1988

A Systematic Reevaluation of the Psychometric Properties of the Assessment Instrument Used in the Development of Evaluation Methods in Dentistry Project

David T. Crandall
Loyola University Chicago

A SYSTEMATIC REEVALUATION OF THE PSYCHOMETRIC PROPERTIES OF THE ASSESSMENT INSTRUMENT USED IN THE DEVELOPMENT OF EVALUATION METHODS IN DENTISTRY PROJECT

by

David T. Crandall

A Thesis Submitted to the Faculty of the Graduate School
of Loyola University of Chicago in Partial Fulfillment
of the Requirements for the Degree of
Master of Arts

August

1988

ACKNOWLEDGMENTS

I would like to express my sincerest gratitude to those who have given me their help, advice, and encouragement in the process of completing this thesis. Special thanks are reserved for Jack A. Kavanagh, Ph.D, who provided the impetus, the professional knowledge, and the guidance for this project. Ronald R. Morgan, Ph.D, who served as the second party on the thesis committee, also deserves my deep gratitude for his time and advice.

I want to express my appreciation to Shobha Srinivasan, Ph.D. for her generous assistance in the data management for this study, and I thank Ms. Joyce Sigmon for her time and effort in helping me review the project literature. Finally, the managers and staff at the Loyola Academic Computing Centers deserve recognition for their kindness, knowledge and assistance.

A particular note of gratitude is extended to the members of the American Dental Association who provided input for this project. This study was completed with the permission and cooperation of the ADA, and it is fully understood that the Assessment Instrument and Evaluator's Manual discussed are confidential. Therefore, permission to use or to view these materials must be obtained from the ADA.

VITA

The author, David Thomas Crandall, is the son of Terrance Michael Crandall and Elizabeth Anne (Klauss) Crandall. He was born October 22, 1964 in Jamaica, New York and raised in Essex County, New Jersey.

His elementary education was obtained in parochial schools of the archdiocese of Newark, New Jersey. His secondary education was completed in May, 1983, at Seton Hall Preparatory School in South Orange, New Jersey.

Mr. Crandall entered the University of Notre Dame in August, 1983, and received a Bachelor of Arts degree in Psychology and Computer Applications in May, 1987.

Mr. Crandall entered the Research Methodology program in the Department of Education at the Loyola University of Chicago in August, 1988. In November, 1988, he became an employee of the Loyola Academic Computing Services. The following January, he was granted an assistantship as a research advisor, and he also received a university fellowship, enabling him to complete the Master of Arts degree in August, 1988.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CONTENTS OF APPENDICES

Chapter I

Introduction

Description of the Original Project

Inception and Purpose.

The W.K. Kellogg Foundation awarded the School of Dental Medicine of the University of Pennsylvania a grant of $1.87 million for a project to begin on February 1, 1982. The project was entitled Development of Evaluation Methods and Computer Applications in Dentistry (DEMCAD), and it was in effect for four and a half years, terminating July 31, 1986. As Morris (1986) states in the "Final Report to the W. K. Kellogg Foundation," "the overall goal of the DEMCAD project was to develop new methods and technologies that can be used by individual dentists and the dental profession to improve the effectiveness and efficacy of the full scope of general practice" (p. 1). To accomplish this endeavor, the project was divided into separate DEM and CAD subprojects, which were devised and developed according to their unique objectives.

Both the DEM and the CAD projects were coordinated through the University of Pennsylvania with the basic premise to develop and test new methods for providing information so that dentists could improve the quality of their professional practices. The DEM component of the project remained under the authority of the University of Pennsylvania with the purpose of developing an objective, practical, and professionally acceptable method for evaluating dental offices through in-office visits. The CAD subproject was delegated to Columbia University, which received $482,600 to create an in-office computer and information system for solo and small group practices to improve fiscal and patient management (Morris, 1986).

The conceptualization and development of the DEMCAD project resulted from the recognition of the dental profession's responsibility to provide quality care appropriate to the needs of patients in light of the limited funding available for health care and the increased competition in the field. Knowledge of the strengths and weaknesses of individual practitioner care, as well as the particular necessities of the clientele, is considered imperative to maintaining high standards in dentistry. The project thesis therefore, centered on the principle that dentistry, as a self-regulating, independent profession, needs to develop its own evaluating system that will best serve the needs of the dentists and their patients (Morris, 1986). The initial step was to establish materials and techniques that would enable professional dentistry to have standardized, practical methods for assessing the dental practice. Perhaps the most effective way to accomplish this was to develop the evaluation system from within the profession.

The evaluation of dental care refers to the systematic use of empirical methods to test the current standards of practitioners across the United States. Prior to establishing tools and procedures to assess the quality of dental care, a definition of quality as it pertains to the field of dentistry was adopted. The definition, taken from Lee & Jones (1933), is as follows: "Quality dental practice is the kind of dentistry practiced by recognized leaders of the dental profession at a given time or period of social, cultural and professional development" (Morris, 1986, p.3).

As a basis for the focus of the testing materials and procedures, the primary objective of the project was to establish answers to these nine fundamental questions, listed by Morris (1986).

1. Can an Assessment Instrument be developed that permits a valid evaluation of private dental practice during a one day visit by one dentist evaluator?

2. Can the Assessment Instrument discriminate between private dental practices that differ in their characteristics?

3. Are the similarities in rural, urban group and urban non-wroup dental practices sufficient that the Assessment Instrument can be used in the evaluation of all practice types?

4. Are dental practices conducted in all geographic areas of the country sufficiently similar that the Assessment Instrument can be used effectively throughout the nation?

5. Can dental practitioners be trained to use the Assessment Instrument in a standardized, disciplined approach that produces comparable results when evaluating comparable dental practices?

6. Can private dental practitioners be recruited to participate in an in-office practice evaluation program?

7. How do private practitioners react to an office evaluation visit?

8. How do evaluators react to conducting evaluation visits to private offices?

9. How much does it cost to conduct an in-office dental practice evaluation program? (p.5)

Development of the Instrumentation.

An Assessment Instrument and Evaluator's Manual were designed to satisfy the project objectives in a manner consistent with the goals of the study. The Instrument was in the form of a 17 page questionnaire to be used by a single dentist evaluator in a one day review of all pertinent facets of private dental practice. This assessment tool was organized into three Dimensions, (Structure, Process, and Outcome) which were originally devised for use in assessing the medical field (Donabedian, 1966), but were adopted to the dental practice by Bailit et al. (1974). Within the context of the three dimensional model, Structure refers to the quality of the dental office and administration as a health care facility, Process refers to the quality of dental procedures in private practice, and Outcome refers to the results of dental treatment and patient satisfaction with the treatment. It should be noted that the Outcome Dimension also has the distinct purpose of validating the Structure and Process portions of the Instrument (Donabedian, 1966).

The Assessment Instrument is organized into a hierarchical structure of subdivisions. Each Dimension is broken down into Components, which are in turn divided into Elements, which are finally made up of Subelements. The three Dimensions do not have the same number of Components, Elements and Subelements, the Components do not have the same number of Elements and Subelements, and the Elements differ in their numbers of Subelements. In total, there are 19 Components and 105 Elements.

It is important to distinguish among test questions, test items, and Subelements with regard to the organization of the scales. On this evaluation tool, a question is equivalent to a test item. Subelements are also items, being the lowest order of the hierarchical structure, but not all items are Subelements. There are items on the Instrument that reside at the Element and Component levels. In other words, some items do not complete the structural arrangement of Dimension, Component, Element, and

Subelement. There is a total of 248 items on the Assessment Instrument, 3 of which are Components, 76 that are Elements, and 169 that are Subelements.

Evaluators.

The dentist evaluators who tested the Assessment Instrument in the private dental offices were 10 general practitioners recruited from across the United States. These evaluators were taught to apply the Instrument using standardized methods in 2, three-day training sessions, one in November of 1983 and the other in January of 1984. A third training session, lasting one and a half days, was scheduled in January of 1985 after the first year of field experience.

Participants.

The professional dentists who participated in the testing procedures were recruited on a volunteer basis. A total of 3,015 letters were sent to dental practitioners across the nation requesting participation in the study. The goal was to recruit 21 or 22 offices from each chosen state. The states in which the dental offices resided were grouped into 14 regions representing separate areas of the United States. A total of 300 general practitioners were tested: Fifty rural offices, 50 urban group practices, and 200 non-group practices.

Testing Procedures.

Testing of the Assessment Instrument on the 300 private practices occurred in the third and forth years of the study. The 10 evaluators each made appointments with selected dentists who had agreed to have their offices reviewed. The evaluation visits each lasted approximately three and a half hours. Along with the data collected from the Instrument, participants' reactions to the evaluation process, evaluators' reactions to performing the office reviews, and project costs were systematically assessed.

Purpose of the Study

The purpose of this study is to reassess the psychometric characteristics of the Assessment Instrument used in the DEM study. The areas on the Instrument needing adjustment will be identified through the application of test construction procedures, and suggestions will be made in reference to corrections that can improve the Instrument as an assessment tool. A careful psychometric reevaluation of the Instrument is imperative because the original research was not performed using current test construction techniques to establish the construct validity or the reliability of the Instrument and the scales. Therefore, it is questionable if the validity and reliability of the Assessment Instrument have been established in the first place.

Based upon the objective of the project, the ex post facto study was designed to test for the internal validity of the Instrument by measuring criterion validity, scale reliability and item validity. Internal validity of the assessment tool is critical; whether or not the test scales are accurately indexing the quality of dental care in the United States is the crux of demonstrating construct validity for the Instrument.

Unfortunately, the establishment of external validity using the test data from the office evaluations is questionable because the sampling procedures employed were inadequate with respect to obtaining a representative cross-section of practitioners. Hence, it is doubtful that the testing results can be generalized to adequately account for regional differences in dental care across the country. The reliability of the Assessment Instrument at reproducing identical results from the testing of a different sample of private practitioners was not previously established and is not the focus of the present study.

The basic problem with the DEM Assessment Instrument therefore, appears to be the lack of systematic empirical procedures used in testing its effectiveness as a psychometric tool. Despite the hierarchical organization of the Instrument, the scales of items at each level of the design were not analyzed. Dimension and Component scores were produced, but there was no discussion of how the scales were formed and no Element, Subelement, or individual item distributions were provided. Although the descriptive statistics for the Component scales were detailed in the original analysis, no validity coefficients were generated, and no item analysis was performed on the scales.

Perhaps the most glaring difficulty with the Assessment Instrument is the non-uniform construction and weighting of the test items and the attempt to create summative scales and subscales with them. No documented reasons were provided for the discrepancies in the number and weighting of scale points for each DEM item or for the assignment of scale and item weights. These problems with the distributions and weighting of items, scales, and subscales cause the formation of summative Dimension scales to be highly suspect, and it is questionable if the Dimension total scores are at all meaningful.

CHAPTER II

Literature Review

In Chapter II, an outline and description of some of the relevant assessment procedures developed to measure quality in the health field are presented. By no means is this an exhaustive review of the literature concerning the establishment and use of assessment techniques and instruments in the achievement of quality control. The articles cited in this chapter focus primarily on the medical and dental disciplines, and the studies discussed center on peer group review in the medical field, patient satisfaction with medical care, and evaluation techniques in dentistry.

Peer Group Assessment

Peer group studies are evaluations of physician decision making based on criteria provided by peer consensus (Anderson & Shields, 1982). Sanazaro (1980) reviewed the development of the analytical techniques employed in peer assessment procedures and came up with two basic features of medical auditing: selecting an important element of performance and comparing the observed level of performance with predetermined criteria or standards. According to Sanazaro (1980), a study by Sheps (1955) "made explicit the view that assessing the quality of hospital care involves application of general principles of measurement and evaluation, especially reliability and validity" (p. 42). Sheps (1955) maintained that quality approval is based on three aspects: assumed prerequisites i.e., facilities, organization, and staff standards, the elements of performance, and the effects of care. The criterion-related validity for these three standards is either normatively or empirically demonstrated, and it must be established for each standard. Lembcke (1956) assessed individual patients using valid criteria and employed independently set standards to evaluate the performance of a complete medical staff (Sanazaro, 1980).

> The papers of Sheps (1955) and Lembcke (1956) described evaluation of hospital-based medical care in operational terms. Scientific validity of criteria was emphasized; standards of good practice were defined as the level of performance observed in a reference group of hospitals. The approach was empirical, descriptive and practical. (Sanazaro, 1980, p. 42).

Donabedian (1966) reformulated the three aspects of quality (Sheps, 1955) and called them structure, process, and outcome. "Structure describes the physical, organizational and other characteristics of the system that provides care of its environment. Process is what is done in caring for patients. Outcome is what is achieved, and improvement usually in health but also in attitudes, knowledge and behavior conducive to future health" (Donabedian, 1987, p. 35). These three dimensions are not mutually exclusive in their interrelations, and their individual relationships to quality require validation. Donabedian (1969, 1986) defined the criteria specifications for the dimensions of quality and devised a system for assessing the criteria.

The review of patient records and the process of auditing criteria based records have been popular methods of assessing physician performance. In several studies, Morehead (1964, 1967, & 1974) described extensive experience with the use of expert physician reviewers to examine medical records (Sanazaro, 1980). This technique was employed in judging the quality of care provided by fellow physicians.

The work of B. C. Payne in the late 1960's and early 1970's determined the approach to auditing medical records used by most hospitals at the time (Sanazaro, 1980). In 1961, Payne (1967) adopted previously developed criteria for auditing medical, gynecologic, and surgical care in reviewing of the accuracy of medical charts. This study was expanded, and Payne (1973) worked with panels of practicing specialists who created sets of criteria for optimal performance with 51 different conditions covering 135 diagnoses. The purpose of the study was to encourage changes in the diagnostic and therapeutic behaviors of the medical staff (Sanazaro, 1980).

Peer group assessment does not occur solely through the collective appraisal of medical records and criterion based record audits. Physicians can rate one another's performance on the quality of health care provided. Anderson and Shields (1982) described a number of studies in which physician decision making was based on process criteria established through peer consensus. In 1953, Peterson, Andrews, Spain, and Greenberg (1956) studied the practice patterns of general practitioners in North Carolina. They observed the physicians over several days and subsequently classified them into one of five categories ranging from excellent to mediocre. Another study performed in North Carolina by Hulka et al. (1979), looked at the quality of ambulatory care. Procedures involved the creation of consensus lists of items that were considered essential and likely to be recorded for the conditions of diabetes, hypertension, general examination and dysuria.

Anderson and Shields (1982) reported that peer group analysis procedures were used in several studies to evaluate areas in the health field such as indicators for admission to hospitals and decision making involved in drug prescribing. Criteria developed through peer consensus were used by Fitzpatrick, Riedel, and Payne (1962) to evaluate several specialties and estimate the proportion of admissions and length of stay in the medical facility as appropriate. Becker et al. (1972) interviewed physicians to assess their prescriptions for five common complaints, illnesses, and drug products. A panel of expert judges evaluated the accuracy of the decisions produced from the interviews.

