6-1979

Establishment of Content Validity and Interrater Reliability for a Presentation Performance Evaluation Instrument: Applicability of a Model

Wanda L. Sterbenz
University of Tennessee, Knoxville

To the Graduate Council:

I am submitting herewith a thesis written by Wanda L. Sterbenz entitled "Establishment of Content Validity and Interrater Reliability for a Presentation Performance Evaluation Instrument: Applicability of a Model." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Food Science and Technology.

Betty L. Beach, Major Professor

We have read this thesis and recommend its acceptance:

Charles A. Chance, Louis A. Ehrcke, Marjorie P. Penfield

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a thesis written by Wanda L. Sterbenz entitled "Establishment of Content Validity and Interrater Reliability for a Presentation Performance Evaluation Instrument: Applicability of a Model." I recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Food Systems Administration.

Betty L. Beach

Betty L. Beach, Major Professor

We have read this thesis
and recommend its acceptance:

Charles A. Shmu

W Rhuke

Marjorie P. Penfield

Accepted for the Council:

Vice Chancellor
Graduate Studies and Research

ESTABLISHMENT OF CONTENT VALIDITY AND INTERRATER RELIABILITY FOR A PRESENTATION PERFORMANCE EVALUATION INSTRUMENT: APPLICABILITY OF A MODEL

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Wanda L. Sterbenz

June 1979

ACKNOWLEDGEMENTS

Sincere appreciation is extended to Dr. Betty L. Beach, Associate Professor of Food Science, Nutrition, and Food Systems Administration, for her guidance and assistance throughout this study.

Appreciation is extended to her committee members: Dr. Marjorie P. Penfield; Dr. Louis A. Ehrcke, Associate Professors of Food Science, Nutrition, and Food Systems Administration; and Dr. Charles A. Chance, Associate Professor of Curriculum and Instruction for their helpful suggestions and critical evaluations of this study. A special thank you to Dr. Mary J. Hitchcock and Ms. Dorothy E. Lyon, FSNFSA, for sharing their expertise which increased her professional knowledge. Special acknowledgement is made of the financial support provided by the Allied Health Traineeship for Administrative Dietetics.

The author wishes to express her appreciation to the panel members: Dr. Frances E. Andrews, Mr. Charles Brooks, Ms. Debbie Burton, Ms. Wanda Dodson, Ms. Beverley Hammonds, Ms. Audrey Hay, Mr. Tom Malone, and Mr. Erskine Smith for their time and assistance during the research project. Special thanks goes to Ms. Susan Duncan and Ms. Sue Larsen for consenting to video taping the presentations. Appreciation is also extended to Dr. William L. Sanders and Mr. Charles Brooks for assistance with the statistical analysis of the research data.

To her family and friends, the author extends her deepest appreciation for their understanding and support throughout graduate school. She would especially like to acknowledge Ms. Donna Morse and Ms. Ann Warren whose friendship and support have been immeasurable.

This thesis is dedicated to her parents, Mr. and Mrs. W. V. Sterbenz for their encouragement and support throughout her many years of education.

ABSTRACT

A model for establishing content validity and interrater reliability for performance evaluation instruments (Fiedler et al., 1979) was examined for applicability in another situation. A Class Presentation Evaluation Instrument was developed for testing the model. The model steps included: 1. examining the evaluation instrument for content validity; 2. revising the instrument to establish content validity; 3. viewing and evaluating a standardized situation for establishment of interrater reliability; 4. calculating item variance and item rateability; 5. calculating intraclass correlation scores; 6. revising the instrument to establish item variance and intraclass correlation at predetermined levels; 7. implementing the instrument; and 8. reviewing the instrument periodically.

Nine dietetic educators with 56.6 years of experience teaching dietetic students, interns, and trainees were selected for the panel of experts. The panel had a total of 32.1 years teaching with the Coordinated Undergraduate Program in Dietetics at The University of Tennessee, Knoxville.

The Class Presentation Evaluation Instrument was developed after the panel selected a format and distinguished between essential and non-essential evaluation criteria. The format selected was similar to an instrument used currently by the program. A prioritized list of 37 behavior statements and frequency of written comments on past presentation evaluations indicated essential evaluation criteria. The instrument had sixteen evaluation items in nine categories. The categories were: planning and organization, introduction, body of

presentation, summary, overall presentation, instructional aids, non-verbal communication, and verbal communication. The first seven categories listed behavior indicators under each and was rated with four graduated narrative descriptors with columns for checking "not applicable" and "not observable" and for writing comments. The last two categories had four and five behavior indicators, respectively, that were rated on a dichotomous scale and had the same columns.

The panel determined content validity by examining each evaluation category and descriptor for clarity, word choice, implied meanings, and consistency with identified competencies. The scale extremes were realistic and attainable by all students. Final content validity was established concurrently when interrater reliability was achieved.

The procedure Fiedler et al. (1979) used for calculating item variance and intraclass correlation, an estimate for interrater reliability, was followed and completed during each trial. Further comparisons of intraclass correlation scores were made by separating rating scales and omitting "hot applicable" and "not observable" responses. Statistical Analysis System (Barr et al., 1976) was selected for determining the mean squares.

In three trials using the same video taped standardized situation, interrater reliability was established. Item variance of 0.30 or lower was obtained for 14 of 16 items possible. Intraclass correlation score was 0.44. The fourth trial was to test the stability of the interrater reliability level achieved using a different standardized situation for the panel to view and evaluate. For total instrument, an intraclass correlation score of 0.69 was obtained and 10 of 16 items had variances

equal or less than 0.30. The panel had improved intraclass correlation scores with each trial for the evaluation categories using the four-point scale. The dichotomous scale evaluation categories and behavior indicators did not improve with each trial. After evaluating each standardized situation, the panel discussed the items with high variances for revising or clarifying the evaluation instrument and for obtaining agreement among each other for rating student performance.

The model provided a systematic process for establishing content validity and interrater reliability for the Class Presentation Evaluation Instrument. Interrater reliability of the instrument was influenced when more than one rating scale was used and "not applicable" and "not observable" columns were available for checking. The model can serve as an effective training tool in acquainting new CUP faculty with expected student performance levels and performance evaluation instruments. Other disciplines concerned with the evaluation of student performance in clinical experiences may benefit from the use of the model.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

CHAPTER I

INTRODUCTION

Health related professions such as dietetics have an obligation to society to provide competent practioners. Professional competence should be at a level in relation to the individual's training and professional experience. The level of competence for dietitians has been determined as the performance of job-related tasks either independently or in cooperation with or under the direction of a dietetic specialist (Loyd and Vaden, 1977). As professional experience increases so should the level of competence. The Coordinated Undergraduate Program in Dietetics, College of Home Economics, at The University of Tennessee, Knoxville has a competency-based education program which provides training in professional settings along with didactic course work. This program increases and broadens student's professional training and experience prior to graduation and entry into the dietetics profession. Essentials for Coordinated Undergraduate Programs in Dietetics (CUP) adopted by The American Dietetic Association provide guidelines for these programs (ADA, 1976).

Coordinated Dietetic Programs are divided into a two-year pre-professional phase consisting of general education requirements and basic sciences and a two-year professional phase emphasizing coordination of didactic study with clinical experiences. Students completing these two phases meet the program's established competencies for entry-level dietitians. Graduates who have completed these requirements and have begun the first professional position are classified as entry-level dietitians (Loyd and Vaden, 1977).

A primary goal of the CUP program at The University of Tennessee, Knoxville has been the application of knowledge to a professional environment. All learning experiences are designed to establish entry-level competencies prior to the student's graduation. Student progress is monitored by Clinical Instructors, Dietetic Coordinators, and didactic faculty to ensure that competency levels are being achieved.

