DOCUMENT RESUME

ED 239 557 HE 016 981

AUTHOR	Friedman, Miriam; And Others
TITLE	Validating Assessment Techniques in an Outcome-Centered Liberal Arts Curriculum: Valuing and Communications Generic Instruments. Final Report, Research Report Number One.
INSTITUTION	Alverno Coll., Milwaukee, Wis.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	80
GRANT	NIE-G-77-0058
NOTE	62p.: Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 1980). For related documents, see HE 016 980-990.
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	College Curriculum; College Students; Communication Skills; Competence; *Competency Based Education; Evaluation Methods; Higher Education; *Language Skills; *Liberal Arts; Moral Values; Outcomes of Education; *Research Methodology; *Student Evaluation; Test Validity; *Value Judgment
IDENTIFIERS	*Alverno College WI

ABSTRACT

The methodology for validating assessment techniques in a performance-based liberal arts curriculum was studied at Alverno College. Two generic instruments for assessing the competencies of "communications" and "valuing" were employed. A generic instrument is defined as one that assesses a competence level across content areas instead of using a large variety of instruments. The assessment system at the college requires students to demonstrate incremental gains while progressing through six sequential levels in each competency area. Twenty students were assessed with the generic communications instrument after 2 years of college; another 20 were assessed upon college entrance. Attention was focused on abilities in speaking, writing, listening, and reading, as well as self-assessments of performance in each mode. Eleven students were assessed with the generic valuing instrument after 2 years of college, while 20 were assessed upon college entrance. Value and moral judgments and decision-making were evaluated using written, oral, and group decision-making modes. Attention was focused on the validity of the assessment technique, along with the validity of the definition of competence. (SW)

VALIDATING ASSESSMENT TECHNIQUES IN AN OUTCOME-CENTERED LIBERAL ARTS CURRICULUM: VALUING AND COMMUNICATIONS GENERIC INSTRUMENTS

Miriam Friedman Marcia Mentkowski
Margaret Earley Georgine Loacker Mary Diez

Office of Research & Evaluation/Valuing Division/Communications Division
ALVERNO COLLEGE

FINAL REPORT TO THE NATIONAL INSTITUTE OF EDUCATION:
RESEARCH REPORT NUMBER ONE

Funded by a grant from the National Institute of Eoucation:
Careering After College: Establishing the Validity of Abilities
Learned in College for Later Success
(NIE-G-77-0058)

Principal Investigators:
Marcia Mentkowski
Austin Doherty
Alverno College
3401 South 39th Street
Milwaukee, Wisconsin 53215

An overview and rationale for our approach to the study of college outcomes, and a summary of the results from the following series of ten research reports, are found in:

Marcia Mentkowski and Austin Doherty. Careering After College: Establishing the Validity of Abilities Learned in College for Later Careering and Professional Performance. Final Report to the National Institute of Education: Overview and Summary. Milwaukee, WI: Alverno Productions, 1983.

Research Reports:

One: Friedman, M., Mentkowski, M., Earley, M., Loacker, G., & Diez, M. Validating Assessment Techniques in an Outcome-Centered Liberal Arts Curriculum: Valuing and Communications Generic Instrument, 1980.

Two: Friedman, M., Mentkowski, M., Deutsch, B., Shovar, M.N., & Allen, Z. Validating Assessment Techniques in an Outcome-Centered Liberal Arts Curriculum: Social Interaction Generic Instrument, 1982.

Three: Assessment Committee/Office of Research and Evaluation. Validating Assessment Techniques in an Outcome-Centered Liberal Arts Curriculum: Insights From the Evaluation and Revision Process, 1980.

Four: Assessment Committee/Office of Research and Evaluation. Validating Assessment Techniques in an Outcome-Centered Liberal Arts Curriculum: Integrated Competence Seminar, 1982.

Five: Assessment Committee/Office of Research and Evaluation. Validating Assessment Techniques in an Outcome-Centered Liberal Arts Curriculum: Six Performance Characteristics Rating, 1983.

Six: Mentkowski, M., & Strait, M. A Longitudinal Study of Student Change in Cognitive Development and Generic Abilities in an Outcome-Centered Liberal Arts Curriculum, 1983.

Seven: Much, N., & Mentkowski, M. Student Perspectives on Liberal Learning at Alverno College: Justifying Learning as Relevant to Performance in Personal and Professional Roles, 1982.

Eight: Mentkowski, M., Much, N., & Giencke-Holl, L. Careering After College: Perspectives on Lifelong Learning and Career Development, 1983.