The assurance of quality medical care received a strong backing when the Professional Standards Review Organization (PSRO) was established in 1970 to monitor medical services and determine if they were necessary and acceptable of professional standards (Anderson & Shields, 1982; Sanazaro, Goldstein, Roberts, Maglott, & McAllister, 1972). That same year, the National Center for Health Services Research and Development (NCHSRD) formed Experimental Medical Care Review Organizations (EMCRO), which served to systematically analyze the content of medical care for patients (Sanazaro et al., 1972). Most EMCRO's adopted criteria proposed by specialty panels and reviewed by general practitioners. Sanazaro et al. (1972) present a table of 15 common diagnosis and the number of EMCRO's that developed criteria for them (p. 1127). The criteria emphasize the process of care, and the two main sources of data for these criteria are insurance claims and medical chart abstracts.

A study that explicitly describes the analysis of five peer review methods in their assessment of quality medical care was performed by Brook and Appel (1973). They began by categorizing the judgments used by physicians in making a decision as either implicit or explicit. Implicit judgments are based on the subjective opinion of the individual; no predetermined criteria were assessed. An explicit judgement involves predetermined criteria established by group agreement. Five peer assessment methods were broken down into three implicit methods and two explicit methods.

The implicit methods entailed the implicit judgments of process, of outcome, and of a combination of the two. For these procedures, the physicians read a detailed, two-page abstract of each case; page one included information about the process of treatment, and the outcome data were on page two. To make the implicit-process judgement, the physician read only page one and decided whether the process of care was adequate. For the implicit-outcome judgement, the physician read page two and stated whether the patient's outcome could have improved if the process had been better. The implicit quality-of-care judgement was based on the physician's conclusion on the overall quality of care.

Criteria were created for the two explicit methods of decision making. The explicit process method had two steps: For each of three medical conditions, the physician was asked to select criteria necessary for good care provision and a favorable outcome. And, seven specialists were chosen for each of the conditions to select the criteria, who also made explicit-process judgments for each of the criteria. The physician made the explicit-outcome judgement by stating what the patient's outcome will be given the particular condition and treatment.

Rather than assess the quality of ambulatory care by using one of the above methods, Hastings, Sonneborn, Lee, Vick, and Sasmour (1980) devised a peer review checklist. A panel of full-time clinicians experienced in quality assessment by unstructured peer review were the participants. Ten ambulatory care medical records were reviewed by each clinician, who recorded relevant observations about the quality of care and listed the criteria used in the judgments. Initially, 59 criteria were considered important. The 10 physicians with the most experience in quality audit by peer assessment assigned weights to the items, and the weights were normalized for each clinician. The criteria were subsequently categorized into six subject areas. The items were analyzed, and 35 were kept. The scale was tested for interrater agreement and intrarater agreement, and the Physician Reexamination method was used to establish the instrument validity. During this procedure, patients were reinterviewed and reexamined, and the findings were correlated with the checklist results.

Patient Assessment

The assessment of patient satisfaction with medical care focuses on the psychological dimension of health care and measures the attitudes that the patient has toward the provider and the care received (Koslowsky, Bailit, & Valluzzo, 1974). Sanazaro (1980) discusses an unpublished study (Sanazaro & Williamson, 1967) in which patients appearing for emergency appointments were interviewed in person prior to being treated by one of eight participating interns. One week later, the patients were again interviewed by phone, and the findings from the interviews were compared to office records in order to identify changes in symptoms, functional status, knowledge and attitudes toward condition and treatment, and concerns over costs. Sanazaro (1980) concluded "the evidence is mounting that patient interviews combined with chart reviews based on valid criteria provide a more complete assessment of physician performance" (p. 51).

Several Patient Satisfaction Questionnaires (PSQ) have been developed to help improve the quality of care through patient input. In a summary of their conceptual work and empirical results from previous studies, Ware, Snyder, Write, and Davies (1983) describe the construction of Form 2, the most comprehensive and reliable version of their PSQ developed in 1976. "The strategy for developing and testing the PSQ focused on improving the reliability and validity of items and multi-item scales" (Ware et al., 1983, p. 248). A taxonomy of the characteristics of patient satisfaction was built to classify the satisfaction measures and assess the content validity of the PSQ. Eight dimensions were formed through a factor analysis of the test items, which are presented in Table 4 of Ware et al. (1983, p. 256).

The original method for selecting PSQ items was through an in-person interview survey testing over 900 items. The results of that method produced Form 1 of the PSQ, but Form 2 was subsequently developed to be shorter and self-administered. From 2 was tested over a four-year period that involved the formulation of the dimension models of patient satisfaction, the construction of dimension measures, empirical tests of the models and measures, and refinements in both areas (Ware et al., 1983). Studies of Form 2 were replicated in four independent field tests.

Items on the Ware, Snyder, and Write (1976b) PSQ were scaled on a 5-point Likert-type response scale with the points ranging from <u>strongly agree</u> to <u>strongly disagree</u>. Several different visual questionnaire formats for the presentation of the criteria were tested. The instrument reliability was established for the individual subscales and for the entire PSQ. Internal consistency reliability measures were obtained with the KR-20 coefficient, and a subgroup of respondents was given the PSQ six months later to establish test-retest reliability.

Several approaches were used to produce the validity indices for the PSQ, and the process of instrument validation is considered by Ware et al. (1983) to be on-going. The content validity of the instrument was indexed by a systematic review of the criteria by expert practitioners. By comparing results across alternate testing methods, Ware et al. (1976b) obtained both convergent and discriminant validity, and a factor analysis of the item and subscale structures helped verify the criteria. Criterion-related validity indices were established through the relation of the PSQ criteria to health and illness behaviors thought to be influenced by individual differences in patient satisfaction.

Hulka et al. (1975) also devised a questionnaire to measure patient satisfaction as an outcome, and it was compared to the Ware et al. (1976c) PSQ in a study by Roberts and Tugwell (1987). Three dimensions of patient satisfaction were developed by Hulka et al. (1975), Professional Competency, Personal Qualities, and Cost/Convenience. The criteria were scaled in two ways, using a Likert-type method and using the Scale Product method, which is a weighting technique. Measures of internal consistency were obtained for the instrument subscales. As previously mentioned, the Ware et al. (1976c) PSQ assessed eight dimensions of satisfaction that were established through the factor analysis of items.

According to Roberts and Tugwell (1987), their article "compares the quality of data obtained from two different questionnaires developed through two different methods: one a more conceptual clinical approach and one through more statistical psychometric methods" (p. 639). Both questionnaires were administered to patients at four and six months post-myocardial infarction. The order of presentation was randomly interchanged for the patients. Results from the questionnaires were analyzed comparing the Hulka et al. (1975) PSQ against the Ware et al. (1976c) PSQ, as well as comparing the two types of scales for the Hulka et al. (1975) instrument against each other. Roberts and Tugwell (1987) support the use of either questionnaire to assess patient satisfaction with medical care.

In a recent study, Matthews and Feinstein (1988) attempted to discover patients' opinions regarding medical care and to use those comments in the construction of a system for the interpersonal exchanges of professional care. Their research methods involved two phases. First, open-ended interviews with hospitalized patients were conducted, during which patients were asked to discuss their positive and negative reactions to the physician's care. These interviews lasted approximately 30 to 60 minutes and were extremely detailed. Second, the patients' comments were organized into categories and then arranged into a taxonomy of desired behaviors in the personal aspects of patient care.

Evaluation Methods in Dentistry

Having detailed some of the developments of quality care assessment in the medical field, this review will turn to the quality evaluation in dentistry. Again, this is not a comprehensive review of the assessment methods created and employed by private dental practitioners, but several important studies involving quality assessment procedures will be referenced.

Friedman and Schoen (1972) performed a study for the purpose of gaining practical experience in auditing dental care by reviewing patient treatment records and radiographs without clinically examining the patients. They designed a form to audit the patient records which allowed for the scoring of treatment categories. The evaluation form was divided into three areas, the patient examination, i.e., history, charting, and radiographs, the assessment of treatment, and the evaluation of the type of procedure followed. These three areas were further subcategorized, and the subcategories and three major subject areas on the instrument were all scored independently of one another. The instrument total score was the average of the subcategory scores for the three subject areas. Criteria for the form were scored with both positive and negative values: 0 = inadequate, 1 = adequate, 2 = good, 3 = excellent, -1 = inadequate due to omission, and -1 to -3 = unnecessary treatment.

In a two-part study, Koslowsky et al. (1974) assessed satisfaction with dental care from the point of view of both the patient and the practitioner. The first instrument devised was a patient satisfaction index. On this questionnaire, patient satisfaction was separated into four dimensions of technical competence, personality, organization of the office, and financial consideration, with each dimension being made up of at least two items. The patient satisfaction scale was a 5-point, Likert-type format ranging from strongly agree to strongly disagree.

Koslowsky et al. (1974) constructed the final form of their questionnaire in two stages. Initially, 57 criteria were presented to 150 participants and item analysis procedures were run to evaluate the items. The criteria were then divided into Form A and Form B of the questionnaire, each having 23 items after one was dropped because of a low item-total correlation. The two forms were presented to dental patients, one prior to treatment and one after treatment. The results were analyzed for scale internal consistency and alternate forms reliability. Criteria with low item-total correlation coefficients were removed from the scales. Table 1 in the Koslowsky et al.

(1974) article shows the final questionnaire form, which has 20 items (p. 190).

Part 2 of the Koslowsky et al. (1974) study involved the creation of an instrument to assess dentist satisfaction. The rationale for this instrument was that the dentist's degree of satisfaction may have an impact on the quality of care. Five dimensions of dentist satisfaction were established: Income and security, intellectual fulfillment, responsibility and independence, working conditions, and accomplishing a goal in life. The number of items in each of these dimensions varied.

The instrument items were scaled with both a 5-point, Likert-type formation and the semantic differential set up. "Dentistry" was chosen as the stimulus word for the semantic differential testing, and the purpose for this procedure was to check the reliability of the Likert scales and to serve as an alternative to the Likert scale items if they had low reliability. The internal consistency of the scales was measured, and the Likert scale items were compared against the semantic differential items. Table 4 in the Koslowsky et al. (1974) article presents the final form of the dentist satisfaction instrument (p. 192). It is composed of 22 items, 17 of which are on a Likert scale and 5 that are evaluated with the semantic differential.

Peer group review was discussed as a common technique for judging quality care in the medical field, particularly through the auditing of patient records and the assessment of decision making by fellow physicians. However, Milgrom, Weinstein, Ratener, and Morrison (1978) studied dental care by requiring 1196 practitioners to perform self-evaluations of quality of their restorations. This study was divided into two phases of self-evaluation procedures. In phase 1, the dentists conducted seven general evaluations of their practice without the presence of patients. A 21 item questionnaire was mailed to the dentists which was composed of two 7-point, Likert-type scales for operative dentistry and crown and bridge, and five 7-point scales on esthetics, tissue health, margin smoothness, contours, and occlusion. Patients of the dentists were then recalled and examined by trained personal or the dentists themselves. The criteria evaluated were taken from the Bailit et al. (1974) study.

Phase 2 of the Milgrom et al. (1978) involved a new sample of patients. The dentists completed a 12 item questionnaire for each patient, which included two self-evaluations of the quality of the operative dentistry and crown and bridge provided to the patient. These criteria were also on a 7-point, Likert-type scale. As in phase 1, the patients were recalled and reexamined on the Bailit et al. (1974) criteria.

The final study of the quality of dental care to be reviewed is by Bailit et al. (1974). This is an article of particular interest because it is similar in concept to the present study being reevaluated. Bailit et al. (1974) conformed to Donabedian's (1969, 1986) process of formulating criteria for quality care. The criteria in this study are normative and they were developed for the common conditions dentists treat. Bailit et al. (1974) also conformed to Donabedian's (1966) dimensions of the structure, process, and outcome of quality care. Process was separated into four components, History & Examination, Diagnosis, Treatment Plan, and Treatment.

The data for the Bailit et al. (1974) instrument were collected through the evaluation of patient records and the clinical examination of patients. The dimensions were broken down into the components of care, which were in turn composed of elements. A specific criteria was created for each element. Bailit et al. (1974) decided to drop the Diagnosis component because direct evidence on diagnoses could not be obtained, and the subject matter was partially covered in the Treatment Plan component. The scale for the criteria is dichotomous: 1 = unsatisfactory, 2 = adequate, and 9 = no decision. Also, a 5-category scale was devised to rate the general quality of each patient's care. (p. 847).

Analysis for the quality assessment instrument involved measures of reliability, validity, variability, and practicality. Two indices of interrater reliability were obtained. Content validity was established through the review and approval of each scale item by a committee of experts. Predictor-criterion correlation coefficients were generated as criterion-related validity indices, and concurrent validity was established by correlating the Treatment Plan and Treatment components. The practicality measure was simply the amount of time required to learn the criteria and evaluate the patient.

CHAPTER III

Method

Organization of Data

The original data from the Instrument testing were received from representatives of the American Dental Association in the form of 16 separate data files on three floppy diskettes on February 3, 1988. These 16 data files were uploaded into a SAS dataset entitled DEMCAD.DATA on the IBM system 3081D mainframe computer at Loyola University of Chicago. Each of the SAS data files was extensively reviewed, and errors were corrected where possible. Nine of the 16 data files contained the data results from the Assessment Instrument testing described in Chapter I.

A second dataset called DEMCAD.SAS was created, and each of the nine raw data files containing the testing results was accessed with a separate SAS control program in the new dataset. The SAS programs assigned names and labels to the item variables in the particular data files accessed. The variables were named according to the Dimension, Component, Element, and Subelement that they represent, and those items not in all levels of the hierarchical structure were named only by their corresponding levels. For example, an item in Dimension II, Component B, Element 3, and Subelement c was named IIB3c. An item in the first Dimension, Component C and Element 4 was called C4 (note that a Roman numeral was not used to indicate a variable in Dimension I, and a Subelement was not specified). This naming convention made it easy to identify items in the analysis results with their corresponding test questions.

The nine SAS control programs together produced names and labels for all 248 DEM test items, using the raw data from the DEMCAD.DATA dataset. These nine programs were then converted onto separate subfiles and stored in a single SAS system file called DEMCAD.SASSYSTM, which held the data values, variable names and variable labels corresponding to the items. Finally, a second SAS system file, DEMCAD.SASYSTEM, was created to merge and store the nine subfiles holding the data and information in DEMCAD.SASSYSTM. At this point, all of the data from the Assessment Instrument could be accessed at once from the DEMCAD.SASYSTEM file. These data, variables and labels were retrieved and used in procedure programs from the DEMCAD.SAS dataset to perform the statistical analysis for the study.

Creation of Scales

The first procedure in analyzing the data from the Assessment Instrument was to generate scale totals, distribution statistics, and graphic representations of all scales, subscales, and items. The SPSS^{x} statistical package was used to access the DEMCAD.SASYSTEM data file in performing these analyses. Initially, a frequency distribution was obtained for each item, along with distribution statistics and a histogram, to indicate how each item is represented across all 300 cases. Element scales were generated as well, with a new variable being created for each one. Each of the variables ELEM1 through ELEM58 is a summation of the scale items, or Subelements, that comprise it. However, 76 of the 105 Elements are actually at the item level, so no summing of Subelements is involved in their totals; their scale score is an individual data point from a test question.

Similar procedures were performed on the Component and Dimension scales. Nineteen new variables, COMP1 through COMP19, were summed across the items composing them. Three of these Components exist at the item level. Dimension variables were created in the same manner, by adding the corresponding items to form DIM1, DIM2, and DIM3. These Component and Dimension scales were also created by adding the lower scales in the structural hierarchy that comprise them, but the totals did not differ from those generated at the item level. Hence, the scale totals at all levels were computed by adding their respective items, and no weighting was used in these initial calculations. All of the scale variables were obtained to get an idea of how the items, Elements, Components, and Dimensions are distributed and for use in other analysis procedures.