A Clinical Instructor and a Dietetic Coordinator are responsible for a small group of students assigned to a particular clinical facility. Fiedler et al. (1979) described this group as the nuclear group and the remaining didactic faculty, Clinical Instructors, students and other professionals as the extended group. The nuclear group Clinical Instructor assisted by the Dietetic Coordinator is responsible for coordinating student activities and assignments, and monitoring these while at the clinical facility. Performance evaluation instruments and checklists are means of monitoring or measuring and evaluating student progress toward competency as a result of these activities and assignments. Student activities and assignments include group discussions, self-instruction modules, written reports, presentations, and video taping. Performance evaluation instruments, checklists, activities, and assignments are also used to identify necessary program revisions.

I. IDENTIFICATION OF PROBLEM

Members of the nuclear and extended groups have been concerned since implementation of the program that the measurement and evaluation of activities and assignments have not been fair and consistent among evaluators. Therefore, in 1975 two educational consultants were retained

to assist program faculty in developing a method for establishing content validity and interrater reliability of performance evaluation instruments. Over a one year period, the performance evaluation instrument, Counseling Checklist, Indirect Patient Care, was developed with content validity and interrater reliability established. Fiedler et al. (1979) developed a model for the process used. The recommendation was made that program faculty determine the generalizability of the model to the development of other types of performance evaluation instruments.

In a group meeting, the Clinical Instructors were asked to identify other evaluation instruments which needed the establishment of content validity and interrater reliability. Two evaluation instruments were suggested, the Class Presentation Checklist and the Case Study Presentation Checklist. The Class Presentation Checklist was used to evaluate the student presentation for staff development in the clinical facility, patient education classes and community education classes. Evaluation of the student presentation measured the degree of competency in the utilization of instructional techniques and materials, communication skills, and applicable subject knowledge. The Case Study Presentation evaluated the level of competency in the student's ability to orally present a patient nutritional care plan and to apply research findings. The Clinical Instructors stated that both instruments had similar evaluation criteria for presentations given by the student.

II. PURPOSE OF STUDY

The Clinical Instructors desired a Class Presentation Evaluation Instrument capable of being used for all student presentations and with

content validity and interrater reliability established by the program faculty. The purpose of this study was to examine the applicability of a model (Fiedler et al., 1979) for developing a Class Presentation Evaluation Instrument and for establishing content validity and interrater reliability for the instrument.

CHAPTER II

REVIEW OF LITERATURE

Performance evaluation instruments are used in learning experiences to provide formative evaluation for dietetic students. The instruments should provide the student with objective, reliable, valid, and useable feedback for changing, improving, or maintaining specific skills, knowledge, or abilities and to reflect the level of competency performed. Other Allied Health Professions have used various techniques and instruments to evaluate student performance in clinical experiences.

I. A MODEL

A model following the process for establishing the content validity and interrater reliability for the Counseling Checklist, Indirect Patient Care (Fiedler et al., 1979) consisted of the following steps: 1. examining the current or developing a new performance evaluation instrument; 2. revising the instrument to establish content validity; 3. viewing a standardized situation for establishment of interrater reliability; 4. calculating item variance; 5. calculating intraclass correlation (r'); 6. revising the instrument to establish item variance and intraclass correlation at predetermined levels; 7. implementing the instrument; and 8. reviewing the instrument periodically.

The instrument developed combined the characteristics of graphic scales, anecdotal records, and checklists. The rating scale was composed

of four graduated narrative behavioral phrases or statements called descriptor blocks. The descriptor blocks were developed to evaluate student competency levels. Each of the four descriptor blocks were expressed in positive terms with the scale extremes being attainable and realistic. Four gradations were selected to avoid central tendency. Columns were provided for checking "not applicable" or "not observable" and writing comments for each evaluation item.

Content validity and interrater reliability were established for the instrument using the expertise of seven dietetic educators. The 7 educators were Clinical Instructors who had been with the program for 1.5 to 4 years and during the 1 year over which the model was developed. Content validity was established by the Clinical Instructors and students examining the instrument for positive, realistic, and attainable evaluation items. Clinical Instructors established interrater reliability by evaluating student performances on video taped standardized situations of student counseling sessions. Item mean, variance, standard deviation, and intraclass correlation, an estimate of interrater reliability, were calculated to determine the degree of agreement for each item and the total instrument. Interrater reliability was achieved for the instrument when the intraclass correlation score was 0.70 or greater, and item agreement was considered high when the item variance was equal or less than 0.30. The Clinical Instructors achieved a 0.72 intraclass correlation score in the last of three trials. Group discussion immediately following each evaluation was considered primary in establishing interrater reliability.

II. PERFORMANCE EVALUATION

Competency-based education programs place the emphasis on the learner and the learning process not the teacher and the teaching process (Bell, 1976; Hart, 1976; Broski et al., 1977). Performance evaluations of students are based on competencies (behaviors or objectives) derived from the knowledge, skills, and attitudes needed to perform in the professional role. Achievement toward competencies is evaluated against performance standards (Conley, 1973; Broski et al., 1977). This information derived from evaluating the student's progress toward competency is necessary for planning the next learning experience by the student and the instructor (Watson, 1976). Both achievement tests and observational instruments can evaluate the student's knowledge and performance (Hughes and Fanslow, 1975). In the nursing program, Conley (1973) stated that some nursing behaviors must be observed to assess competency.

Tape recordings of nursing students in clinical activities were used in a performance evaluation technique (McGrane, 1975). The recordings provided more information for the evaluator for evaluating student performance in clinical activities. Communication skills of patient-nurse interaction were improved when audio-tapes were used by students for self-assessment and instructor-assessment of activities at a psychiatric hospital (Topf, 1969).

Video tapes have been employed by various professions to evaluate student performance in clinical facilities. Student teacher performance (Crosby, 1977), medical student's diagnostic ability (Barrows and Abrahamson, 1964), physical therapy student's patient examination skill

(May, 1978), and nursing student's clinical performance (Frejlach and Corcoran, 1971) were video taped then evaluated by both the student and instructor. A workshop was conducted on performance evaluation for nursing educators (Hayter, 1973). Staged video tapes depicting three levels of nursing students' performance in a laboratory setting facilitated the discussion among the workshop participants. Video taping students' clinical performance for self-evaluation and instructor-evaluation was recommended as a learning tool and as an aid for evaluating clinical performance.

III. PERFORMANCE EVALUATION INSTRUMENTS

Performance evaluation instruments both measure and evaluate the student activity. Erickson and Wentling (1976) differentiated between measurement and evaluation by describing measurement as a data and information collection process; and evaluation as a judgement of a student performance which demonstrates knowledge, understanding, skills, or feelings. An analogy of performance appraisal was given as a yardstick of performance (Jones, 1977; Remmers, 1963). To adequately evaluate student performance, the rater should use an instrument which is clearly defined, easy-to-use, and organized (Chance, 1978).

Procedures for Development of an Evaluation Instrument

The development of an evaluation instrument for assessing nursing student clinical performances resulted in the following recommendations: analyze course objectives and state specific behaviors for achievement; countercheck behaviors by analyzing anecdotal records; review the instrument with students and faculty to achieve adequate, clear, and realistic

behavior items; develop instrument for ease of use; and discuss and agree with the faculty on the levels of performance (Mortiz and Sexton, 1970). The development of a Six-Dimension Scale of Nursing Performance followed a similar procedure for establishing content, structure, validity, and reliability (Schwirian, 1978).

Scales and Formats

In occupational education programs, Erickson and Wentling (1976) listed three scales and formats of observation instruments for performance evaluation: checklists, numerical scales, and graphic scales. Lien (1976) included anecdotal records to the three mentioned.

Checklists. Instruments which list behaviors, skills, or activities and are checked off by the evaluator when performed or accomplished are characteristics of checklists (Lien, 1976; Remmers, 1963; Erickson and Wentling, 1976). An example of an item on a checklist is: "Did the student establish eye contact with each member of the audience?" Then, the evaluator checks a "yes" or "no" column.

Numerical scales. The degree of achievement of specific behaviors, skills, or activities is assigned a corresponding number to the degree of performance displayed. An advantage of the numerical scale is the repeatability of the scale to rate a number of different behaviors or objects. This scale provides more efficient use of rater time and instrument space (Erickson and Wentling, 1976).