Nine: Mentkowski, M., DeBack, V., Bishop, J., Allen, Z., & Blanton, B. Developing a Professional Competence Model for Nursing Education, 1980.

Ten: Mentkowski, M., O'Brien, K., McEachern, W., & Fowler, D. Developing a Professional Competence Model for Management Education, 1982.

ABSTRACT

Two studies test methodology for validating assessment techniques in a performance-based liberal arts curriculum. Alverno College has a system-wide performance based curriculum, with an assessment process that requires students to demonstrate incremental gains while progressing through six sequential levels in each of eight competences. The eight competences are integrated with the con[illegible] in each discipline. Students are required to attain each competence level in sequence to demonstrate cummulative achievement. The two studies assess the effects of instruction on patterns of student response using instruments created to ensure cross-college credentialing on the same instruments. Both instruments are "generic," that is, general criteria are integrated with criteria specific to the way the ability appears in the discipline in which the instrument is used. Studies of two generic instruments, assessing level 4 of the competences of Communications and Valuing are reported here.

Twenty students performed on the generic Communications instrument after two years in college; another twenty performed upon entrance to college. They demonstrated abilities in four modes of communication: speaking, writing, listening and reading, providing data on student performance across different modes of the same competence. The student is also asked to self-assess her performance in each mode on the same criteria on which she is judged by the assessor(s). Eleven students performed on the generic Valuing instrument after two years in college; another twenty performed upon entrance to college. Students demonstrated value and moral judgments and decision-making through written, oral and group decision-making modes. Students also self-assess their performance.

In the Valuing study, the instruction group performed significantly better than the no instruction group. Data from the instruction group provided support for the validity of the cumulative hierarchical nature of the competence. The no instruction group did not show any consistent cumulative or sequential patterns. Overall, the instruction group demonstated clusters of relationships among scores on the criteria and the no instruction group appeared to perform in a randomly scattered manner, indicating effectiveness of instruction. In the Communications study, students with no instruction demonstrated a wider range of variability in performance as compared to the instruction group, who showed a less dispersed pattern. Student performance varies with the mode of communication. The instruction group performed significantly better particularly on the upper levels of the four communication modes. The different patterns of the interrelationships of student performance across the four modes are seen in relation to the levels. Students who had instruction can better self-assess their performance.

The study methodology reflects our current pattern analysis approach rather than using score analysis, correlational analysis or an item analysis approach alone. The interpretation of the results and the methodology developed have implications for similar programs which are seeking out new methods to establish construct as well as content validity of complex assessment techniques used in performance-based curricula in higher education.

ACKNOWLEDGEMENTS

We acknowledge the following faculty who assisted in designing the study, collecting the data, scoring the instruments or interpreting the results:

Patricia Burns
Mary Diez
Austin Doherty
Margaret Earley
Rita Eisterhold
Rosemary Hufker
Anne Huston
Patricia Hutchings
Everett Kisinger
Gertrude Kramer
Georgine Loacker
Marcia Mentkowski
Marlene Neises
Jean Schafer
Christine Trimberger

Eunice Monroe assisted in the data management.

Paper presented at the Annual Meeting of the American Educational Research Association
Boston, April 1980.

VALIDATING ASSESSMENT TECHNIQUES IN AN OUTCOME-CENTERED LIBERAL ARTS CURRICULUM: VALUING AND COMMUNICATIONS GENERIC INSTRUMENTS

INTRODUCTION

Some liberal arts colleges have recently been responding to a growing concern for the adequacy of students' professional and career preparation by specifying the outcomes or abilities critical for future effective performance. These colleges have also taken the next step and created curricula to develop these abilities in each student in such a way that they can be expected to transfer to work settings after college.

Such "outcome-centered" colleges focus on assessing performance as well as knowledge as a key to bridging the gap between college and career. They have developed more nontraditional assessment techniques to capture both the learning and performance of these broad abilities to enable faculty to judge the extent to which these competences have b [illegible] ed

The purpose of this paper is to explore the issues re [illegible] validation of these more nontraditional assessment techniques, and to illustrate, empirically, some ways in which such validation studies may proceed. Validation of these techniques is particularly important since the learning that results from the use of performance-based assessment techniques are often an intrinsic part of the objectives and methods of competence-based curricula (King, 1979). Further, validation of assessment techniques can be a cornerstone in establishing the validity of the abilities learned in college for later careering (Mentkowski & Doherty, 1977, 1983).