Replication of Results

To verify the DEM scale distributions reported in the initial analysis, the scale scores from the Assessment Instrument were originally summed for all 300 cases. However, the simple summation of test items could not be used in replicating the results. Instead, the 19 Component scales were added together after being weighted according to procedures described in the DEM Evaluator's Manual. The weighting procedures are not uniform across the Components-- only 11 of them are assigned a weighted scale total, and the scale value of some Components is reduced through multiplication by a fraction, while others are increased through multiplication by a positive integer. This table illustrates the operations and constants used in weighting the Component scales:

DEM COMPONENT WEIGHTS

VARIABLE	LABEL	WEIGHT
COMP1	Facilities	Multiply by 0.5
COMP2	Personnel	Multiply by 2.0
COMP4	Administration	Multiply by 2.0
COMP5	Practice Management	Multiply by 3.0
COMP6	Radiographic Eval.	Multiply by 0.5
COMP7	Data Collection	Multiply by 0.5
COMP8	Diagnosis	Multiply by 2.0
COMP9	Treatment Plan	Multiply by 4.0
COMP10	Treatment	Multiply by 0.5
COMP11	Steril.-Infec. Cntl.	Multiply by 2.0
COMP12	Patient Management	Multiply by 2.0

Distribution Statistics.

The weighted Components were used in reproducing the total score, Dimension, and Component statistics, as well as those for the Element Treatment, represented in Tables 39, 40, 41, 42, and 43 of Morris (1986). The total score was computed by adding the Structure, Process, and Outcome Dimension scores, which were in turn produced by adding the weighted Components comprising them. As the table indicates, no weighting coefficients were used with the scale scores of the Outcome Components. Like the Component they make up, the Elements of Treatment were multiplied by 0.5 to reflect the weighting of their scale. A SAS statistical procedure provided the scale means, standard deviations, and coefficients of variation for these distributions.

Graphic Distributions.

The graphic distributions of the total scores and Dimension totals from Figures 6, 7, 8, and 9 of the Morris (1986) report were also reproduced through the summation of the weighted Component scales. Using the SPSS^{X} program to generate these distributions, the specified interval width of each frequency histogram was matched to that of the original graph so the two sets of results could be adequately compared. All of the actual Tables and Figures that were replicated from the Morris (1986) report are presented in Appendix B.

Criterion-Related Validity Indices

Product moment correlation coefficients were generated with an SPSS^{X} procedure to establish criterion validity for the Assessment Instrument. For this Instrument, the criterion validity coefficients identify how well the scores on the testing areas relate to those in the outcome section (Ghiselli, Campbell, & Zedeck, 1981, Ch. 10). Since the Outcome Dimension was intended as a validity index, the weighted Components of the Structure and Process Dimensions were correlated with the weighted Outcome Components. These validity coefficients show whether or not the Components of Structure and Process relate to any of the Outcome

Components, and if so, the specific strength and direction of their relationships.

The Structure, Process, and Outcome Dimensions were correlated with one another to determine the nature and size of their relationships. These Dimensions are the summations of the weighted Components forming them. The validity coefficient for the Process and Outcome Dimensions is of particular importance, since the quality of dental procedures should be indicative of the treatment results. Also, the three Dimensions should all have some degree of relation to each other because they individually assess the construct quality (Donabedian, 1966).

Item Analysis

The item analysis of the DEM scales involved two procedures. First, the SPSS^{X} statistical package provided an assessment of the internal consistency of the items on the Component scales to indicate the reliability index for each Component scale. Second, SAS programs were used in comparing each item against the quartiles of the smallest scale on which it falls to discover how well the items discriminate on their criterion scales.

Scale Reliability.

In determining scale reliability, a Cronbach's coefficient alpha was generated for each non-weighted Component scale. Three Component variables, Patient Education, Patient Disability, and Completion of Treatment could not be evaluated because they exist at the item level. The desired results for this procedure are a high inter-item correlation coefficient and a high coefficient alpha value for each scale. A high inter-item correlation on a scale signifies that all scales are simultaneously measuring the same thing, and a high coefficient alpha indicates that there is internal consistency among the items, and the scale reliably assesses the score of an individual. The results of this procedure point out which items, if any, should be removed from their respective scales because they lower the scale reliability.

Item Validity.

The purpose of an item validity index is to assess how well the scale items correlate with the criterion they measure by assessing the correspondence of the item scores to their scale scores (Anastasi, 1976, Ch. 8). To do this, the second item analysis procedure generated a discrimination index for each DEM item. Each item was compared against the ranked percentages of the lowest scale on which it falls in the hierarchical structure. In other words, if an item is a Subelement, it was compared against an Element scale, if the item is an Element, it was cross-tabulated with a Component scale, and if the item is a Component, it was compared against a Dimension scale.

First, the scores on each comparison scale were ranked, and the rankings were divided into 25th percentiles, or quartiles. Each item on the scale was then tabulated against the scale percentages to see how its scores distribute in the quartiles. This information was represented in the form of a table that compares the item scores from all 300 dentists to the scale quartiles from all cases. Ideally, those dentists who scored in the lower 25% on the scale should have consistently low scores on the individual scale items, and the ones scoring in the upper 25% on the scale should have high item scores.

In the tabular results, perfectly discriminating items have their scores falling in the cells along the diagonal of the table, indicating that the 300 cases scoring in a particular quartile on the criterion scale scored the same way on that item. A poorly discriminating item has its scores falling in the off-diagonal cells, demonstrating that the way individuals scored on the item is not indicative of how they tended to score on the criterion scale. The upper and lower quartiles are of primary interest because they provide the numbers of individuals who scored high and low on the test item. If these amounts are heavily skewed toward either end of the distribution or grouped in the middle, the item is not adequately discriminating on the scale, and it therefore does not adequately measure the criterion.

Factor Analysis

In addition to the item analysis procedures, a factor analysis was performed using SAS to generate a principle components analysis on the 19 non-weighted Components. This was done to evaluate how well the items comprising each Component relate to one another and to produce any underlying factors that make up the separate Component scales. If the DEM Components can be subdivided according to theoretical factors, the results will point out which scale items load onto each particular factor, and judgments can be made as to what the factors represent. The axes of the Components were orthogonally rotated to aid in the locating of factors.

CHAPTER IV

Results

Replication of Results

Distribution Statistics.

The initial step in reproducing the results as reported and discussed in the original study was to come up with the statistics from Tables 39, 40, 41, 42, and 43 in Morris (1986), which are presented in Appendix B. Scale means, standard deviations, and coefficients of variation were produced in each of these tables, as well as an individual's possible score on the scale and the scale mean as a percentage of the possible score. Since these last two statistics are derived, they are not of primary concern here; the replication of means, standard deviations, and coefficients of variation was the focus of this procedure. All of the scales and variables used are the ones previously developed to review the Dimension, Component, Element, and total score distributions.

In short, the first four sets of statistics for the DEM total scores, the Dimension scores, and the Component scores were all successfully reproduced with the exception of rounding differences. Tables 1 - 4 in Appendix A match their corresponding tables from the Morris (1986) report in Appendix B. The final set of statistics that were reproduced are the average scores for the Elements of the Treatment Component. The rationale for generating these numbers is that the Treatment Component in Table 3 is the largest of any DEM Component scale score. Therefore, Treatment contributes very highly to the total score for each case. Despite the size of the Treatment mean, this Component score has been divided by two prior to analysis, and each of the four Element scores was similarly weighted before generating the statistics. Table 5 in Appendix A shows that these statistics were accurately reproduced from the originals in Table 43, Appendix B, again taking into account the rounding of decimal places.

Graphic Distributions.

The graphic representations of the total score and Dimension distributions are attempts at replicating the analyses in Figures 6, 7, 8, and 9 from Morris (1986), which are presented in Appendix B. Each of the four new graphs is a histogram with the interval width equal to its corresponding distribution from the initial results. In general, these graphs are approximate representations of the original ones, but none of them is exact enough to be considered an accurate reproduction.

Figure 1 in Appendix A is the distribution of total scores across the 300 dentist participants. The interval width is 21 and the number of intervals is 20, both of which match the original graph. However, the lower real limit and upper real limit of Figure 1 are 369 and 789, versus those of 389 and 788 in Figure 6 of Appendix B. One of the intervals in Figure 6, 569 - 588, is incorrectly computed, having a width of only 20. These discrepancies in the results explain the differences in the graphic distributions of total scores.

The replication of Figure 7 in Appendix B, the distribution of Structure scores for each case, is displayed in Figure 2 of Appendix A. The interval width for both graphs is seven, but the distribution in Figure 2 has 18 intervals while the one in Figure 7 has 20. The lower and upper real limits of these distributions also do not match. For Figure 2, they are 78.5 and 204, and they are 84 and 200 in the original graph. Three of the 20 intervals in Figure 7, 125 - 130, 148 - 153 and 171 - 176 have incorrect widths of six instead of seven, helping to account for the inability to adequately duplicate the histogram.

An attempt to reproduce the distribution of Process scores in Figure 8 of Appendix B is represented in Figure 3, Appendix A. Each figure has an interval width of 16 and 20 intervals in total. The lower real limit and upper real limit of Figure 3 are 170.7 and 488, versus 187 and 481 for the original graph. Many improperly calculated intervals explains these distinct differences. Six intervals in Figure 8, 217 - 231, 261 - 275, 305 - 319, 349 - 363, 393-407 and 437 - 451 are erroneous, causing the originally reported distribution limits to be incorrect.

The distribution of Outcome scores for all cases, Figure 9 in Appendix B, was the last graph to be reproduced, and the results were not successfully duplicated. The interval width for each histogram is four, but the number of intervals in Figure 4 of Appendix A is 16, while Figure 9 has 20. The new graph has lower and upper real limits of 79.9 and 142.2, which do not equal those of 82 and 140 in the original. The intervals of 100 - 102 and 120 - 122 each have a width of three, not four, further demonstrating that the initial graph is incorrect.

Criterion-Related Validity Indices

The first set of procedures to establish Instrument validity shows how the structure of the dental facility relates to the outcome of treatment by correlating the Structure Components with the Outcome Components. Table 6 in Appendix A presents the correlation coefficients of the significantly related Structure and Outcome Components. Only 14 of 28 possible relationships were found to be significant at the .05 level, and all but one of them were weakly associated. For purposes of this study, a weak relation is roughly a validity coefficient of 0.00 to 0.40, a moderate association is between 0.40 and 0.70 and a strong one is between 0.70 and 1.00.

The single moderate correlation in this procedure is that of Administration with Completion of Treatment. Three of the seven Outcome Components, Patient Oral Hygiene, Periodontal Disease, and Completion of Treatment, are best at relating to the four Structure Components. In fact, they each correlate significantly with all Structure Components, except for Patient Oral Hygiene, which does not relate to Personnel.

Testing the relationships among the Process and Outcome Components was the most important procedure in establishing criterion-related validity for the Instrument. These validity coefficients indicate how the process of private dental practice relates to the treatment results. There are eight Process Components and seven Outcome Components which combined for 56 possible relationships, 34 of which were significant at the .05 level. Table 7 in Appendix A provides a list of the 34 significant associations among the Process and Outcome Components. Twenty five of them are weak relations, while eight are moderate, and one, the correlation of Treatment Plan with Completion of Treatment, is strong.

The four Outcome Components of Patient Oral Hygiene, Patient Education, Periodontal Disease, and Completion of Treatment have the greatest number of significant correlations with the Process Components. Respectively, they account for six of eight, six of eight, eight of eight, and seven of eight significant relationships, which is 27 out of the 34 that occurred. Patient Disability, an Outcome Component at the item level, yielded the worst results, producing no significant associations with the Components of Process.

The intercorrelation of the Structure, Process, and Outcome Dimensions provided indices of how strongly the separate Dimensions related among themselves. Of particular interest are the coefficients for the Structure with Outcome and the Process with Outcome pairings, since the Outcome Dimension serves as a validity index. All three of the relationships were found to be significant at the .05 level: Structure with Process = .5002, Structure with Outcome = .4056 and Process with Outcome = .7996. The two correlations involving Structure were moderate, and a desired strong relationship existed for Process and Outcome.

Item Analysis

Scale Reliability.

Each Component scale on the DEM Assessment Instrument was evaluated with respect to the homogeneity of its items. The reliability procedure indexes the scale internal consistency for each Component by providing individual item and total scale coefficients. The item coefficient indicates what the internal consistency reliability of the scale would be if the item were deleted (SPSS^{X} User's Guide, 1986, Ch. 45). For the purpose of this study, a scale alpha level of less than 0.70 is considered too low for a reliable set of scale items, a coefficient alpha of 0.70 to 0.80 is moderate and needs some improvement, and a coefficient higher than 0.80 indicates good scale internal consistency.

Table 8 in Appendix A details the DEM Components with their corresponding internal consistency coefficients and inter-item correlation coefficients. Notice that all Component scales, with the exception of Treatment Plan, had low inter-item correlation coefficients, signifying poor item homogeneity (Anastasi, 1976, Ch. 8; Ghiselli et al., 1981, Ch. 13). Three of the 19 Components, Patient Education, Patient Disability, and Completion of Treatment are items, and therefore could not be measured with this item analysis procedure.

The interpretation of the results from the reliability assessment procedure focuses on the relation of the individual items to the total scale. Each item on the Component scale is judged by whether or not the scale alpha level would be raised if that item was removed. In other words, if the presence of a question on the scale lowers its internal consistency, then that item is not assessing the same thing as the other items on the scale.

A low item-total correlation coefficient and a low squared multiple correlation coefficient for an item are also determinants of a poor scale item. The item-total correlation coefficient indexes how well the item relates to the scale total score, and thus the other items. The squared multiple correlation coefficient indicates the amount of variability in the total score explained by the item. Normally, the alpha if-item-deleted is higher than the scale alpha when the item has poor internal consistency characteristics.

A list of all the DEM items that appear to detract from their respective Component scale reliabilities is presented in Table 9 of Appendix A. Notice that not every item has an alpha if-item-deleted score that exceeds its scale alpha value. These are items with fairly high alpha if-item-deleted coefficients coupled with low multiple correlation indices and low item-total correlation coefficients. They should therefore be reviewed for possible removal from the scale.

The results of the reliability procedure are broken down by the Components with low, moderate, and high internal consistency coefficients. Seven of the 16 scales analyzed require substantial revision of their items due to an unacceptable alpha level. Three of these Components, Patient Oral Hygiene, Recall, and Periodontal Disease, have alarmingly low internal consistency coefficients, and they all belong to the Outcome Dimension (see Table 8, Appendix A). Patient Oral Hygiene has only three scale items, one being an Element which detracts from the scale alpha level. Recall is completely made up of two Elements, neither of which can be deleted, so the scale is extremely unreliable. Of the four Elements on the Periodontal Disease scale, one must be removed, further exemplifying that the scales with very few items tend to do a poor job at measuring with reliability.