Graphic scales. Observation instruments using this scale have been referred to as descriptive or Likert-type scales (Matell and Jacoby,

1971). The instrument contains a statement or item stem followed by word descriptors in a line. These descriptors replace the numbers used in numerical scales. For example, the scale may reflect levels of superior, excellent, good, or poor. Erickson and Wentling (1976) reported that consistency among raters increased when graphic scales were used in place of numerical scales. This consistency has been attributed to the word descriptions being an easier means of classifying the observed behavior. A glossary or a guide can accompany the graphic scale as a reference for finer description of each category and for training raters. The extent of the descriptions would depend on the experience of the raters and familiarity with the behavior being rated (Erickson and Wentling, 1976).

The evaluation form for evaluating dietetic student's clinical performance at The Ohio State University was discussed by Johnson and Hurley (1976). Competencies were translated into a graphic scale with five degrees of achievement. No numerical ranks were listed on the form to eliminate the appearance of giving or receiving a grade. The purpose of the instrument was to show progress toward a competency during the student's involvement in the coordinated program. The student was expected to meet the lowest criteria when entering the program in the Junior year and progress to the highest level by completion of the Senior year.

Anecdotal records. Student performance can be evaluated by recording comments or narrative descriptions of observed behavior on paper. This method is time consuming, inefficient, and subjective in evaluating student behavior (Lien, 1976).

Combining scales and formats. An instrument combining the checklist, graphic, and anecdotal scales was developed by Tower and Vosburgh (1976) to measure student performance in a clinical setting. The instrument had five gradations with "not observed" and "not applicable" columns. Raters could comment on student performance for each evaluation item. A training session for the raters, using ten-minute video tapes and a glossary of descriptors, allowed for discussion and clarification of the instrument.

A combined checklist-rating scale was developed for evaluating physical therapy student's clinical performance (Kern and Mickelson, 1971). Five categories and "no opportunity to observe" were used to evaluate the student's progress. This form provided a means of evaluating the student and effectiveness of the program.

Critical Incident Technique. A nine-point scale was developed by Fruin and Campbell (1977) to evaluate dietitian's performance in observed incidents. A vertical scale was employed listing expected and acceptable behavior in descriptive form at the top of the scale with minimum acceptable behavior at the bottom. The mid-point illustrated neither effective nor ineffective behavior occurring. The authors stated that this type of scale could be used by evaluators who were not involved in the development of the scale. Ingalsbe and Spears (1979) gave guidelines and definitions used in developing the Critical Incident Technique for dietetic students in a management course at Kansas State University. The researchers stated that the technique provided a more objective and efficient method of determining performance effectiveness.

This technique has been used for assessing nursing student performance (McGuire, 1968). LaDuca et al. (1978) used a similar technique to develop the Professional Performance Situation Model for student nurses.

Evaluation Criteria

Performance evaluation criteria for determining effective presentations include verbal and non-verbal skills. These skills can be developed and strengthened as Lee (1974) discussed and demonstrated with an evaluation instrument for evaluating student-teacher relationships for student teachers. Non-verbal communication helps reinforce verbal messages and to establish the atmosphere of the training room. Criteria of non-verbal communication include: eye contact, facial expression, gestures, tone of voice, appearance, and position in the classroom. Evaluation criteria for student teachers at The University of Tennessee, Knoxville, include both verbal and non-verbal communication competencies (Butefish, 1978). Topf (1969) used a Communication Skills Checklist composed of effective and ineffective behavior in initiating the interaction, questioning, and listening for nursing students at clinical facilities.

All evaluation criteria should be objective, valid, and reliable (Hughes and Fanslow, 1975; Lein, 1976; Erickson and Wentling, 1976). MacKay (1974) stated that evaluation of student behavior should be based on some acknowledged or shared criteria. The criteria or goals must be realistic, and students should be potentially capable of achieving these goals for a specific level of preparation.

Rating should be based solely on the student performance as discussed by Hughes and Fanslow (1975) and Erickson and Wentling (1976).

This requires clearly defined descriptions of behaviors to be observed. Finer descriptions would minimize the tendency to subjectively evaluate a performance. Vosburgh et al. (1976) reported a method of minimizing subjective evaluation. Raters were trained to use the observation instrument and to understand the descriptors by using a glossary. Developing instrument guidelines and applying these in rater training sessions reduced raters discrepancies and improved interrater reliability (McGuire, 1968; White et al., 1971).

An evaluation instrument is valid if the intended objectives or competencies are being measured. In developing an instrument to evaluate dietetic student's competencies, Chambers and Hubbard (1978a, 1978b) used an eight member panel to judge the relevancy of each item in relation to the competency being measured. Interrater reliability was also established at 0.75 and 0.69 for the two evaluation forms using Kuder-Richardson 20 formula. Hughes and Fanslow (1975) suggested that a level of 0.85 is appropriate for observational instruments.

Tinsley and Weiss (1975) suggested using intraclass correlation as an estimate for interrater reliability. Interrater agreement measures the consistency of evaluator's ratings when the rate-rerate method of determining reliability and agreement is used.

CHAPTER III

PROCEDURE

A model for establishing content validity and interrater reliability (Fiedler et al., 1979) of performance evaluation instruments was tested for applicability in the development of a student presentation evaluation instrument. Student presentations in the Coordinated Undergraduate Program in Dietetics at The University of Tennessee, Knoxville (UTK) include staff development sessions, patient education classes, community education sessions and case studies. An instrument which could be used for all presentations for fair and consistent evaluation among Clinical Instructors, Dietetic Coordinators, and other CUP faculty was developed prior to establishment of content validity and interrater reliability.

I. PANEL OF EXPERTS

Nine dietetic educators were selected as the panel of experts. Criteria for selection were prior experience teaching dietetic students, interns, or trainees and proximity to campus. Attendance to all research sessions was mandatory. The panel was composed of seven Clinical Instructors and two CUP faculty members who were responsible for evaluating student presentations within the program. One Clinical Instructor and all Dietetic Coordinators did not participate in the study due to scheduling conflicts.

Panelists had 56.6 years of experience with various dietetic education programs teaching dietetic students in traditional or

coordinated programs, internships, and/or traineeships. The panel had 32.1 years teaching experience with the UTK coordinated program (Table 1). Panelists mean number of years with the dietetics program was 3.6 with a range of 0.5 to 7 years. Three panelists had been with the program for two or less years. Panelist C was a Dietetic Coordinator for three years prior to becoming a Clinical Instructor.

II. DEVELOPMENT OF EVALUATION INSTRUMENT

Student presentation checklists used in the program were not considered effective by faculty members and students in measuring competency levels. The program needed a flexible but time efficient instrument for evaluating all presentations and to serve as a guide for student self-evaluation and self-improvement. A new instrument which met the needs of the program was developed.

Instrument Format

Three evaluation formats were developed and given to the panelists. Table 2 summarizes the evaluation categories, behavior indicators, rating scales, and other characteristics of each format. All formats included columns for checking "not applicable," "not observable," and writing comments for each evaluation category or item.

Format 1. The item stem was a simple behavioral statement. Each item stem was evaluated on a graphic two-point rating scale: satisfactory or needs improvement. The item stems were placed under two categories, personal characteristics and presentation with appropriate subdivisions. Subdivisions of the personal characteristics category were

Table 1--Panel members years of experience as dietetic educators by type of program.

Panelist Code	Years of Experience				Total
	Traditional Undergraduate	CUP (UTK)	Internship	Traineeship	
A	8	3	--	--	11
B	3	7	7	--	17
C		4.5	1	1.5	7
D		1.6	--	--	1.6
E		3.5	2	--	5.5
F		2	--	--	2
G		6	--	--	6
H		0.5	--	--	0.5
I		4	2	--	6
Total	11	32.1	12	1.5	56.6
$\overline{X}$		3.6			6.3

Table 2--Summary of format characteristics of presentation evaluation instruments.