The faculty of outcome-centered or competence-based programs are concerned with validity issues. They are concerned with the quality of the learning process, including the assessment techniques, and with the extent to which learning outcomes measured are the result of instruction (internal validity). They are also interested in how these outcomes measured by assessment techniques compare with what is possible for students to achieve—both in regard to outcomes credentialed and to the more "intangible" outcomes of college often thought to be related to future success. Further, colleges want to know, do the abilities learned in college impact graduates' future performance (external validity)? However, questions of external validity often follow questions regarding a program's internal validity. Thus, the reliability and validity of the techniques of assessment play a crucial role in any validity studies undertaken by outcome-centered colleges.

As much as researchers may be tempted to apply existing theoretical validation models to these assessment techniques <u>in toto</u>, the methodological constraints embedded in outcome-centered programs require nontraditional validation strategies. The unique characteristics of these programs impact the assumptions underlying commonly accepted methodological approaches for establishing validity. Clearly then, in order to establish the internal validity of assessment techniques and the constructs underlying them in a competence-based program, one needs to develop methods that are derived from the holistic, complex nature of outcome-centered curricula. Cronbach (1971) notes that investigations used for construct validation should be purposeful rather than haphazard; performance data should be interpreted within a given theoretical framework. Thus, nontraditional approaches to establishing the validity of assessment techniques—and the consequent internal validity of the program—demands an all-encompassing view of outcome-centered programs and instrument characteristics and requires rethinking and re-evaluating existing validation methods.

The measurement techniques employed in competence-based education are derived from the program characteristics. Gamson (1979) identified three important similarities among competence-based programs in higher education: (1) Educational outcomes reflect successful functioning in life roles; (2) Instructional time is independent of the achievement of educational outcomes; and (3) Certification of achievement of outcomes is reasonably objective and verifiable.

Three measurement implications can be derived from these similarities in program characteristics: (1) Measuring successful functioning in life roles calls for performance assessment techniques rather than paper and pencil tests;. (2) There are no absolute statem [illegible] made [illegible] validity of instruments or instruction at a certain point in time, because added instructional time subsequently alters students' performance; and (3) Assessment techniques are most often criterion-referenced, since students are credentialed or certified according to a specified set of criteria.

These general descriptions of program and measurement characteristics are a beginning for rethinking and re-evaluating existing instrument validation strategies. As we have worked to establish the validity of techniques in one su outcome-centered program, we have come to realize that we must also understand the specific theoretical framework underlying this program—even though there are some similarities to other competence-based programs. The way in which the faculty works together to develop instruments, curricula and program improvements, must also be taken into account. There is no substitute for "trying out" various methods for validating instruments, and then presenting the results to faculty, who ask questions of clarification, suggest directions, and spell out "what they want to know" about their instruments.

For us, developing a conceptual framework for the validation of assessment techniques in an outcome-centered curriculum demanded that it be applicable across competences, disciplines, and instruments. Our validation model wrs derived from the following sources:

- The competence conceptual model defined by faculty
- The assessmer [illegible] igned to measur [illegible] students' performance
- The character [illegible] s of the [illegible] essment techniques
- Ongoi [illegible] validation studies submitted to the faculty for critique
- Faculty questions

The following sections provide a brief glimpse of each of the above named sources, and results in a description of the questions that guided the validation strategies used to validate two instruments--the "empirical illustrations" that follow.

Competence Conceptual Model

The theoretical and pedagogical framework of eight competences as defined by Alverno faculty constituted our frame of reference.[1] The Alverno curriculum centers on student competence as outcomes. Students

[1] By framework we refer to the competence conceptual model as outlined in Liberal Learning at Alverno College, 1976.

are required to demonstrate mastery of eight competences:

—Effective communications ability

—Analytical capability

—Problem solving ability

—Valuing in a decision making context

—Effective social interaction

—Effectivenes in individual/environment relationships

—Responsible involvement in the contemporary world

—Aesthetic responsiveness

The conceptual framework underlying the competences is defined as Generic, Developmental and Holistic. The assessment techniques are created to follow these concepts. Consequently, faculty design instruments according to the following questions:

—To what extent can the student demonstrate the same ability in a variety of settings (Generic)?

—To what extent does e student demonstrate a progressive learning pattern (Developmental)?

—To what extent does the student demonstrate integration of competences in a single performance given that the competences are inseparable parts of the whole person (Holistic)?