The four other Components with low internal consistency coefficients fall within the 0.50 to 0.70 range, as shown in Table 9, Appendix A. The Practice Management Component has a fairly low alpha level for a 16 Subelement scale, which is supported by the fact that five of those items need to be reviewed. The Patient Management scale is another one with low internal consistency. On this scale, 2 of the 10 items, which are Elements, appear to lower the scale alpha. The only Structure Component with low scale reliability is personnel, which is made up of ten items. Three of them are questionable, and they all exist at the Element level. The last Component scale with low internal consistency is Treatment. Both Subelements and Elements comprise this eight item scale, but only the Elements are problematic--all three should be considered for removal.

The DEM scales with moderate alpha levels can potentially be improved by dropping the poor items. Scale alpha coefficients range between 0.70 and 0.80 on these five Components (see Table 8, Appendix A), and except for the Diagnosis Component, there are substantially more items comprising them. The Sterilization-Infection Control Component has 20 Elements, seven of which need review. The Equipment scale is made up of 59 Subelements and 12 of them are questionable. Similarly, Administration has 9 out of a possible 34 Subelements that lower the scale alpha level, so it is apparent that the Components with many items can have their reliability improved without a considerable reduction in the number of test questions.

Another scale with a combination of Subelements and Elements demonstrates that the Elements provide the reliability problems. Data Collection has 14 items, two of which are Elements that lower the alpha level, while only one Subelement should be deleted. This scale has an unusually high variance of 184.21, which may result from the Subelement-Element combinations or differences in the number of points on the scale items; some questions have four choices while others have two, and they all reflect the summation of five patient records. The Diagnosis Component also has a very high scale variance of 144.98, and its coefficient alpha is fairly good. However, the scale consists of only three items, and one of them must be removed, leaving a two-item scale. Even if a high internal consistency coefficient results in this case, a scale with so few items is suspect.

The four remaining Component scales have good scale reliability indices of 0.80 or higher (see Table 8, Appendix A). Facilities, the first Component on the testing Instrument, is made up of 30 Subelements, and eight of them should be deleted. The items on this scale have either dichotomous or 4-point scales. This discrepancy may result in the high scale variance of 51.02 and the lack of internal consistency. Patient Satisfaction demonstrates fairly good scale reliability with only 2 of 17 questionable items, which are Elements. Despite the strong internal consistency of the Radiographic Evaluation Component, having only 3 of 11 Subelements requiring scrutiny, its scale variance is extremely elevated at 140.73. This phenomenon is due to the continuous nature of the Element scales and the differences in their possible range of scores.

Table 8 in Appendix A shows that the DEM scale with the highest coefficient alpha is Treatment Plan. The one characteristic this scale has that differs from the others is an inter-item correlation coefficient of 0.8431, indicating a very high degree of item homogeneity. However, several problems with this scale render it doubtful that the alpha level accurately represents the scale reliability. The fact that only four Elements comprise the Component suggests that the scale internal consistency would normally be low, and even though the individual items have good alpha if-item-deleted coefficients, one of them is still subject to review (see Table 9, Appendix A). The scale variance of 54.88 is also quite inflated.

__Item Validity__.

The second item analysis procedure evaluated each DEM item on whether or not it adequately measures the criterion by distinguishing between the individuals who scored high and low on the scales. Based upon the results, they have been categorized by whether their discrimination ability is good, questionable, or poor. Good items, presented in Table 10 of Appendix A, clearly distinguish among different dentists' performance on the scale, and poor items simply do not. Those items with questionable discrimination ability are not particularly bad discriminators, but they fall short of accurately and discretely differentiating among individuals on the scale, and therefore should be reviewed (see Table 11, Appendix A).

In general, very few DEM items, only 35, adequately show how the 300 dental practitioners differ in their level of quality with respect to a particular criterion. Fifty four items are questionable discriminators and need to be carefully scrutinized before assuming good item validity characteristics. More than half of the DEM items are blatantly poor discriminators, which is evident in the way they fail to distribute properly in the quartiles of their criterion scales.

Factor Analysis

The results of the factor analysis procedure do not shed any light on how the Component items relate to one another. No Component generated a small enough number of factors that were identifiable and theoretically explainable. Components that are made up of a few items obviously produced only one or two factors, but this was expected and does not aid the interpretation of how the test questions combine to measure a construct. The Component scales with many items, such as Facilities and Equipment, produced 10 to 20 factors, most of which equally shared items with one or two others. When the number of factors for a Component was reduced for logical interpretability, they explained merely 25% to 35% of the scale variability.

In sum, a factor analysis of the DEM Components demonstrated that their items tap into many of the same things without helping to define the theoretical contents of the components, so no definitive factors were discovered.

CHAPTER V

Discussion

Replication of Results

The ability to reproduce the results of the DEM study is very important in demonstrating the credibility of the present study. Accurate replication of the original statistics would show that the data received from the American Dental Association, as well as the versions of the Assessment Instrument and Evaluator's Manual being analyzed, were involved in the Instrument testing procedures. Also, when identical results are reproduced, it is easier to understand how the data were initially analyzed.

Distribution Statistics.

The successful duplication of the distribution statistics for the DEM scale total scores and Dimension scores verified the accuracy of the data and analysis procedures used in this study. As previously mentioned, the simple addition of scale items was not sufficient in matching the scale totals and statistics. Correct replication of these numbers was contingent upon the application of the weighting scheme for the Component scales.

The use of scale weights in computing the Component distribution statistics was not immediately apparent. In fact, the first few attempts at verifying the DEM results were unsuccessful because the weighting procedures were not explained or justified in the Assessment Instrument, in the Evaluator's Manual, or by Morris (1986). The only indication that total scale values were being manipulated was the "How to Score" directions for each Component in the Evaluator's Manual, where the person scoring the Instrument results was instructed to multiply or divide the scale score by a constant. Again, no general explanation or theoretical reasons for altering the total score values of the Components were provided, and there was not a discussion of why particular Components were assigned their weights. It seems that the scales were arbitrarily assigned constants which would increase or decrease their total score value and distribution statistics in comparison with other scales.

Given the preexisting differences in scale scores due to varying numbers of items and their possible values, it is not clear why certain Components were augmented and others diminished in importance. Some Component scores are larger than others without the use of scale weighting prior to weighting, so the assignment of weights may have been an attempt to make up for uneven item values resulting from the differences in subscale points. In other words, various Components are composed of dichotomously scored items while others have 4-point items or continuously scored items. Multiplying a dichotomous scale by three would make its total equivalent to that of a 4-point scale, but this logic was not employed in the Instrument analysis. Also, the Components have varying numbers of items, which is a weighting factor.

Evidence that the Component weighting scheme was not devised to account for differences in item scores is provided by scales such as Facilities, which has Subelements scaled across two scale points as well as Subelements on 4-point scales. The Facilities total score was divided by two, thus reducing the effect of a dichotomous item to 0.5 for a _yes_ score and a 4-point item to a possible score of 2. Since both dichotomous and 4-point scales have a low value of 0, the criterion scale itself becomes artificially skewed because one end cannot be altered.

The absence of weighting procedures for the Outcome Components suggests that the Structure and Process Components were weighted because they are considered more important than the Outcome validity scales. No justification for this possible explanation was provided in by Morris (1986), but it is apparent that the author deemed the Structure and Process Dimensions as more indicative of the quality of dental care than the Outcome Dimension.

The Morris (1986) report also failed to clarify the rationale behind reporting the statistics for the Elements of the Treatment Component. Table 5 in Appendix A shows that Treatment has the highest mean of all Components, which is directly related to the fact that its Elements are represented by a large range of values. For example, the Endodontic and Periodontic Elements have dichotomous scale values of 0 or 25 and Oral Medicine has a value of 0 or 10. Restorative and Dies, which are comprised of Subelements, have continuous ranges of at least 0 - 60 and 0 - 30 respectively. Judging from Table 5, it is evident that a total score on this Component could exceed 100 even after the scale score was divided by 2.

The apparent value of the Treatment Component is that it documents the completeness of patient records. Why the Elements receive such high scores was not explained, nor was the relative importance of Endodontic and Periodontic records over the Oral Medicine records. In terms of the Instrument total score, the presence of Endodontic records was viewed as 25 times more important than any item in the Equipment Component. Reasons supporting the extreme importance of the Treatment Component other than completeness of patient records, which is also covered in the Data Collection Component, were not given.

Table 1 in Appendix A shows the possible scores for the Structure, Process, and Outcome Dimensions. Structure represents 25% of the total score, Process is 58% of the total and Outcome represents 17% of the total. These percentages directly result from the application of component weights; they do not reflect the percentages of the total score each Dimension occupies based on its number of scale items. The Structure Dimension is composed of 133 items which is 53% of the total, Process has 86 items, 35% of the total, and Outcome is made up of 29 items representing 12% of the total. These figures reveal a reversal in the relative importance of the Structure and Process Dimensions on the scale total scores. This shift in importance was not defended, and there was no explanation of why the Dimensions have vast differences in their numbers of items.

Graphic Distributions.

The obvious problem existing with the original distribution graphs is the discrepancies in the interval widths for each of them. Simply put, the widths of the intervals were calculated incorrectly at intermittent points in the distributions. The cause of these errors is most likely related to the fact that the number of intervals for all four histograms was held constant at 20. It is not clear if the interval width and the number of intervals were simultaneously forced or if pre-setting the number of intervals alone produced the problem, since the procedures used in the process were not detailed in Morris (1986). Even if the holding the number of intervals constant did result in errors, it does not explain why there are 20 intervals in both sets of results for the total scores and Process score distributions, and the original graphs still have incorrect interval widths. Logically, the number of intervals in these histograms should be 20 since the replicated results produced the proper amount of intervals necessary to accurately duplicate the specified interval width. Although the errors in the DEM results cannot be completely accounted for, it is concluded that the histograms appearing in Figure 1 through Figure 4 in Appendix A are the accurate representations of the scale distributions.

Summary of Replication Results.

The Component weighting procedures make the interpretation of the results particularly difficult. Clearly, the practical and theoretical reasons for assigning scale weights must be detailed and aptly defended before conclusions about the scale statistics and distributions can be drawn. There are no intuitive reasons apparent that justify the use of Component weights other than to make the Process score more significant in the totals. If this is in fact the purpose of the weighting system, then it should be properly documented.

Also of concern is the adequate use of procedures in establishing the weighted Component scores, which are subsequently applied in the calculations of the Dimension and total scores. Arguably, the individual item weights, which are similarly not defended, pre-weight the scales they comprise, as do the differing numbers of items on the scales.

Criterion-Related Validity Indices

The statistical comparison of the Structure Components with the Outcome Components produced generally poor correlations, as shown in Table 6, Appendix A. The fact that only half of the correlations were significant, and all but one of them is a weak relation reinforces this judgement. Administration and Completion of Treatment had the only moderate relationship in the set of comparisons, which was expected since a patient most likely completes the dental treatment prior to establishing the detailed records required in the Patient Related Records Element (see Assessment Instrument).

The purpose of this procedure is to show that if the structural components of dental practice, i.e., Facilities, Personnel, Equipment, and Administration are considered crucial factors in the quality of care, then they should relate well with the outcome measures, which has not been demonstrated. This problem is most likely due to the construction procedures for the Outcome Dimension more than the Structure Dimension (McAuliffe, 1979).

The Standards for Educational and Psychological Testing (1985) specifies that "all criteria measures should be described accurately, and the rationale for choosing them as relevant criteria should be explicit" (p. 16). There were no specific Outcome measures designed to relate to the four different Structure Components, so no particular associations were expected. Therefore, it is not absolutely clear why the Structure Components consistently related only to Patient Oral Hygiene, Periodontal Disease, and Completion of Treatment. These correlations were probably a result of the patient records kept in the dentist's office. All three of the significant Outcome Components are somehow made up of or related to the compilation of patient records, which reflect the structure of dentistry. It could be stated then, that the significant relationships are probably bogus since they apparently rely upon the mere presence of patient records and not the theoretical importance of how the structure of dental practice leads to a favorable outcome. A possible interpretation of these results would be that a favorable outcome is a complete and documented sequence of treatments, not the successful effects of the treatment.

The Component correlations among the Process and Outcome scales were better than those among the Structure and Outcome scales, yet less than two-thirds were significant. A strong association existed between Treatment Plan and Completion of Treatment (see Table 7, Appendix A), but this relationship is intuitive since a patient is unlikely to have a series of Treatment Plan records without first completing the treatment.

Four Outcome Components, Patient Oral Hygiene, Patient Education, Periodontal Disease, and Completion of Treatment were best at relating to the Process scales. They accounted for 27 of the 34 significant correlations and seven of the eight moderate ones. Out of these eight moderate relationships, five involved the Periodontal Disease Component, which related fairly well with Radiographic Evaluation, Data Collection, Diagnosis, Treatment Plan, and Treatment. In fact, Periodontal Disease was significantly correlated with all Components from Structure and Process.

A review of the scoring procedures for Periodontal Disease in the Evaluator's Manual, revealed that the numbers were based directly on the question scores from the Data Collection, Diagnosis, Treatment Plan, and Treatment Components from Process. Obviously, a significant relationship between Periodontal Disease and the other Components will result from this circularity in scoring, so these Components cannot be considered independent of one another. The development of scales using such scoring procedures is not the proper way to establish criterion validity. As stated in the Standards (1985), the "criteria should be determined independently of the predictor test scores" (p. 16).

The Patient Oral Hygiene Component posed a similar problem; scores from the Treatment Component were used to generate the Patient Oral Hygiene totals (see Evaluator's Manual). Patient Education, which related to six of the eight Process Components, also exhibits the circularity of using data points from the Components it serves to validate. The total score for Patient Education is formed from the summation of correct item responses on the patient questionnaires. Morris (1986) does not state whether the questions summed are the same as those in the Patient Management Component, but Patient Education and Patient Management were significantly associated (see Table 7, Appendix A).

Completion of Treatment is one of the three Components at the item level on the Assessment Instrument. However, this item was multiplied by 3 prior to being recorded as a Component score and was not assigned a Component weight. Technically, the Completion of Treatment item was given 3 points for each completed treatment. Neither the reason for this unorthodox method of weighting the item, and hence the Component, is known, nor is the rationale for altering the weighting procedure.

The one moderate relationship that did not involve the four Components previously discussed is the correlation of Patient Management with Patient Satisfaction (see Table 7, Appendix A). Both of these scales are in the form of a questionnaire (see Assessment Instrument). Patient Management deals with the quality of treatment the patient receives in all areas of the visit, and Patient Satisfaction records the patients' response to the dentist's care. Therefore, this relationship is expected.

No significant relationships were discovered between the Patient Disability Component in Outcome and any of the Structure and Process Components. Foremost, Patient Disability is an item, which means there is little substance in its power to correlate with entire scales. Secondly, the Patient Disability score is simply the number of hours lost due to dental emergencies (see Evaluator's Manual). The Standards (1985) maintain that "the technical quality of all criteria should be considered carefully" (p. 16), yet it is unclear how this subject matter relates to the structure and process of dentistry, and it is safe to conclude that Patient Disability is unacceptable as an Outcome measure.

Although all of the Dimensions were significantly related to each other, which is desirable if they measure the same construct, their validity coefficients are questionable. The size of the correlation coefficient for

Process and Outcome seems quite high in comparison to those between their individual Components. Table 7 in Appendix A shows that only two-thirds of these Components were related, few of which are better than weak associations, yet the Dimension coefficient is strong. This phenomenon cannot be explained easily; it may be an effect of the Component weighting scheme or poorly constructed subscales and items. Overall, none of the three Dimension coefficients are as high as they should be if each one is legitimately assessing quality.

