accuracy is within

5 percentage points of the true mean 95% of the time, a maximum sample size of 89 encounters is required. This estimate is based on the largest standard deviation observed in presentation accuracy (24%). If the standard deviation for all cases in 1987 is employed, a sample size of 23 encounters per patient would be required. An alternate means of monitoring standardized patient accuracy is clearly required.

For Multi-Site Standardized Patient-Based Evaluation of Competence

1. A university which has had no previous experience with standardized patients should only be included as a test site after they have had one opportunity to 'set-up' the evaluation procedure required and receive feedback. If all universities are novice, a 'dry-run' evaluation procedure would be recommended before using the data resulting from the evaluation for academic, licensure or research purposes.

2. When patients are being recruited and trained for the same problem at two or more evaluation sites, a common protocol should be used in the training procedure. In addition, a videotape of the recommended training procedure should be provided with examples of the patient's response in a number of clinician encounters.

RECOMMENDED AREAS OF FUTURE RESEARCH

1. The method of measuring the accuracy of patient presentation should be altered to include:

    1) the conditions in which clinical features of the problem are to be provided;

    2) the measurement of inappropriate data provided by the patient in the relationship to the diagnosis, management or clinical actions to be taken with the problem or the provision of information on other health problems/symptoms which may alter the course of action taken with the problem.

2. In the interest of identifying a standard minimum threshold for the accuracy of patient presentation, clinical features to be presented in a case should be weighted by their relative importance to competence scores. It was evident that errors in the presentation of some clinical features were more important than others for decisions in data collection, diagnosis and management.

3. The relationship between patient accuracy and competence score should be re-evaluated using an independent rater for the evaluation of student actions and the revised method of accuracy score calculation. This question could be addressed using the videotapes sampled for the 1987 and 1988 evaluation. The contribution of patient accuracy relative to cases and students in competence score should then be calculated.

4. Standardized patients included in this study were a select subset of all individuals who may have volunteered or been recruited for the role. There is evidence that the trainer who has one year of experience is better at selecting patients who will be more accurate. The factors which enter into this selection decision need to be elucidated and empirically evaluated. A protocol for standardized patient recruitment and selection can then be established.

5. Verification of the results of these two studies, through replication, is recommended. In view of the difficulties in gaining precise and unbiased estimates of effects in observational studies, an experimental design approach would be advised.

CONCLUSIONS AND JUSTIFICATION

Conclusion #1

The reliability of standardized patient raters in this study was below the conventionally accepted minimum for research and evaluation studies of .8 or .9.

Justification

Rater reliability was studied in 44 standardized patients from 2 universities using a sample of 456 videotaped encounters. Two hundred and fifty-two items were rated, 8-29 items for each case and rater. The average intra-class correlation coefficient among pairs of raters was r=.41. The average kappa for the rating of individual items was k=.45. For 36% of items, agreement was slight to poor (k= 0 to .2). Raters were responsible for 22% of the variance in student score.

Conclusion #2

The reliability of standardized patient raters is not influenced by training location.

Justification

The average intra-class correlation coefficient for raters trained in different universities was r=.42 and the average kappa for item rating was k=.4. For raters trained in the same university the average intra-class correlation coefficient was r=.41 and the average kappa for item rating was k=.4. The major limitation in this study relevant to this conclusion is that only two universities were studied.

Conclusion #3

Standardized patient raters in different universities may bias competence score by virtue of systematic differences in rating.

Justification

Standardized patients from Southern Illinois University rated the same student encounters, on average, 7% lower than standardized patients from the University of Manitoba. This was associated with a trend for standardized patients from Southern Illinois to fail more students on each case than patient raters from Manitoba.

Conclusion #4

Standardized patients who are more accurate in their presentation of the problem are also more apt to be reliable raters.

Justification

There was a significant linear relationship between the intra-class correlation coefficient calculated for within rater agreement and accuracy score for patient presentation. Patients who had accuracy scores of 100% had an intra-class correlation coefficient of r=.68. Patients with accuracy scores less than 70% had an intra-class correlation coefficient of r=.35.

Conclusion #5

Better reliability is demonstrated for the rating of patient management items.

Justification

The content of the item being rated was the only characteristic of the rating form which was associated with the observed agreement for an item. The limitation in this analysis is that all rating form characteristics were strongly correlated. In order to provide unbiased estimates of the effects of judgement type, item number and item ambiguity on observed agreement, an experimental design is recommended.

GUIDELINES FOR FUTURE APPLICATION

For Single and Multiple Site Evaluation of Clinical Competence

1. The number of training sessions for standardized patient raters should likely be increased. Sessions should include opportunities to rate a series of example encounters and review the results of rating discrepancies. Special attention should be devoted to training for rating patient communication skills and teaching.

2. The reliability of standardized patient raters should be pre-tested prior to use in the evaluation procedure.

3. Selecting patients who are more likely to be accurate may result in more consistent standardized patient raters.

4. Since systematic differences in rating exist between different raters presenting the same case, students should be randomly allocated to patients within case and the practice of failing or remediating students on the basis of their performance with one case reconsidered.

For Multiple Site Evaluation of Clinical Competence

1. To control bias created by systematic differences in standardized patient raters in different evaluation sites, a standard protocol for training and rating form use should be developed for each case used in the evaluation. This should be coupled with a standard set of

videotape encounters which are used in the training process. A 'gold standard' for rating these encounters should be established. For the rating of actions on history, physical examination and management this standard should be set by a sample of faculty and/or practitioners. For patient communication and teaching, this standard should be set by a sample of real patients who have the health problem being presented in the case.

2. Pre-testing raters in different settings prior to the implementation of the evaluation would be advised to detect systematic differences in rating. If present, re-training or adjustment in the analysis should be considered.

3. For the written aspects of performance which are measured in each case, systematic differences between faculty raters in different evaluation sites could be avoided by having one common set of raters for all test sites.

RECOMMENDED AREAS OF FUTURE RESEARCH

1. The findings in this study and those in the study of medical faculty by Newble (1980) indicate that some individuals are more reliable raters than others. The characteristics which distinguish these raters are unknown. Future research efforts targeted at identifying these characteristics would be of practical value in the development and application of procedures to measure clinical competence.

2. In this study, attributes of the form used to rate encounters were strongly correlated. A balanced experimental design would allow the independent contribution of each of these factors to rater agreement to be estimated.

3. Evidence to support the validity of the standardized patient's ratings of history and physical examination to that of medical faculty (the 'gold standard') has been reported. Evidence to support the validity of standardized patient rating of patient communication and teaching is required. The appropriate 'gold standard' for these areas would be the ratings of real patients with a similar problem to that being

presented.

4. Factors which may contribute to systematic differences between raters in different sites need to be pursued.

5. The impact of the length of time the patient to rate performance on rater agreement needs to be evaluated in order to establish an optimal climate for rating during the measurement procedure.

6. The reliability of raters who are being used to score the written aspects of each case needs to be evaluated. It is an additional source of case-confounded measurement error which could act to inflate the number of cases required for a stable estimate of clinical competence in the evaluation.

FUTURE AREAS FOR RESEARCH IN CLINICAL COMPETENCE

There are an infinite number of issues which would be valuable to address in the pursuit of a reliable, valid and efficient method of clinical competence measurement. In this section three issues will be highlighted.

The Conceptualization of Competence

In Chapter 1, clinical competence was treated as one attribute which is possessed by the individual to varying degrees for certain types of clinical situations. The measurement of clinical competence is based on this presumption. Factor analysis of clinical competence measures uniformly suggests that there are a number of independent attributes of an individual which are being measured (Maatsch,1983; Arnold,1984; Verhulst,1986; Klass,1988). The relative importance of each of these attributes for the various categories of health outcome is unknown. If a clinical competence measure is being used to draw inferences about individuals who will be ineffective or unsafe in practice, then the relative importance of these attributes to health outcome will need to be understood. An empirical basis for sampling and scoring attributes of competence could then be developed.

Licensure examinations have been used to identify individuals who are apt to be unsafe in professional practice. This inference is based on the

assumption that a single continuum exists. At one end of the continuum, are individuals who may do harm to patients, the outcome being worse than would have occurred through the natural course of the disease. In the middle of the continuum are individuals who are ineffective, they do no harm and no good. The natural course of the patient's illness remains unchanged. At the other end of the continuum are individuals who are effective, the services they render result in improvements over the natural course of the illness. It has been assumed that scores from licensure examinations identify those individuals who will do harm.

This conceptualization of a single continuum of competence is likely too simplistic. Individuals who are more apt to intervene in the management of a situation may be more apt to render a better outcome than the natural course of the illness but may be similarly more apt to do harm. This situation suggests that at least two continuums exist, the probability of doing harm and the probability of doing good. The additional issue of the costs incurred by the provider in rendering service is of growing concern in societies which possess third party payment for health services. Should a costly provider who is ineffective be licensed to practice? A better understanding needs to be gained about the nature of the relationship between these three attributes.

The Measurement of Clinical Competence

The optimal method of specifying the domain of situations from which a sample will be drawn needs further study. A number of attributes of the clinical situation appear to have an influence on provider performance. Of these attributes only age, gender and presenting symptom/diagnosis have been used to characterize the sample frame. The effect of additional attributes of the clinical situation which have been identified in the literature on clinical competence measures requires further study. These attributes include the structure of the presenting problem and the socio-economic status and ethnicity of the patient.

The most appropriate method for specifying the standards of performance which should be expected of a practitioner in a clinical situation needs to be identified. At present, there is no standard protocol for specifying what aspects of competence are to be measured in clinical situations and

how performance standards should be established. Different studies define competence in different ways, measure different groups of abilities and employ different standards of performance in the same situation. There is no commonly accepted nomenclature which allow these differences to be characterized. This limits the comparisons which can be made among studies which, in turn, limits the progress which can be made in our understanding of clinical competence and it's relationships.

Clinical Competence and Its Relationships

The basic assumption which provides the rationale for clinical competence evaluation has not been evaluated. The relationship between competence, performance and health outcome needs to be studied. The relationship of clinical competence to other factors which may influence daily practice performance requires investigation. If the provider's ability to deliver services is a minor determinant of the quality of their daily practice performance, alternate methods of protecting the public from ineffective or harmful providers will need to be identified.

CONCLUSIONS

This thesis provides a significant contribution to our understanding of the measurement properties of the standardized patient method. The results of this thesis provide verification of the assumption that standardized patients can be trained to provide an accurate reproduction of a real patient case. It also provides evidence to disprove the commonly held belief that all patients will be standardized after participation in a conventional training process.

Up until this time, there has been no empirical basis for patient selection, training or use. This thesis contributes the first empirically derived group of factors which can be used to enhance the accuracy of patient presentation. Important attributes of the case, patient, training process, and measurement process were identified.

Sources of bias attributable to standardized patients who are used in the measurement of clinical competence in two or more university settings have not been previously investigated. This thesis provides the first estimate of systematic differences in the presentation and rating of a case by standardized patients who have been trained and used in two university settings. These results have implications for licensing bodies who are considering the use of multi-centre standardized patient-based approaches to the evaluation of competence.

Finally this thesis provides a more precise estimate of the reliability of standardized patient raters. The contribution of raters to score variance is significant. Strategies for improving rater reliability are suggested. Improvements in this area may result in a reduction of the number of cases (or test time) which must be used to gain a stable estimate of competence.

The major advantage of the standardized patient is the ability to control variance attributable to patients in the estimation of competence and performance. This technique provides a means of estimating the relationship between competence and performance with the same problem under a variety of different practice circumstances. Relationships between competence and other determinants of performance can then be studied.

The results of Study 1 and Study 2 indicate that standardized patients can be accurate in their presentation of the problem. Certain individuals are more apt to be accurate than others and certain training practices can improve accuracy. Efforts to improve standardized patient accuracy will result in the availability of a powerful methodological tool which could be used in a variety of research and evaluation areas.

The results from Study 3 indicate that the reliability of standardized patient raters would have to be improved before they are used for research or evaluation. Attention to rater selection, training and pre-testing may resolve this difficulty.

REFERENCES

Adler,K(1977) "Simulated Patients for Communication Research." Journal of Medical Education 52: 151.

Ambulatory Paediatric Association(1984) Educational Guidelines for Training in General/Ambulatory Paediatrics. University of Washington: author.

American Board of Internal Medicine(1979) "Clinical Competence in Internal Medicine." Annals Int.Med. 90: 402.

Anderson,J.L.(1979) "A Practical Approach to Teaching About Communication with Terminal Cancer Patients." Journal of Medical Education 54: 823.

Anderson,K.K. and Meyer,T.C.(1978) "The Use of Instructor-Patients to Teach Physical Examination Techniques." Journal of Medical Education 53: 897.

Andrew,B.(1977) "The Use of Behaviourial Checklists to Assess Physical Examination Skills." Journal of Medical Education 52: 589.

Arnold,D.J.(1970) "Cholecystectomies in Ohio: Results of a Survey of Ohio Hospitals." American Journal Surgery 119: 714.

Arnold,L.,Willoughby,T.L. and Calkins,E.V. (1984) "Understanding the Clinical Performance of Physicians: A Factor Analysis Approach." Journal of Medical Education 59: 590.

Barnes,H.V., Albanese,M., Schroeder,J. and Reiter,S.(1978) "Senior Medical Students Teaching the Basic Skills of History and Physical Examination." Journal of Medical Education 53: 432.

Barrows,H.S., Patek,P.R. and Abrahamson,S.(1968) "Introduction of the Living Human Body in Freshman Gross Anatomy." British Journal of Medical Education 2(1): 33.

Barrows,H.S.(1968) "Simulated Patients in Medical Teaching." Canadian Medical Association Journal 98: 674.

Barrows,H.S.(1971) "Simulated Patients." Springfield,Illinois: Charles C.Thomas.

Barrows, H.S. and Mitchell, D.M.(1975) "An Innovative Course in Undergraduate in Neuroscience: Experiment in Problem-Based Learning with "Problem Boxes". Br. J. Med. Ed. 9(4): 223.

Barrows,H.S.,Neufeld,V.R.,Norman ' .R. and Feightner,J.W.(1978) An Analysis of the Clinical Methods of Medical Students and Physicians: Final Report to the Ontario Ministry of Health. Toronto.

Barrows,H.S. and Tamblyn,R.M.(1980) "Problem-Based Learning: An Approach to Medical Education." New York: Springer Publishing Co.
Barrows,H.S.(1987) "Simulated (Standardized) Patients and Other Human Simulations." Chapel Hill,North Carolina: Health Sciences Consortium.

Bashook,P.G. and Lloyd,J.S.(1985) "A Topological Paradigm of Physician Performance and Competence." Proceedings of the Annual Conference on Research in Medical Education: 272.

Battista,R.N.,Williams,J.I. and McFarlane,L.A.(1986) "Determinants of Primary Care Medical Practice in Adult Cancer Prevention." Medical Care 24(3): 216.

Behrens,A., Barnes,H.V., Gerber,W.L., Albanese,M., Matthes,S. and Cangelosi,A. (1979) "A Model for Teaching Sophmore Medical Students the Essentials of the Male Genital-Rectal Examination." Journal of Medical Education 54: 585.

Berard,M.(1989) Personal Communication, Medical Council of Canada, Ottawa.

Bergman,D.A. and Beck,A.L.(1986) "The Impact of Clinical Appearance on Paediatric Residents' Assessment of the Febrile Infant." Proceedings of the Annual Conference of Research in Medical Education: 135.

Berwick,D.M. and Thibodeau,L.A.(1983) "Receiver Operating Characteristic Analysis of Diagnostic Skill." Medical Care 21(9): 876.

Bland,J.M. & Altman,D.G. (1986) "Statistical Methods for Assessing Agreement Between Two Methods of Measurement". The Lancet February 8: 307.

Bloom,B.S.(1956) A Taxonomy of Educational Objectives: Handbook 1-The Cognitive Domain. New York: Longmans,Green.

Bloom,B.[ed](1956) Taxonomy of Educational Objectives: Cognitive Domain. New York: David MacKay.

Bloom,B.S. and Peterson,O.L.(1973) "End Results,Costs and Productivity of Coronary Care Units." New England J.Med. 288: 72.

Bordage,G.(1982) "The Etiology of Diagnostic Errors:Process or Content? An Exploratory Study." Proceedings of the Annual Conference of Research in Medical Education: 171.

Brook,R.H. and Williams,K.N.(1976) "Evaluation of the New Mexico Peer Review System." Medical Care 14 (Suppl): 1.

Broski,D.,Alexander,D.,Brunner,M.,Chidely,M.,Finney,W.,Johnson,C., Karas,B. and Rothenberg,S.(1977) "Competency-Based Curriculum Development: A Pragmatic Approach." Journal of Allied Health, Winter: 38.

Burg,F.D.(1982) "Documentation of Continuing Competence in Paediatrics: Why and How." Adv. Paediatrics 29: 369.

Burri,A, McCaughan,K., Barrows,H.(1976) "The Feasibility of Using the Simulated Patient to Evaluate Clinical Competence of Practising Physicians in a Community." Proceedings Annual Conference of Research in Medical Education 15: 295.

Butterworth,J.A. and Reppart,E.H.(1960) "Auscultatory Acumen in the General Medical Population." Journal American Medical Association 174: 114.

Callaway,S.,Bosshart,D.A. and O'Donell,A.A.(1977) "Patient Educators in Teaching Patient Education Skills to Family Practice Residents." The Journal of Family Practice 4(4): 709.

Carmines,E. & Zeller.R.(1979) Reliability and Validity Assessment. In: Sullivan(Ed) Quantitative Applications in the Social Sciences. Beverley Hills: Sage University Paper.

Carroll,J.G., Schwartz,M.W. and Ludwig,S.(1981) "An Evaluation of Simulated Patients as Instructors for Teaching Medical Interviewing Skills". Journal of Medical Education 56: 522.

Carroll,J.G. and Hutchins,E.B.(1978) "Teaching Interviewing Skills to Medical Students: An Integrated Approach". Paper presentation Annual Meeting American Psychological Association. August 31, Toronto.

Cherkin,D.C.,Rosenblatt,R.A.,Hart.L.G.,Schneeweiss,R. and Logerfo,J.(1987) "The Use of Medical Resources By Residency-Trained Family Physicians and General Internists." Medical Care 25(6): 455.