Format	Evaluation Criteria		Rating Scale	Other Characteristics
	Category	Behavior Indicator		
1	Personal Characteristics a-Non-Verbal Communication b-Verbal Communication Presentation a-Planning and Organizing b-Content and Delivery c-Instructional Aids	Item stem is a behavior statement stated in simple terms	Graphic Scale 2-point: Satisfactory; Needs Improvement	Columns for Not Applicable, Not Observable, and Comments
2	Same as Above	Same as Above	Numerical Scale 4-point: Poor to Excellent; or Never to Always	Same as Above
3	Planning and Organization Introduction Body of Presentation Summary Overall Presentation Participation Instructional Aids Non-Verbal Communication Verbal Communication	Narrative Descriptor Blocks	Graphic Scale 4 Graduated Levels denoting: Did Not Meet Criteria; Met Minimal Criteria; Acceptable, Needs Improvement, Met All Criteria	Same as Above

non-verbal communication and verbal communication. Subdivisions of the presentation category were planning and organization, content and delivery, and instructional aids.

Format 2. Similar item stems, categories, and subdivisions for evaluation criteria were used. The rating scale was a four-point numerical scale which could denote poor, fair, good, and excellent levels or never, occasionally, frequently, and always levels.

Format 3. The style and format of the Counseling Checklist, Indirect Patient Care used currently in the program was followed. Evaluation criteria were written as narrative descriptor blocks using short sentences or phrases. The four graduated descriptor blocks represented one of these rating levels: did not meet criteria; met minimal criteria; acceptable, needs improvement; and met all criteria. The rating scale was used for each category. Broad behavior categories, such as planning and organization, were placed to the left of the four corresponding descriptor blocks.

Evaluation Criteria

A list of 37 behavior statements was compiled by the researcher from program competencies, other presentation evaluation instruments, and a training manual (Tracey, 1968). The statements were grouped by the categories of presentation, instructional aids, verbal communication, and non-verbal communication. All panelists received a list for prioritizing each statement according to importance using 1 to 37 with 1 representing highest priority. Each number was used only once. One week later, seven lists were collected and ranked totals were determined for each

statement. Ranked totals ranged from 20 to 211 within a possible range from 7 to 259. The statements were reorganized by ranked totals to determine priority by category (Table 3). The list was divided into thirds to better reflect priority rankings. Student planning and organization which considered the audience needs was ranked most important. Other items within the presentation category were sequence; student's subject knowledge; statement of purpose and objectives; emphasis of main points; and effective introduction and summary. Under the category of instructional aids, panelists placed selection and development of aids to complement the presentation and be appropriate for the audience. Under the verbal communication category, student vocabulary, sentence structure, and illustrations appropriate for the audience were placed.

The researcher reviewed evaluators' written comments on the Class Presentation Checklist for the past three years. Comments were tallied for Case Study presentations, Staff Development sessions, and Community Education classes (Table 4). The comments were placed under four categories: presentation; instructional aids; verbal communication, non-verbal communication. Examples of comments placed under these categories were: appropriate subject for audience for presentation; small visuals for instructional aids; pronouniciation of words for verbal communication; and eye contact for non-verbal communication. The mean number of comments for each category under the three types of student presentations was determined.

Seventy-seven percent of the behaviors in the top-third of the prioritized list and approximately one-half of the summarized written

Table 3--Categorization of prioritized behavioral statements for evaluation of student presentations by a panel of dietetic educators.

Priority Level	Category	Number of Statements	Percent of Category Statements in Each Priority Level
Top One-Third	Presentation	10	50
	Instructional Aids	2	40
	Verbal Communication	1	17
	Non-Verbal Communication	0	0
Middle One-Third	Presentation	6	30
	Instructional Aids	0	0
	Verbal Communication	3	50
	Non-Verbal Communication	3	50
Lower One-Third	Presentation	4	20
	Instructional Aids	3	60
	Verbal Communication	2	33
	Non-Verbal Communication	3	50

Table 4--Summary of written comments on the class presentation checklist used for three types of presentations for the period 1975-78.[a]

Type of Presentation	Category	Number of Evaluations Reviewed	Number of Comments	Mean Number Comments Per Evaluation
Case Study		25		
	Presentation		35	1.4
	Instructional Aids		20	.8
	Verbal Communication		34	1.4
	Non-Verbal Communication		16	.6
Staff Development		20		
	Presentation		59	3.0
	Instructional Aids		11	.6
	Verbal Communication		15	.8
	Non-Verbal Communication		23	1.2
Community Education		15		
	Presentation		51	3.4
	Instructional Aids		8	.5
	Verbal Communication		8	.5
	Non-Verbal Communication		8	.5

[a]Coordinated Undergraduate Program in Dietetics, College of Home Economics, The University of Tennessee, Knoxville, 37916.

comments were classified in the presentation category. Panelists were given the results of the prioritized list and summary of comments. Discussion resulted in identification of criteria considered essential for evaluating class presentations. Suggestions made during the discussion were noted. A presentation instrument was developed by the researcher using the format selected, evaluation criteria identified as essential by the panel, and suggestions received.

III. ESTABLISHMENT OF CONTENT VALIDITY

The developed Class Presentation Evaluation Instrument was given to the panelists in a group meeting. The instrument format and rating scale were explained by the researcher. The panel examined and discussed individual evaluation criteria under each category and for each item for clarity, word choice, implied meanings, and consistency. To facilitate the discussion, portions of a video taped student presentation was shown. Individual categories and behavior indicators were revised following the group discussion and the revised instrument returned to the panelists. Final content validity and interrater reliability was achieved concurrently when interrater reliability was established.

IV. STANDARDIZED SITUATION

Repeated viewing and evaluation of a standardized student presentation provided the basis for establishing content validity and interrater reliability for the instrument. The viewing of the same presentation without any deviations was achieved by video taping a student presentation. Fiedler et al. (1979) referred to a video taped student performance as a standardized situation.

Prior to video taping, application was made and approval granted by the Committee for Review of Research Involving Human Subjects. Protection measures for the research participants included anonymity during the project video taping and in publications. Prior to video taping at a clinical facility, participants (student, employees, and Clinical Instructors) were requested to sign a Model Release Form (Appendix A) after the purpose of the video taping and the potential uses of the video tape were explained by the researcher. If an employee did not wish to be taped, the person was placed outside the camera area and not included on the video tape. If a student or a Clinical Instructor objected, the tape was not used in the study. All participants signed the Model Release Form.

Standardized Situation 1

For Standardized Situation 1 an actual student presentation was filmed instead of role playing to lend authenticity to the situation. The situation was a 10-minute staff development session on "Fire Safety" conducted at a local clinical facility. The presentation was video taped on a three-fourths inch video cassette tape cartridge using a Sony black and white camera mounted on a stationary tripod. The student, audience, training room facilities, and instructional aids were filmed using a zoom lens for close, medium, and long shots to enable panelists to more effectively evaluate student performance.

Standardized Situation 2

A video taped patient education class for diabetic patients was selected from the program's tape files as Standardized Situation 2 to

determine the stability of the interrater reliability level achieved with Standardized Situation 1.

A Clinical Instructor who had taped a majority of the program's video tapes filmed the student presentation at the clinical facility. The technical quality of the tape was not evaluated by video tape technicians since the camera person had been responsible for CUP program video taping the last four years.

IV. ESTABLISHMENT OF INTERRATER RELIABILITY

Establishing interrater reliability consisted of following a six-step process called a trial. The steps were: listening to introductory remarks by the researcher; viewing the standardized situation; evaluating the student presentation with the instrument; recording the panelist's responses; calculating interrater reliability level; discussing items in disagreement based on item variance, item rateability, and intraclass correlation; and revising the instrument. Four trials were conducted over a one month period.

Panelists were given copies of the revised Class Presentation Instrument for evaluating the standardized situation. The researcher gave introductory remarks describing type of presentation, clinical facility, training room environment, and identification of the audience, i.e. dietary employees. Panelists were encouraged to keep interaction to a minimum during the viewing and evaluation of the performance. The situation was viewed on a 19-inch diagonal television screen.