The Component considered to be the key outcome of patient care is Patient Satisfaction (see Evaluator's Manual). The question raised here is whether or not the patient questionnaire was properly constructed to assess the outcome of treatment. None of the standard psychometric testing procedures used in development of the assessment instruments for the Koslowsky et al. (1974), Hulka et al. (1975), or Ware et al. (1976a, 1976b, & 1976c) studies were cited by Morris (1986), and no other references to patient satisfaction scale construction were discussed. Besides, this Component was significantly associated with only five of the 12 Structure and Process Components, which considerably diminishes its efficacy as an outcome measure.

Summary of Validity Results.

The results from correlational procedures suggest that there are no adequate indices of criterion validity for the DEM Assessment Instrument. The problems most likely stem from the poor criterion validity indices in the Outcome Dimension (McAuliffe, 1979). To begin with, the Outcome criteria were not separately and operationally defined so they would be exterior to the Structure and Process Components to which they should relate. Sheps (1955) claims that the validity of each of the three standards of quality appraisal was independently established. Donabedian (1966) likewise maintains that although the three dimensions are interrelated, their individual validities must be established with the construct quality. Instead, the Outcome criteria for the DEM study were artificially developed from Process item scores, so they were not individually validated. McAuliffe (1979) states that "contrary to the current practice, outcome measures must be empirically validated just as process measures must, for outcome measures of quality are not obviously valid" (p. 124).

The circularity of using the subtotals from Process Component scoring sheets to create totals for the Outcome Components produced an interdependence of Dimensions and Components. Hence, the outcome indices of Periodontal

Disease, Patient Oral Hygiene, and Patient Education are, in effect, still processes. In sum, the Outcome Dimension must be a set of empirically established criterions relating to the theoretically based and psychometrically sound procedures used to test quality care.

Item Analysis

The purpose of the item analysis procedures is to assess the item properties of the DEM criteria. The reliability procedure establishes the degree to which the Component scale items intercorrelate and thus jointly measure the intended construct. A scale with highly intercorrelated items is considered homogeneous because the items all measure the same thing. Therefore, as Ghiselli et al. (1981) point out, an item should be chosen on the basis of its high, positive intercorrelations with the other scale items to maximize the scale reliability (Ch. 13).

The item validity procedure does not look to establish a homogeneous grouping of items. The purpose of this procedure is to demonstrate how well each item relates to an external criterion, not the other items. For this reason, it is desirable to create heterogeneous scales of items that index a specific, empirically established criterion instead of a particular characteristic or construct. Anastasi, (1976) claims that the best items have the highest association with the criterion and the lowest relation to the scale total score (Ch. 8). To achieve maximum validity, the scale items should have little or no relation to one another and a high correlation with the criterion (Anastasi, 1976, Ch. 8; Ghiselli et al., 1981, Ch. 13). The discrimination analysis accomplishes this by showing that good items have the same scoring patterns as their scales. In other words, the items correlate with the criterion the scale was designed to measure.

A problem arises when selecting items based on these two procedures. Obviously, test items cannot be chosen to maximize both scale homogeneity and item heterogeneity. This difficulty is addressed by Ghiselli et al. (1981), who maintain that the way to simultaneously maximize scale reliability and item validity is to construct several subtests, each with high internal consistency reliability, and correlate these subtests with an external criterion (Ch. 13). For now, both item analysis techniques will be reviewed and their implications for the Instrument scales and items will be discussed.

Scale Reliability.

The fact that all but one of the Component scales had a low inter-item coefficient suggests that the DEM Components do not have homogeneous items. This is understandable given that the Components are made up of Elements which could also be assessing unique aspects of the Component construct. Although the Elements would better assess the homogeneity of the Subelements, the Component scales are more complete because 58 of the DEM Elements are items due to the unsystematic construction of the subscales within the structure of the Instrument. Despite the low inter-item correlations for the Component scales, varying levels of the coefficient alpha were generated.

Three of the Component scales, Patient Oral Hygiene, Recall, and Periodontal Disease, had such low reliability indices that they cannot be improved without complete scale reconstruction (see Table 8, Appendix A). The small number of items comprising each of these scales is immediately evident. The combination of very few items and a low alpha coefficient renders it impossible to increase scale reliability by dropping an item. This procedure cannot be performed with the Recall Component anyway, because it has only two items. A decent scale should have substantially more test questions to adequately measure the criterion. The four other Components with low internal consistency have more items, between 8 and 16, but not enough to delete the increased amount that detract from their scales. Due to the relatively large amount of test items that would have to be dropped, these scales also require major reconstruction to achieve a sound level of reliability.

Judging from the results, the most significant problem of the DEM Component scales with low internal consistency is that they are not composed of enough items to be accurately measuring the construct. In general, Component scales with more items tend to have an elevated coefficient alpha. Therefore, the greater the number of items on a scale, the more fully it assesses the quality of dental care as it pertains to the individual Component. Items cannot be arbitrarily added to scales to improve the reliability index, a procedure that is counter-intuitive to evaluating items on their ability to fit well on a scale.

Another difficulty with the low reliability Components is the frequency of problematic items at the Element level. Of the seven Components with low internal consistency, only one, Practice management, is composed of Subelements. The other six are either all Elements or a combination of Subelements and Elements. The explanation for this problem is uncertain. Based on the hierarchical arrangement of the Instrument, Elements are theoretically more important than Subelements, which are equivalent to test items. Interrupting the scoring pattern of the scales and subscales by creating items at various levels of the hierarchy may produce difficulties when items from the various levels are added together or compared.

Most of the poor Component scales are in the Outcome Dimension, where there are few Subelements, and the items are predominantly continuous in nature. This instability in forming items mostly through counting patient records probably results in poor item consistency across the Component scale. Specific problems with the construction of subscales, like the formation of continuous items and weighted questions, will be further detailed later in the discussion.

On the scales where Elements and Subelements are both at the item level, the Elements were more problematic. For example, the Patient Oral Hygiene Component has three items -- the only questionable one was the Element. Treatment consists of eight items, three of which require removal from the scale, and they are the only Elements (see Table 9, Appendix A). These findings reinforce the supposition that the violation of the hierarchical structure of the Instrument adds to the scale problems. However, the frequency of Elements that lower scale reliability may also be a byproduct of other more serious difficulties with the scale and subscale construction.

Components with a moderate reliability rating can more easily be improved and therefore do not require complete revision. The advantage of these scales is that they have a sufficient number of items that the problematic ones can be removed without seriously altering the scale. Likewise, the Components with high internal consistency typically have many items and few that need to be removed from the scale.

Treatment Plan, the Component with the highest alpha level on the Instrument, is a special case which will be further reviewed.

Some of the Components with moderate alpha levels are made up exclusively of Elements, but the problem of Elements detracting from the internal consistency of a Component is not as prevalent when many of them compose the scale. However, when the Sterilization-Infection Control Component is compared to Equipment and Administration, it is evident that the Components composed of Subelements have a much better ratio of poor to good scale items. Seven of the 20 Elements on the Sterilization-Infection scale need revision while 12 of 59 and 9 of 34 Subelements should be deleted from the Equipment and Administration Components respectively (see Table 9, Appendix A).

Analysis of the Data Collection Component, which has two Elements and 12 Subelements, showed that both of the Elements should be dropped while only one Subelement needs scrutiny. Apparently, the use of Elements to constitute scale items reduces the internal consistency of the Components even when the coefficient alpha indicates fair reliability. Again, the cause for the instability of Elemental items when they are combined with Subelements is probably related to the way the items were created and how they were scaled.

Several of the DEM Components have a curiously high scale variance. Interestingly, these Components produced either moderate or high internal consistency indices in the analysis, so the large variances are not necessarily damaging to the scales. The probable cause for the elevated variances is the formation of items and the construction of subscales, which concerns the level of the scale Elements. For example, Data Collection is a combination of Elemental and Subelemental items, and its scale variance is 184.21. More important is the way these items were made; the Elements are the addition of five 4-point items prior to being scored, and the Subelements are the summation of five 2-point items before being recorded (see Evaluator's Manual). Technically, the actual data points added to form the items are in a lower level of the scale hierarchy. This is fine for Elemental items, since the data points could be considered Subelements, but the data summed to make Subelements are really sub-Subelements.

The Diagnosis, Facilities, and Radiographic Evaluation Components all have high scale variances and possess either continuous items, incompatible subscales, Elemental items, or some combination of these, which is further evidence that the elevated scale variances are related to problems in the construction of items. Diagnosis is made up of Elements that are summed across five patient records, Facilities has both 2-point and 4-point Subelement scales, and Radiographic Evaluation is composed of Subelements that are the addition of data points to form a continuous subscale. These scale configurations are displayed in the Assessment Instrument and described in the Evaluator's Manual.

The Treatment Plan Component had the highest coefficient alpha on the Assessment Instrument, but it is not sufficient to claim that it is a reliable scale. The Treatment Plan scale reported an inter-item correlation of approximately 85%, meaning the four items are sharing most of the scale variability. When this occurs, the scale items are extremely homogeneous, but there may be an alternative reason for these high results in light of the other scale characteristics. There are only four items existing at the Element level on the scale, and they consist of the data from five patient records (see Evaluator's Manual). The scale variance is also inflated at 54.88.

These problems with Treatment Plan may result from discrepancies in the scoring procedures or simply the setup of the dental facilities. Since the scoring is contingent upon the presence of patient records, many practitioners would have equal scores if dentists, by nature of the practice, tend to either have or not have the four sets of records in the office. If this pattern arose in the testing process, it would account for the interdependence the Sequencing, Completeness, Appropriateness, and Implementation Elements, and the consistency of Component scores across participants.

Item Validity.

Each of the DEM items was compared against the lowest scale it falls on because that particular scale is the most immediate criterion on the Instrument for the item. Unfortunately, all of the item criterions are not at the same level of the hierarchical structure, which would allow for comparison among the items. However, this point is practically moot since so few of the items demonstrate adequate validity. A poorly discriminating item does not index the criterion because its scoring pattern is different from that of the scale. This problem is most likely a function of the way the test items were written, as evinced by some of the more common patterns of item distributions and their corresponding tabulations against the criterion scale. Although the reasons for such failure to distinguish properly between high and low performance are not definite, some are clearly evident. For instance, dichotomous items which do not discriminate well typically have very high scores or very low scores. An example is the Subelement Laboratory, which has 291 yes answers and 9 no answers. Obviously, most dentists easily gain a point on this question.

The poorly discriminating items with 4-point scales tend to have negatively skewed distributions with many scores falling in the number 2 category on the 0 - 3 scale.

For example, the item Appearance has 244 dentists scoring a 2, while a total of only 56 dentists scored in the other three categories. These 4-point scales can have exclusively high or low scores as well. Items that have continuously scored scales, such as totals from patient records, tend to have either high or low scores or a bimodally distributed range of scores, where the individuals fall most frequently on the scale end-points. A bimodal distribution is depicted by the Pre-existing Dental Treatment item, on which 127 dentists scored a 0, and 84 scored a 5, with substantially fewer individuals who fell in the 1 - 4 range.

Summary of Item Analysis Results.

The interpretation of the item analysis results is difficult because the Assessment Instrument requires a balance of scale reliability and item validity. The disruption of the structural hierarchy of the scales produces many problems with the assessment of scale internal consistency and item discrimination ability. It is conceivable that if the Components were made up of fairly equal numbers of Elements and the Elements were similarly composed of Subelements, the item analysis would be more successful. With a stable structure, the Elements could be measured on the homogeneity of the Subelements and these subscales could be treated as heterogeneous predictors of the Component criterions, as discussed by Ghiselli et al. (1981, Ch. 13). The Components scales could then be employed to generate the criterion validity coefficients for the Instrument.

Presently, there are too many Elements at the item level to accomplish the combination of maximizing scale reliability and item validity. Not only are there Elements which cannot be tested for internal consistency because they are items, but there are several Components that cannot be used as criterions since they are at the item level.

Along with the uneven construction of the DEM scales and subscales, the scale reliability and item validity problems are byproducts of the formation of the items themselves. The differences in subscale types that are combined to make the Component scales certainly detracts from the item homogeneity. Also, the subjective formation of scale points is reflected in the poor discrimination abilities of the items. These problems will be further detailed later in this discussion.

The low internal consistency indices for the Components do not support the establishment of content validity for the Assessment Instrument. Scales with high internal consistency reliability demonstrate that they measure the intended construct (Ghiselli et al., 1981, Ch. 10). Judging from the item analysis results then, the selection of DEMCAD items did not reinforce the content that was intended to be assessed. According to Anastasi (1976), "content validity is built into a test from the outset through the choice of appropriate items" (p. 135). Problems with the Component scale reliability relate to the previously mentioned difficulties resulting from the configuration of Elements and Subelements. Therefore, the most dependable way to establish the content validity of this Instrument is through the systematic examination of the content to determine if it adequately covers the domain in question (Anastasi, 1976, Ch. 6). If a subsequent item analysis is performed, Ghiselli et al., (1981) recommend cross validation procedures to verify the content validity and internal consistency of the scales (Ch. 13). Also suggested is the use of multiple expert judges to rate the content of the test items (Ch. 10).

Factor Analysis

The factor analysis results should not have produced multiple factors for a Component if it is considered to have internal consistency reliability. Because all of the Component scales generated many factors with a high degree of shared items, it can be concluded that the Components do not exhibit item homogeneity. These results reinforce the findings reported for the reliability procedure; a perfectly homogeneous scale would have all of its items loaded on to a single factor representing the hypothetical construct being assessed.

Problems with the Scales

In light of the previously mentioned scale reliability and item validity problems, it is appropriate to define and discuss the flaws in scale construction that have been cited. The most serious difficulty with the Assessment Instrument is the formation of Component scales by combining subscales having different numbers of scale points. When this occurs, a Component is made up of various Subelements and/or Elements that do not share the same basic scale construction. For instance, the Facilities Component is comprised of seven Elements, six of which have Subelements on a 4-point subscale, and one that has Subelements on a dichotomous subscale. The effect of this discrepancy is to favorably weight the 4-point Subelements, scored 0 - 3, over the dichotomous ones, scored 0 or 1. Therefore, all of the items in the Support Rooms/Areas Element are worth half of any other item in Facilities. They are also of less value than most other items on the Instrument. This same process happens in the Practice Management Component, where there are both 4-point and 2-point subscales. However, a Component weight is assigned which exaggerates the differences produced by the individual item weights.

The effect of the incompatibilities in the subscales on the Instrument is to render the total scores for Components and Dimensions almost impossible to interpret. Without documentation or explanations of why certain items are weighted over others, the reasons one dentist scored higher than another on a Component are not evident, and the Dimension scores are practically meaningless. Because no defense for the differential weighting of test items was provided in Morris (1986), it is assumed that all DEM items were originally intended to have equal value on the Instrument, barring the effect of the Component weights. In reality, this is simply not the case.

The techniques involved in the creation of many items augment the problem of incompatible subscales and the weighting of test items. In order to generate item scores in almost all of the Process and Outcome Components, five or six patient records were summed and the total was considered the item score. This procedure was not uniform across the Instrument scales, for it was employed to form items at the Component, Element, and Subelement levels. For example, each Subelement of the Radiographic Evaluation Component has a range of 0 - 5 because a point is given for each record considered satisfactory by the evaluator. In contrast, the items in the Data Collection Component are Elements on a 4-point scale, so each one has a range of 0 - 15 since five patient records are reviewed (see Assessment Instrument &

Evaluator's Manual).