Chong,J.P.,Neufeld,V.R.,Oates,M.J. and Secord,M.(1984) "The Selection of Priority Problems and Conditions." Annual Conference Research in Medical Education: 17.

Clute,K.F. (1963) The General Practitioner:A Study of Medical Education and Practice in Ontario and Nova Scotia. Toronto: Toronto University Press.

Cohen,J.(1960) "A Coefficient of Agreement for Nominal Scales". Educational and Psychological Measurement 20(1): 37.

Cohen,S.J.,Weinberger,M.,Hui,S.L.,Tierney,W.M. and McDonald,C.J.(1985) "The Impact of Reading on Physicians' Nonadherence to Recommended Standards of Medical Care." Soc.Sci.Med. 21(8): 909.

Cohen,R., Rothman,A., Ross,J., Keystone,J., Kulesha,D. MacInnes,A., McLeary,P. et.al. (1988) " A Comprehensive Assessment of Graduates From Foreign Medical Schools". Internal Report, University of Toronto.

Cohn,K.H.(1985) "Misadventures in Surgical Residency: Analysis of Mistakes During Training." Current Surgery July-August:278.

Committee on Health Care Resources in the Veterans Administration(1977) Assembly of Life Sciences,National Research Council: Health Care for American Veterans. Washington,D.C.: National Academy of Sciences.

CFPC (College of Family Physicians of Canada)(1974) Educational Objectives for Certification in Family Medicine. Toronto:author.

College of Family Physicians of Canada(1981) Educational Objectives for Certification in Family Medicine. Toronto,Ont.:author.

Colton,T.(1974) Statistics in Medicine. Boston: Little, Brown and Co.

Coulehan,J.H.(1984) "Dissecting the Clinical Art." The Pharos Fall:21.

CREOG (Council on Resident Education in Obstetrics and Gynecology) (1980) Educational Objectives for Residents in Obstetrics and Gynecology (second edition) Chicago,Ill.: author.

Cronbach,L.J.(1970) Essentials of Psychological Testing (third edition) New York: Harper & Row.

Curry,L.(1985) "Postgraduate Training Route and Content of Subsequent Practice." Can. Family Physician 31: 1417.

DaRosa,D.A., Mazur,J. and Markus,J.(1982) "Assessing Patient Evaluation Skills: A Structured Vs. A Traditional Approach". Journal of Medical Education 57: 472.

Dawson-Saunders,B., Verhulst,S.J., Marcy,M. and Steward,D.E.(1987) "Variability in Standardized Patients and It's Effect on Student Performance." Proceedings International Conference: Further Developments in Assessing Clinical Competence. Ottawa,Can. June 27-30.

Dawson-Saunders,B.,Mast,T.A.,Finch,W.T.,Konrad,H.R. and False,J.R.(1984) "Content Knowledge and Problem-Solving Skill in Reviewing Medical Charts." Medical Education 18: 31.

D'Costa,A.(1986) "The Validity of Credentialing Examinations." Evaluation and the Health Professions 9(2): 137.

De Graff,E.,Post,G.J.and Drop,M.J.(1987) "Validation of a New Measure of Clinical Problem-Solving." Medical Education 21: 213.

Des Raj (1972) The Design of Sample Surveys. New York: McGraw-Hill Inc.

DiMatteo,M.R. and Hays,R.(1980) "The Significance of Patients' Perceptions of Physician Conduct." J.Community Health 6: 18.

DiMatteo,M.R.,Prince,I.M. and Taranta,A.(1979) "Patients' Perceptions of Physicians' Behaviour:Determinants of Patients' Commitment to the Therapeutic Relationship." J. Community Health 4: 280.

Donabedian,A.,Wheeler,J.P. and Wyszewianski,L.(1982) "Quality,Cost and Health; An Integrative Model." Medical Care 20(10): 975.

Duff,R.S.,Cook,C.D.Wanerka,G.R.,Rowe,D.S. and Dolan,T.F.(1972) "Uses of Utilization Review to Assess the Quality of Paediatric Inpatient Care." Paediatrics 49: 169.

Eisele,C.W.,Slee,V.N. and Hoffman R.G.(1956) "Can the Practice of Internal Medicine be Evaluated?" Ann. Intern. Med. 44: 144.

Eisenberg,J.,Mackie,A.,Kahn,L. and Perkoff,G.T.(1974) "Patterns of Paediatric Practice by the Same Physician in a Pre-Paid and Fee-For-Service Setting." Clin. Paediatric. 13: 352.

Eisenberg,J.M.(1986) Doctor's Decisions and the Cost of Medical Care." Ann Arbor, Michigan: Health Administration Press Perspectives.

Elstein,A.S.,Schulman,L.S. and Sprafka,S.A.(1978) Medical Problem-Solving: An Analysis of Clinical Reasoning. Cambridge: Harvard University Press.

Engel,W.,Seime,R.,Powell,V. and D'Alessandri,R.(1987) "Clinical Performance of Interns After Being On Call." Southern Medical Journal 80(6):761.

Engel,I.M., Resnick,P.J. and Levine,S.B.(1976) "The Use of Programmed Patients and Videotape in Teaching Medical Students to Take a Sexual History." Journal of Medical Education 51: 425.

Epstein,A.M. and McNeil,B.J.(1985) "The Effects of Patient Characteristics on Ambulatory Test Ordering." Soc.Sci.Med. 21(10): 1071.

Erviti,V.F.,Templeton,B.,Bunce,J.V. and Burg,F.D.(1980) "The Relationships of the Paediatric Resident Recording Behaviour Across Medical Conditions." Medical Care 28(10): 1020.

Evans,S. and Curtis,P.(1983) "Using Patient Simulators to Teach Telephone Communication Skills to Health Professionals". Journal of Medical Education 58: 894.

Evans,R.G.(1984) Strained Mercy:The Economics of Canadian Health Care. Toronto: Butterworths.

Evans,G.E.,Haynes,R.B.,Birkett,N.J.,Gilbert,R.,Taylor,D.W.,Sackett,D.L., Johnston M.E. and Hewson,S.A.(1986) "Does a Mailed Continuing Education Program Improve Physician Performance?" J.A.M.A. 255(4): 501.

Falvo,D.R. and Smith,J.K.(1983) "Assessing Residents' Behaviourial Science Skills: Patients' Views of the Physician-Patient Interaction." Journal Family Practice 17(3): 479.

Feightner,J.W. and Norman,G.R.(1978) "Computer-Based Problems as a Measure of the Problem-Solving Process-Some Concerns about Validity." Annual Conference on Research in Medical Education.

Fernow,L.C.,Mackie,C.,McColl,I. and Rendall,M.(1978) "The Effect of Problem-Oriented Medical Records on Clinical Management Controlled for Patient Risks." Medical Care 26(6): 476.

Fleiss,J.L.(1981) Statistical Methods for Rates and Proportions. New York: John Wiley & Sons.
Fletcher.R.H.,Fletcher,S.W. and Wagner,E.H.(1982) Clinical Epidemiology--the Essentials Baltimore,MD: Williams & Wilkins.

Frazer,N.B. and Miller,R.H.(1977) "Training Practical Instructors (Programmed Patients) to Teach Basic Physical Examination." Journal of Medical Education 52: 149.

Frederiksen,N.(1984) "Implications of Cognitive Theory for Instructions in Problem Solving." Review of Educational Research 54(3): 363.

Froelich,R.E.(1969) "A Course in Medical Interviewing". Journal of Medical Education 44: 1165.

General Medical Council(1984) "Serious Professional Misconduct-Points from the GMC Working Party." Lancet,November 10: 1105.

Georgopoulus,B. and Mann,F.(1962) The Community General Hospital. New York: McMillan.

Gerritsma,J.G. and Smal,J.A.(1986) "Decision-Making of Internists and Family Physicians in the Netherlands." Annual Conference of Research in Medical Education: 197.

Gliva,G(1980) Personal Communication. Director Simulated Patient Program, McMaster University, Hamilton.

Godkins,T.R., Duffy,D., Greenwood,J. and Stanhope,W.D.(1974) "Utilization of Simulated Patients to Teach the 'Routine' Pelvic Examination." Journal of Medical Education 49: 1174.

Gonnella,J.S.,and Hojat,M.(1983) "Relationship Between Performance in Medical School and Postgraduate Competence." Journal of Medical Education 58: 679.

Goran,M.J.,Williamson,J.W. and Gonnella,J.S.(1973) "The Validity of Patient Management Problems." Journal Medical Education 48: 171.

Graham,J.B. and Paloucek,F.P.(1963) "Where should Cancer of the Cervix be Treated?" American J. Obstetrics and Gynecology 87: 405.

Greenberg,L.W. and Jewett,L.S.(1985) "The Impact of Two Teaching Techniques on Physicians' Knowledge and Performance." Journal of Medical Education 60: 390.

Greenfield,S.,Nadler,M.A.,Morgan,M.T. and Shine,K.I.(1977) "The Clinical Investigation and Management of Chest Pain in the Emergency Department: Quality Assessment by Criteria Mapping." Medical Care 15: 898.

Greene,R.(1976) Assuring Quality in Medical Care: The State of the Art. Cambridge,Mass.: Ballinger.

Haertel,E.(1986) "The Valid Use of Student Performance Measures for Teacher Evaluation." Educational Evaluation and Policy Analysis 8(1): 45.

Hannay,D.R.(1980) "Teaching Interviewing With Simulated Patients". Medical Education 14: 246.

Harasym,P.,Baumber,J.,Bryant,H.,Fundytus,D.,Preshaw,R.,Watanabe,M. and Wyse,G. (1980) "An Evaluation of the Clinical Problem-Solving Process Using Simulation Technique." Medical Education 14: 381.

Harden,R., Stevneson,M., Downie, W. and Wilson, G.(1975) "Assessment of Clinical Competence Using Objective Structured Examinations". British Med. J. 1: 447.

Held,P.,Lindberg,B. and Swedberg,K.(1984) "Audibility of Third Heart Sound in Relation to its Frequency, Amplitude, Delay form the Second Heart Sound and Experience of the Observer." Amer. J. Cardiology 53: 1169.

Hemenway,D. and Fallow,D.(1985) "Testing for Physician Induced Demand with Hypothetical Cases." Medical Care 23(4): 344.

Hennan,B.(1984) "Measuring the Complexity of Clinical Problems." Journal of Medical Education 39: 488.

Holmes,S.(1986) "Comments on this Special Issue." Evaluation and the Health Professions 9(2): 131.

Holzman,G.B., Singleton,D., Holmes,T.F. and Maatsch,J.L.(1977) "Initial Pelvic Examination Instruction: The Effectiveness of Three Contemporary Approaches." American Journal Obstetrics and Gynecology September 15: 124.

Ho Ping Kong,P.H.H.(chair),Neufeld,V.R.,Hart,I.and Dauphinee,D.(1987) "Symposium:The Evaluation Of Clinical Competence." Annals RCPSC 20(5): 361.

Houston,W.R and Warner,A.R.(1977) "The Competency-Based Movement: Origins and Future." Educational Technology June: 14.

Hulka,B.S.,Kupper,L.L. and Cassel,J.C.(1976) "Physician Management in Primary Care." Amer, J. Public Health 66: 1173.

Hull,J.(1979) "Factors Influencing Styles of Medical Practice." Medical Care 27(7): 718.

Jason,H., Kagan,N., Werner,A. Elstein,A.S. and Thomas,J.B.(1971) "New Approaches to Teaching Basic Interviewing Skills to Medical Students". American Journal Psychiatry 127(10): 1404.

Jennett,P.A.,Laxdal,O.E.,Hayton,R.C.,Klaasen,D.J.,Swanson,R.W., Wilson,T.W.,Spooner,J.,Mainprize,G.W. and Wickett,R.E.(1988) "The Effects of C.M.E. upon Family Physician Performance in Office Practice: A Randomized Controlled Study." accepted for publication in Medical Education 22.

Kagan,N.Krathwol,D.R.,Goldberg,A.D. et.al.(1967) Studies in Human Interaction. East Lansing Publication Services: Michigan State University.

Kahn,L.,Wirth,P. and Turner,J.K.(1977) "The Influence of a Change in Practice Setting on Paediatric Activity:A Case Study." Paediatrics 59: 69.

Kane,M.(1982) "The Validity of Licensure Examinations" American Psychologist 37(8): 911.

Kane,M.(1987) "Is Predictive Validity the Gold Standard or Is It the Holy Grail of Examinations in the Professions". Invited Address, American Educational Research Association. Washington.

Kennedy,M.M.(1987) "Inexact Sciences:Professional Education and the Development of Expertise ." In: Rothkopf, E.Z.(ed.) Review of Research in Education 14. Washington: American Educational Research Association.

Kerr,M.G., Templeton,A.A. and Parboosingh,J.(1977) "Simulated Patients as a Learning Resource in the Study of Reproductive Medicine." Medical Education 11: 374.

Klass, D.J.(1989) Personal Communication. Associate Dean Undergraduate Education, Faculty of Medicine, University of Manitoba, Winnipeg.

Klass,D.J., Kopelow,M., Hassard,T., Tamblyn,R.M. & Schnabl,G.(1988) The Comprehensive Clinical Examination: An Evaluation of Two Years of Inter-University Results." Annual Conference of Research in Medical Education: 324.

Kleinbaum,D. & Kupper,L.(1978) Applied Regression Analysis and Other Multivariable Methods. Massachusetts. Duxbury Press.

Kroboth,F.J., Kapoor,W., Brown,F.H., Karpf,M. and Levey,G.S.(1985) "A Comparative Trial of the Clinical Evaluation Exercise." Arch. Intern. Med. 145: 1121.

Kroencke,K., Hanley,J.F., Copley,J.B., Matthews,J.I., Davis,C.E., Foulks,C.J and Carpenter,J.L.(1987) "Improving House Staff Ordering of Three Common Laboratory Tests." Medical Care 25(10): 928.

Kuder,J.M., Vilmain,J.A. and Demlo,L.K.(1987) "Exploring Physician response to Patients' Extramedical Characteristics." Medical Care 25(9): 882.

IaDuca,A., Taylor,D.D. and Hill,I.K.(1984) "The Design of a New Licensure Examination." Evaluation and the Health Professions 7(2):115.

Lamont,C.T. and Hennan,B.K.(1972) "The Use of Simulated Patients in the Certification Examination in Family Medicine." Journal of Medical Education 47:789.

Landis,J.R. & Koch,G.G.(1977) "The Measurement of Observer Agreement for Categorical Data" Biometrics 33:159.

Larsson,U.S., Saljo,R. and Aronsson,K.(1987) "Patient Doctor Communication on Smoking and Drinking:Lifestyle in Medical Consultations." Soc. Sci.Med. 25(10): 1129.

LaSor,B.(1979) "The Use of Simulation in Teaching Psychiatric Nursing." Canadian Nurse October:36-338.

Leff,M., Moore,V.M., Zakus,G., Whiteside,V. and Rotnem.D.(1979) "Interviewing the Adolescent Patient: An Educational Program for Health Professionals". Journal of Medical Education 54:899.

Lewis,C.E. and Hassanein,R.S.(1970) "Continuing Medical Education: An Epidemiological Evaluation." New England Journal of Medicine 285:254.

Levenkron,J., Greenland,P. and Bowley,N.(1987) "Using Patient Instructors to Teach Behavicurial Counselling Skills." Journal of Medical Education 62: 665.

Lichstein,P.R. and Nieman,L.Z.(1985) "Diagnosing Technique Problems in Interviewing Patients". Journal of Medical Education 60: 566.

Lincoln,R., Layton,J. and Holdman,H.(1978) "Using Simulated Patients to Teach Assessment." Nursing Outlook May: 316.

Lindsay,M.I., Hermans,P.E., Nobrega,F.T. and Ilstrup,D.M.(1976) "Quality of Care I Outpatient Management of Acute Bacterial Cystitis as the Model". Mayo Clin. Proc. 51: 307.

Lindsay,M.I., Nobrega,F.T., Offord,K.P., Carter,E.T., Rutherford,B.D., Kennel,A.J. and Mankin,H.T.(1977) "Quality of Care Assessment II. Outpatient Medical Care Following Hospital Dismissal After Myocardial Infarction." Mayo Clin. Proc. 52: 220.

Lockyer,J.M., Parboosingh,J.T., McDougall,G.M. and Chugh,U.(1985) "How Physicians Integrate Advances into Clinical Practice." Mobius 5(2): 5.

Lomas,J. and Haynes,R.B.(1988) "A Taxonomy and Critical Review of Tested Strategies for the Application of Clinical Practice Recommendations: From "Official to Individual" Clinical Policy." Battista,R.N. and Lawrence,R.S.(eds) Implementing Preventive Services. Amer.J.Prev.Med. (suppl) 4(4):77.

Love,D.W., Wiese,J. Henson,R.E. and Parker,C.L.(1978) "Teaching Interviewing Skills to Pharmacy Residents". American Journal of Hospital Pharmacy 35: 1073.

Luft,H.S.(1981) Health Maintenance Organizations: Dimensions of Performance. New York: Wiley Inter-Science.

Maatsch,J.L., Huang,R., Downing,S.M. and Barker,D.(1983) Predictive Validity of Medical Speciality Examinations-A Final Report for Nat'l Cen. Health Serv.Res.(grant#HS 02038-04) Michigan State University: Office of Medical Education Research and Development.

Maguire,G.P., Clarke,D. and Jolley,B.(1977) "An Experimental Comparison of Three Courses in History-Taking Skills for Medical Students." Medical Education 11: 175.

Maguire,P.(1976) "The Use of Patient Simulation in Training Medical Students in History-Taking Skills." Medical and Biological Illustration 26: 91.

Mandelbaum,L.(1987) "Periodic Competency Review: An Assessment and Policy Model." Evaluation and the Health Professions 10(3): 342.

MARN(Manitoba Association of Registered Nurses)(1984) Entry to Practice Report. Winnipeg, Manitoba: author.

Markham,B., Gessner,T., Warburton,S.W. and Sadler,G.(1979) "Medical Students Become Patients: A New Teaching Strategy." Journal of Medical Education 54: 416.

McAuliffe,W.E.(1978) "Studies of Process-Outcome Correlations in Medical Care Evaluations: A Critique." Medical Care 26(11): 907.

McGuire, C.H. and Babbott,D.(1967) "Simulation Technique in the Measurement of Problem-Solving Skills". Journal of Instructional Measurement 4(1).