After rating the student performance, the panelists were requested to assign a numerical value of one to four to the corresponding

rating scale descriptor blocks. "Not applicable" and "not observable" columns were assigned values of five and six respectively. Dichotomous rating scale items were arbitrarily assigned numerical values of two and three. Panelist's rating for each item was recorded and item mean, variance, and standard deviation calculated on the Interrater Reliability Response Form (Appendix B).

Intraclass correlation scores (r'), an estimate of interrater reliability, was calculated for each trial following the formula (Fiedler et al., 1979):

$$r' = \frac{s_b^2 - s_w^2}{s_b^2 + (n - 1)s_w^2}$$

The scores were determined for the total instrument (Fiedler et al., 1979) and by rating scales (Sanders, 1979) with each considering the influence of "not applicable" and "not observable" responses by different methods.

Total instrument r' was calculated by determining the item mean using only primary responses, and standard deviation, item variance, $\sum x$, $\sum x^2$, and $(\sum x)^2$ using primary and column responses for a constant (N). Primary responses were the panelists ratings on one of the rating scales and checks for "not applicable" and "not observable" were column responses that indicated the panelists inability to rate an item. Values of 5 were assigned to column responses if the item mean was 2.5 or higher and 0 was assigned if the item mean was lower. Item variance of 0.30 or lower reflected rater agreement for each item. Discussion of each item and instrument for clarification or revision was necessary when item variance

was greater than 0.30 and intraclass correlation scores were lower than 0.70 (Fiedler et al., 1979).

Intraclass correlation scores were calculated by separating rating scales and omitting column responses. Group 1 was the responses from the 4-point scale and Group 2 was the dichotomous scale responses. The influence of column responses was considered by determining the item rateability of each item. Item rateability was a percent of panelists responses rated or not rated on the rating scale. With the uneven number of responses, mean squares were determined using Statistical Analysis System (Barr et al., 1976) on an IBM 360 computer.

Total instrument r' and by rating scale r' were compared by trials. The item variance and item rateability was determined for each item and trial.

CHAPTER IV

RESULTS AND DISCUSSION

The Class Presentation Evaluation Instrument was developed to measure student achievement toward three terminal competencies; utilization of instructional strategies; selection of instructional techniques and materials; and utilization of communication skills (CUP, 1978). The panel selected by consensus the formats and distinguished between essential and non-essential evaluation criteria for inclusion in the instrument. Content validity of the instrument and interrater reliability were established by a panel of dietetic educators after four trials viewing and evaluating two standardized situations.

I. PRESENTATION EVALUATION INSTRUMENT

Three evaluation formats were developed and presented to the panel. Advantages and disadvantages of each format's characteristics were discussed prior to selection.

The panel favorably viewed Format 1 for the personal characteristics category but not for the presentation category. Personal characteristics were considered to be important and could be satisfactorily evaluated with a "yes or no" rating. Written comments would clarify the rating when needed for the personal characteristics category. Panelists were concerned that no degrees of performance in the presentation category was available for checking and that excessive written comments necessary for documenting this category would be time consuming.

Format 2 was not selected because of the potential translation of a numerical rating into a letter grade by both students and faculty. Therefore, the student would not receive the formative evaluation benefit from the rating scales' four levels.

The panel selected Format 3 to maintain consistency with program performance evaluation instruments. Advantages of maintaining a consistent format were considered that students would have less difficulty in understanding evaluations received and would be more inclined to use the evaluation results positively for improving future presentation performance. The instrument would be time efficient and would effectively measure student competency levels once evaluation categories and descriptors were identified and refined.

An instrument was developed by the researcher following panelists' suggestions for format and evaluation criteria (Appendix C). The instrument had 16 evaluation items in 9 categories. Behavior indicators were listed under each category.

II. ESTABLISHMENT OF INTERRATER RELIABILITY

Content validity and interrater reliability were determined for the Class Presentation Evaluation Instrument in three trials with one standardized situation. A fourth trial tested the stability of the interrater reliability level achieved using a different standardized situation. Intraclass correlation scores were determined for the total instrument and by rating scales for each trial.

Results of Trials

Trial 1. Intraclass correlation score for the total instrument

was 0.10 (Table 5) and by rating scales, Group 1 was 0.01 and Group 2 was 0.03 (Table 6). All gradations of the rating scales were used. In Group 1, the panel rated all items with the planning and organization category receiving the same ratings. Group 2 had total panel agreement for rating except for two items. Item rateability was split for these two items (Table 7). New words had five panelists rating the item and four not rating the item. Nine of 16 evaluation items had variances of 0.30 or less (Table 5). Item variances for all items and trials are listed in Appendix D (Table 8). Five of the seven items with variances above 0.30 were rated on the four-point scale. These items were: introduction, body of presentation, summary, overall presentation, and participation. With these items, panelists disagreed as to what level of performance was to be demonstrated for each competency. The two remaining items with high variances were eye contact under non-verbal communication and new words under verbal communication. The frequency of occurrence for rating the items was discussed. Clarification for the use of the "not applicable" and "not observable" columns were given.

The panel suggested a few word changes for the descriptors. These changes were made prior to Trial 2. Panelists felt that by discussing the evaluation items a consensus of expected performance levels for each category was beginning to be achieved.

Trial 2. With this trial, total instrument r' was improved to 0.47 and by rating scales, Group 1 improved to 0.24 and Group 2 to 0.31. The fourth gradation descriptor was not used for rating any items. Panelists ratings by rating scale were in the second and third gradations with one response in the first gradation. Item variances of 0.30 or less improved by three items, leaving four evaluation items to be discussed. These four items had been discussed during Trial 1, i.e. overall presentation, participation,

Table 5--Measures of interrater reliability for total instrument for panel members evaluating student presentations in standardized situations.

Standardized Situation Number	Trial Number	Number of Panelists	Intraclass Correlation (r')	Number of Items (N = 16) with Variance $\leq$ 0.30
1	1	9	.10	9
	2	9	.47	12
	3	9	.44	14
2	4	9	.69	10

Table 6--Measures of interrater reliability by rating scales for panel members evaluating student presentations in standardized situations.[a,b]

Standardized Situation Number	Trial Number	Number of Panelists	Intraclass Correlation (r')	
			Group 1[c]	Group 2[d]
			N = 7	N = 9
1	1	9	.01	.03
	2	9	.24	.31
	3	9	.55	.07
2	4	9	.57	.03

[a]Values assigned 0, 1 dichotomous items.

[b]'Not applicable' and 'not observable' not calculated.

[c]Four point rating scale items.

[d]Dichomotous rating scale items.

Table 7--Percent of items rated by panelists evaluating student presentations in standardized situations by rating scales, evaluation items, and trials.

Rating Scale	Instrument Item Number	Evaluation Items	Trials (%) 1	2	3	4
Four-Point Scale	1	Planning and Organization	--[a]	--	--	--
	2	Introduction	--	--	--	--
	3	Body of Presentation	--	--	--	--
	4	Summary	--	--	--	--
	5	Overall Presentation	--	--	--	--
	6	Participation	--	89	--	33
	7	Instructional Aids	--	--	--	--
Dichotomous Scale	8	Non-Verbal Communication				
		a-Appearance	--	--	--	--
		b-Performance	--	--	--	--
		c-Confidence	--	--	--	--
		d-Eye Contact	89	67	44	22
	9	Verbal Communication				
		a-Vocabulary	--	--	--	--
		b-Speech	--	--	--	--
		c-Articulation	--	--	--	--
		d-Voice	--	--	--	--
		e-New Words	56	33	56	--

[a] --=100%.

eye contact, and new words. Two items had lowered variance and two had raised variance from Trial 1. Panelists disagreement as to the level of performance for these evaluation items was pursued in the discussion. The overall presentation category needed a statement in the descriptors for accuracy of information. The instrument was changed to include this suggestion. Through discussion, performance levels for the participation category were clarified and item rateability solved. Evaluation items, eye contact and new words, were further discussed for the frequency in relation to rating. Item rateability for eye contact was changed by two panelists. Six panelists indicated that new words could not be rated, an increase of two from Trial 1.