As previously stated, the data points added to make the items were not considered part of the Instrument scale hierarchy, but they are actually one level below the subscale they form. The data making up the Subelements then, consist of a lower order in the hierarchical system, while the data forming the Elemental items are equivalent to Subelements, and the data added to form the Component items are on the Element level. The obvious problem here is that actual data points were given a priori scale weights depending on what type of item they constitute.

The use of multiple patient records produced similar problems in the creation of the Patient Satisfaction, Patient Oral Hygiene, and Patient Management scales. For these Component scores, eight patient questionnaires were reviewed, and the total score for each was calculated by dividing the patient's score on the questionnaire by the number of questions answered, thus producing a decimal value (see Evaluator's Manual). These decimal values were very troublesome to analyze and interpret, especially in the item validity procedures. Again, none of the scoring techniques involving the summation of patient records were adequately defended in Morris (1986).

A scoring procedure that gave considerably large values to scale items was the assignment of points to reflect the absence or presence of patient records. The Endodontic and

Periodontic Elements of the Treatment Component were scored 0 if there were no patient records and 25 if the records were in the office. The question here is why are these items so important to the quality of professional dentistry that they receive such extreme values? In all of the Structure Components, the absence or presence of the criterion was scored 0 or 1, which means that an office with a complete set of Endodontic records received the same amount of points as one having 25 of the required pieces of dental equipment.

This example accentuates the seriousness of the scaling problems involved in comparing items on the DEM Assessment Instrument. Also of concern is the legitimacy of the theoretical comparisons among the items residing on the same Component scale. For instance, all of the dental accoutrements in the Equipment Component were scaled with equal value. It is conceivable however, that certain pieces of equipment are more important to a dentist's office than others. To illustrate, having a periodontal probe may be more crucial to a dentist than having a modern style chair, yet these important differences were not reflected in the scoring. In the same way, items from different Components are more indicative of quality care than others. An obvious example is a dentist acquires three points for having "unusually attractive and well cared for" grounds in the Facilities Component, but only receives one point for having "proper venting for fumes" in the Equipment Component (see Assessment Instrument).

The last major problem with the DEM scale items is the arbitrariness of the scale points, particularly in the Structure Dimension. The 4-point items on the Instrument have extremely subjective labels, which are at best, ordinal in nature. The scales of the Facilities Component illustrate this point well. Theoretically, it cannot be claimed that there is an equal distance between the scale values. For example, the choices for the Subelement "filing" (see Assessment Instrument) are: (0) files spread in multiple areas, (1) inconvenient to access, (2) conveniently accessible, and (3) separate filing area. Although these scale points obviously do not exhibit interval distances, it is questionable if there is even an ordinal difference between points 2 and 3.

The effect of the inconsistent labeling of scale points showed in the discrimination analysis. Many of the items were negatively skewed because the zero points on the items tend to be extreme. And, the items with 4-point scales had a high frequency of answers in the number 2 category, suggesting that the questions were being written according to an answer pattern. In other words, the wording of the scale points has resulted in the items being not evenly distributed.

This problem is also revealed in the questionnaires that assess the response of the evaluators to participating in the project (Morris, 1986, pp. 30-34). In fact, one or two of these questions cannot be considered to be at the ordinal, much less interval level. For instance, the choices for the question "What was your reaction to the second year of the project compared to the first?" are: (a) I did not enjoy evaluating offices, (b) I enjoyed the first year more, (c) no difference between first and second year, and (d) I enjoyed the second year more. Morris (1986) does not state if these answers were scored 0 - 3, like the 4-point scales in the Instrument. It appears that the questions were merely judged by the frequency count for each answer. The point is that no numerical difference exists between choices (b) and (d), so weighted values should not be applied to the scale points. These examples of scale-point labels clearly illustrate that more care must be taken in the systematic formation and writing of the scale items.

Suggestions for Instrument Revision

Criteria and Scaling.

Adequate revision of the DEM Assessment Instrument rests upon establishing its psychometric properties through scale and item adjustment. Clearly, the process must begin at the item level, for reliable scales cannot exist without the proper establishment of their criteria. Judging by the results of the item analysis procedures, the criteria for the DEM scales and subscales should be reexamined. The article by Donabedian (1986) is a comprehensive review of the descriptive characteristics of criteria and their formation process. Since the Assessment Instrument is organized according to Donabedian's (1966) dimensions, it would be advantageous to continue with the theoretical and systematic development of the components of those dimensions.

The greatest problem with the DEM items is that they have been scored on different types of subscales, so they cannot be added together to form meaningful Component and Dimension scale scores. Bailit et al. (1974) avoided this difficulty by scoring all of their criteria on dichotomous subscales, thus enabling the formation of components and dimensions through item summation. Dichotomous scoring procedures can be used for both quantitative and qualitative criteria. This is important, because the Structure criteria tend to be quantitative in nature while Process criteria are mostly qualitative, and those in Outcome can be either quantitative or qualitative.

The criteria developed by Bailit et al. (1974) were qualitatively scaled with points of unsatisfactory and adequate, as well as a no decision category. A possibility for a quantitative scale would be not present and present, categories which are suited to the items in the Structure Dimension. For the Equipment subscales in the Structure Dimension with the "available and in good repair" item (see Assessment Instrument), a separate "equipment condition" scale could be set up with points of "not available/working" and "available/working." Then, each item would be scored on two scales: the first one assessing whether or not the piece of equipment is in the office, and the second one evaluating its working condition. Of course, the weighting of these two-scaled Elements should be defended. If all DEM items have the same scoring system, then the scales and subscales will have logical meaning, and item and scale weighting procedures can be easily explained and justified.

Reliability and Validity Issues.

The problems with the Instrument reliability and validity can only be solved by systematically applying the procedures necessary to show that the Assessment Instrument does indeed accurately and reliably measure quality. It must be demonstrated that the Instrument produces the same results when representative samples of practitioners are tested in different regions of the country. Ware et al. (1976a, 1976b) achieved instrument reliability by replicating their results in four separate field tests. Similar procedures must be taken with the DEM Assessment

Instrument, especially if the results are to be generalized to private practitioners in the United States. Testing of representative samples is crucial to establishing reliability, which means that techniques other than the ones used for the DEM study must be employed in random sample selection.

Although content validity can be reinforced by internally consistent subscales, which have not resulted from the analysis, the way to verify the content being evaluated is through the use of expert judges to agree upon the criteria for the Instrument (Anastasi, 1976, Ch. 6; Ghiselli et al., 1981, Ch. 10). Bailit et al. (1974) relied upon the agreement of separate practitioner panels of experts to determine the criteria for the study on the quality of dental care. Ware et al. (1983) also discussed the systematic review of the patient satisfaction criteria by experts. Because the DEM study assesses a nation-wide population, it is recommended that practitioner panels in different regions of the country evaluate the criteria and compare their decisions.

The poor results from the criterion-related validity analysis suggest that the Outcome Components need restructuring. The solutions to criteria formation and item scaling previously discussed apply to the Outcome scales and subscales, because this Dimension requires special attention for revision. As McAuliffe (1979) stated, the outcome validities need to be empirically established. Although the outcomes serve to validate the structure and process of dental practice, they themselves must demonstrate that they are measures of quality care. McAuliffe (1979) points out that poor process validities may directly result from invalid outcomes, so the outcome measure must first be examined before making statements about the correlational procedures between process and outcome. Techniques for increasing outcome validity involve statistical adjustments, examining patterns of care, using statistically derived cutoffs for acceptable outcome rates, discounting poor outcome indices, and focusing on tracers or sentinel outcomes that are known to be relatively pure measures of quality (McAuliffe, 1979). McAuliffe also discusses the observation that "however promising the techniques may be, none has yet been shown to be both practical and effective" (p. 132).

Since the Patient Satisfaction Component is considered the key indicator of outcome on the Assessment Instrument, and it has been shown to be a poor scale, the use or adaptation of one of the previously discussed Patient Satisfaction Questionnaires is recommended. Both Hulka et al. (1975) and Ware et al. (1976c) developed PSQ's that were extensively tested for their psychometric properties. Roberts and Tugwell (1987) performed a conceptual and statistical comparison of these two questionnaires and found them acceptable evaluators of patient satisfaction. Koslowsky et al. (1974) devised a pair of instruments to assess both patient and practitioner satisfaction in the dental field. Not only are the criteria relevant to the DEM study, but the dentist satisfaction questionnaire could be an added outcome measure not previously considered.

Along with content and criterion-related validity, it would be beneficial to index both convergent and discriminant validity for the Instrument. In other words, demonstrate that the Component criteria relate to other criteria that they should and do not correlate with criteria having nothing to do with the construct being assessed. Bailit et al. (1974) achieved concurrent validity through the correlation of the Treatment Plan and Treatment components, and Ware et al. (1983) demonstrate both types of validity by comparing their scales with several alternative testing methods.

Structural Hierarchy.

Finally, it has been emphasized that the radical organization of the Instrument structural hierarchy significantly adds to the scaling difficulties. If at all possible, the Dimensions, Components, and Elements should have fairly equal numbers of items, and all items should exist at the Subelement level. This would greatly facilitate the homogeneity of the Element scales which could then serve as the heterogeneous elements of the Components. The reason for this process, as discussed by Ghiselli et al. (1981), is to establish reliable subscales that validly assess the Component criteria (Ch. 13). Hence, a meaningful Dimension score could be obtained by summing the individual Component scores, and the total score for a case would logically be the addition of the Dimension scores.

Summary of Findings

The adaptation of Donabedian's (1966) dimensions of structure, process, and outcome to the evaluation of quality practice in professional dentistry was a good foundation for the DEM project. However, it is apparent that current, established psychometric techniques were not employed in the Assessment Instrument construction, and the criteria were not developed according to Donabedian's (1969) format.

The replication of the results from the DEM project verified that the data analyzed in the present study were identical to the original data. It has been concluded that the total score and Dimension distributions reported in the present study were correct and the histograms from the original study were erroneous.

The poor intercorrelations of the Structure and Process Components with the Outcome Components indicated that criterion-related validity for the Instrument is practically non-existent. As previously stated, this problem most likely results from the weakness of the Outcome Dimension criteria. None of the Component weights were adequately explained or defended, making the interpretation of analyses results extremely difficult.

Item analysis procedures focused on scale reliability and item validity. Although the Components with many items demonstrated scale homogeneity, the internal consistency reliability of the DEM Component scales was generally low. The DEM items exhibited very poor discrimination abilities, and therefore did not relate well to their scale criteria.

The problems mentioned thus far probably resulted from two major test-construction flaws: The hierarchical structure of the Assessment Instrument was not followed appropriately throughout its construction, and summative scales were created by adding non-uniform items. Strict adherence to the hierarchy of the Instrument would facilitate the testing of Elements for scale reliability and then assessing their ability to discriminate on their Components scales. The Instrument would then be composed of homogeneous Element subscales that are themselves heterogenous predictors of the Component criteria, thus establishing both scale reliability and item validity.

In order for the items to be added to form a total score, they must all have uniform scales. This means the scales and subscales in all three Dimensions must be scored on the same point system. The differences in the DEM scale points produced weighted items, and no explanation for the weighting was provided. Presently, the accurate interpretation of the DEM results is impossible because the total scores, the Dimension scores, and some of the Component scores are the combination of different scale types. Also, attention must be given to the assignment of scale points so that they approximate interval-level scales; some of the subscales on the Assessment Instrument were not even at the ordinal level.

Since the original sample of 300 private dentist practitioners was not representative, and the internal consistency of the Component scales was low, the content validity for the Assessment Instrument has not been properly established. The fact that no replication studies have been performed with the Assessment Instrument also leaves the Instrument reliability in question. At best, it could be stated that the DEM Instrument has face validity.

The statistical and analytical procedures used in the present study were traditional psychometric testing procedures that have been clearly and accurately demonstrated to produce correct results. Alternate techniques to analyze the DEM data in future projects could also involve more recently established techniques such as cluster analysis, multiple regression, and multivariate analysis. Factor analysis, a multivariate procedure, has been employed in the present study and could also be used in further analyses.

In sum, the DEM Assessment Instrument requires substantial revision in order to be used as an accurate evaluation tool. If the intention is to employ the Instrument as a self-evaluation tool, then the present form of the Assessment Instrument could be adapted using the procedures and analyses previously suggested. However, if the Instrument is for the evaluation of private dentist practitioners across the nation, it would be best to begin with newly established criteria that have been tested by experts from different regions of the United States and to fully reconstruct an assessment instrument based on sound, psychometric procedures to be tested on a very large, representative sample.

References

American Educational Research Association, American Psychological Association, & National council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington DC: American Psychological Association.

Anastasi, A. (1976). Psychological testing (4th ed.). New York: Macmillan.

Anderson, O. W., & Shields, M. C. (1982). Quality measurement and control in physician decision making: State of the art. Health Services Research, 17(2), 125-155.

Bailit, H., Koslowsky, M., Grasso, J., Holtzman, S., Levine, R., Valluzzo, P., & Atwood, P. (1974). Quality of care: Development of standards. Journal of the American Dental Association, 89, 842-853.

Becker, M. H., et al. (1972). Correlates of physicians' prescribing behavior. Inquiry, 9, 30-42.

Brook, R. H., & Appel, F. A. (1973). Quality-of-care assessment: Choosing a method for peer review. New England Journal of Medicine, 288(25), 1323-1329.

Donabedian, A. (1966). Evaluating the quality of medical care. Milbank Memorial Fund Quarterly, 44(3), 166-203.

Donabedian, A. (1969). A guide to medical care administration. Vol. II: Medical care appraisal. American Public Health Association. U.S. Public Health Services (contract No. PH108-66-153). 1-12.

Donabedian, A. (1982). Explorations in quality assessment and monitoring: Vol. II. The criteria and standards of quality. Ann Arbor: Health Administration Press.

Donabedian, A. (1986). Criteria and standards for quality assessment and monitoring. Quality Review Bulletin, 12(3), 99-107.

Fitzpatrick, T., Riedel, D., & Payne, B. (1962). Character and effectiveness of hospital use. In W. J. McNerney et al. (Eds.), Hospitals and medical economics: A study of population services, costs, methods of payment, and controls (Vol. I, p. 361). Chicago: Hospital Research and Education Trust.

Friedman, J. W., & Schoen, M. H. (1972). Audit of quality dental care: A pilot study. Journal of Public Health Dentistry, 32(4), 214-224.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). Measurement theory for the behavioral sciences. San Francisco: W. H. Freeman & Co.

Hastings, G. E., Sonneborn, R., Lee, G. H., Vick L., & Sasmour, L. (1980). Peer review checklist: Reproducability and validity of a method for evaluating the quality of ambulatory care. American Journal of Public Health, 70(3), 222-228.

Hulka, B. S., et al. (1975). Correlates of satisfaction and dissatisfaction with medical care: A community perspective. Medical Care, 13(8), 648-58.

Hulka, B. S., Romm, J., Parkerson, G. R., Russell, I. T., Clapp, E., & Johnson F. S. (1979). Peer review in ambulatory care: use of explicit criteria and implicit judgments. Medical Care, 17, 1-73.