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

Morehead, M.A. and [illegible] [illegible] A Study of the Quality of Hospital Care Secured by a Sample of Teamster Families in New York City. New York: Columbia University School of Public Health and Administrative [illegible]

Morehead, M.A. [illegible] Quality of Medical Care Provided by Family Physicians as Related to Their Education, Training and Methods of Practice. New York: Health Insurance Plan of [illegible]

Morehead, M.A. and Donaldson R. 1974 "Quality of Clinical Management of Disease in Comprehensive Neighborhood Health Centers." Medical Care 12: 301.

Morehead, M.A., Donaldson R. and Seravalli M.R. 1971 "Comparisons Between OEO Neighborhood Health Centers and other Health Providers." Amer. J. Public Health 61: 1294.

Mushlin, A.I. and Appel, F.A. 1976 "Extramedical factors in the Decision to Hospitalize Medical Patients." Amer. J. Public Health 66(2): 170.

Muzzin, L.J. and Hart, L. (1985) "Oral Examinations." Neufeld, V.R. and Norman, G.R. (eds) Assessing Clinical Competence. New York: Springer.

Naftulin, D.H. and Andrew, B.J. (1975) "The Effects of Patient Simulation on Actors." Journal of Medical Education 50: 87.

Nelman, L.Z., Vernon, M.S., Holbert, D. and Boyett, L. (1988) "Training and Validating the Use of Geriatric Simulated Patients". Annual Conference on Research in Medical Education: 154.

Neufeld,V.R.(1985) "Written Examinations(chapter 6)." Assessing Clinical Competence,Neufeld,V.R. and Norman,G.R.(eds). New York: Springer.

Newble,D.I., Hoare,J. and Sheldrake,P.F.(1980) "The Selection and Training of Examiners for Clinical Examinations." Medical Education 14: 345.

Newble,D.(1988) "Eight Years' Experience with a Structured Clinical Examination." Medical Education 22: 200.

Newble,D., Elmslie,R. and Baxter,A.(1978) "A Problem-based Criterion—Referenced Examination of Clinical Competence." Journal of Medical Education 53: 720.

Newble,D., Hoare,J. and Elmslie,R.(1981) "The Validity and Reliability of a New Examination of the Clinical Competence of Medical Students." Medical Education 17: 165.

Newble,D.I. and Swanson,D.B.(1988) "Psychometric Characteristics of the Objective Structured Clinical Examination." Medical Education (in press).

Nightengale,S.D.(1988) "Risk Preference and Admitting Rates of Emergency Room Physicians." Medical Care 26(1): 84.

Nobrega,F.T., Morrow,G.W., Smoldt,R.K. and Offord,K.P.(1977) "Quality Assessment in Hypertension: Analysis of Process and Outcome Methods." New England Journal of Medicine 296(3): 145.

Norcini,J.J., Swanson,D.B., Grosso,L.J. and Webster,G.D.(1985) "The Reliability, Validity and Efficiency of Multiple Choice Question and Patient Management Item Formats in the Assessment of Clinical Competence". Medical Education 19: 238.

Norcini,J.J., Webster,G.D., Grosso,L.J., Blank,L.L. and Benson,J.A.(1987) "Ratings of Residents Clinical Competence and Performance on Certification Examination." Journal of Medical Education 62: 457.

Norcini,J.J., Meskauskas,J.A., Langdon,L.O. and Webster,G.D.(1986)"An Evaluation of a Computer Simulation in the Assessment of Clinical Competence." Evaluation and the Health Professions 9(3): 286.

Norcini,J.J., Lipner,R.S., Benson,J.A. and Webster,G.D.(1985) "An Analysis of the Knowledge Base of Practising Internists as Measured by the 1980 Recertification Examination." Annals of Internal Medicine 102: 385.

Norcini,J.J.(1988) Personal Communication. American Board of Internal Medicine, Philadelphia.

Norman,G.R., Feightner,J.W., Tugwell,P., Muzzin,L.J., and Guyatt,G.(1983) "The Generalizability of Measures of Clinical Problem-Solving." R.I.M.E. Proceedings, A.A.M.C. meeting, Washington, November: 110.

Norman,G.R., Neufeld,V.R., Walsh,A., Woodward,C.A. and McConvey,G.A.(1985) "Measuring Physician Performance by Using Simulated Patients." Journal of Medical Education 60: 925.

Norman,G.R., Brooks,L.R., Allen,S.W. and Rosenthal,D.(1988) "Retrieval Factors in Medical Expertise: Improvement Independent of Stable Knowledge-." Conference, American Educational Research Assoc., New Orleans, March.

Norman,G.R., Tugwell,P. and Feightner,J.W.(1982) "A Comparison Of Resident Performance on Real and Simulated Patients." Journal of Medical Education 57: 708.

Norman,G.R., Williams,L. and Swanson,D.(1985) "Simulation in Health Sciences Education." Journal of Instructional Development 11.

Norman, G.R.(1988) "Problem-solving skills, Solving Problems and Problem--based Learning". Medical Education 22: 279.

Nowotony,R.E. and Grove,D.I.(1982) "Description of an Examination for the Objective Assessment of History-Taking." Medical Education 16: 259.

Nunnally,J.(1978) Psychometric Theory. New York: McGraw Hill Book Co.

Owen,A. and Winkler,R.(1974) "General Practitioners and Psychosocial Problems: An Evaluation Using Pseudo-Patients." Medical Journal of Australia 2: 393.

Page,G.G.and Fielding,D.W.(1980) "Performance on PMP's and Performance in Practice: Are They Related?" Journal of Medical Education 55: 529.

Palmer.R.H.(1976) "Definitions and Data." In: Greene,R.(ed) Assessing Quality in Medical care: The State of the Art. Cambridge, Mass.: Ballinger.

Palmer.R.H. and Reilly,M.C.(1979) "Individual and Institutional Variables which may serve as Indicators of Quality of Care." Medical Care 27(7):693.

Payne,B.C., Lyons.T.F. and Neuhaus,E.(1984) "Relationships of Physician Characteristics to Performance Quality and Improvement." Health Services Research 19(3):307.

Payne,B.C. and Lyons,T.F.(1972) Method of Evaluating and Improving Personal Medical Care Quality:Episode of Illness Study and Office Care Study. Ann Arbor:University of Michigan School of Medicine.

Peterson.O.L., Andrew,L.P., Spain,R.S. and Greenberg,B.G.(1956) "An Analytic Study of North Carolina Medical Practice." Journal of Medical Education 31(Part 2): 31.

Peterson,O.L. and Barsamian,E.M.(1976) Medical Care Chart Book(sixth ed.) Ann Arbor: University of Michigan.

Petrusa,E., Blackwell,L., Rogers,C., Saydjari,C., Parcel,S. and Guckian,J.(1987) "An Objective Measure of Clinical Competence." American Journal of Medicine 83: 34.

Petrusa,E.(1988) "Collaborative Project to Improve the Evaluation of Clinical Competence." Final Report to the National Fund for Medical Education.

Pierce,J.C. and Downing,S.M.(1982) "The Evolution of Competency in Internal Medicine:The Validity of a Clinical Performance Measure." Annual Conference on Research in Medical Education: 205.

Pierce,J.C. and Downing,S.M.(1982) "The Evolution of Competency in Internal Medicine:The Validity of a Clinical Performance Measure." Annual Conference on Research in Medical Education:105.

Piscano,N.J.,Veloski,J.J.,Brucker,P.C. and Gonnela,J.S.(1986) "Defining the Content of Board Certification Examinations." Annual Conference on Research in Medical Education:205.

Platt,F.W. and McMath,J.C.(1979) "Clinical Hypocompetence:The Interview." Annals Internal Medicine 91: 898.

Popham,W.J.(1978) "Measurement Requisites for Competency Assurance in the Health Professions." Evaluation and the Health Professions 1(1):9.

Putnam,R.W. and Curry,I.(1985) "Impact of Physician Care Appraisal on Physician Behaviour in the Office Setting." Can.Med.Assoc.J. 132: 1025.

Quirk,M. and Letendre,A.(1986) "Teaching Communication Skills to First Year Medical Students". Journal of Medical Education 61: 603.

Ramsay,D.L. and Benimoff,F.(1981) "The Ability of Primary Care Physicians to Recognize the Common Dermatoses." Arch.Dermatology 117: 620.

Renaud,M., Beauchemin,J., LaLonde,C., Poirier,H. and Berthiaume,S.(1980) "Practice Settings and Prescribing Profiles: The Simulation of Tension Headaches to General Practitioners Working in Different Practice Settings in the Montreal Area." Amer. J. Public Health 70(10): 1068.

Rethans,J.J.E. and van Boven,C.P.A.(1987) "Simulated Patients in general Practice: A Different Look at the Consultation." British Medical J. 294::809.

Rhee,S.(1976) "Factors Determining the Quality of Physician Performance in Patient Care." Medical Care 14: 733.

Rhee,S.(1977) "Relative Importance of Physicians Personal and Situational Characteristics for the Quality of Medical Care." Journal Health Social Behaviour 18: 10.

Rhee,S.(1986) "U.S. Graduates vs. Foreign Graduates:Are there Differences in Practice Performance?" Medical Care 24(3): 248.

Rhee,S.,Lyons,T.F. and Payne,B.C.(1979) "Patient Race and Physician Performances: Quality of Medical Care, Hospital Admissions and Patient Stay." Medical Care 27(7): 737.

Rhee,S., Luke,R.D. and Culverwell,M.B.(1980) "Influence of Client/Colleague Dependence on Physician Performance in Patient Care." Medical Care 28(8): 829.

Robb,K. and Rothman,A.(1985) "The Assessment of Clinical Skills in General Internal Medicine Residents: A Comparison of the Objective Structured Clinical Examination to a Conventional Oral Examination." Annals of the Royal College of Physicians and Surgeons of Canada 18: 235.

Roemer,M.I. and Gartside,F.(1973) "Effect of Peer Review in Medical Foundations on Qualifications of Surgeons." Health Serv. Rep. 88: 808.

Rosenblatt,R.A. and Moscovice,I.S.(1984) "The Physician as Gatekeeper Determinants of Physicians' Hospitalization Rates." Medical Care 22(2):150.

Rosenfeld,L.S.(1957) "Quality of Medical Care in Hospitals." Amer. J. Public Health 47: 856.

Rubenstein,R., Niccolini,R. and Zora,J.(1979) "The Use of Live Simulation in Teaching the Mental Status Examination to Medical Students." Journal of Medical Education 54: 663.

Sanazaro,P.J. and Worth,R.M.(1978) "Concurrent Quality Assurance in Hospital Care-Report of a Study of Private Initiative in PSRO." New England J. Medicine 298: 1171.

Sanazaro,P.J. and Williamson,J.W.(1968) "A Classification of Physician Performance in Internal Medicine." Journal of Medical Education 43: 389.

Sanson-Fisher,R.W. and Poole,A.D.(1980) "Simulated Patients and the Assessment of Interpersonal Skills." Medical Education 14: 249.

SAS (1985) SAS User's Guide: Statistics 5th edition. North Carolina: author.

Saywell,R.M.,Studnicki,J.Bean,J.A. and Ludke,R.L.(1980) "A Performance Comparison: USMG-FMG House Staff Physicians." Amer.Journ.Public Health. 70(1):23.

Scherger,J.E., Gordon,M.J., Philips,T.J. and LoGerfo,J.P.(1980) "Comparison of Diagnostic Methods of Family Practice and Internal Medicine Residents." Journal of Family Practice 10: 95.

Schnabl,G., Hassard,T., Kopelow,M. and Klass,D.(1989) "Effect of Rater Variability on Student Scores". Winnipeg:University of Manitoba Faculty of Medicine.

Senior.J.R.(1976) Toward the Measurement of Competence in Medicine. Philadelphia: National Board of Medical Examiners.

Sheehan,T.J., Husted,S.D.R., Candee,D., Cook,C.D. and Bargen,M.(1980) Evaluation and the Health Professions 3(4):393.

Sibley,J.C., Sackett,D.L., Neufeld,V.R., Gerrard,B., Rudnick,V. and Fraser,W.(1982) "A Randomized Trial of Continuing Medical Education." New England Journal of Medicine 306(9): 511.

Sparling,J.F.(1962) "Measuring Medical Care Quality:A Comparative Study: Part 1:Hospitals." N.Y. State Med. Journ. 36(7): 56.

Staff of the Stanford Centre for Health Care Research(1974) Study of the Institutional Differences in Post-Operative Mortality. Springfield VA: Assembly of Life, National Academy of Sciences, National Research Council (Contract# PH 43-63-65).

Stapleton,J.F.and Zwerneman,J.(1965) "The Influence of an Intern-Resident Staff on the Quality of Private Patient Care." J.A.M.A. 194: 137.

Starfield,B. and Scheff,D.(1972) "Effectiveness of Paediatric Care:The Relationship Between Process and Outcome." Paediatrics 49(4): 547.

Stillman,P.L., Ruggill,J.S., Rutala,P.J. and Sabers,D.L.(1980) "Patient Instructors as Teachers and Evaluators." Journal of Medical Education 55: 186.

Stillman,P.L., Ruggill,J.S. and Sabers,D.L.(1978) "The Use of Practical Instructors to Teach and Evaluate a Complete Physical Examination." Evaluation and the Health Professions 1: 49.

Stillman,P.L., Rutala,P.J., Nicholson,G., Sabers,D.L. and Stillman,A.(1982) "Measurement of the Clinical Competence of Residents Using Patient Instructors." Annual Conference on Research in Medical Education: 111.

Stillman,P.L., Burpeau-DiGregorio,M., Nicholson,G., Sabers,D. and Stillman,A.(1983) "Six Years of Experience Using Patient Instructors to Teach Interviewing Skills." Journal of Medical Education 58: 942.

Stillman,P.L. and Swanson,D.B.(1987) "Ensuring the Competence of Medical Graduates Through Standardized Patients." Archives of Internal Medicine 147: 1049.

Stillman,P.L., Swanson,D.B., Smee,S. et.al.(1986) "Assessing the Clinical Skills of Residents with Standardized Patients." Annals of Internal Medicine 105: 762.

Stillamn,P.L., Regan,M. and Swanson,D.B.(1987) "Impact of Several Variables on Physical Examination Skills of Medical Students." Journal of Medical Education 62: 937.

Stillman,P.L., Swanson,D.B. et.al.(1986) "Psychometric Characteristics of Standardized Patients for the Assessment of Clinical Skills." Final Report to the American Board of Internal Medicine Committee on Research & Development.

Stross,J.K.(1983) "Maintaining Competency in Advanced Cardiac Life Support Skills." J.A.M.A. 249(24): 3339.

Swanson,D.B. and Stillman,P.L.(in press) "Use of Standardized Patients for Teaching and Assessing Clinical Skills." Evaluation & the Health Professions.

Swanson,D.B.(1988) "A Measurement Framework for Performance-Based Tests." Proceedings International Conference on Assessment of Clinical Competence. Ottawa.

Swanson,D.B., Norcini,J. and Grosso,L.(1987) "Assessment of Clinical Competence: Written and Computer-Based Simulations." Assessment and Evaluation in Higher Education 12: 220.

Swanson,D.B. & Norcini,J (in press) "Factors Influencing the Reproducibility of Tests Using Standardized Patients". Teaching and Learning in Medicine.

Tamblyn,R.M. and Barrows,H.S.(1980) "An Initial Evaluation of Learning Units to Facilitate Problem-Solving and Self-Directed Study (the Portable Patient Problem Pack)." Medical Education 14: 394.

Tamblyn,R.M., Lewis,K.E. and Murray,R.P.(1985) "Increasing the Objectivity of Measuring Clinical Problem-Solving in Patient Situations." Schimdt, H.G. and deVolder,M.L.(eds) Tutorials in Problem-Based Learning. Maastricht, Netherlands:Van Gorcum.

Templeton,B., Best,A. Samph,T. & Case,S. (1978) "Short Term Outcomes Achieved in Interviewing Medical Students". Internal Report. National Board of Medical Examiners.

Thibodeau,L.A. and Berwick,D.M.(1980) " Variation in Rates of Diagnosis of Otitis Media." Journal of Medical Education 55: 1021.

Tinning,F.(1974) "An Experimental Study Investigating the Effects of Real and Simulated Clinical Training On Psychomotor, Affective and Cognitive Variables During Real Clinical Performance of First Year Osteopathic Medical Students." Annual Conference Research In Medical Education.

Trussel,R.E.,Morehead,M.A. and Erhlick,J.(1962) The Quantity, Quality and Costs of Medical Care Secured by a Sample of Teamster Families in the New York Area. New York: Columbia University School of Public Health and Administrative Medicine.

Tugwell,P.(1979) "A Methodological Perspective on Process Measures of the Quality of Medical Care." Clinical and Investigative Medicine 2(2,3): 113.

Van der Vleuten,C.P. and Van Luyk,S.J.(1987) "Decomposition of the OSCE's: Some Methodological Considerations and Empirical Findings." Proceedings International Conference on Assessing Clinical Competence, Ottawa.

Van der Vleutan,C.P. & Swanson,D. (manuscript in preparation) "Assessment of Clinical Skills with Standardized Patients: The State of the Art". Maastricht, The Netherlands, University of Limburg.

Vayda,E.J.(1976) "The Use of Simulated Clients as an Aid to Teaching Social Work Practice." Proceedings Canadian Association of Social Workers, Quebec City. May.

Veloski,J.,Herman,M.W.,Gonella,J.S.,Zelenik,C. and Kellow,W.F.(1979) "Relationships Between Performance in Medical School and the First Postgraduate Year." Journal of Medical Education 54: 909.

Verhaak,P.F.M.(1986) "Variations in the Diagnosis of Psycho-social Disorders: A General Practice Observation Study." Soc.Sci.Med. 23(6): 595.

Verhulst,S.J., Colliver,J.A., Paiva,R.E. and Williams,R.G.(1986) "A Factor Analysis Study of Performance of First Year Residents." Journal of Medical Education 61: 132.

Wakeford,R.E., Norman,G.R. and Belton A.(1986) "Something Old Something New:The Certification Examination of the U.K. Royal College of General Practitioners." Annual Conference of Research in Medical Education: 211.