Assessment System

The Alverno assessment system requires students to demonstrate incremental gains while progressing through six sequential levels in each of the eight competences. Students are required to attain competence levels in sequence and to demonstrate cumulative mastery. Multiple assessors, multiple contexts and multiple modes of assessment add to the quality assurance of the assessment process. The competences are integrated with the concepts within various disciplines.

Characteristics of Assessment Techniques

Since the assessment system requires a wide network of assessment techniques, faculty members individually and jointly design, evaluate and revise instruments[1] within disciplines, departments and interdisciplinary competence divisions. The instruments are designed to:

—measure the learning objectives for the competence level

—elicit the full nature of the ability

—provide opportunities to integrate content and competence at an appropriate level of sophistication

—measure the integration of the competence with other relevant competences

—use a production task rather than a recognition task

—use an assessment mode similar to the ability as usually expressed rather than an artificial mode

—allow for the judgment of performance against public and explicit criteria, by the assessor(s) and by the student in a self-assessment

—allow for administration external to the learning situation

—provide diagnostic, structured feedback to the student on her strengths and weaknesses

—provide evidence for credentialing student's performance

Ongoing Validation Studies

The process of conducting separate, independent validation studies led us toward development of a more general validation framework applicable across competences, disciplines and instruments. Findings from validation studies establish new frameworks for subsequent studies by broadening the scope and generating a "pool" of validation methods.

[1] By instrument we mean the set of criteria employed to evaluate student performance while reacting to a specific stimulus.

Faculty Questions

Our attempts to modify existing validation methods or to create new ones was characterized by a field approach. Efforts to develop validation strategies were actually directed in large part by faculty questions that arose during faculty involvement with (1) the process of refining and revising the instruments, (2) responding to results from initial ongoing studies, and (3) brainstorming questions in specially designed group sessions. Some of the questions generated were:

—Are the instruments created to assess what we want to assess?

—Do students perform successfully in the assessment process as the result of learning experiences they complete?

—Is student's performance on a particular technique following instruction a true representation of what she has learned and can do?

—How do the competences differ and how do you get at the difference?

—How best can we use group data from student performance on our instruments to test out assumptions about the complex nature of a given competence?

—How best do we assess whether the competence levels are truly of sequential complexity?

—How can we best describe the patterns of a student's performance across time as she progresses through the competence levels?

—How do we describe or chart increasingly complex gains in student's performance?

These and other similar questions indicated that the faculty are interested in the validity of the instrument criteria and the extent to which the instruments measure the effectiveness of instruction. Faculty are even more interested in the construct validity of each of the eight competences. They wish to achieve greater insight into the underlying meaning of each of the eight dimensions and the developmental, cumulative nature of the competences. These faculty questions provided direction for the creation of our validation framework and confirmed our earlier objectives for establishing the internal validity of the competences and assessment techniques (Mentkowski & Doherty, 1977):

- Establish the validity of the techniques used to assess students' behavioral performance of the competences by adapting or developing validation strategies appropriate for use with nontraditional assessment techniques;
- Compare student performance across and within competences to further refine the nature of the competences and their interrelationships; and
Examine the relationships between student performance and external criterion measures.

The Empirical Illustrations

The two empirical studies described in this paper respond to these faculty concerns for validity. We first establish the validity of the assessment technique by demonstrating the effects of instruction on different student response in an instructed, uninstructed group comparison which also allows for criteria evaluation. Then we establish the construct validity of the meaning and developmental, cumulative nature of a competence to contribute to further exploration of the underlying meaning of the ability.

Establishing the Validity of the Assessment Technique

Establishing the validity of an assessment technique means conducting [illegible] and determining the extent to which the instrument actually measures what it intends to measure, that is, the effects of instruction.

[illegible] questions that guide such analyses are:

- Does the instrument discriminate between the performance of an instructed and uninstructed group of students?
- To what extent do the criteria discriminate between instructed and uninstructed students in the same direction as does the students' complete performance?

3. Is variability in student performance within the instructed group different from variability within the uninstructed group? What are the directions of such differences within a single competence level? Across competence levels? Can we specify a desired variability pattern given the competence learning objectives?

Establishing the Validity of the Meaning and Developmental Nature of the Competence

Then we establish the construct validity of the meaning and developmental, cumulative nature of the competence. A competence is defined as an ability that can be broken open into several components and specified developmentally[1] at each of six sequential levels. For us, establishing the construct validity of a competence means verifying the expert judgment or interpretation[2] of student performance against the competence as defined by faculty, and exploring the meaning of the competence in light of the empirical data. This means we must investigate the relationship among competence criteria before and after instruction and examine the extent to which the competence definition can account for all the demonstrated behavior.