Koslowsky, M., Bailit, H., & Valluzzo. (1974). Satisfaction of the patient and provider: Evaluation by questionnaire. Journal of Public Health Dentistry, 34(3), 188-194.

Lembcke, P. A. (1956). Medical auditing by scientific methods: Illustrated by major female pelvic surgery. Journal of the American Medical Association, 162, 646-655.

Matthews, D. A., & Feinstein, A. R. (1988). A review of systems for the personal aspects of patient care. The American Journal of Medical Science, 295(3), 159-171.

McAuliffe, W. E. (1979). Measuring the quality of medical care: Process versus outcome. Milbank Memorial Fund Quarterly/Health and Society, 57(1), 118-152.

Milgrom, P., Weinstein, P., Ratener, P., & Morrison, K. (1978). Dentists' self-evaluations: Relationship to clinical performance. Journal of Dental Evaluation, 42(4), 180-185.

Morehead, M. A., Donaldson, R. S., & Burt, S. (1964). A study of the quality of hospital care secured by a sample of teamster family members in New York City. New York: Columbia University School of Public Health and Administrative Medicine.

Morehead, M. A. (1967). The medical audit as an operational tool. American Journal of Public Health, 57, 1643-1656.

Morehead, M. A., & Donaldson, R. (1974). Quality of clinical management of disease in comprehensive neighborhood health centers. Medical Care, 12, 301-315.

Morris, A. L. (1986). Development of evaluation methods and computer applications in dentistry: Final report to the W. K. Kellogg foundation: Development of evaluation methods. Unpublished manuscript.

Payne, B. C. (1967). Continued evolution of a system of medical care appraisal. Journal of the American Medical Association, 201, 536-540.

Payne, B. C. (1973). From performance measures to utilization review to quality assurance. In Regional medical programs service, quality assurance of medical care (pp. 241-260). (DHEW Publication No. HSM 73-7021). Washington DC: U.S. Government Printing Office.

Peterson, O. L., Andrews, P., Spain, R. S., & Greenberg, B. G. (1956). An analytical study of North Carolina general practice, 1953-1954. Evanston IL: Association of American Medical Colleges.

Roberts, J. G., & Tugwell, P. (1987). Comparison of methods determining patient satisfaction with medical care. Health Services Research, 22(5), 637-654.

Sanazaro, P. J., Goldstein, R. L., Roberts, J. S., Maglott, D. B., & McAllister, J. W. (1972). Research and development in quality assurance. New England Journal of Medicine, 287(22), 1125-1131.

Sanazaro, P. J. (1980). Quality assessment and quality assurance in medical care. Annual Review Public Health, 1, 37-68.

Sheps, M. C. (1955). Approaches to the quality of hospital care. Public Health Report, 70, 877-886.

SPSS Inc. (1986). SPSS^{X} user's guide (2nd ed.). Chicago: Author.

Ware, J. E., Snyder, M. K., & Write, W. R. (1976a). Development and validation of scales to measure patient satisfaction with health care services: Volume 1 of final report part a: Review of literature, overview of methods, and results regarding construction of scales. Springfield VA: National Technical Information Service. (NTIS No. PB 288-329).

Ware, J. E., Snyder, M. K., & Write, W. R. (1976b). Development and validation of scales to measure patient satisfaction with health care services: Volume 1 of final report b: Results regarding scales constructed from the patient satisfaction questionnaire and measures of other health care perceptions. Springfield VA: National Technical Information Service. (NTIS No. PB 288-330).

Ware, J. E., Snyder, M. K., & Write, W. R. (1976c). Development and validation of scales to measure patient satisfaction with health care services. Carbondale IL: University of Southern Illinois.

Ware, J. E., Snyder, M. K., Write, W. R., & Davies, A. R. (1983). Defining and measuring patient satisfaction with medical care. Evaluation and Program Planning, 6, 247-263.

APPENDIX A

Table 1

Descriptive Statistics for the Total Score and Dimension Scores for all 300 Cases

	Mean Score	Standard Deviation	Coefficient of Variation
Total Score	589.512	83.502	14.165
Structure	139.613	20.266	14.516
Process	337.166	62.384	18.502
Outcome	112.733	11.044	9.796

Table 2

Descriptive Statistics for the Components of Structure

	Mean Score	Standard Deviation	Coefficient of Variation
Structure	139.613	20.266	14.516
Facilities	20.837	3.517	17.139
Equipment	42.390	4.995	11.784
Personnel	37.920	8.185	21.589
Administration	38.467	9.488	24.666

Table 3

Descriptive Statistics for the Components of Process

	Mean Score	Standard Deviation	Coefficient of Variation
Process	337.166	62.384	18.502
Practice Mgt.	40.350	8.186	20.288
Radiographic Eval.	15.242	5.931	38.916
Data Collection	23.540	6.786	28.829
Diagnosis	64.053	24.082	37.596
Treatment Plan	28.013	29.632	105.779
Treatment	87.488	11.535	13.184
Ster.-Infec. Cntl.	25.633	6.987	27.257
Patient Mgt.	52.846	2.911	5.508

Table 4

Descriptive Statistics for the Components of Outcome

	Mean Score	Standard Deviation	Coefficient of Variation
Outcome	112.733	11.044	79.900
Pat. Satisfaction	47.860	1.781	3.721
Pat. Oral Hygiene	13.620	2.732	20.059
Pat. Education	10.543	2.105	19.962
Pat. Disability	8.320	1.921	23.088
Periodontal Dis.	12.410	3.814	30.730
Completion of Tmt.	5.720	6.053	105.829
Recall	14.260	3.188	22.359

Table 5

Descriptive Statistics for the Elements of the Treatment Component

	Mean Score	Standard Deviation	Coefficient of Variation
Treatment	87.488	11.535	13.184
Restorative	52.680	7.746	14.703
Endodontic	8.708	3.436	39.460
Periodontic	11.750	2.974	25.307
Oral Medicine	4.400	1.628	36.989
Dies	9.950	2.547	25.595

Table 6

Correlation Coefficients of the Significantly Related Structure and Outcome Components

Structure	Outcome					
	Patient Satisfaction	Patient Hygiene	Patient Education	Periodontal Disease	Completion Treatment	Recall
Facilities		.2111		.1447	.2381	
Equipment		.1672		.1510	.1446	
Personnel	.1089		.1147	.2635	.1878	.1479
Administration		.2116		.2993	.4037	

Table 7

Correlation Coefficients of the Significantly Related Process and Outcome Components

Structure	Outcome					
	Patient Satisfaction	Patient Hygiene	Patient Education	Periodontal Disease	Completion Treatment	Recall
Practice Mgt.				.2464	.2488	.1380
Radiograph. Eval.		.1651	.1732	.4258	.2609	
Data Collection	.1343	.1585		.4574	.5650	
Diagnosis	.1138	.1770	.1523	.5447	.2783	
Treatment Plan			.1217	.5037	.9525	
Treatment	.1646	.3866	.1206	.5188	.2284	
Ster-Infec. Cntl.		.2312	.1487	.2479	.2125	
Patient Mgt.	.6845	.1199	.1077	.1704		.0983

Table 8

Internal Consistency Coefficients and Mean Item-Total Correlation Coefficients for all Components

Dimension Component	Scale Alpha	Mean Item-Total Correlation
Structure		
Facilities	.8103	.1194
Equipment	.7193	.0443
Personal	.6416	.1541
Administration	.7962	.0926
Process		
Practice Mgt.	.5292	.0544
Radiographic Eval.	.8670	.3482
Data Collection	.7255	.1873
Diagnosis	.7487	.4933
Treatment Plan	.9555	.8341
Treatment	.6719	.2026
Ster.-Infec. Cntl.	.7182	.1135
Pat. Management	.6050	.1754
Outcome		
Pat. Satisfaction	.8502	.2720
Pat. Oral Hygiene	.0367	.1119
Periodontal Disease	.4485	.2399
Recall	.3023	.1295

Table 9

Items that Detract from the Internal Consistency of their Component Scale

Component Problematic Item	Scale Alpha	Alpha If-Item-Deleted
Facilities	.8103	
Access for Handcpd.		.8097
Size Reception Room		.8101
Educational Material		.8168
Filing		.8091
Shielding		.8133
Sterilization		.8093
Laboratory		.8103
Lavatory		.8111
Darkroom		.8109
File Room		.8103
Recovery Room		.8105
Equipment	.7193	
Dry Heat		.7228
Closed Storage-Instmts.		.7214
Lead Apron		.7193
X-ray Avail-Good Repair		.7195
Enamel Chisels		.7197
Scalers/Currettes		.7194
Optical Loops		.7218
Polishing Lathe		.7192
Eye Protection		.7205
Hair Protection		.7288
Sleeve Protection		.7296
Mercury Spill Cntl.		.7207
Personal	.6416	
Appearance		.6328
Demeanor		.6349
Longevity		.6467

Table 9 (Continued)

Items that Detract from the Internal Consistency of their Component Scale

Component Problematic Item	Scale Alpha	Alpha If-Item-Deleted
Administration	.7962	
BP Recording		.7967
Med. Alert on Chart		.7960
Documenting New Pat.		.7973
Progress Notes		.7980
Informed Consent		.7980
Lab Prescrip. Forms		.7969
Referral Forms		.7986
Appt. Sched. Cards		.7962
Prescrip. Forms		.7966
Practice Management	.5292	
Neat/Legible		.5256
Lunch Scheduled		.5303
Open Time Within 2 Wks.		.5381
Special Hours		.5434
Recalls Scheduled		.5369
Radiographic Eval.	.8670	
Films Correct Freq.		.8804
Films Are Dated		.8686
Maxilla-Mandible		.8925
Data Collection	.7255	
Organiz. Pat. Records		.7399
Legibility of Records		.7346
Progress Notes		.7289
Diagnosis	.7487	
Carious Lesions		.8124

Table 9 (Continued)

Items that Detract from the Internal Consistency of their Component Scale

Component ---------------------- Problematic Item	Scale Alpha	Alpha If- Item-Deleted
Treatment Plan	.9555	
Appropriateness		.9557
Treatment	.6719	
Endodontic		.6916
Periodontic		.6808
Oral medicine		.6810
Ster.-Infec. Cntl.	.7182	
Heat Steril. Used		.7186
Instruments Scrubbed		.7172
Ster. Instrmts. Packaged		.7186
Antibact. Soap Used		.7235
Paper Towels Used		.7198
Head Covers Used		.7274
Face Masks Worn		.7216
Pat. Management	.6050	
Avg. Waiting Time		.5981
Import. of Teeth Health		.5962
Pat. Satisfaction	.8502	
Dr Explains Tmt. Procs.		.8510
Qual. Care Imp. to Me		.8496
Pat. Oral Hygiene	.0367	
Report on Pat. Exams		.3676
Periodontal Disease	.4485	
Treatment		.6191

Table 10

The Adequately Discriminating Items and their Criterion Scales from the Item Validity Analysis

Item Level	Item	Criterion Scale
Subelement		
	Access for Handicapped	Office Setting
	X-ray Room	Support Rooms
	Staff Lounge	Support Rooms
	Fiber Optics Handpiece	Treatment Support
	Optical Loops	Treatment Support
	Extraoral Film Equipment	Office Support
	Soldering Torch	Office Support
	Hair Protection	Environ./Hazard Cntl.
	Sleeve Protection	Environ./Hazard Cntl.
	Blood Pressure Recording	Patient Related Recs.
	Head/Neck Soft Tissue Exam	Patient Related Recs.
Treatment Plan		Patient Related Recs.
	Personnel Manual	Admin. Protocols
	Staff Job Descriptions	Admin. Protocols
	Organized-Operatories	Appointment Book
	Accoms. for Emergencies	Appointment Book
	Special Hours	Appointment Book
	FMXR - 14 Periapicals	Admin. Considerations
	Exposure Density/Contrast	Radiographic Technique

Table 10 (Continued)

The Adequately Discriminating Items and their Criterion Scales from the Item Validity Analysis

Item Level Item	Criterion Scale
Subelement	
Angulation Vert/Horiz	Radiographic Technique
Processing Technique	Radiographic Technique
Bone - Maxilla/Mandible	Diagnostic Value
Bone - Interdental	Diagnostic Value
Teeth - Interproximal	Diagnostic Value
Physician's Name	Completeness of Recs.
Element	
Ster. Instrums. Packaged	Ster.-Infec. Cntl.
Glutaraldehyde Disinfect.	Ster.-Infec. Cntl.
Antibacterial Soap	Ster.-Infec. Cntl.
Head Covers	Ster.-Infec. Cntl.
Asstnt. No Contamination	Ster.-Infec. Cntl.
Light Handles Disinfect.	Ster.-Infec. Cntl.
Switches/Cntls. Disinfect.	Ster.-Infec. Cntl.
Hoses/Couplings Disinfect.	Ster.-Infec. Cntl.
Avg. Waiting Time	Patient Management
Component	
Patient Education	Outcome

Table 11

The Questionably Discriminating Items and their Criterion Scales from the Item Validity Analysis

Item Level	
Item	Criterion Scale
Subelement	
Educational Material	Reception Room
Number of Tmt. Rooms	Treatment Rooms
Panorex Unit	X-ray Equipment
Periodontal-Surgical	Instruments
Automatic Film Proc.	Office Support
Casting Machine	Office Support
Inter-Offc. Comm. Syst.	Office Support
Computer	Office Support
Nitrous Oxide	Patient Support
Eye Protection	Environ. Hazard Cntl.
Scrap Amalgam Storage	Environ. Hazard Cntl.
Smoke Alarms	Environ. Hazard Cntl.
Ventilation	Environ. Hazard Cntl.
Element	
Numbers Admin. Support	Personnel
Training Care Supt. Pers.	Personnel
Hygienist	Personnel
Longevity	Personnel
Continuing Educ. Staff	Personnel

Table 11 (Continued)

The Questionably Discriminating Items and their Criterion Scales from the Item Validity Analysis

Item Level Item	Criterion Scale
Subelement	
Medical Alert on Chart	Patient Related Recs.
Recording-Occlusal Anls.	Patient Related Recs.
Referral Forms	Patient Related Recs.
Emergency Phone Service	Admin. Pat. Care Systs.
Protocol - Admin/Staff	Admin. Protocols
Protocol - Pat. Support	Admin. Protocols
Office Philosophy	Materials for Patient
Recall Instructions	Materials for Patient
Daily Schedules	Receptionist Appt. Cntl.
Staff Meetings	Personnel Management
In-Service Training	Personnel Management
Element	
Organiz. of Pat. Recs.	Data Collection
Legibility of Recs.	Data Collection
Subelement	
Dental History	Completeness of Recs.
Preexisting Dental Tmt.	Completeness of Recs.
Periodontal Disease	Completeness of Recs.
Treatment Plan	Completeness of Recs.