Ware,J.E., Brook,R.H., Rogers,W.H., Keeler,E.B., Davies,A.R., Sherbourne, C.D., Goldberg,G.A., Camp P. and Newhouse,J.P.(1987) Health Outcomes for Adults in Prepaid and Fee-For-Service Systems of Care. Santa Monica, CA.: Rand Health Insurance Experiment Series.

Webster,G.D., Shea,J.A., Norcini,J.J.,Grosso,L.J. and Swanson,D.B.(in press) "Comparison of Methods for Scoring Patient Management Problems: Use of External Criteria to Validate Scores."

Weed,L.(1969) Medical Records,Medical Education and Patient Care. Cleveland: Case Western Reserve University.

Werner,A. and Schneider,J.M.(1974) "Teaching Medical Students Interaction Skills". New England Journal of Medicine 290: 1232.

Whatling,T. and Wodak,E.(1979) "The Simulated Client in Social Work Training." International Social Work 22(2): 33.

Wigton,R.S.(1980) "Factors Important in the Evaluation of Clinical Performance of Internal Medicine Residents." Journal of Medical Education. 55(3): 423.

Williams,R., Barrows,H., Verhulst,S., Colliver,J., Marcy,M. and Steward,D. (1987) "Direct, Standardized Assessment of Clinical Competence." Medical Education 21: 482.

Williamson,P.M.(1975) "The Adoption of New Drugs by Doctors Practising in Group and Solo Practice." Soc.Sci.Med. 9: 233.

Wilson-Pessano,S.R. "Factors Underlying the Cognitive Difficulty of Patient Care Episodes." Annual Conference on Research in Medical Education: 158.

Winer,B. (1971) Statistical Principles in Experimental Design (2nd Ed.). New York: McGraw Hill Book Co.

Winicoff,R.N., Coltin,K.L., Morgan,M.M., Buxbaum,R.C. and Barnett,G.O. (1984) "Improving Physician Performance Through Peer Comparison Feedback." Medical Care 22(6): 527.

Wolf,F.M., Allen,N.P., Cassidy, J.Y., Axim,B.R. and Davis W.K. "Concurrent and Criterion-Referenced Validity of Patient Management Problems." Annual Conference on Research in Medical Education: 116.

Wolf,F.M., Woolliscroft,J.O., Calhoun,J.G. and Boxer,G.J.(1987) "A Controlled Experiment in Teaching Students to Respond to Patients' Emotional Concerns". Journal of Medical Education 62: 26.

Woodward,C.A., Norman,G.R. and Stillman,P.(1983) "Simulated Patients in Evaluation of Medical Education." Proceedings Annual Conference in Medical Education: 238.

Woodward,C.A., McConvey,G.A., Neufeld,V.R., Norman,G.R. and Walsh,A.(1985) "Measurement of Physician Performance by Standardized Patients." Medical Care 23(8): 1019.

Wooliscroft,J.O., Stross,J.K. and Silva,J.(1984) "Clinical Competence Certification: A Critical Appraisal." Journal Medical Education 59: 800.

Yankauer,A. and Allaway,N.C.(1958) "An Analysis of Hospital Neonatal Mortality rates in New York State." Amer. J.Dis. Child 95: 240.

Yergin,J.,Flood,A.B.,LoGerfo,J.P. and Dieher,P.(1987) "Relationship Between Patient Race and the Intensity of Hospital Services." Medical Care 25(7): 592.

Young,M.J., Fried,L.S., Eisenberg,J., Hershey,J. and Williams,S.(1987) "Do Cardiologists Have Higher Thresholds for Recommending Coronary Arteriography than Family Physicians?" Health Services Research 22(5): 623.

Young,E.A., Weser,E., McBride,H.M., Page,C.P.and Littlefield,J.H.(1983) "Development of Core Competencies in Clinical Nutrition." Journal American Diet. Assoc. 82(5): 482.

Appendix 1: Accuracy Checklists for 1987

ACCURACY CHECKLIST - CASE 1

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
High blood pressure for 15 years with no symptoms	0	1	2	3	4	5	6
Controlled with Dyazide	0	1	2	3	4	5	6
Symptoms began last few weeks	0	1	2	3	4	5	6
Now short of breath on exertion	0	1	2	3	4	5	6
Now tired, no energy	0	1	2	3	4	5	6
Now "woozy" all the time	0	1	2	3	4	5	6
Dull, achy headache every day	0	1	2	3	4	5	6
Sees MD regularly for blood pressure check & prescription refill	0	1	2	3	4	5	6
Last MD appt. 3-6 months ago	0	1	2	3	4	5	6
Today/yesterday B.P. taken & found to be 220/130 led to this appointment	0	1	2	3	4	5	6
Looks tired	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 2

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq
Burning pain	0	1	2	3	4	5	6
Pain in chest and stomach	0	1	2	3	4	5	6
About 1 1/2 year history	0	1	2	3	4	5	6
Occurs after eating	0	1	2	3	4	5	6
Occurs when bending/ lying down	0	1	2	3	4	5	6
Pain wakes him up at night	0	1	2	3	4	5	6
Foul tasting material comes out of mouth	0	1	2	3	4	5	6
Antacids no longer help much	0	1	2	3	4	5	6
Milk helps	0	1	2	3	4	5	6
No pain with exercise	0	1	2	3	4	5	6
No sweating	0	1	2	3	4	5	6
Drinks about 4 cups coffee in A.M.	0	1	2	3	4	5	6
Drinks beer daily - makes worse	0	1	2	3	4	5	6
Smokes	0	1	2	3	4	5	6
No change in bowel/ bladder habits	0	1	2	3	4	5	6
No blood in stool	0	1	2	3	4	5	6
Worried problem is his heart	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Family history of heart disease	0	1	2	3	4	5	6
On physical exam tenderness in epigastrium (high up in abdomen below lower end of the breast bone)	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 3

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
3 or 4 week history	0	1	2	3	4	5	6
Pain in lower back (dull ache)	0	1	2	3	4	5	6
Worse in last week or two	0	1	2	3	4	5	6
Constant all day	0	1	2	3	4	5	6
Worse on movement (or bending) or getting up after sitting	0	1	2	3	4	5	6
Burning on urination	0	1	2	3	4	5	6
Trouble starting flow/ dribbles	0	1	2	3	4	5	6
Gets up 3 or 4 times at night to urinate	0	1	2	3	4	5	6
Urine is dark & has strong foul odour	0	1	2	3	4	5	6
No loss of control (may find a drop or two on underwear, but no incontinence)	0	1	2	3	4	5	6
No bowel troubles	0	1	2	3	4	5	6
Aspirin no longer helps	0	1	2	3	4	5	6
Lost 10-15 lbs. in last few months without dieting	0	1	2	3	4	5	6
Slight pain or tender-ness on palpation of back (demonstration)	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 4

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Sore throat for 2 days	0	1	2	3	4	5	6
Getting worse ("the worst sore throat I've ever had")	0	1	2	3	4	5	6
Fever confirmed - temperature taken	0	1	2	3	4	5	6
Had a chill once	0	1	2	3	4	5	6
Difficulty swallowing	0	1	2	3	4	5	6
Home from school for 2 days	0	1	2	3	4	5	6
No contact with persons known to him/her who have a similar problem	0	1	2	3	4	5	6
No other problems (eyes ears, chest, etc.)	0	1	2	3	4	5	6
Isn't eating (too hard to swallow)	0	1	2	3	4	5	6
Feels very sick	0	1	2	3	4	5	6
Tenderness when neck is palpated (demonstrated)	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 5

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA**	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq
Aware BP elevated	0	1	2	3	4	5	6
Father hypertensive	0	1	2	3	4	5	6
Drinks 3-4 cups coffee	0	1	2	3	4	5	6
Drinks 2-3 beer daily (more on weekends)	0	1	2	3	4	5	6
Smokes	0	1	2	3	4	5	6
Exercise minimal	0	1	2	3	4	5	6
Reluctant but ultimately agrees to try lifestyle changes	0	1	2	3	4	5	6
Offers excuses for his lifestyle	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 6

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Abdominal pain started in morning	0	1	2	3	4	5	6
Has gotten worse all day	0	1	2	3	4	5	6
Now sharp	0	1	2	3	4	5	6
Started central, moved to RLQ	0	1	2	3	4	5	6
Hurts to move	0	1	2	3	4	5	6
Vomitted (can't keep anything down)	0	1	2	3	4	5	6
Nausea	0	1	2	3	4	5	6
Hasn't missed a menstrual period	0	1	2	3	4	5	6
On birth control (hasn't had unprotected intercourse)	0	1	2	3	4	5	6
Hasn't had appendectomy	0	1	2	3	4	5	6
Never had general anesthetic	0	1	2	3	4	5	6
No allergy	0	1	2	3	4	5	6
No change in bowel habits	0	1	2	3	4	5	6
No change in bladder habits	0	1	2	3	4	5	6
No vaginal discharge	0	1	2	3	4	5	6
Looks in pain	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Moves gingerly/ tentatively	0	1	2	3	4	5	6
Tenderness in RLQ	0	1	2	3	4	5	6
Rebound tenderness in RLQ	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 7

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
3 day history	0	1	2	3	4	5	6
Short of breath	0	1	2	3	4	5	6
Getting worse	0	1	2	3	4	5	6
Has had morning cough for 10 years with white sputum	0	1	2	3	4	5	6
Cough now worse sputum yellow-green	0	1	2	3	4	5	6
No blood in sputum	0	1	2	3	4	5	6
Can usually walk 1/2 block or 1/2 flight of stairs before needing to stop	0	1	2	3	4	5	6
Hot with shaking chills yesterday	0	1	2	3	4	5	6
Past 24 hrs: Sharp chest pain - left chest & underarm & left side of back	0	1	2	3	4	5	6
Left chest wall tenderness	0	1	2	3	4	5	6
Has had pneumonia before	0	1	2	3	4	5	6
Currently smokes 1 ppd	0	1	2	3	4	5	6
Takes Ventolin if needed	0	1	2	3	4	5	6
Doesn't regularly take pills for breathing	0	1	2	3	4	5	6
No allergies	0	1	2	3	4	5	6
Heart problem history: occasional executional angina	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Looks short of breath	0	1	2	3	4	5	6
Winces when coughs or breathes deep	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 8

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Seizure lasted less than 5 minutes	0	1	2	3	4	5	6
Arms & legs involved (movement, jerked, straightened, stiffened)	0	1	2	3	4	5	6
Eyes rolled back	0	1	2	3	4	5	6
Back arched	0	1	2	3	4	5	6
Child appeared "unconscious" during episode	0	1	2	3	4	5	6
Child fell asleep afterwards	0	1	2	3	4	5	6
Child woke up before reaching hospital	0	1	2	3	4	5	6
Child playing in playpen prior to seizure	0	1	2	3	4	5	6
Had fever: Yesterday 104 - Today 102	0	1	2	3	4	5	6
Gave Tylenol	0	1	2	3	4	5	6
Not vomitting before episode	0	1	2	3	4	5	6
No diarrhea before episode	0	1	2	3	4	5	6
No recent head injury	0	1	2	3	4	5	6
No history of pica	0	1	2	3	4	5	6
No exposure to toxic substance	0	1	2	3	4	5	6
Normal pregnancy and delivery	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA**	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq
Mother indicates that she thinks child has normal development	0	1	2	3	4	5	6
Rolled over 3-4 months	0	1	2	3	4	5	6
Sat up at 6 months	0	1	2	3	4	5	6
Now pulling self up	0	1	2	3	4	5	6
No othe illness since birth	0	1	2	3	4	5	6
Father had seizure as baby	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 9

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Disoriented to time	0	1	2	3	4	5	6
Poor recent memory	0	1	2	3	4	5	6
Inability to perform more than very simplist arithmetic calculation	0	1	2	3	4	5	6
Appears nervous/ worried/vague	0	1	2	3	4	5	6
<u>Son provides or confirms if asked:</u>							
Unsure mother's medication is taken properly	0	1	2	3	4	5	6
Mother not better on new medication	0	1	2	3	4	5	6
Poor recent memory	0	1	2	3	4	5	6
Concerned re father's ability to cope with situation	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 11

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Headache, like a band squeezing her head	0	1	2	3	4	5	6
Off & on for 4 yrs, worse in last 3 or 4 months (more often & lasts longer)	0	1	2	3	4	5	6
Dull throbbing pain - almost constant	0	1	2	3	4	5	6
Light hurts her eyes (makes it worse) no other vision problems	0	1	2	3	4	5	6
No other vision problems	0	1	2	3	4	5	6
Occasionally nauseated with headache (no vomitting)	0	1	2	3	4	5	6
No weakness	0	1	2	3	4	5	6
Sometimes headaches wake her up	0	1	2	3	4	5	6
Neck and back feel tense	0	1	2	3	4	5	6
Tried Tylenol/Aspirin but didn't help	0	1	2	3	4	5	6
On no medications	0	1	2	3	4	5	6
Otherwise healthy	0	1	2	3	4	5	6
Mother had similar headaches	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Social History							
Married with 4 children	0	1	2	3	4	5	6
Husband steadily employed	0	1	2	3	4	5	6
Some financial stress	0	1	2	3	4	5	6
Stress from her own job	0	1	2	3	4	5	6
Husband abuses chemicals	0	1	2	3	4	5	6
Husband has abused her (Not hit her head)	0	1	2	3	4	5	6
Husband doesn't abuse children	0	1	2	3	4	5	6
Patient has been treated in Emergency Room for abuse	0	1	2	3	4	5	6
Left husband about 4 months ago because of abuse, but returned after 1 month	0	1	2	3	4	5	6
Husband hasn't abused her since return	0	1	2	3	4	5	6
Mother is alcoholic	0	1	2	3	4	5	6
Mother has abused prescription meds given for headaches in past	0	1	2	3	4	5	6
Mother has history of depression	0	1	2	3	4	5	6
Willing to accept help for problems but doubtful of husband's willingness (not willing to leave husband at this point)	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 12

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Baby well till 3 days ago	0	1	2	3	4	5	6
Diarrhea - watery, green-yellow, about 8-10 times daily, no blood	0	1	2	3	4	5	6
Vomiting (no blood, no bile)	0	1	2	3	4	5	6
Refusing to nurse	0	1	2	3	4	5	6
Feels hot since yesterday	0	1	2	3	4	5	6
Decreased voiding (fewer diapers)	0	1	2	3	4	5	6
Has become lethargic	0	1	2	3	4	5	6
Has only been breastfed							
Parents healthy	0	1	2	3	4	5	6
Sister had similar illness	0	1	2	3	4	5	6
Mother had full term pregnancy and normal delivery	0	1	2	3	4	5	6
Mother looks and sounds worried	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 13

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Lost about 40 lbs in last year	0	1	2	3	4	5	6
Not feeling well for about a year	0	1	2	3	4	5	6
About 6 months ago, became increasingly thirsty and passed more urine	0	1	2	3	4	5	6
Diabetes diagnosed 6 months ago	0	1	2	3	4	5	6
Started on insulin after diagnosis	0	1	2	3	4	5	6
Started on diet after diagnosis	0	1	2	3	4	5	6
Regular medical follow-up	0	1	2	3	4	5	6
Since insulin no more problems with thirest or urine	0	1	2	3	4	5	6
Because of continued weight loss, doctor increased calories in diet	0	1	2	3	4	5	6
Because of continued weight loss, doctor increased insulin	0	1	2	3	4	5	6
Follows diet but finds it difficult to eat so much at times	0	1	2	3	4	5	6
Weight loss continues to present	0	1	2	3	4	5	6
Aching calves, almost constant	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq
Numbness & tingling in feet & toes	0	1	2	3	4	5	6
Quit smoking 6 months ago after more than 70 pack years	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 14

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq
Looks tired	0	1	2	3	4	5	6
Sounds vague	0	1	2	3	4	5	6
About 2 month history	0	1	2	3	4	5	6
Fever	0	1	2	3	4	5	6
Night Sweats	0	1	2	3	4	5	6
Weight loss	0	1	2	3	4	5	6
Sore throat	0	1	2	3	4	5	6
Headache	0	1	2	3	4	5	6
Loose stools/diarrhea	0	1	2	3	4	5	6
Stiff neck	0	1	2	3	4	5	6
Swollen lymph nodes	0	1	2	3	4	5	6
Fatigue	0	1	2	3	4	5	6
Decreased appetite	0	1	2	3	4	5	6
Not better	0	1	2	3	4	5	6
Has taken/takes lots of medications	0	1	2	3	4	5	6
Has had tests	0	1	2	3	4	5	6
Heterosexual	0	1	2	3	4	5	6
Some past problem with heart	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed
CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 15

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Patient has difficulty speaking initially	0	1	2	3	4	5	6
Patient wheezes/ indicates cannot breathe	0	1	2	3	4	5	6
Patient scratching at limbs	0	1	2	3	4	5	6
Nurse says it could be anaphylactic shock	0	1	2	3	4	5	6

	**UA	Not Ordered	Yes	No
If epinephrine/ adrenaline administered wheezing decreases	___		_____	
If epinephrine/ adrenaline ordered, nurse asks for dosage	___		_____	
If epinephrine/ adrenaline administered nurse reports increased BP	___		_____	

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

ACCURACY CHECKLIST - CASE 16

Student Number _____ Patient Simulator _____ Rater _____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq
Alert, cooperative	0	1	2	3	4	5	6
Initial Symptoms (Began about 2 months ago)							
Tired, lost appetite	0	1	2	3	4	5	6
Urine darker	0	1	2	3	4	5	6
Stools grayish (clay)	0	1	2	3	4	5	6
Occasional nausea & vomiting	0	1	2	3	4	5	6
Generalized itching	0	1	2	3	4	5	6
Skin yellow colour	0	1	2	3	4	5	6
Then hospitalized first time	0	1	2	3	4	5	6
First Hospitalization							
Dr. said liver enlarged & tender	0	1	2	3	4	5	6
Got little "bruises" on body	0	1	2	3	4	5	6
Told she had jaundice	0	1	2	3	4	5	6
Second Hospitalization (Referred to a second doctor because she was still hospitalized and yellow)							
Gallstones found on sonogram	0	1	2	3	4	5	6
Told by MD problem not gallstones	0	1	2	3	4	5	6
Told she had hepatitis	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
No specific treatment & discharged	0	1	2	3	4	5	6
Third Hospitalization (current, one month after 2nd)							
Tires easily	0	1	2	3	4	5	6
No chills, fever	0	1	2	3	4	5	6
Urine normal	0	1	2	3	4	5	6
Medications							
Prior to 1st hospitalization was taking female hormone supplement (i.e. oral contraceptives or premarin)	0	1	2	3	4	5	6
Microfurantoin (i.e. Furdantine or Macrodantin) for U.T.I.	0	1	2	3	4	5	6
Phenothiazine (i.e. Compazine or Stemitil) for nausea	0	1	2	3	4	5	6
All meds discontinues during 1st admission	0	1	2	3	4	5	6
Other							
Does not abuse alcohol	0	1	2	3	4	5	6
No IV drug abuse	0	1	2	3	4	5	6
No contact with anyone known to have hepatitis	0	1	2	3	4	5	6
Never had a blood transfusion	0	1	2	3	4	5	6

** UNABLE TO ASSESS - should only be checked if observer is unable to judge due to technical problems - e.g. poor sound, video quality, etc.