Trial 3. Panelists commented that the Trial 2 panel discussion influenced the ratings. Total instrument r' decreased to 0.44 and by rating scales, Group 1 r' increased to 0.55 and Group 2 decreased to 0.07. Panelists disagreement on rating dichotomous items, eye contact and new words, influenced r' for the total instrument and Group 2. Item rateability and item variance reflected that the panel was almost evenly split on rating the items on the rating scale. The panelists agreed on rating 14 of 16 evaluation items. Group 1 items had item variances of 0.28 or lower. The high variances of items, eye contact and new words, were based on disagreement as related to the frequency of occurrence of rating one of the two points on the scale and checking the "not applicable" or "not observable" columns. The panel through consensus determined a satisfactory level of interrater reliability for using the instrument with Standardized Situation 1 except for eye contact and new words had been achieved.

Trial 4. The stability of the level of the interrater reliability achieved in three trials of evaluating a staff development standardized situation was tested by having the panelists view and evaluate Standardized Situation 2. The panel rated 10 items with a variance of 0.30 or lower and obtained r' of 0.69 for the total instrument. All variances were 0.78 or lower indicating item agreement was not as distant as with the previous three trials. The high variance for the item, eye contact, panelists attributed to the video tape. By rating scales, r' for Group 1 was 0.57 and Group 2 was 0.03.

III. GENERAL DISCUSSION

Panelists tended to mark the middle to the lower gradations on the Class Presentation Evaluation Instrument. The top gradation was written to be realistic and reasonable for performance in relation to the student's training. Some panelists found this a difficult concept to accept, reflecting the need to adjust thinking and attitudes toward using the instrument scale.

The use of the "not applicable" and "not observable" columns did influence the level of intraclass correlation. Fiedler (1979) stated that during the development of the Counseling Checklist, Indirect Patient Care the panelists did not use these columns during the final trials as agreement for rateability of performance had been achieved during the one year of developing the model.

Interrater reliability was achieved through formal discussion of evaluation criteria among the panelists. The evaluation of authentic standardized situations initiated the discussion for determining acceptable levels of student performance among panelists. The 0.69 intraclass

correlation score for the total instrument was achieved over a one month period. Further comparison of the panelists responses indicated that interrater reliability for evaluation items rated on the four-point scale improved with each trial. The dichotomous scale items did not consistently improve, reflecting that agreement of item rateability for these evaluation items was not present among the panelists.

CHAPTER V

CONCLUSIONS, RECOMMENDATIONS, AND SUMMARY

I. CONCLUSIONS

Following the model (Fiedler et al., 1979), content validity and interrater reliability were established for the Class Presentation Evaluation Instrument. The model served as a guide for viewing and evaluating standardized student presentations which initiated group discussions for determining various levels of acceptable performance. The model can assist the evaluators in achieving consensus and consistency of rating students presentations when performance levels may change due to needs or emphasis within the program or CUP faculty may be new and untrained.

The procedure for calculating item variance and intraclass correlation (r') (Fiedler et al., 1979) for total instrument provided a quick, on-site method for identifying specific items which needed further discussion for clarification and possible item revision. Further comparison of data emphasized the importance of agreement among raters as to item rateability in establishing interrater reliability. The item rateability indicated whether there was agreement among panelists as to the ability to rate items. Whereas, item variance reflected how closely panelists agreed for rating different levels of performance. Evaluation items with a low or high percent of rateability indicated most panelists agreed to the rating but mid-percentages indicated the panelists were almost evenly split. Therefore, the question of rateability for each

evaluation item should be discussed by the panel as to why some panelists could rate the item and others could not. This should be resolved before proceeding on to a new trial. Item variance facilitated the discussion for distinguishing among the levels of performance. Knowing both item rateability and item variance can assist the panel with discussion of the instrument and student performance levels.

The standardized situations filmed at the clinical facility produced adequate video tapes for viewing and evaluating student performances. The authentic situation generally revealed the training room environment and the handling of the situation by the student. This element would not be illustrated with a role played or staged situation.

The placement of behavior indicators to be considered for each category assisted the panelists in evaluating the performance quickly. The panelists commented that often an evaluator's guide was not available when the student was being evaluated. The incorporation of the guide with the evaluation instrument would also be beneficial to the student and faculty when reviewing the evaluation.

II. RECOMMENDATIONS

Content validity and interrater reliability were established for the Class Presentation Evaluation Instrument by seven Clinical Instructors and two CUP faculty members. Content validity and interrater reliability must be extended to other members of the nuclear group. The transferability of content validity and interrater reliability for student presentations should be determined for patient education and community education sessions. These standardized situations should be authentic

presentations conducted at a clinical facility. The video tape should include a complete picture of facility conditions which may influence student performance. A complete picture is needed for panelists to evaluate the standardized situations realistically.

A systematic process is provided by the model for establishing content validity and interrater reliability for performance evaluation instruments. The trials provide the opportunity for evaluators to discuss performance levels by viewing student presentations on video tapes. The procedure for calculating r' for total instrument are simple to complete with a hand calculator and are instrumental in directing the panel discussion. However, the rateability of each evaluation item should be considered as well as item variance in achieving interrater agreement. Intraclass correlations scores by rating scales can identify which scale items need further discussion and revision.

The UTK coordinated program should implement the model's process for establishing content validity and interrater reliability for all performance evaluation instruments used by the program. The periodic review should be conducted on a yearly basis since the CUP faculty may emphasize different competency levels due to changing program or students needs.

The model could serve as an effective training tool. Following the model steps new CUP faculty would become acquainted with expected student performance levels and performance evaluation instruments. Students could also benefit from participating with the CUP faculty in the training session by becoming more familiar with the evaluation instrument and expected performance and competency levels. The use of

the model could benefit other disciplines concerned with the evaluation of student performance in clinical experiences.

III. SUMMARY

Faculty with the Coordinated Undergraduate Program in Dietetics at The University of Tennessee, Knoxville have been concerned that student performance has not been fair and consistent among evaluators. The Class Presentation Checklist and Case Study Checklist were identified by Clinical Instructors as needing establishment of content validity and interrater reliability. An instrument reflecting commonalities among all student presentations and requiring minimal time for completion was desired by the CUP faculty. The purpose of this study was to examine the applicability of a model (Fiedler et al., 1979) for developing a Class Presentation Evaluation Instrument and establishing content validity and interrater reliability for the instrument. The steps followed were: review current or a new performance evaluation instrument for content validity; revise evaluation items or instrument; use instrument to evaluate a standardized situation; calculate intraclass correlation scores, item variance, and item rateability; revise evaluation items or instrument; implement instrument; and review periodically.

The instrument was developed from program competencies, panelists' suggestions, selected formats, and identified essential evaluation criteria. Three formats with different rating scales, one currently used in the program, were given to the panel for selection. A list of 37 behavioral statements were prioritized and evaluators' written comments on old presentation checklists for a three year period

were tallied. Both the prioritized list and tallied comments determined the evaluation criteria for the instrument.

The instrument was composed of sixteen evaluation items in nine categories. Under each category, two to five behavior indicators were listed as a guide for evaluating the categories. The first seven categories: planning and organization, introduction, body of presentation, summary, overall presentation, participation, and instructional aids were rated on a scale of four graduated narrative descriptors. The last two categories: non-verbal and verbal communication had four and five behavior indicators, respectively, that were written as behavioral statements and rated on a dichotomous scale. The instrument had columns for checking "not applicable" and "not observable" and writing comments.

Panelists participated in four trials using two standardized situations. In the first three, a staff development situation was viewed and evaluated and the fourth was a patient education class. The fourth was to test the stability of the interrater reliability level achieved with the first situation. All trials were held one week apart.