Table 11 (Continued)

The Questionably Discriminating Items and their Criterion Scales from the Item Validity Analysis

Item Level Item	Criterion Scale
Element	
Sequencing	Treatment Plan
Appropriateness	Treatment Plan
Implementation	Treatment Plan
Subelement	
Surface	Restorative Treatment
Anatomic Form	Restorative Treatment
Retention	Dies - Examination
Element	
Paper Towels Used	Ster.-Infec. Cntl.
Dr. No Contamination	Ster.-Infec. Cntl.
Asstnt. Wash Hands	Ster.-Infec. Cntl.
Dr. Wash Hands	Ster.-Infec. Cntl.
Dr. Discusses Costs	Patient Management
Feel Good - Mouth Appearnc.	Patient Satisfaction
Feel Good - Mouth Health	Patient Satisfaction
Subelement	
Frequ. - Tooth Brushing	Report - Pat. Questnrs.
Frequ. - Flossing	Report - Pat. Questnrs.

Table 11 (Continued)

The Questionably Discriminating Items and their Criterion Scales from the Item Validity Analysis

Item Level Item	Criterion Scale
Element	
Report on Patient Exam	Patient Oral Hygiene
Diagnosis	Periodontal Disease
Component	
Completion of Treatment	Outcome

Figure 1

Distribution of Total Scores for All 300 Cases

```
Count    Midpoint  One Symbol Equals Approximately  .80 Occurrences

  3       379.50   *:**
  1       400.50   * .
  2       421.50   *** .
  8       442.50   *******:**
 12       463.50   ***********:***
 17       484.50   ****************:****
 14       505.50   ******************    .
 22       526.50   ***************************:
 37       547.50   ********************************:*************
 27       568.50   ********************************** .
 19       589.50   ************************             .
 27       610.50   ********************************** .
 29       631.50   ********************************:***
 20       652.50   *************************  .
 19       673.50   **********************:*
 15       694.50   ****************:**
 12       715.50   ***********:***
  9       736.50   *******:***
  2       757.50   *** .
  5       778.50   **:***
                   I....+....I....+....I....+....I....+....I....+....I
                   0         8         16        24        32        40

                                  Frequencies
```

Figure 2

Distribution of Structure Scores for All 300 Cases

```
Count     Midpoint  One Symbol Equals Approximately  1.00 Occurrence

    3       82.00   :**
    2       89.00   *:
    3       96.00   ***.
    9      103.00   *******:*
    7      110.00   *******       .
   22      117.00   *********************:
   32      124.00   ******************************:*
   46      131.00   *************************************:********
   36      138.00   ************************************     .
   32      145.00   ********************************        .
   39      152.00   *********************************:*****
   26      159.00   *************************:
   20      166.00   *****************:**
   18      173.00   **********:*******
    3      180.00   ***  .
    1      187.00   * .
    0      194.00   .
    1      199.50   *
                   I....+....I....+....I....+....I....+....I....+....I
                   0        10        20        30        40        50

                                  Frequencies
```

Figure 3

Distribution of Process Scores for All 300 Cases

```
Count   Midpoint  One Symbol Equals Approximately  .60 Occurrences

    1     178.70  *:
    3     194.70  ***:*
    3     210.70  ***** .
    7     226.70  **********:*
   12     242.70  ***************:****
   13     258.70  **********************.
   22     274.70  ******************************:******
   30     290.70  **************************************:***********
   22     306.70  *************************************       .
   29     322.70  ************************************************ .
   26     338.70  *******************************************
   26     354.70  *********************************************     .
   23     370.70  **************************************     .
   29     386.70  ************************************:***********
   14     402.70  ***********************     .
   11     418.70  ******************   .
   17     434.70  **************:*************
    5     450.70  ******** .
    5     466.70  *****:**
    2     480.10  **:
                 I....+....I....+....I....+....I....+....I....+....I
                 0         6        12        18        24        30

                                     Frequencies
```

Figure 4

Distribution of Outcome Scores for All 300 Cases

```
Count  Midpoint   One Symbol Equals Approximately  1.00 Occurrence

    1     81.90   :
    1     85.90   *.
    5     89.90   ****:
    9     93.90   *********.
   17     97.90   *****************.
   31    101.90   **************************:****
   40    105.90   ***********************************:****
   45    109.90   *****************************************:***
   37    113.90   *************************************      .
   36    117.90   ************************************   .
   26    121.90   **************************     . *
   22    125.90   ********************:*
   14    129.90   ************:*
   13    133.90   ******:******
    2    137.90   **.
    1    140.20   :
                  I....+....I....+....I....+....I....+....I....+....I
                  0         10        20        30        40        50

                                      Frequencies
```

APPENDIX B

TABLE 39

AVERAGE TOTAL AND DIMENSION OFFICE EVALUATION SCORES OF 300 DEMCAD OFFICES

	Mean Score	Standard Deviation	Coefficient of Variation	Possible Score	Mean as % of Possible
Total	590	83.6	14.2	884	66.7
Structure	140	20.3	14.5	219	63.9
Process	338	62.6	18.5	516	65.5
Outcome	113	11.0	9.8	149	75.8

TABLE 40

AVERAGE SCORES FOR COMPONENTS OF STRUCTURE

	Mean Score	Standard Deviation	Coefficient of Variation	Possible Score	Mean as % of Possible
Structure	140	20.3	14.5	219	63.9
Facilities	21	3.6	17.0	32	65.6
Equipment	42	5.0	11.8	59	71.2
Personnel	38	8.3	21.9	60	63.3
Administration	39	9.5	24.7	68	57.4

TABLE 41

AVERAGE SCORES FOR COMPONENTS OF PROCESS

	Mean Score	Standard Deviation	Coefficient of Variation	Possible Score	Mean as % of Possible
Process	338	62.6	18.5	516	65.6
Practice Management	40	8.2	20.2	60	66.7
Radiographic Interpretation	16	5.9	38.2	28	57.1
Data Collection	24	6.8	28.6	45	53.3
Diagnosis	64	24.2	37.9	90	71.1
Treatment Plan	28	29.5	106.1	80	35.0
Treatment	88	11.6	13.2	113	77.9
Sterilization/Infection Control	26	7.0	27.5	40	65.0
Patient Management	53	3.0	5.6	60	88.3

TABLE 42

AVERAGE SCORES FOR COMPONENTS OF OUTCOME

	Mean Score	Standard Deviation	Coefficient of Variation	Possible Score	Mean as % of Possible
Outcome	113	11.0	9.8	149	75.8
Patient Satisfaction	48	1.8	3.8	51	94.1
Patient Oral Hygiene	14	2.7	19.8	21	66.7
Patient Education	11	2.1	20.0	18	61.1
Patient Disability	8	1.9	23.1	9	88.9
Periodontal Disease	12	3.9	31.1	19	63.2
Completion of Treatment	6	6.1	105.8	15	40.0
Recall	14	3.2	22.4	16	87.5

TABLE 43

AVERAGE SCORES FOR ELEMENTS OF TREATMENT COMPONENT

	Mean Score	Standard Deviation	Coefficient of Variation	Possible Score	Mean as % of Possible
Treatment	88	11.6	13.2	113	77.9
Restorative	52.5	7.7	14.7	67.5	77.8
Endodontic	8.7	3.4	39.0	12.5	69.6
Periodontic	11.7	2.9	24.8	12.5	93.6
Oral Medicine	4.4	1.6	36.4	5.0	88.0
Dies	9.9	2.5	25.3	15.0	66.0

FIGURE 6

DISTRIBUTION OF TOTAL SCORES — POSSIBLE 884

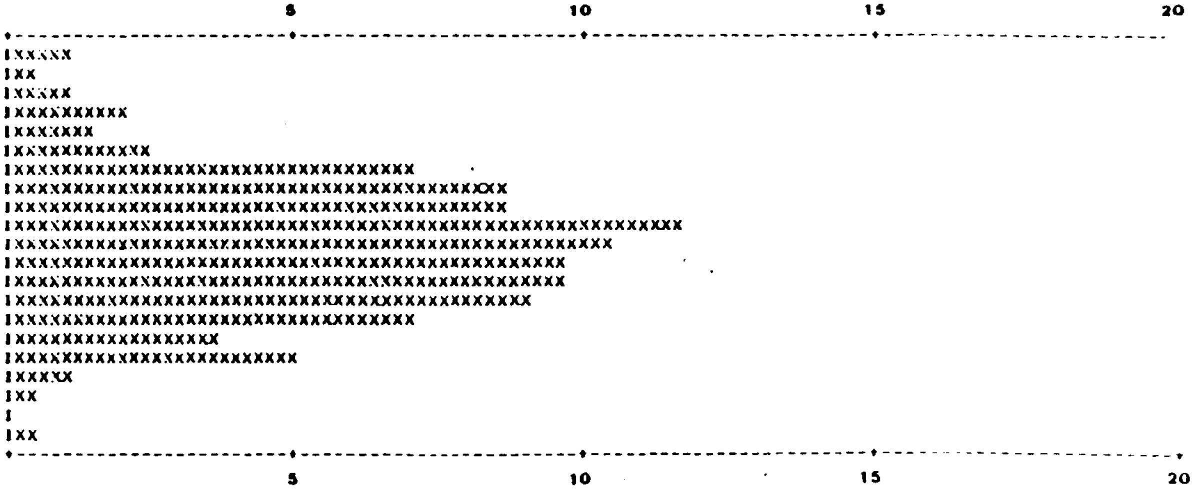
```
                                                  5                   10                  15                  20
AT LEAST         370.000
BUT NOT OVER:     FREQ       %     +------------------------+-------------------------+------------------------+-------------------------+
    389             3       1.0    IXXXXX
    409             1       0.3    IXX
    429             2       0.7    IXXX
    449             7       2.3    IXXXXXXXXXXXXX
    469             7       2.3    IXXXXXXXXXXXXX
    489            17       5.7    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    509            16       5.3    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    529            22       7.3    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    549            20       6.7    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    569            29       9.7    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    588            31      10.3    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    608            17       5.7    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    628            24       8.0    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    648            26       8.7    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    668            21       7.0    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    688            16       5.3    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    708            13       4.3    IXXXXXXXXXXXXXXXXXXXXXXXXX
    728            14       4.7    IXXXXXXXXXXXXXXXXXXXXXXXXXX
    748             7       2.3    IXXXXXXXXXXXXX
    768             2       0.7    IXXX
    788             5       1.7    IXXXXXXXXX
                                   +------------------------+-------------------------+------------------------+-------------------------+
    TOTAL         300     100.0                             5                        10                       15                        20
```

Percent with Score Increment

FIGURE 7

DISTRIBUTION OF STRUCTURE SCORES — POSSIBLE 219

AT LEAST	79.000	
BUT NOT OVER:	FREQ	%
84	3	1.0
90	1	0.3
96	3	1.0
102	6	2.0
107	4	1.3
113	7	2.3
119	21	7.0
125	26	8.7
130	26	8.7
136	35	11.7
142	31	10.3
148	29	9.7
153	29	9.7
159	27	9.0
165	21	7.0
171	11	3.7
176	15	5.0
182	3	1.0
188	1	0.3
194	0	00.0
200	1	0.3
TOTAL	300	100.0



Percent with Score Increment

FIGURE 8

DISTRIBUTION OF PROCESS SCORES — POSSIBLE 516

AT LEAST	173 000	
BUT NOT OVER:	FREQ	%
187	1	0.3
202	3	1.0
217	2	0.7
231	8	2.7
246	8	2.7
261	9	3.0
275	18	6.0
290	25	8.3
305	22	7.3
319	23	7.7
334	29	9.7
349	21	7.0
363	26	8.7
378	20	6.7
393	22	7.3
407	20	6.7
422	11	3.7
437	12	4.0
451	12	4.0
466	4	1.3
481	4	1.3
TOTAL	300	100.0

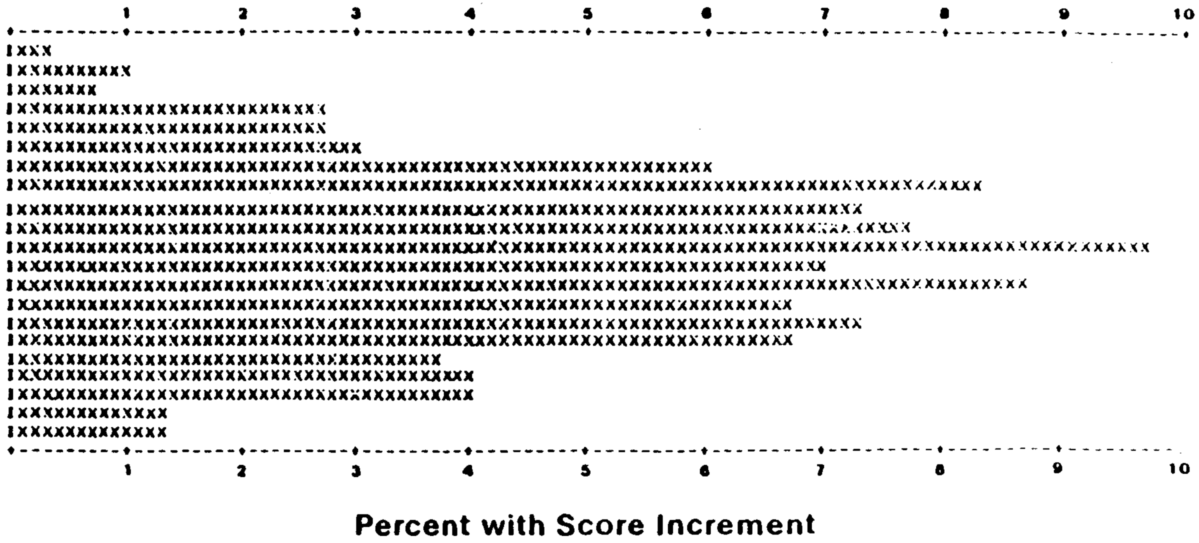


FIGURE 9

DISTRIBUTION OF OUTCOME SCORES — POSSIBLE 149

```
AT LEAST      80.0000                                   5                       10                       15                       20
BUT NOT OVER:   FREQ       %   +------------------------+------------------------+------------------------+------------------------+
      82           1      0.3  IXX
      85           1      0.3  IXX
      88           1      0.3  IXX
      91           4      1.3  IXXXXXXX
      94           6      2.0  IXXXXXXXXXX
      97          10      3.3  IXXXXXXXXXXXXXXXXX
     100          10      3.3  IXXXXXXXXXXXXXXXXX
     102          20      6.7  IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     105          23      7.7  IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     108          33     11.0  IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     111          35     11.7  IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     114          30     10.0  IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     117          25      8.3  IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     120          19      6.3  IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     122          19      6.3  IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     125          19      6.3  IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     128          16      5.3  IXXXXXXXXXXXXXXXXXXXXXXXXXXX
     131          11      3.7  IXXXXXXXXXXXXXXXXXX
     134          10      3.3  IXXXXXXXXXXXXXXXXX
     137           5      1.7  IXXXXXXXX
     140           2      0.7  IXXX
                               +------------------------+------------------------+------------------------+------------------------+
   TOTAL         300    100.0
                                                        5                       10                       15                       20
```

Percent with Score Increment

APPROVAL SHEET

The thesis submitted by David T. Crandall has been read and approved by the following committee:

Dr. Jack A. Kavanagh, Director
Professor, Counseling and Educational Psychology, Loyola

Dr. Ronald R. Morgan
Associate Professor and Director of School Psychology Program
Counseling and Educational Psychology, Loyola

The final copies have been examined by the director of the thesis and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the thesis is now given final approval by the committee with reference to content and form.

The thesis is therefore accepted in partial fulfillment of the requirements for the degree of Master of Arts.

10/18/88
Date

Jack A. Kavanagh
Director's Signature