INCORRECT - information provided by simulated patient is not consistent with item listed

CORRECT - information provided by simulated patient is consistent

Appendix 2: Standardized Patient Information Form
Exam-Quality of Performance Rating Form

APPENDIX 2 (part 1)

SIMULATED PATIENT INFORMATION FORM

INSTRUCTIONS:

This form is to be completed by each simulated patient who is participating in the November-December, 1987 Comprehensive Evaluation of medical school graduates. The form should be completed after your training sessions have ended but before the evaluation exercise begins. The information that you provide will remain confidential. It will be used to help us improve the training of simulated patients for this examination.

Please return your completed forms to Gail or Vera (Room S204 Faculty of Medicine).

STANDARDIZED PATIENT INFORMATION FORM

CODE

YOUR NAME_____ 12 45 7

NAME OF YOUR SIMULATED PATIENT CASE_____

YOUR AGE_____(YRS) 9 10

SEX ___ FEMALE___ MALE___ (check the correct category) 12

A. FIRST HOW EXPERIENCED ARE YOU AS A SIMULATOR?

		(circle the correct response)		
1.	I have simulated this patient case before	YES	NO	14
	If YES: Indicate the number of times you have simulated this case before:_____times.			16 17
2.	I have simulated other patient cases before	YES	NO	19
	If YES: Indicate the number of other patient cases you have simulated:_____cases.			21 22
3.	I have been involved in role-playing before	YES	NO	24
4.	I have had experience as an actor/actress	YES	NO	26
5.	I have had training as an actor/actress	YES	NO	28

B. SOME INFORMATION ON YOUR EXPERIENCE WITH THE MEDICAL PROBLEM YOU ARE SIMULATING FOR THIS EVALUATION

		(circle the correct response)		
1.	I personally have or have had the medical problem I am simulating.	YES	NO	30
2.	I have had some of the symptoms/ experiences of the patient I am simulating	YES	NO	32

(circle the correct response)

3. I know someone who has the problem I am simulating. YES NO 34

4. I can imagine what it would have been like to be this patient. Not Very Well / Fairly Well / Very Well - 36

C. FINALLY SOME INFORMATION ON YOUR TRAINING FOR THIS SIMULATION

(circle the correct response)

1. Indicate the number of training sessions you have had for this simulation. 1 2 3 4 38

2. Indicate the number of times a physician was present at your training sessions. 1 2 3 4 - 40

3. How well do you think you will be able to simulate:

	With Difficulty		Some Difficulty		No Difficulty		
history of this patient's problem	1	2	3	4	5		42
the physical examination findings	1	2	3	4	5	N/A	44
the patient's emotional state	1	2	3	4	5		46

4. Using the patient checklist, how well do you think you will be able to record the actions taken by the student? 1 2 3 4 5 48

5. Do you have any suggestions for improving your training as a simulated patient?
Suggestions: _____ 50-60

_____

_____

_____

THANK-YOU VERY MUCH FOR TAKING THE TIME TO COMPLETE THIS FORM
PLEASE RETURN IT TO GAIL OR VERA IN ROOM S204 (FACULTY OF MEDICINE BUILDING)

APPENDIX 2 (Part 2)

Simulator's Name _____ Station Name _____

Student #: _____

Please circle <u>your assessment</u> of how well you rate your performance of your simulated role with the previous student which was video taped.

	Very Well				Did Poorly
I performed my role:	5	4	3	2	1

Appendix 3: Standardized Patient Training Record

SIMULATED PATIENT TRAINING RECORD

INSTRUCTIONS:

The simulated patient training record is to be completed by the patient trainer after the training sessions for the patient have been completed. A training record should be competed for each individual who was trained as a simulated patient for the Comprehensive Clinical Evaluation (November-December, 1987). The information provided will be used to evaluate factors which may contribute to the accuracy and consistency of patient presentation. Ultimately, it will aid in the refinement of the technique for multi-centre application.

THE SIMULATED PATIENT TRAINING RECORD

CODE

TRAINER_____ 1

PHYSICIAN ASSISTING_____ 3 4

SIMULATOR NAME_____ 6 7

PATIENT CASE NAME_____ 9 10

SOURCE OF CASE S.I.U.____ U OF M____ 12

	(circle the correct response)				
TRAINING TAPE AVAILABLE(if source=U OF M)	YES			NO	14
NUMBER OF TRAINING SESSIONS	1	2	3	4	16
NUMBER TRAINING SESSIONS WITH M.D. ASSISTANCE	1	2	3	4	18

APPROXIMATE TOTAL TIME SPENT IN TRAINING_____(hours) 20 21

YOUR IMPRESSIONS ABOUT THE SIMULATED PATIENT YOU HAVE TRAINED

Could you predict the ability of this individual to accurately and consistently present the case you have trained them for at this time.

A. THE PATIENT'S HISTORY	With Difficulty		Some Difficulty		No Difficulty	
1.This patient will be able to accurately present the important fetaures of the history.	1	2	3	4	5	23
2.This patient will be able to consistently present the patient from student to student.	1	2	3	4	5	25
B. PHYSICAL FINDINGS						
3.This patient will be able to accurately portray the physical findings of this case.	1	2	3	4	5	27
4.This patient will be able to consistently portray the physical from student to student.	1	2	3	4	5	29

	With Difficulty		Some Difficulty		No Difficulty	
C. THE PATIENT'S AFFECT						
5. This patient will be able to provide a realistic portrayal of this patient's affect.	1	2	3	4	5	31
6. This patient will be able to provide a consistent portrayal of this patient's affect.	1	2	3	4	5	33
D. THE PATIENT'S ABILITY TO ACCURATELY RECORD STUDENT ACTIONS						
7. This patient will be able to accurately record the student's actions during the patient encounter using the patient checklist.	1	2	3	4	5	35

ADDITIONAL COMMENTS ON THE TRAINING SESSIONS OR PATIENT SIMULATOR

_____ 37 50

_____

_____

_____

Appendix 4: Interpersonal Skills Checklist

INTERPERSONAL SKILLS RATING SCALE

Station Name _____ Student No: _____

Simulator's Name _____

Below are listed a number of statements that describe a variety of ways that one person could feel or behave in relation to another person. Please consider each statement with respect to whether you think it is true or not true about your relationship with the student doctor in the interview you have just had. Mark each statement in the right margin according to how strongly you feel it is true or not true. Please mark **every one**. Write in +3, +2, +1 or -1, -2, -3 to stand for the following answers:

+3.	Yes, I strongly feel that it is true	-1:	No, I feel that it is probably untrue, or more untrue than true
+2	Yes, I feel it is true	-2.	No, I feel it is not true.
+1	Yes, I feel that it is probably true, or more true than untrue.	-3	No, I strongly feel that it is not true.

		ANSWER
1.	The student doctor wanted to understand how I saw things.	___
2	The student doctor may have understood my words but he/she did not see the way I felt	___
3	I was able to explain my problem to the student doctor as fully as I needed to	___
4	The student doctor nearly always knew exactly what I meant.	___
5	The student doctor looked at what I did from his/her own point of view	___
6	The student doctor explained things so that now I know what is wrong with me .	___
7	The student doctor usually sensed or realized what I was feeling .	___
8	The student doctor's own attitudes toward some of the things I did or said prevented him/her from understanding me . . .	___
9	The student doctor explained what treatment, tests or other follow up is going to happen	___

(DO NOT TURN PAGE OVER)

10. Sometimes the student doctor thought I felt a certain way because that's the way he/she feels. . . . . ___

11. The student doctor realized what I meant even when I had difficulty in saying it. ___

12. The student doctor gave me the opportunity to express my feelings or ideas in planning treatment, tests or follow up. . . . . ___

13. The student doctor usually understood the whole of what I meant. . . . . ___

14. The student doctor just took no notice of some things that I thought or felt . . ___

15. The student doctor spoke in language I didn't always understand. . . . ___

16. The student doctor appreciated exactly how the things I experienced felt to me ___

17. At times the student doctor thought that I felt a lot more strongly about a particular thing than I really did. . . . . ___

18. The student doctor gave me the opportunity to ask questions. . . . . ___

19. The student doctor did not realize how sensitive I was about some of the things we discussed. . . . . ___

20. The student doctor understood me. . . . . ___

21. The student doctor was not as thorough as he/she should have been . . . ___

22. The student doctor's response to me was usually so fixed and automatic that I didn't really get through to him/her. . . . . ___

23. When I was hurt or upset the student doctor was able to recognize my feelings exactly, without becoming upset too. . . . . ___

24. I feel satisfied with the medical care that I received. . . . . _

Appendix 5: Accuracy Checklists for 1988

ACCURACY CHECKLIST - CASE 2

Student Number____ Patient Simulator_____ Rater_____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq.
Demeanor							
1. Looks depressed	0	1	2	3	4	5	6
Presenting Problem							
In hospital for:							
2. "Take down" (closure) of colostomy	0	1	2	3	4	5	6
3. Has chronic & productive cough	0	1	2	3	4	5	6
4. No change in cough or sputum	0	1	2	3	4	5	6
5. Short of breath on exertion	0	1	2	3	4	5	6
6. Pain in abdomen on coughing	0	1	2	3	4	5	6
7. No chest pain	0	1	2	3	4	5	6
Past History							
8. High blood pressure for approx 5 yrs	0	1	2	3	4	5	6
9. Diabetic for approx 5 yrs.	0	1	2	3	4	5	6
10. Bronchiectasis for 10-15 yrs.	0	1	2	3	4	5	6
11. Cardiac palpatations in past	0	1	2	3	4	5	6
12. Palpatations now only on rare occasions	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Past Surgery							
13. Previous procedures for urological/ gynec. problems	0	1	2	3	4	5	6
14. Previous surgery for gallbladder removal	0	1	2	3	4	5	6
15. Colon resection for adenocarcinoma 3 mos ago	0	1	2	3	4	5	6
16. Trouble weaning ventilator with last surgery	0	1	2	3	4	5	6
Current Medications							
17. Theo-Dur	0	1	2	3	4	5	6
18. Micronase	0	1	2	3	4	5	6
19. Lasix	0	1	2	3	4	5	6
20. Home Oxygen	0	1	2	3	4	5	6
21. Digoxin	0	1	2	3	4	5	6
22. Inhaler	0	1	2	3	4	5	6
Important Negatives							
23. No prior heart attack	0	1	2	3	4	5	6
24. Not short of breath lying down	0	1	2	3	4	5	6
25. No foot/ankle swelling	0	1	2	3	4	5	6
26. No change in bowel habits	0	1	2	3	4	5	6
27. No blood in stools	0	1	2	3	4	5	6
28. Doesn't smoke, drink	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Family History							
29. Father & grandfather had bronchiectasis	0	1	2	3	4	5	6
30. Mother had diabetes and stroke	0	1	2	3	4	5	6

ACCURACY CHECKLIST - CASE 4

Student Number_____ Patient Simulator_____ Rater_____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq.
Demeanor							
1. Looks/acts anxious	0	1	2	3	4	5	6
History of Presenting Problem							
2. Awoke this AM with paralysed legs	0	1	2	3	4	5	6
3. Urinary frequency yesterday	0	1	2	3	4	5	6
4. Not urinated today	0	1	2	3	4	5	6
5. Tingling,warm sens. on left side body 3 weeks ago	0	1	2	3	4	5	6
6. Feeling is still different on left side	0	1	2	3	4	5	6
7. Lost sight in right eye 4 wks ago	0	1	2	3	4	5	6
8. Given prednisone	0	1	2	3	4	5	6
9. Vision improved	0	1	2	3	4	5	6
10. Not menstruated in 6 weeks	0	1	2	3	4	5	6
11. Has had unprotected intercourse	0	1	2	3	4	5	6
12. Afraid she is pregnant	0	1	2	3	4	5	6
13. Single,lives with roomate	0	1	2	3	4	5	6

ITEM		COULDN'T EVALUATE		COULD EVALUATE			
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Physical Findings							
14. Paralysis legs & lower abdominals	0	1	2	3	4	5	6
15. Legs flaccid	0	1	2	3	4	5	6
16. Absent superficial abdominals	0	1	2	3	4	5	6
17. Bilateral babinskis	0	1	2	3	4	5	6
18. No DTRs in legs	0	1	2	3	4	5	6
19. No touch to umbilicus	0	1	2	3	4	5	6
20. No pin to nipple line on right &	0	1	2	3	4	5	6
21. one inch above nipple on left	0	1	2	3	4	5	6
22. No vibration to sternum	0	1	2	3	4	5	6
23. No temperature to clavicles	0	1	2	3	4	5	6
24. No position sense to hips	0	1	2	3	4	5	6
25. Sacral sparing to all modalities	0	1	2	3	4	5	6
26. Reduced strength left triceps (arm cannot be extended against gravity)	0	1	2	3	4	5	6

ACCURACY CHECKLIST - CASE 6

Student Number_____ Date Videotape: Month___ Day___ Rater_____.

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM		COULDN'T EVALUATE			COULD EVALUATE			
		UA	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq.
Mother's Demeanor								
1.	Appears anxious	0	1	2	3	4	5	6
Questions Asked By Mother if Information Not Provided by Student								
2.	What does x-ray show	0	1	2	3	4	5	6
3.	Will infant be OK	0	1	2	3	4	5	6
4.	Could something else show up/happen	0	1	2	3	4	5	6
5.	How will she know if a problem occurs at home	0	1	2	3	4	5	6
6.	How can future problems be prevented	0	1	2	3	4	5	6
History of Accident								
7.	Child fell down basement steps	0	1	2	3	4	5	6
8.	Did not lose consciousness	0	1	2	3	4	5	6
9.	Sister might have left baby gate open	0	1	2	3	4	5	6
Related History and Lifestyle								
10.	No previous visits to ER or MD for accidents	0	1	2	3	4	5	6
Two near accidents:								
11.	Baby walker collapsed once	0	1	2	3	4	5	6

ITEM		COULDN'T EVALUATE		COULD EVALUATE				
		UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
12.	Cord to electric appliance pulled by child	0	1	2	3	4	5	6
13.	Don't always use carseat	0	1	2	3	4	5	6

ACCURACY CHECKLIST - CASE 8

Student Number_____ Date Videotape: Month___ Day___ Rater_____
Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq.
Present Problem							
1. Periodic pelvic pain	0	1	2	3	4	5	6
2. Started approx. 1 yr ago	0	1	2	3	4	5	6
3. Begins 2-3 days before period	0	1	2	3	4	5	6
4. Continues for 2-3 days of flow	0	1	2	3	4	5	6
5. Pain increasing over past 6 mos.	0	1	2	3	4	5	6
6. Has missed work on occasion b/c of pain	0	1	2	3	4	5	6
7. Pain on intercourse for past few months	0	1	2	3	4	5	6
8. Intercorse pain worse 2-3 days before period	0	1	2	3	4	5	6
9. Loose stools during period	0	1	2	3	4	5	6
Menstrual & Past History							
10. Menstruation began age 13	0	1	2	3	4	5	6
11. Periods occur every 28-30 days	0	1	2	3	4	5	6
12. Periods last 5-6 days	0	1	2	3	4	5	6
13. Took oral contraceptive for 3 yrs.	0	1	2	3	4	5	6
14. Stopped BC pill 2 years ago	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
15. Has been trying to get pregnant for 18 mos	0	1	2	3	4	5	6
16. Has intercourse every 2-3 days	0	1	2	3	4	5	6
17. Ruptured ovarian cyst 5 yrs ago	0	1	2	3	4	5	6
18. Ruptured appendix at age 6	0	1	2	3	4	5	6
19. Husband has no chronic or communicable disease	0	1	2	3	4	5	6

ACCURACY CHECKLIST - CASE 11

Student Number_____ Date Videotaped: Month____ Day_____ Rater_____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE			
	UA	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq.
Husband's Demeanor							
1. Little eye contact	0	n/a	n/a	3	n/a	5	n/a
2. Turns to wife to answer questions about his feelings	0	1	n/a	3	n/a	5	n/a
3. Will go along who counselling if asked directly	0	1	n/a	3	4	5	6
Presenting Problem							
4. Trouble swallowing food for 4-6 wks.	0	1	2	3	4	5	6
5. Occasionally regurgitates food	0	1	2	3	4	5	6
6. Lost 20 lbs. in 4-6 weeks	0	1	2	3	4	5	6
7. Lost 5 lbs. since last visit	0	1	2	3	4	5	6
8. Pain under ribs for 4-6 weeks	0	1	2	3	4	5	6
9. No change in bowel habits	0	1	2	3	4	5	6
10. Sleep interrupted	0	1	2	3	4	5	6
11. Has been nervous, withdrawn	0	1	2	3	4	5	6
12. Meds=Ventolin	0	1	2	3	4	5	6

Social History

13. Previously in construction business with father	0	1	2	3	4	5	6
14. Declared bankruptcy 4 years ago	0	1	2	3	4	5	6
15. Father angry at sons ever since	0	1	2	3	4	5	6
16. Sees father only on special occasions	0	1	2	3	4	5	6
17. Husband upset about relationship with father	0	1	2	3	4	5	6

Social History

18. Wife concerned about marriage	0	1	2	3	4	5	6
19. Wife concerned about husband's heavy drinking for past 3-4 yrs	0	1	2	3	4	5	6
20. Husband drinks 10-12 beers/day	0	1	2	3	4	5	6
21. Few close friends	0	1	2	3	4	5	6
22. No children	0	1	2	3	4	5	6