Interrater reliability levels were calculated and compared using intraclass correlation scores and two methods of considering the influence of the columns of "not applicable" and "not observable" with the panel's rating scales responses. The total instrument intraclass correlation score was based on the panel's primary responses and assigning a value of 5 or 0 to the column responses. The item variance indicated rater agreement on each item. The intraclass correlation score by rating scale was based on omitting the column responses and using

only the primary responses. Item rateability was determined as a percentage of panelists rating or not rating an evaluation item.

Trials 1 through 3 achieved an intraclass correlation score of 0.44 for the total instrument and 0.69 for Trial 4. For the first three trials, 14 items of 16 possible and 10 of 16 for the fourth trial had item variances of 0.30 or lower. Intraclass correlation scores by rating scale showed improvement with each trial for the 4-point scale. The dichotomous scale did not consistently improve scores for each Trial. Item rateability was 100 percent for all items and all trials except twice for the 4-point scale. The dichotomous scale had item rateability for all except seven times. This indicated the need for item revision in the dichotomous scale.

The model provided a systematic process for establishing content validity and interrater reliability for a performance evaluation instrument. Interrater reliability of the instrument was influenced when 1) two or more rating scales were employed and 2) "not applicable" and "not observable" columns were used. Calculation of intraclass correlation scores were performed by two methods that considered the two influences.

LIST OF REFERENCES

LIST OF REFERENCES

ADA. 1976. "Essentials for Coordinated Undergraduate Programs in Dietetics." The American Dietetic Association, Chicago.

Barr, A. J., Goodnight, J. H., Sall, J. P., and Helwig, J. T. 1976. Statistical Analysis System. SAS Institute Inc., Raleigh, N.C.

Barrows, H. S. and Abrahamson, S. 1964. The programmed patient: A technique for appraising student performance in clinical neurology. J. Med. Educ. 39: 802.

Bell, C. G. 1976. Role- vs. entry-level competencies in competency-based education. J. Am. Dietet. Assoc. 69: 133.

Broski, D., Alexander, D., Brunner, M., Chidley, M., Finney, W., Johnson, C., Karas, B., and Rothenberg, S. 1977. Competency-based curriculum development: A pragmatic approach. J. Allied Health. 6: 38.

Butefish, W. L. 1978. "Student Teaching Handbook." The University of Tennessee. College of Education. Knoxville, Tennessee.

Chambers, M. J. and Hubbard, R. M. 1978a. Assessing achievement for minimum academic competency. II. Validity and reliability. J. Am. Dietet. Assoc. 73: 31.

Chambers, M. J. and Hubbard, R. M. 1978b. Assessing achievement for minimum academic competency. I. Instrument development. J. Am. Dietet. Assoc. 73: 27.

Chance, C. A. 1978. Private communication. The University of Tennessee, Knoxville, Tennessee.

Conley, V. C. 1973. "Curriculum and Instruction in Nursing." Little, Brown and Company, Boston.

Crosby, M. H. 1977. Teaching strategies: A microteaching project for nurses in Virginia. Nurs. Res. 26: 144.

CUP. 1978. Terminal and enabling competencies. Unpublished paper. The University of Tennessee, Knoxville, Tennessee.

Erickson, R. C. and Wentling, T. L. 1976. "Measuring Student Growth Techniques and Procedures for Occupational Education." Allyn and Bacon, Inc., Boston.

Fiedler, K. M. 1979. Private communication. Case Western Reserve University, Cleveland, Ohio.

Fiedler, K. M., Beach, B. L., and Hayman, J. 1979. Dietetic performance evaluation: Establishment of validity and reliability. Submitted for publication, J. Am. Dietet. Assoc.

Frejlach, G. and Corcoran, S. 1971. Measuring clinical performance. Nurs. Outlook. 19: 270.

Fruin, M. F. and Campbell, J. P. 1977. Developing behaviorally anchored scales for rating dietitians' performance. J. Am. Dietet. Assoc. 71: 111.

Hart, M. 1976. Competency-based education. J. Am. Dietet. Assoc. 69: 616.

Hayter, J. 1973. An approach to laboratory evaluation. J. of Nurs. Educ. 12(4): 17.

Hughes, R. P. and Fanslow, A. 1975. Evaluation: A neglected area of competency-based education. J. Home Econ. 67(5): 23.

Ingalsbe, N. and Spears, M. C. 1979. Development of an instrument to evaluate critical incident performance. J. Am. Dietet. Assoc. 74: 134.

Johnson, C. A. and Hurley, R. S. 1976. Design and use of an instrument to evaluate students' clinical performance. J. Am. Dietet. Assoc. 68: 450.

Jones, L. 1977. In place of performance appraisal. Educ. and Training. 19(1): 28.

Kern, B. P. and Mickelson, J. M. 1971. The development and use of an evaluation instrument for clinical instruction. Phys. Therapy. 51: 540.

LaDuca, A., Engel, J. D., and Risley, M.E. 1978. Progress toward development of a general model for competence definition in health professions. J. Allied Health. 7: 149.

Lee, E. C. 1974. Interpersonal communication skills. Unpublished paper. Emory University, Atlanta, Georgia.

Lien, A. J. 1976. "Measurement and Evaluation of Learning." 3rd ed. Wm. C. Brown Company Publishers, Dubuque, Iowa.

Loyd, M. S. and Vaden, A. G. 1977. Practioners identify competencies for entry-level generalist dietitians. J. Am. Dietet. Assoc. 71: 510.

MacKay, R. C. 1974. Evaluation of faculty and students. . . A means towards fuller communication and greater productivity. J. of Nurs. Educ. 13(1): 3.

Matell, M. S. and Jacoby, J. 1971. Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. Educ. and Psych. Meas. 31: 657.

May, B. J. 1978. Competency based evaluation of student performance. J. Allied Health. 7: 232.

McGrane, H. F. 1975. Tape recorded evaluation: A method of teaching. J. of Nurs. Educ. 14(1): 11.

McGuire, C. H. 1968. Testing in professional education. Rev. of Educ. Res. 28(1): 54.

Mortiz, D. A. and Sexton, D. L. 1970. Evaluation: A suggested method for appriasing quality. J. of Nurs. Educ. 9(1): 17.

Remmers, H. H. 1963. Rating methods in research on teaching. In: Gage, N. L., ed. 1963. "Handbook of Research on Teaching." Rand McNally and Company, Chicago.

Sanders, W. L. 1979. Private communication. The University of Tennessee, Knoxville, Tennessee.

Schwirian, P. M. 1978. Evaluating the performance of nurses: A multidimensional approach. Nurs. Res. 27: 347.

Tinsley, H. E. A. and Weiss, D. J. 1975. Interrater reliability and agreement of subjective judgements. J. of Couns. Psych. 22: 358.

Topf, M. 1969. A behavioral checklist for estimating the development of communication skills. J. of Nurs. Educ. 8(4): 29.

Tower, J. B. and Vosburgh, P. M. 1976. Development of a rating scale to measure learning in clinical dietetics. I. Theoretical considerations and method of construction. J. Am. Dietet. Assoc. 68: 440.

Tracey, W. R. 1968. "Evaluating Training and Development Systems." American Management Association, Inc., New York.

Vosburgh, P. M., Tower, J. B., Peckos, P. S., and Mason, M. 1976. Development of a rating scale to measure learning in clinical dietetics. II. Pilot test. J. Am. Dietet. Assoc. 68: 446.

Watson, D. R. 1976. Coordination of classroom and clinical experience. J. Am. Dietet. Assoc. 69: 621.

White, J. L., Wenberg, B. G., Camisconi, J. S. 1971. Evaluation of medical dietetic graduates. J. Am. Dietet. Assoc. 58: 516.

APPENDICES

APPENDIX A

MODEL RELEASE FORM

In consideration of value received, receipt of which is hereby acknowledged, I give the College of Home Economics, The University of Tennessee, Knoxville, the right to copyright and make use of photographs, audio tapes, video tapes, or film through any media, for educational or research purposes as deemed fit by the photographer, research director, project director, or their agent. I do not desire to examine or inspect the finished product or the use to which it may be applied. The production and all its rights now belong to the photographer, research director, project director and/or their agents.