ACCURACY CHECKLIST - CASE 12

Student Number___ Patient Simulator_____ Rater_____

Circle the number for each item that best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq.
Presenting Problem							
1. Lightheaded for approx. 1 wk.	0	1	2	3	4	5	6
2. Precipitated by getting up after leaning forward	0	1	2	3	4	5	6
3. Went to ER 3 days ago & told it was low blood pressure	0	1	2	3	4	5	6
4. Told to stop BP pills(chlorthalidone)	0	1	2	3	4	5	6
5. Told 3 mos. ago to stop pills b/c of low BP and low potassium	0	1	2	3	4	5	6
6. Symptoms persisted so resumed taking pills	0	1	2	3	4	5	6
7. Has had similar, less severe symptoms for 2 mos.	0	1	2	3	4	5	6
8. Had occasional episodes for 4 yrs.	0	1	2	3	4	5	6
9. Was told due to low potassium due to BP pills	0	1	2	3	4	5	6
Significant Negatives							
10. Not short of breath	0	1	2	3	4	5	6
11. No chest pain	0	1	2	3	4	5	6
12. No indigestion	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
13. No change in bowels	0	1	2	3	4	5	6
14. No tarry stools	0	1	2	3	4	5	6
Other Problems							
15. Chronic stuffy nose (not seasonal)	0	1	2	3	4	5	6
16. Frequency and dribbling on urination	0	1	2	3	4	5	6
17. Intermittent aching in legs for 10 yrs.	0	1	2	3	4	5	6
18. Legs ache at night & with exercise	0	1	2	3	4	5	6
19. Arthritis in knees	0	1	2	3	4	5	6
20. Stiff neck with pain into left shoulder	0	1	2	3	4	5	6
21. Rectal bleeding few times a year noted on wiping	0	1	2	3	4	5	6
22. Stopped work b/c of arthritis and lightheadedness	0	1	2	3	4	5	6
23. Retired fireman	0	1	2	3	4	5	6
24. Father & sister had MI	0	1	2	3	4	5	6
Medications							
25. Entrophen for arthritis and leg pain	0	1	2	3	4	5	6
26. Chlorthalidone	0	1	2	3	4	5	6

ACCURACY CHECKLIST - CASE 13

Student Number_____ Date Videotape: Month___ Day___ Rater_____

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq.
Appearance							
1. Nervous,edgy	0	n/a	n/a	3	n/a	5	n/a
Present Problem							
2. Afraid she has brain tumour	0	1	2	3	4	5	6
3. Classmate died of tumour	0	1	2	3	4	5	6
4. Sudden onset numbness in hands	0	1	2	3	4	5	6
5. Had headache at same time	0	1	2	3	4	5	6
6. Had dizziness at same time	0	1	2	3	4	5	6
7. Occurred when driving	0	1	2	3	4	5	6
8. Lasted 20 minutes	0	1	2	3	4	5	6
9. Taken to ER-nothing found	0	1	2	3	4	5	6
10. Symptoms recurred a few days later	0	1	2	3	4	5	6
11. Happend in movie theatre with boyfriend	0	1	2	3	4	5	6
12. Taken to ER a 2nd time-nothing found	0	1	2	3	4	5	6
13. Has had brief, milder episodes over past week	0	1	2	3	4	5	6

ITEM	UA**	COULDN'T EVALUATE		COULD EVALUATE			
		Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Past & Family History							
14. No previous past episodes	0	1	2	3	4	5	6
15. Often feels anxious	0	1	2	3	4	5	6
16. Concerned about own health	0	1	2	3	4	5	6
17. Father had heart problem	0	1	2	3	4	5	6
18. Not told initially b/c family didn't think she could cope	0	1	2	3	4	5	6
Social History							
19. Ambivalent about continuing university	0	1	2	3	4	5	6
20. Ambivalent about continuing relationship with boyfriend	0	1	2	3	4	5	6
21. Doesn't smoke	0	1	2	3	4	5	6
22. Doesn't drink	0	1	2	3	4	5	6

ACCURACY CHECKLIST - CASE 14

Student Number_____ Patient Simulator_____ Rater_____.

Circle the number for each item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA	Not Asked	NotCorrect Examined	Incorrect Spon.	To Inq.	Spon.	To Inq.
Presenting Problem							
1. Fatigued-able to do less for about 3 mos.	0	1	2	3	4	5	6
2. Incresingly tired for past 2 wks.	0	1	2	3	4	5	6
3. Pain in right lower back for 1 wk.	0	1	2	3	4	5	6
4. Pain radiates to upper abdomen and groin	0	1	2	3	4	5	6
5. Attributes pain to lifting railway ties	0	1	2	3	4	5	6
6. Pain worse in past 4-5 days.	0	1	2	3	4	5	6
7. Less appetite for approx. 4 days	0	1	2	3	4	5	6
8. Feels full after 3 bites	0	1	2	3	4	5	6
9. Drinks mainly cool drinks	0	1	2	3	4	5	6
10. No nausea/vomiting	0	1	2	3	4	5	6
11. Chills & hot spells over past week	0	1	2	3	4	5	6
12. Feels feverish/not taken temperature	0	1	2	3	4	5	6
13. Urine darker over 3-4 days	0	1	2	3	4	5	6
14. No change in stools	0	1	2	3	4	5	6
15. No blood in stools	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Past History							
16. Had constipation	0	1	2	3	4	5	6
17. Now uses laxatives	0	1	2	3	4	5	6
18. Renal stones approx. 10 years ago	0	1	2	3	4	5	6
19. Hospitalized 1 yr ago for possible MI (was dx as indigestion)	0	1	2	3	4	5	6
20. Stopped smoking 1 year ago	0	1	2	3	4	5	6
21. Dosen't drink	0	1	2	3	4	5	6

ACCURACY CHECKLIST - CASE 15

Student Number_____ Date Videotape: Month___ Day_____ Rater_____

Circle a number for each item listed which best describes the patient response in the student-patient encounter.

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq.
History							
1. No problems with pregnancy	0	1	2	3	4	5	6
2. No problems with delivery	0	1	2	3	4	5	6
3. Rolled at 4 months	0	1	2	3	4	5	6
4. Sat at 6 months	0	1	2	3	4	5	6
5. Walked at 11 months	0	1	2	3	4	5	6
6. Talked at 2 years	0	1	2	3	4	5	6
7. Attends kinder-garden	0	1	2	3	4	5	6
8. No behavior problems	0	1	2	3	4	5	6
9. Has no recurrent or chronic illnesses	0	1	2	3	4	5	6
10. Mild constipation	0	1	2	3	4	5	6
11. Has not kept record of height and weight	0	1	2	3	4	5	6
Family History							
12. Parents are of average hgt & wgt	0	1	2	3	4	5	6
13. No family diseases	0	1	2	3	4	5	6

ACCURACY CHECKLIST - CASE 16

Student Number_____ Patient Simulator_____ Rater_____

Circle the number fro eacxh item which best describes the patient's response in the student-patient encounter.

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq.
Presenting Problem							
1. Awoke with headache	0	1	2	3	4	5	6
2. Trouble speaking (words garbled)	0	1	2	3	4	5	6
3. Inco-ordinated on right side	0	1	2	3	4	5	6
4. Improved since this morning	0	1	2	3	4	5	6
5. No similar symptoms in past with headache	0	1	2	3	4	5	6
Headache And Related History							
6. Has headaches since approx. age 17	0	1	2	3	4	5	6
7. Occur 1-2/month	0	1	2	3	4	5	6
8. Are pounding & severe	0	1	2	3	4	5	6
9. Usually on left side occasionally on right	0	1	2	3	4	5	6
10. Last up to 1-2 days	0	1	2	3	4	5	6
11. Nausea & occasional vomiting with headache	0	1	2	3	4	5	6
12. Aggravated by light and noise	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
Precipitated or worsened by:							
13. Stress	0	1	2	3	4	5	6
14. Pregnancy	0	1	2	3	4	5	6
15. Birth control pills	0	1	2	3	4	5	6
16. Fatigue	0	1	2	3	4	5	6
17. No fever	0	1	2	3	4	5	6
18. Relieved by rest & tylenol	0	1	2	3	4	5	6
19. Mother has migrane	0	1	2	3	4	5	6
20. Not on BC pill	0	1	2	3	4	5	6
21. Doesn't smoke	0	1	2	3	4	5	6
22. Uses Tylenol PRN	0	1	2	3	4	5	6
Social History							
23. Some incr. stress lately (ie. new baby, return to work, helping friend with business)	0	1	2	3	4	5	6
Physical Examination							
24. Right arm drifts downward slightly when eyes are closed and arms extended	0	1	2	3	4	5	6

ACCURACY CHECKLIST - CASE 19

Student Number_____ Patient Simulator_____ Rater_____

Circle the number which best describes the patient's response in the student-patient encounter. For this case the response could be provided by the husband or patient.

ITEM	COULDN'T EVALUATE			COULD EVALUATE			
	UA	Not Asked	Not Examined	Correct Spon.	Incorrect To Inq.	Incorrect Spon.	Incorrect To Inq.
Appearance							
1. Walks with a broad-based gait	0	1	2	3	4	5	6
2. Posture-slightly stooped	0	1	2	3	4	5	6
History							
3. Difficulty with memory for 3-4 yrs.	0	1	2	3	4	5	6
4. Memory & concent'n getting gradually worse	0	1	2	3	4	5	6
5. Needs some help in daily activities	0	1	2	3	4	5	6
6. Can eat without assistance	0	1	2	3	4	5	6
7. Personality change less interested in usual activities	0	1	2	3	4	5	6
8. Hypertension for some time	0	1	2	3	4	5	6
9. Takes Aldoril for hypertension	0	1	2	3	4	5	6
Important Negatives							
10. Denies depression	0	1	2	3	4	5	6
11. Appetite & weight unchanged	0	1	2	3	4	5	6
12. No cold intolerance	0	1	2	3	4	5	6
13. No head injury	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	Correct To Inq.	Incorrect Spon.	Incorrect To Inq
14. No excessive drinking	0	1	2	3	4	5	6
15. No encephalitis/ meningitis(brain fever)	0	1	2	3	4	5	6
16. No use of tranquillizers/hypnotics	0	1	2	3	4	5	6
17. No history stroke	0	1	2	3	4	5	6
18. No syncope	0	1	2	3	4	5	6
19. No incontinence	0	1	2	3	4	5	6
20. No problem sleeping	0	1	2	3	4	5	6
Important Negatives							
21. No suicidal thoughts	0	1	2	3	4	5	6
Mental Status							
Orientation to Person:							
22. Knows her name	0	1	2	3	4	5	6
23. Knows husband's name	0	1	2	3	4	5	6
Orientation to Time:							
24. Can't identify year correctly	0	1	2	3	4	5	6
25. Knows month	0	1	2	3	4	5	6
26. Can't identify day of the month	0	1	2	3	4	5	6
Orientation to Place:							
27. Can identify city	0	1	2	3	4	5	6
28. Can't identify hospital	0	1	2	3	4	5	6

ITEM	COULDN'T EVALUATE		COULD EVALUATE				
	UA**	Not Asked	Not Examined	Correct Spon.	To Inq.	Incorrect Spon.	To Inq
Memory and calculation:							
29. Makes 2 or more errors in subtracting 7 from 100	0	1	2	3	4	5	6
30. Can't recall a word given to her after a minute or more	0	1	2	3	4	5	6

Appendix 4: Interpersonal Skills Checklist

Appendix 6: Student Rating Forms for 1987

Case 1

Interaction # _____ Rater's Name _____

Place a checkmark next to each action taken by the examinee:

*UA

PHYSICAL EXAMINATION

Ophthalmoscope Examination performed correctly ___ ___

Ophthalmoscope Examination performed incorrectly ___ ___

Ophthalmoscope Examination not performed
(Criteria: dark setting, focusing, proper examination of retina) ___ ___

Examine (listen) heart
(Criteria: Stops from breathing, listens at 2 sites) ___ ___

Examine (listen) lungs
(Criteria: Listening to lower base of lungs on both sides) ___ ___

Examine (listen) abdomen
(Criteria: feels for aeorta, listens for murmurs at 3 sites) ___ ___

Check extremities
(Criteria: checks for swelling) ___ ___

Check for pulses
(Criteria: feels and listens neck and groins, feels both ankles) ___ ___

COMMUNICATION OF FINDINGS

Indicate to me that:

"Your present condition is ....." (relates headaches, fatigue, wooziness to findings) ___ ___

"Your blood pressure is too high/higher than usual" ___ ___

"Your blood pressure can be managed and treated" (reassurance or positive note" ___ ___

"There is a need for more tests and treatment" ___ ___

***UA - UNABLE TO ASSESS - should be checked only if simulator is unable to assess because he/she is watching on video and not in room.**

Case 2

Examiner's Name _____ Rater's Name _____

Date _____ Simulator's Name _____

Place a checkmark next to each of the following history findings which the examinee has checked with you and which you have provided to the examinees.

	*UA
Chest pain usually occurs after meals	___ ___
Pain occurs sometimes when I stoop over or lay down flat	___ ___
I have foul tasting material that comes up into my throat and mouth	___ ___
Pain gets better if I drink milk	___ ___
I have tried antacids and they used to work - but they do not work well anymore	___ ___
I have about 4 cups of coffee every morning	___ ___
I do not have the pain when I exert myself	___ ___

PATIENT EDUCATION:

Have indicated to me that my chest pain is probably not related to heart problems but may be a problem of digestion and of the esophagus	___ ___
Recommends that I take antacids	___ ___
Prescribes Tagamet (Cimetidine) and/or Zantac (Ranitidine)	___ ___
Recommends to reduce coffee intake	___ ___
Recommends to reduce alcohol intake	___ ___
Recommends to reduce aspirin intake	___ ___
Recommends to reduce smoking	___ ___
Recommends to take small and frequent feedings	___ ___
Recommends no eating at bedtime	___ ___
Recommends elevation of the head of the bed	___ ___

Case 2 (cont.)

*UA

OTHER ACCEPTABLE ANSWERS:

Lose weight ___ ___

Recommend further tests before treatment ___ ___

Barium studies ___ ___

EKG ___ ___

*UA - UNABLE TO ASSESS - should only be checked if simulator is unable to assess because he/she is watching on video and not in room.

Case 3

Examiner's Name _____

Rater's Name _____

Time and Date _____

Check all of the items below that this student talked to you about during this session.

	*UASATISFACTORY	UNSATISFACTORY	
asked about urinary frequency	___	___	___
pain with urination	___	___	___
slow-weak dribbling stream	___	___	___
hesitancy getting flow started	___	___	___
blood in the urine	___	___	___
Nocturia - how often patient gets up at night to urinate	___	___	___
description of back pain - worse with movement	___	___	___
appetite	___	___	___
weight loss	___	___	___

*UA - UNABLE TO ASSESS - should only be checked if simulator is unable to assess because he/she is watching on video and not in room.

Case 4

Observer _____ Examinee _____

Date & Time _____Total Time for Encounter _____

Check all of the items below that this student talked to you about during this session.

HISTORY OF PRESENT ILLNESS

		*UA
Fever	__	__
Shakes and/or chills with fever	__	__
Difficulty swallowing	__	__
Tender neck (lymph glands)	__	__
Swollen neck (lymph glands)	__	__
Cough	__	__
Ear aches	__	__
Eye inflammation (conjunctivitis)	__	__
Other enlarged lymph nodes	__	__
Other symptoms	__	__
Any allergies	__	__

FAMILY HISTORY

Other contacts with similar problem	__	__
History of Rheumatic fever in family	__	__
Children younger than 12 who live with you	__	__

PHYSICAL EXAMINATION

Inspection of throat	__	__
Palpitation of lymph nodes (above collar bone, behind & in front of ear & in back of neck)	__	__
Inquire whether lymph nodes in neck are tender when being palpated	__	__

Case 4 (cont.)

		*UA
Palpitation of axillary (arm pits) inguinal (public area) and epitrochlar (above elbow) lymph node	___	___
Palpate spleen	___	___
Visual inspection of eyes for conjuctivities	___	___

***UA - UNABLE TO ASSESS - should be checked only if simulator is unable to assess because he/she is watching on video & not in room**

Case 5

Examinee _____

Evaluator/patient _____

Time & Date _____

Check all of the items below that this student talked to you about during this session.

	*UASATISFACTORY		UNSATISFACTORY
lose weight	___	___	___
stop smoking	___	___	___
exercise program	___	___	___
low cholesterol diet	___	___	___
low sodium	___	___	___
limit caffeine	___	___	___
check blood pressure	___	___	___
limit alcohol	___	___	___

***UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.**

Case 6

Examinee's Name _____

Rater's Name _____

Time and Date _____

Place a checkmark by each question asked.
One point will be awarded for each question asked.

	*UA	
Asked about previous episodes of pain	___	___
Asked what factors make pain better or worse	___	___
Asked about appetite	___	___
Asked about unusual vaginal discharge or vaginal bleeding	___	___
Asked about burning while passing water or urinary frequency	___	___
Asked whether patient has ever had anaesthetic before and if there were problems	___	___
Student listened to the abdomen	___	___
Student checked various spots on abdomen to see where it hurts the most	___	___
Student did rebound, or asked patient to cough, or suck in and blow out tummy, to jump up, or to get on or off the examination table	___	___
Student requested and read the rectal and vaginal exam report	___	___

***UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.**

Case 7

Interaction # _____ Rater _____

*UA

Place a checkmark by each question asked. __ __

When did your shortness of breath start? __ __

Has it changed since that time? __ __

Do you have a history of chronic cough and sputum? __ __

Has your sputum changed in amount and character? __ __

Have you had shortness of breath before? __ __

Have you had pneumonia? __ __

Student asked about the following associated symptoms:

Have you ever had blood in your sputum? __ __

Is the pain in your side worse when taking a deep breath? __ __

Have you had fever and/or chills? __ __

Student quantifies smoking Hx (Must ask how much and how long to receive credit). __ __

Are you currently taking medication(s)? __ __

Are you allergic to any medications? __ __

Place a check by each part of the examination performed:

Student examines neck for Tracheal position from the front. __ __

Student palpates (feels) chest wall for tenderness. __ __

Student performs excursion of the lung anteriorly (puts hands on rib cage and asks you to take deep breath or has you hold your breath and presses rib cage) __ __

Student percusses (taps) chest (at least on right side) __ __

Student percusses (taps) back when you are breathing normally. __ __

Student listened to chest. __ __

Student listens to back. __ __

Student palpates (feels) for supraclavicular adenopathy (lumps above collar bone). __ __

*UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.

Case 8

Examinee _____ Patient/Evaluator _____

Date and Time _____

Instructions: On this page you will find statements regarding the medical care provided by the examinee. Please indicate the degree to which you agree or disagree with each statement by circling the number which best represents your feelings and opinions. Please provide constructive suggestions to the examinee on the back of this page.

COMMUNICATION

The examinee explain my child's medical problem so that I understand the problem clearly. 6 5 4 3 2 1 0 *UA

The examinee explained his/her plan for managing my problem so that I know what is being done and why. 6 5 4 3 2 1 0 *UA

The examinee spoke in a clear and coorganized fashion. 6 5 4 3 2 1 0 *UA

PERCEPTION OF PROFESSIONAL SERVICE

The examinee took my feelings and preferences into consideration. 6 5 4 3 2 1 0 *UA

The examinee treated me with respect. 6 5 4 3 2 1 0 *UA

The examinee was not as thorough as should be. 6 5 4 3 2 1 0 *UA

The examinee did the best possible to keep me from worrying unnecessarily. 6 5 4 3 2 1 0 *UA

I'm very satisfied with the medical care I received. 6 5 4 3 2 1 0 *UA

***UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.**

Case 9 (son)

Examinee _____

Evaluator/Patient _____

Time and Date _____

Indicate with a check which of the following occurred in the medical student's interaction with you.

	*UA	
Student made clear that the cause of the problem is uncertain and that any of the following might be part of the problem (one point each for the following):	__	__
_____ dementia		
_____ depression		
_____ medication related		
Student cross checked and confirmed patient's answers with son.	__	__
Student stated an intent to involve the husband (son's father) in the handling of the case.	__	__
Student explained that there are alternatives available for management other than nursing homes, e.g., home help, medicine adjustment.	__	__
Student suggested aids for helping mother remember and manage medications (e.g., day-of-the-week pill boxes).	__	__
Student discussed plans for further evaluation of mother's problems (laboratory work).	__	__
Student clearly communicated and discussed treatment/ management instructions (re: drugs).	__	__
Student discussed plan for future appointments (or gave written note).	__	__
Student involved son in decisions for future plans (negotiated with son regarding plans).	__	__

***UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.**

Case 9 (mother)

Examinee _____

Evaluator/patient _____

Time and Date _____

	*UASATISFACTORY	UNSATISFACTORY	
Evaluate the student's interview with you according to the following criteria:	___	___	___
Used simple words that I could understand.	___	___	___
Patience. Gave enough time for answers. Did not show annoyance regarding inconsistency, etc.	___	___	___
Sympathy: Acknowledged that I was emotionally upset. Offered encouragement, etc.	___	___	___
Performs mental status exam by doing the following (all 3 to be performed satisfactorily to receive credit):	___	___	___
a. orientation (person, place, time)	___	___	___
b. memory: repeat 3 objects, recall 3 objects	___	___	___
c. calculations	___	___	___
Asks pertinent questions regarding depression - sleep habits, eating habits, appetite, level of energy, suicidal ideation.	___	___	___
Persistency. Asks important questions several ways, several times. Given the patient a chance to answer before posing to son.	___	___	___
Cross checks and confirms patient's answers with son.	___	___	___

***UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.**

Case 10

Place a checkmark by each question asked.

HISTORY

*UA

Asked about location of pain. ___ ___

Asked about temporal pattern of pain (continuous or intermittent) ___ ___

Asked if pain radiates and to where. ___ ___

Asked about time of onset of pain (how long have you had pain?) ___ ___

Asked about numbness, tingling, weakness. ___ ___

Asked about problems with bladder or bowel control. ___ ___

Asked about activities immediately prior to onset of pain (what were you doing?) ___ ___

Asked about exacerbating factors. ___ ___

Asked about alleviating factors. ___ ___

Asked if pain had become worse, better or no change since first began. ___ ___

Asked about previous history of back pain. ___ ___

Asked if award of congenital back problem. ___ ___

Asked me what I had done to treat pain, including medicines, specifically: ___ ___

Dose _____ Frequency _____

*UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.

Case 11

PROTOCOL FOR SIMULATORS ASSESSING STUDENTS BY VIDEO

You will be observing a number of interactions between a medical student and a simulated patient, and completing a checklist on each interaction.

Each interaction is numbered 1,2,3...

Watch the interactions in numerical order.

The tape will indicate when each interaction begins and ends.

When an interaction is finished, stop the tape.

You have 5 minutes to complete the checklist. Please ensure that the checklist you use has the same identifying number as the interaction you are assessing.

Please do not complete the checklist while watching each interaction, but only at the end of it.

If you are unable to judge an item on the checklist because you are observing on video rater than in the room, you should check the "unable to assess" column. However, try to complete as many items as possible.

Case 11

Examinee's Name _____ Rater's Name

Time and Date _____

HISTORY

Place a checkmark by each question asked.

	*UA	SATISFACTORY	UNSATISFACTORY
Asked about prior occurrency of headache.	___	___	___
Asked about sequency of events prior to onset of headache.	___	___	___
Asked about temporal pattern of headache (onset, frequency, duration, & change over time).	___	___	___
Asked about pattern of location of headache (location, radiation, movement over time).	___	___	___
Asked about intensity of headache.	___	___	___
Asked about quality of pain (throbbing, pressure, etc.).	___	___	___
Asked about associated symptoms of headache (nausea, vomiting, photophobia, prodromes, rhinitis).	___	___	___
Asked about aggravating and relieving factors.	___	___	___
Asked about relationship to menses.	___	___	___
Asked about past medical history.	___	___	___
Asked about currently allergies.	___	___	___
Asked about medications currently used.	___	___	___
Asked about family history of headache.	___	___	___
Asked about psychosocial history:	___	___	___
Present work situation	_____	___	
Present family situation	___	___	___
Present financial status	___	___	___
Present marital situation	___	___	___

Case 11 (cont.)

PHYSICAL EXAM:

Place a checkmark next to each item performed.

	*UA	SATISFACTORY	UNSATISFACTORY
Examined disc	___	___	___
Palpated temples	___	___	___
Palpated trapezius muscle	___	___	
Palpated neck/posterior occipital muscles	___	___	___
Checked neck range of motion	___	___	
Assessee TMJ status (mouth open/close; fingers in ears)	___	___	___
Checked ears	___	___	___
Checked mouth	___	___	___
Requested clenched teeth	___	___	___

*UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.

Case 12

Examinee _____

Patient/Evaluator _____

Date and Time _____

Evaluate the student's interview with you according to the following criteria:

	*UA	Done Incorrectly	Done Correctly
The student:			
introduced him/herself to the mother	___	___	___
has eye contact with the mother	___	___	___
asks question in a direct manner	___	___	___
uses simple language	___	___	___
proceeds through the history in an organized manner	___	___	___

***UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.**

Case 13

Place a checkmark next to each action taken by the examinee.

		*UA
Palpate thyroid gland	___	___
Check pulse	___	___
Check or ask about hand sweating or tremor	___	___
LYMPH NODE		
Neck	___	___
Armpit	___	___
Groins	___	___
LUNGS		
Auscultate	___	___
Percuss	___	___
Palpate abdomen	___	___
Examine lower limbs	___	___
Request to do a rectal and genital exam	___	___

*UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.

Case 14

Examinee's Name _____ Rater's Name _____

Date and Time _____ Simulator's Name _____

Place a checkmark next to each action taken by the examinee

		*UA
HISTORY		
Lost 40 lbs. in the last few weeks	___	___
Does not eat well and have poor appetite lately	___	___
Swelling in the last two weeks near the axilae and groin area	___	___
Antibiotics do not help fever	___	___
Anemia	___	___
Night sweats	___	___
Fever for the last 2 weeks	___	___
History of Tuberculosis	___	___
Pets at home - Animals in contact with	___	___
Sexual orientation	___	___
Recent dental work	___	___
Recent travel	___	___
PHYSICAL EXAMINATION		
Lymph Nodes	___	___
Examine axillary area	___	___
Examine supraclavicular area	___	___
Examine groin area		
Examine epitrochlear area	___	___
Abdomen	___	___
Palpate liver and spleen	___	___
Cardiopulmonary	___	___
Listen to lungs	___	___
Palpate heart	___	___
Auscultate heart	___	___
Liver	___	___
Examine extremities	___	___

***UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.**

Case 15

Interaction # _____ Rater's Name _____

Place a checkmark against each item mentioned by the examinee

		*UA
Acknowledge to nurse that it could be anaphylactic reaction to ampicillin	___	___
Spoke to patient/attempt at history	___	___
Calls for help (emergency team; 911; resident)	___	___
Checks pulse (nurse gives results)	___	___
Checks blood pressure	___	___
Listens to chest	___	___
Stops ampicillin	___	___
Orders oxygen	___	___
Orders adrenaline: (Nurse should ask for specific dosage)	___	___
___ Inadequate dose (_____)	___	___
___ Correct dose range (.3 - .5cc of 1:1000)	___	___
___ FATAL DOSE (> 1mg.)	___	___
Orders antihistamine	___	___
Orders steroid	___	___
Orders inhales sympathomimethic dilator (if adrenaline is administered & working)	___	___
Orders aminophylline	___	___
Orders investigations (any of the following: CBC, Blood urea/creatinine, blood sugar, electrolytes, blood culture, blood gases, ECG, chest X-ray)	___	___
Orders "mast" pants	___	___
Reassures patient and explains what is happening	___	___

***UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.**

Case 16

Examinee's Name _____Patient/Evaluator _____

Date and Time _____

Instructions: On this page you will find statements regarding the medical care provided by the examinee. Please indicate the degree to which you agree or disagree with each statement by circling the number which best represents your feelings and opinions. Please provide constructive suggestions to the examinee on the back of this page.

COMMUNICATIONS

The examinee explained by medical problem so that I understand the problem clearly 7 6 5 4 3 2 1 *UA

The examinee explained his/her plan for managing my problem so that I know what is being done and why 7 6 5 4 3 2 1 *UA

The examinee spoke in a clear and organized fashion 7 6 5 4 3 2 1 *UA

PERCEPTION OF PROFESSIONAL SERVICE

The examinee took my feelings and preferences into consideration 7 6 5 4 3 2 1 *UA

The examine treated me with respects 7 6 5 4 3 2 1 *UA

The examinee was not as thorough as should be 7 6 5 4 3 2 1 *UA

The examinee did the best possible to keep me from worrying unnecessarily 7 6 5 4 3 2 1 *UA

I'm very satisfied with the medical care I received 7 6 5 4 3 2 1 *UA

***UA - UNABLE TO ASSESS - should be check if simulator is unable to assess because he/she is watching on video and not in room.**

Appendix 7: Chapter 7: Tables A7.1 to A7.8

TABLE A7.1 ACCURACY SCORES FOR PATIENTS WITH AND WITHOUT MISSING DATA FOR PATIENT ATTRIBUTES

Patient Category	N	Percent Accuracy Score (s.d.)	P-Value (t-test)
Missing Data on Patient Attributes	9	96.4% (8.2%)	01
No Missing Data on Patient Attributes	29	93.0% (8.7%)	

Legend: Missing Data: one or more values of the predictor variables for patient attributes are missing.

N: number of patients

TABLE A7.2 THE PERCENTAGE OF TIMES ACCURACY ITEMS COULD NOT BE RATED BY CASE AND PATIENT

Cases With 2 Patients	Number Encounters	Number of Items/ Encounter	Total	Percent Missing Patient #1 N	Mean	S.D.	Patient #2 N	Mean	S.D.	Prob
1	25	19	475	10	22.1	(6.5)	15	20.7	(7.3)	.62
2	21	30	630	16	24.4	(10.7)	5	22.7	(12 3)	.79
3	22	15	330	14	19.0	(11.4)	8	22.5	(11.8)	.51
4	20	26	520	9	58.1	(9.7)	11	56.3	(6.7)	.63
6	30	13	390	15	36.9	(13.7)	15	48.7	(12.2)	01
8	23	19	437	9	31.0	(13.3)	14	35.7	(9.9)	.34
10	22	31	682	13	11.2	(7.1)	9	17.2	(9.3)	.10
12	28	26	728	12	44.2	(8.6)	10	48.5	(8.1)	.25
13	24	22	528	11	30.2	(7.9)	13	32.5	(10.8)	55
14	22	21	462	9	26.9	(28.4)	13	30.4	(7.4)	.68
16	26	24	624	14	41.7	(17.4)	12	43.1	(12.7)	.82
18	21	8	168	5	20.0	(14.3)	16	25.8	(15.5)	46
19	19	30	570	9	51.1	(16.8)	10	54.0	(13.4)	68
20	22	11	242	8	7.5	(4.6)	14	19.3	(14.9)	.01

Cases With > 2 Patients				Pt#1	Pt#2	Pt#3	Pt#4	Pt#5	Pt#6	Prob
11	27	22	N	9	7	6	2	3		.17
			Mean	40.4	33.1	40.2	22.7	33.3		
			S.D.	10.0	11.0	10.1	0	6.9		
			Total	594						
15	28	13	N	3	2	3	6	2	12	34
			Mean	53.8	61.5	35.9	33.3	34.6	46.8	
			S.D.	40.0	10.9	4.4	10.5	5.4	18.1	
			Total	364						
Overall	374	330		7368		33.54 (17.6)			0001	

Legend: Prob: the value resulting from an independent t test (for 2 patients) of F test (for >2 patients) of mean differences between patients within each case. For overall, the p-value resulting from a one-way ANOVA for the null hypothesis that there are no differences in the percentage of accuracy items missing for different cases.

Total: the number of accuracy items to be rated with each encounter times the number of encounters rated

Percent Missing: the percentage of times accuracy items could not be rated

TABLE A7.3 THE PERCENT OF TIMES ACCURACY ITEMS NOT EVALUATED FOR EACH OF THE POTENTIAL PREDICTOR VARIABLES

Predictor Variables	Total Accuracy Items (# items * # encounters)	Number	Percent of Times Accuracy Items Not Evaluated (s.d.)
Group 1			
A) Case Attributes			
Number of Items/Case			
8-15	1472	5 cases	30.9% (19.0)
16-20	912	2 cases	27.3% (11.2)
21-25	2692	3 cases	31.2% (15.8)
26-31	2292	4 cases	44.6% (16.6)
Item Type			
History	6424	290 items	33.3% (18.0)
Physical	714	27 items	48.2% (34.6)
Affect	230	13 items	5 9% (20.8)
B) Patient Attributes			
Age			
20-29 years	1603	8 patients	39.8% (16.1)
30-39 years	1565	10 patients	36.1% (17.4)
40-49 years	590	3 patients	27.6% (17.1)
50-59 years	260	1 patient	48.5% (8.1)
60-69 years	1673	7 patients	23.7% (15.8)
>70 years	1018	3 patients	32.0% (20.2)
Missing	32	6 patients	
Gender			
Male	2277	11 patients	29.8% (16.8)
Female	5059	21 patients	34.7% (18.0)
Missing	32	6 patients	

Predictor Variables		Total Accuracy Items (# items * # encounters)	Number	Percent of Times Accuracy Items Not Evaluated (s.d.)
Previous Experience				
Acting	Yes	4147	21 patients	33.9% (17.5)
	No	3189	11 patients	31.4% (18.1)
	Missing	32	6 patients	
Simulation	Yes	3598	16 patients	33.3% (19.1)
	No	3738	12 patients	32.6% (16.1)
	Missing	32	6 patients	
Health Prob	Yes	3100	14 patients	30 3% (16.4)
	No	4236	18 patients	35.8% (18.6)
	Missing	32	6 patients	
Vicarious	Yes	2638	15 patients	33.9% (17.3)
	No	4698	16 patients	32.7% (18.1)
	Missing	40	7 patients	
Understand	Well	4279	17 patients	35.0% (16 3)
	Fair	2722	13 patients	29 6% (18.7)
	Not	335	1 patient	53.2% (21.0)
	Missing	35	7 patients	
Group 2				
A) Patient Attributes				
Patient Confidence				
51-75%		1112	5 patients	35.6% (16.4)
76-100%		6224	25 patients	32.0% (17.6)
Missing		41	8 patients	
B) Training Attributes				
Trainer Confidence				
51-75%		2582	11 patients	32.1% (18.0)
76-100%		4754	16 patients	32.7% (17.6)
Missing		55	11 patients	
Training Length				
# Sessions	1	1277	9 patients	26.7% (18.7)
	2	4002	18 patients	31.5% (16.7)
	3	2057	5 patients	46.3% (13.2)
Missing		32	6 patients	

Predictor Variables		Total Accuracy Items (# items * # encounters)	Number	Percent of Times Accuracy Items Not Evaluated (s.d.)
Hours	1	815	7 patients	25.7% (18.9)
	2	1448	7 patients	24.0% (15.0)
	3	3016	13 patients	35.2% (16.4)
	4	2057	5 patients	46.3% (13.2)
	Missing	32	6 patients	
MD Assistance (# sessions)	0	872	6 patients	17.6% (13.2)
	1	630	10 patients	28.7% (15.3)
	2	3165	13 patients	38.6% (16.2)
	3	2669	6 patients	49.6% (10.3)
	Missing	32	6 patients	
Group 3				
A) Procedural Attributes				
Number of Sessions Done that Day	1-3	2308		33.8% (19.5)
	4-6	2002		33.2% (16.6)
	7-10	3058		33.5% (16.9)
Time Since Training				
1 week		2940		33.5% (17.0)
2 weeks		4428		33.6% (18.1)
B) Encounter Attributes				
Patient Confidence in Performance	1-2	54		65.8% (13.0)
	3-4	4106		32.2% (16.6)
	5	709		31.3% (19.3)
Student Performance Interpersonal Skills				
0-20		0		0
21-40		0		0
41-60		320		38.6% (19.1)
61-80		6427		33.3% (17.7)
81-100		320		32.5% (16.0)
Student Performance Data Collection				
0-20		0		0
21-40		0		0
41-60		1078		39.6% (18.1)
61-80		5448		32.1% (17.6)
81-100		0		0
Overall		7368		33.54% (s.d.=17.6%)