Date: _____

Signature of Model: _____

Witness: _____

Coordinated Undergraduate Program in Dietetics, 1977.

APPENDIX B

INTERRATER RELIABILITY RESPONSE FORM

Item	Response 1	2	3	4	5	6	7	s^2	$\bar{x}$	s	n	$\sum x$	$\sum x^2$	$(\sum x)^2$
1														
2														
3														
4														
5														
6														
7														
8														
9														
10														
11														
12														
13														
14														
15														
16														
17														
18														
19														
20														
21														
22														
23														

k = number of items = _____; n = number of subjects: ________ | ____ N | ____ $\sum\sum x$ | ____ $\sum\sum x^2$ | ____ $\sum(\sum x)^2$

* *

Within SS = $SS_w = 1/n(n\sum\sum x^2 - \sum(\sum x)^2)$ = 1/ (-) = 1/ (-) = ________

$df_w = kn - k$ = _____ = _____ $s_w^2 = \frac{SS_w}{df_w}$ = _____ = _____

* *

Between SS = $SS_b = 1/N(k\sum(\sum x)^2 - (\sum\sum x)^2)$ = 1/ (-) = 1/ (-) = ________

$df_w = k - 1$ = _____ = _____ $s_b^2 = \frac{SS_b}{df_b}$ = _____ = _____

* *

Interrater Reliability = (intraclass correlation) = $\frac{s_b^2 - s_w^2}{s_b^2 + (n-1)s_w^2}$ = ________ = ________

Fiedler et al., 1979.

APPENDIX C

CLASS PRESENTATION EVALUATION INSTRUMENT

Observer _____ Student _____

Date _____ Topic _____

Please indicate with a checkmark above the descriptor in each category you feel best describes the presentation.

Category and Behavior Indicators	Descriptors					
1. Planning and organization - type - audience's needs - knowledge of subject	Made little attempt for in-depth planning and organization. Made little attempt to consider audience's needs. Demonstrated confused knowledge of subject.	Attempted, but somewhat lacked in-depth planning and organization. Attempted to consider audience's needs. Demonstrated limited knowledge of subject.	Demonstrated in-depth planning and organization; some specific areas needed further "polishing." Attempted to meet audience's needs. Demonstrated basic knowledge of subject.	Demonstrated careful complete, imaginative planning and organization. Met audience's needs. Demonstrated thorough understanding of subject.	Not Applicable Not Observable	Comments
2. Introduction - type - purposes and objectives - audience's attention	Gave no definable introduction. Made little attempt to state purposes and objectives. Had little of the audience's attention.	Gave too brief/lengthy introduction. Attempted to state purposes and objectives. Had most of the audience's attention.	Gave adequate introduction. Stated purposes and objectives. Had audience's attention.	Gave effective and imaginative introduction. Stated purposes and objectives. Had audience's attention.	N.A. N.O.	

Category and Behavior Indicators	Descriptors					
3. Body of Presentation - main points - sequence	Made little attempt to emphasize, support/ reinforce main points. Made little attempt to proceed in logical sequence.	Attempted to emphasize support-reinforce main points. Attempted to proceed in logical sequence.	Good attempt to emphasize, support/reinforce main points. Good attempt to proceed in logical sequence.	Emphasized, supported and reinforced main points throughout presentation. Proceeded in logical sequence.	N.A. N.O.	Comments
4. Summary - type - main points	Gave abrupt closing. Restated minimally main points. Introduced new points.	Gave too brief/lengthy summary. Attempted to restate main points.	Gave adequate summary. Good attempt to restate main points.	Gave effective summary and restated main points concisely.	N.A. N.O.	
5. Overall presentation - purposes and objectives - information - accuracy - flow	Met some of presentation's purposes and objectives. Gave incorrect or vague information. Gave a choppy presentation.	Met most of presentation's purposes and objectives. Gave partially correct and clear information. Gave a somewhat smooth presentation.	Fulfilled presentation's purposes and objectives. Gave correct and clear information. Gave a smooth presentation.	Fulfilled presentation's purposes and objectives. Gave correct and clear information. Gave a smooth presentation.	N.A. N.O.	
6. Participation - audience and student - feedback (nonverbal and verbal)	Encouraged minimal participation. Noticed, but uncertain how to handle audience's feedback.	Attempted to encourage participation. Noticed but made little attempt to handle audience's feedback.	Encouraged participation. Noticed and made good attempt to handle audience's feedback.	Encouraged participation. Noticed and handled audience's feedback well.	N.A. N.O.	
7. Instructional Aids - selection and development - number - size - clarity	Selected/developed inappropriate aids to meet objectives/ audience's background. Detracted from presentation (too many, too few, too confusing.)	Attempted to select/ develop aids to meet objectives/audience's background. Vaguely enhanced presentation (too few, too many.)	Good attempt to select and develop aids to meet objectives/audience's background. Enhanced presentation.	Good selection and imaginative development of aids to meet objectives and audience's background. Enhanced presentation.	N.A. N.O.	

Category	Behavior Indicators	Needs Improvement	Acceptable	N.A.	N.O.	Comments
8. Nonverbal Communication	a. Appearance reflected pride in self and served as a model for neatness, cleanliness, and being well-groomed.					
	b. Performance was presented in a well-balanced, courteous, poised, enthusiastic manner.					
	c. Presenter conveyed confidence, interest in subject, and a sense of humor.					
	d. Eye contact encompassed the entire audience.					
	e. Other					
9. Verbal Communication	a. Vocabulary, sentence structure, and illustrations used were appropriate for the audience.					
	b. Speech conveyed interest and enthusiasm; used appropriate emphasis.					
	c. Articulation and enunciation were clear and correct.					
	d. Voice had appropriate variety in rate, pitch, and volume.					
	e. New words, terms, or ideas were explained.					
	f. Other					

10. Additional Comments

Signature of Student: _____

Signature of Observer: _____

APPENDIX D

Table 8--Measures of item variance among panel members evaluating student presentations in standardized situations.

Rating Scale	Instrument Item Number	Category	Trials (%) 1	2	3	4
Four-Point Scale	1	Planning and Organization	.00	.11	.00	.36
	2	Introduction	.94	.00	.00	.61
	3	Body of Presentation	.78	.20	.25	.50
	4	Summary	.61	.25	.11	.11
	5	Overall Presentation	.45	.50	.11	.44
	6	Participation	1.03	1.00	.11	.78
	7	Instructional Aids	.28	.25	.28	.00
Dichotomous Scale	8	Non-Verbal Communication				
		a-Appearance	.00	.00	.00	.00
		b-Performance	.00	.25	.28	.11
		c-Confidence	.25	.25	.28	.25
		d-Eye Contact	.86	1.00	1.50	.78
	9	Verbal Communication				
		a-Vocabulary	.00	.00	.20	.25
		b-Speech	.11	.28	.28	.28
		c-Articulation	.25	.20	.25	.11
		d-Voice	.20	.20	.25	.25
		e-New Words	1.75	1.44	1.75	.28

VITA

Wanda L. Sterbenz was born in Wichita, Kansas on January 2, 1952. She attended public schools in Kansas and graduated from El Dorado High School, El Dorado, Kansas in May, 1970. She attended Kansas State University, was a member of the Coordinated Undergraduate Program in Dietetics in the College of Home Economics, and received her Bachelor of Science in Home Economics in May, 1974.

As a Registered Dietitian, her professional employment has included positions in Illinois, Kansas, and North Carolina.

Wanda completed her Master of Science degree in Food Systems Administration in the College of Home Economics from The University of Tennessee, Knoxville in June, 1979. She was the recipient of an Allied Health Traineeship in Administrative Dietetics.

The author is a member of The American Dietetic Association, the state and local dietetic associations, the Consultant Dietitians Practice Group in Health Care Facilities, and the Society for Nutrition Education.

Wanda is the daughter of Mr. and Mrs. W. V. Sterbenz of El Dorado, Kansas.