TABLE A7.4 THE LINEAR RELATIONSHIP BETWEEN ACCURACY OF PATIENT PRESENTATION AND OVERALL COMPETENCY SCORE BY CASE FOR THE 1987 AND 1988 PATIENT STUDENT SAMPLES

Case	Number Encounters	Mean Accuracy Score Patient 1	Mean Accuracy Score Patient 2	Mean Student Score Patient 1	Mean Student Score Patient 2	B	P-Value (HO:B=0)
1987 Cases							
2	17	99.4%	89.9%	71.6%	68.7%	-.03	.94
3	17	90.3%	97.3%	55.6%	54.5%	.04	.86
4	19	70.6%	69.3%	53.4%	50.7%	-.04	.51
5	18	88.2%	100.0%	81.8%	77.8%	.00	.95
6	15	98.0%		76.3%		-.95	.37
7	16	90.9%	79.9%	70.5%	79.6%	-.10	.74
8	15	98.8%	100.0%	54.6%	54.2%	-.62	.70
9	15	83.5%	94.6%	-	-		
11	17	98.2%	100.0%	57.0%	63.4%	1.15	.26
12	17	86.9%	84.1%	89.9%	75.1%	-.23	.85
13	16	69.5%	73.4%	70.8%	69.4%	-.17	.41
14	18	89.9%	87.9%	72.9%	70.5%	.84	.15
15	20	95.2%	98.5%	59.0%	67.5%	.35	.17
1987 & 1988 Cases							
'87 #1 &	37	83.1%	78.0%	77.3%	82.4%	-.09	.68
'88 #20		81.3%	81.2%	67.7%	68.3%		
'87 #16 &	39	99.7%	97.7%	60.8%	75.6%	-.77	.18
'88 #10		100.0%	97.6%	68.0%	68.3%		
1988 Cases (Patients=2/case)							
1	25	99.4%	89.4%	67.5%	69.1%	-.14	.07
2	21	93.0%	97.8%	67.9%	69.5%	-.12	.54
3	22	99.5%	96.8%	68.4%	68.3%	- 19	.43
4	20	85.0%	93.9%	69.6%	67.4%	-.18	.12
6	30	99.4%	94.2%	68.7%	68.8%	-.02	.83
8	23	99.1%	98.3%	68.1%	67.5%	-.08	.76
12	28	87.7%	89.1%	67.2%	69 6%	- 24	10
13	24	93.5%	92.2%	69.1%	67.7%	22	.16
14	22	84.1%	83.8%	67.7%	68.0%	- 04	.75
16	26	98.7%	94.8%	68.6%	68 6%	.16	.19
18	21	100.0%	100.0%	68.2%	68.0%	0	
19	19	98.1%	96.4%	68.7%	67.8%	.00	.95

TABLE A7.4 THE LINEAR RELATIONSHIP BETWEEN ACCURACY OF PATIENT PRESENTATION AND OVERALL COMPETENCY SCORE BY CASE FOR THE 1987 AND 1988 PATIENT STUDENT SAMPLES

Case	Number Encounters	Mean Accuracy Score Patient 1	Mean Accuracy Score Patient 2	Mean Student Score Patient 1	Mean Student Score Patient 2
1988 Cases (Patients > 2/case)					
11	27				
Accuracy Score		98.8%	100.0%	97.2%	100.0%	100.0%	
Student Score		66.9%	68.4%	68.4%	66.1%	68.4%	
Beta=-.19 (p=.45)							
15	28						
Accuracy Score		77.8%	79.2%	88.4%	93.7%	70.1%	84.3%
Student Score		64.5%	68.8%	69.6%	68.5%	69.7%	68.7%
Beta=-.04 (p=.75)							

Legend: The estimate of beta and the P-value is derived from the linear regression of student competence score on patient accurary score

TABLE A7.5 THE LINEAR RELATIONSHIP BETWEEN ACCURACY OF PATIENT PRESENTATION AND DATA COLLECTION SCORE BY CASE FOR THE 1987 AND 1988 PATIENT STUDENT SAMPLES

Case	Number Encounters	Mean Accuracy Score		Mean Student Score		B	P-Value [HO:B=0]
		Patient 1	Patient 2	Patient 1	Patient 2		
1987 Cases							
2	17	99.4%	89.9%	88.0%	77 1%	1.0	.14
3	17	90.3%	97.3%	88.2%	83.3%	-.16	.64
4	19	70.6%	69.3%	74.6%	62.5%	.11	.52
5	18	88.2%	100.0%	76.1%	76.2%	.01	.97
6	15	98.0%		68.6%		.37	.81
7	16	90.9%	79.9%	73.0%	86.3%	-.82	.00
8	15	98.8%	100.0%	52.5%	51.7%	-1.09	.70
9	15	83.5%	94.6%	43.8%	38.6%	- 16	60
11	17	98.2%	100.0%	62.1%	75.0%	33	82
12	17	86.9%	84.1%	-	-		
13	16	69.5%	73.4%	80.4%	68.9%	03	91
14	18	89.9%	87.9%	67.6%	68.9%	.38	37
15	20	95.2%	98.5%	41.1%	58.6%	.46	19
1987 & 1988 Cases							
'87 #1 &	37	83.1%	78.0%	84.1%	86.4%		
'88 #20		81.3%	81.2%	64.0%	69 6%	-.13	61
'87 #16 &	39	99.7%	97.7%	-	-		
'88 #10		100.0%	97.6%	66.9%	66 6%	- 74	28
1988 Cases (Patients=2/case)							
1	25	99.4%	89.4%	63.8%	70.0%	-.17	.24
2	21	93.0%	97.8%	66.8%	69.5%	- 45	.17
3	22	99.5%	96.8%	65.5%	69.2%	-.46	.27
4	20	85.0%	93.9%	70.6%	64.9%	- 45	.02
6	30	99.4%	94.2%	69.0%	66.3%	- 14	32
8	23	99.1%	98.3%	68.7%	65.5%	-.42	36
12	28	87.7%	89.1%	64.3%	70.3%	- 13	59
13	24	93.5%	92.2%	71.3%	63.9%	25	35
14	22	84.1%	83.8%	68.5%	66.5%	- 22	31
16	26	98.7%	94.8%	67.7%	66.2%	20	.4
18	21	100.0%	100.0%	64.9%	67.7%	0	
19	19	98.1%	96.4%	66.7%	67.2%	-.43	04

TABLE A7.5 THE LINEAR RELATIONSHIP BETWEEN ACCURACY OF PATIENT PRESENTATION AND DATA COLLECTION SCORE BY CASE FOR THE 1987 AND 1988 PATIENT STUDENT SAMPLES

Case	Number Encounters	Mean Accuracy Score Patient 1	Mean Accuracy Score Patient 2	Mean Student Score Patient 1	Mean Student Score Patient 2
1988 Cases (Patients > 2/case)					
11	27				
Accuracy Score		98.8%	100.0%	97.2%	100.0%
Student Score		66.6%	64.7%	67.9%	74.0%
Beta= -.76 (P=.04)					
15	28				
Accuracy Score		77.8%	79.2%	88.4%	93.7%
Student Score		63.5%	70.0%	64.1%	65.8%
Beta= -.07 (p=.43)					

	Pt#1	Pt#2	Pt#3	Pt#4	Pt#5	Pt#6
11						
Accuracy Score	98.8%	100.0%	97.2%	100.0%	100.0%	
Student Score	66.6%	64.7%	67.9%	74.0%	62.9%	
Beta= -.76 (P=.04)						
15	28					
Accuracy Score	77.8%	79.2%	88.4%	93.7%	70.1%	84.3%
Student Score	63.5%	70.0%	64.1%	65.8%	75.7%	67.8%
Beta= -.07 (p=.43)						

Legend: The estimate of beta and the P-value is derived from the linear regression of student competence score on patient accurary score

TABLE A7.6 THE LINEAR RELATIONSHIP BETWEEN ACCURACY OF PATIENT PRESENTATION AND DIAGNOSIS SCORE BY CASE FOR THE 1987 AND 1988 PATIENT STUDENT SAMPLES

Case	Number Encounters	Mean Accuracy Score		Mean Student Score		B	P-Value (H0:B=0)
		Patient 1	Patient 2	Patient 1	Patient 2		
1987 Cases							
2	17	99.4%	89.9%	88.3%	88.1%	.29	73
3	17	90.3%	97.3%	52.9%	57.6%	.67	.23
4	19	70.6%	69.3%	90.0%	80.0%	-.02	92
5	18	88.2%	100.0%	92.5%	90.0%	.26	.46
6	15	98.0%		100.0%		0	
7	16	90.9%	79.9%	100.0%	100.0%	0	
8	15	98.8%	100.0%	57.9%	54.3%	-4.6	24
9	15	83.5%	94 6%	-	-		
11	17	98.2%	100.0%	89.3%	87.5%	2.1	35
12	17	86.9%	84.1%	57.7%	72.3%	.36	.76
13	16	69.5%	73.4%	60.0%	68.9%	.18	69
14	18	89.9%	87.9%	82.5%	77.9%	38	72
15	20	95.2%	98.5%	72.2%	77.2%	19	.58
1987 & 1988 Cases							
'87 #1 &	37	83.1%	78.0%	90.0%	82.4%	.14	71
'88 #20		81.3%	81.2%	70.7%	68 3%		
'87 #16 &	39	99.7%	97.7%	56.4%	72 2%	-.81	49
'88 #10		100.0%	97.6%	70.5%	66.3%		
1988 Cases (Patients=2/case)							
1	25	99.4%	89.4%	64.7%	69.9%	27	19
2	21	93.0%	97.8%	68.7%	71 6%	03	96
3	22	99.5%	96.8%	71.3%	69.5%	66	31
4	20	85.0%	93.9%	69.3%	70.6%	16	.60
6	30	99.4%	94.2%	70 5%	69.2%	.06	80
8	23	99.1%	98.3%	71 6%	65.2%	1 06	.22
12	28	87.7%	89.1%	66.4%	68.7%	- 47	24
13	24	93.5%	92.2%	71.6%	68 0%	22	.64
14	22	84.1%	83.8%	67.7%	69.5%	17	.63
16	26	98.7%	94.8%	68.7%	71 5%	.08	.83
18	21	100.0%	100.0%	71.4%	68 0%	0	
19	19	98.1%	96.4%	70.8%	66.0%	24	50

TABLE A7.6 THE LINEAR RELATIONSHIP BETWEEN ACCURACY OF PATIENT PRESENTATION AND DIAGNOSIS SCORE BY CASE FOR THE 1987 AND 1988 PATIENT STUDENT SAMPLES

Case	Number Encounters	Mean Accuracy Score		Mean Student Score	
		Patient 1	Patient 2	Patient 1	Patient 2
1988 Cases (Patients > 2/case)					
11	27	Pt#1	Pt#2	Pt#3	Pt#4	Pt#5	Pt#6
Accuracy Score		98.8%	100.0%	97.2%	100.0%	100.0%	
Student Score		64.8%	66.2%	69.6%	63.4%	73.5%	
Beta=.05 (p=.95)							
15	28						
Accuracy Score		77.8%	79.2%	88.4%	97.7%	70.1%	84.3%
Student Score		67.9%	64.9%	71.1%	66.0%	67.2%	68.3%
Beta= -.23 (p= 15)							

Legend: The estimate of beta and the P-value is derived from the linear regression of student competence score on patient accurary score

TABLE A7.7 THE LINEAR RELATIONSHIP BETWEEN ACCURACY OF PATIENT PRESENTATION AND MANAGEMENT SCORE BY CASE FOR THE 1987 AND 1988 PATIENT STUDENT SAMPLES

Case	Number Encounters	Mean Accuracy Score		Mean Student Score		B	P Value (H0:B 0)
		Patient 1	Patient 2	Patient 1	Patient 2		
1987 Cases							
2	17	99 4%	89.9%	70.0%	67.9%	.54	37
3	17	90.3%	97.3%	38.4%	38.9%	09	89
4	19	70.6%	69.3%	100.0%	100.0%	0	
5	18	88.2%	100.0%	85.0%	95.0%	- 06	89
6	15	98.0%		91.7%		-1 38	53
7	16	90.9%	79.9%	30.0%	25.0%	-.23	70
8	15	98.8%	100.0%	-	-		
9	15	83 5%	94.6%	-			
11	17	98.2%	100 0%	16.9%	33.0%	1 03	61
12	17	86.9%	84.1%	-	-		
13	16	69.5%	73 4%	-	-		
14	18	89.9%	87.9%	-	-		
15	20	95.2%	98.5%	46.7%	51.4%	23	51
1987 & 1988 Cases							
'87 #1 &	37	83 1%	78.0%	-	-	00	97
'88 #20		81.3%	81.2%	59 3%	53.5%		
'87 #16 &	39	99.7%	97.7%	-	-	22	80
'88 #10		100.0%	97.6%	55.8%	54.3%		
1988 Cases (Patients=2/case)							
1	25	99 4%	89.4%	55 7%	54.0%	- 18	25
2	21	93.0%	97.8%	54.9%	56.3%	56	22
3	22	99.5%	96.8%	57.7%	53 9%	20	72
4	20	85.0%	93.9%	54.4%	57.0%	42	10
6	30	99.4%	94.2%	55.8%	57.1%	09	58
8	23	99.1%	98.3%	57 8%	53.4%	28	62
12	28	87.7%	89.1%	56.7%	53.6%	- 66	02
13	24	93.5%	92.2%	55.1%	57.9%	51	11
14	22	84.1%	83 8%	55.1%	55.1%	19	51
16	26	98.7%	94.8%	56.4%	56 0%	37	17
18	21	100.0%	100.0%	56.0%	55.3%	0	
19	19	98.1%	96.4%	57.3%	55.3%	.09	72

TABLE A7.7 THE LINEAR RELATIONSHIP BETWEEN ACCURACY OF PATIENT PRESENTATION AND MANAGEMENT SCORE BY CASE FOR THE 1987 AND 1988 PATIENT STUDENT SAMPLES

Case	Number Encounters	Mean Accuracy Score Patient 1	Mean Accuracy Score Patient 2	Mean Student Score Patient 1	Mean Student Score Patient 2
1988 Cases (Patients > 2/case)					
11	27				
Accuracy Score		98.8%	100.0%	97.2%	100.0%
Student Score		51.3%	58.6%	56.2%	46.5%
Beta= -.06 (p=.91)					
15	28				
Accuracy Score		77.8%	79.2%	88.4%	93.7%
Student Score		56.5%	48.9%	53.4%	55.2%
Beta= -.19 (p=.12)					

	Pt#1	Pt#2	Pt#3	Pt#4	Pt#5	Pt#6
Accuracy Score	98.8%	100.0%	97.2%	100.0%	100.0%	
Student Score	51.3%	58.6%	56.2%	46.5%	60 3%	
Accuracy Score	77 8%	79.2%	88.4%	93.7%	70 1%	84.3%
Student Score	56.5%	48.9%	53.4%	55.2%	58.4%	55.5%

Legend: The estimate of beta and the P-value is derived from the linear regression of student competence score on patient accurary score

TABLE A7.8 THE LINEAR RELATIONSHIP BETWEEN ACCURACY OF PATIENT PRESENTATION AND INTERPERSONAL SKILLS SCORE BY CASE FOR THE 1987 AND 1988 PATIENT STUDENT SAMPLES

Case	Number Encounters	Mean Accuracy Score		Mean Student Score		B	P-Value [HO:B 0]
		Patient 1	Patient 2	Patient 1	Patient 2		
1987 Cases							
2	17	99 4%	89.9%	-	-		
3	17	90 3%	97 3%	-	-		
4	19	70.6%	69.3%	-	-		
5	18	88.2%	100 0%	-	-		
6	15	98.0%		-			
7	16	90.9%	79.9%	-	-		
8	15	98.8%	100.0%	72.1%	73.5%	-1.04	53
9	15	83.5%	94.6%	-	-		
11	17	98 2%	100.0%	-	-		
12	17	86.9%	84.1%	100.0%	97.2%	- 33	43
13	16	69.5%	73.4%	-	-		
14	18	89.9%	87.9%	-	-		
15	20	95.2%	98.5%	-	-		
1987 & 1988 Cases							
'87 #1 &	37	83.1%	78.0%	-	-	09	62
'88 #20		81.3%	81 2%	71.7%	73.1%		
'87 #16 &	39	99.7%	97.7%	68 0%	78.6%	- 50	51
'88 #10		100.0%	97.6%	71.9%	74 7%		
1988 Cases (Patients=2/case)							
1	25	99.4%	89 4%	73.1%	73 8%	- 22	09
2	21	93.0%	97 8%	71.8%	75 8%	46	20
3	22	99.5%	96.8%	72 9%	71 6%	30	49
4	20	85.0%	93.9%	75 2%	70.2%	32	14
6	30	99.4%	94.2%	72.5%	75.2%	.01	92
8	23	99.1%	98.3%	69.9%	74.2%	-.51	25
12	28	87.7%	89 1%	71.6%	75 5%	17	48
13	24	93.5%	92.2%	73.2%	71 6%	.32	27
14	22	84.1%	83.8%	72 7%	72.6%	-.12	.58
16	26	98.7%	94.8%	73.9%	72 1%	24	26
18	21	100.0%	100.0%	69.9%	72 9%	0	
19	19	98.1%	96.4%	74.7%	70.3%	-.00	98

TABLE A7.8 THE LINEAR RELATIONSHIP BETWEEN ACCURACY OF PATIENT PRESENTATION AND INTERPERSONAL SKILLS SCORE BY CASE FOR THE 1987 AND 1988 PATIENT STUDENT SAMPLES

Case	Number Encounters	Mean Accuracy Score Patient 1	Mean Accuracy Score Patient 2	Mean Student Score Patient 1	Mean Student Score Patient 2
1988 Cases (Patients > 2/case)					
11	27	Pt#1	Pt#2	Pt#3	Pt#4	Pt#5	Pt#6
Accuracy Score		98.8%	100.0%	97.2%	100.0%	100.0%	
Student Score		72.3%	73.8%	73.5%	69.3%	67.9%	
Beta= -.20 (p=.65)							
15	28						
Accuracy Score		77.8%	79.2%	88.4%	93.7%	70.1%	84.3%
Student Score		66.4%	71.7%	73.8%	75.1%	73.3%	73.5%
Beta= .03 (p= 74) R							

Legend The estimate of beta and the P-value is derived from the linear regression of student competence score on patient accurary score