The questions that guide the attempt to establish construct validity are derived from the generic, developmental and holistic nature of competence definition:

[1] We use the word developmental to imply sequential levels of an ability so specified for pedagogical reasons. They are not cognitive-developmental "stages."

[2] The majority of the assessment techniques employed in the college yield inferential data generated by expert judgment. Within the Alverno learning process, irrespective of where students are assessed, their proficiency at a competence level is evaluated by faculty who share the same understanding of the meaning of a given competence and use similar criteria. Even off-campus assessors participate in training workshops and adopt Alverno's competence definition and criteria as a basis for their judgment. In establishing construct validity, we were actually attempting to establish the validity of assessors' interpretation of student performance against a set of competence criteria. As Messick (1975) notes, our task is to validate not a test but an interpretation of data arising from a specific procedure.

1. Can we identify improved associations among separate components of a competence within an instructed group of students as compared to an uninstructed group?

2. Can we identify definite patterns of response in an instructed group of students which are different from patterns of response in an uninstructed group?

3. Do clusters of performance form the unidimensional ability specified in the competence definition?

4. To what extent can we attribute differences between instructed and uninstructed students to the sequential or cumulative nature of the competence levels?

5. Does the attainment of one component of a competence facilitate attainment of another component?

6. What are the prerequisite skills needed to acquire new abilities in a given competence?

Developing Validation Strategies

Because our evolving validation model is derived from the internal framework of the program characteristics, we pay close attention to the corresponding implications for measurement we outlined earlier (see p. 3):

1) measuring performance rather than paper and pencil tests,

2) measuring student's progress as a function of instructional time, and

3) criterion referenced measurement techniques.

Alverno faculty measure performance in action rather than just on paper and pencil tests; they measure a student's progress as a function of instructional time, and they use criterion referenced measurement techniques.

Student performance on assessment techniques is examined by a variety of strategies. We establish the validity of assessors' interpretation of student performance against the competence criteria or definition, and then establish inter-rater reliability of assessor judgments. We investigate patterns of performance, their differences and similarities, within and between contrasted groups (instructed and uninstructed). We also attempt to establish the validity of the sequential, developmental and cumulative nature of the competence levels. As for the time dimension, i.e. the rate of competence attainment, our validation strategies focus on the range of individual differences, direction in variability changes, magnitude of instructional effect, reduction in student variability while progressing upward on the mastery continuum, and establishing baseline for entering students. The use of criterion referenced assessment techniques direct us toward an analysis of relationships among criteria, criteria evaluation and identifying possible cutoff points for credentialing.

In our view, a valid assessment technique in a competence-based framework will show evidence of reduced variability in the instructed group. Since mastery for Alverno students is not viewed as an all-or-none outcome but as a continual process, we consider an instrument at least partially valid if the instructed group shows a decrease in variability. An instructed group should perform significantly better on the entire instrument, as well as on the individual competence criteria. If the magnitude of the instructional intervention accounts for at least 25% of the variation in the instructed group. we accept this as evidence that improved performance is not due only to individual differences, but also to the effectiveness of the learning experiences.

We expect that instruction will improve associations among components of a given competence whereas an uninstructed group will show weak associations. Improved associations should form clusters of abilities which will conform to the definition of the competence. We also expect that the instrument is measuring the unidimensional abilities of a single competence. Is the instrument valid if it measures other abilities as well? We arrived at the term "improved associations" while selecting a construct validation technique that identifies relationships among variables, i.e. the behavioral manifestation of the ability under study (Payne, 1975, p. 113). In a comparative analysis (instructed vs. uninstructed group) patterns of relationships can then be identified separately within the two groups, and then compared.

Let us suppose that the components of analytic ability are to observe, to make logical inferences and to draw relationships. Faculty who wish to educate toward those components or skills design a curriculum which will enhance the development of analytic abilities. Learning objectives can then be verified against actual student performance. Students who complete the learning sequence are expected to demonstrate "improved associations" among variables which measure observation skills and the ability to make logical inferences and to draw relationships, whereas students who just entered the program will demonstrate random associations among the skills under study. Furthermore, the instructed group will show a clustering of the three components, forming the analysis competence.

Since the competences are also defined developmentally in a pedagogical competence model, we expect that the uninstructed group will not demonstrate a coherent sequence and will deviate from the one specified. If a competence is developmental, instructed students will demonstrate the prescribed sequence of the competence levels and cumulative mastery.

By employing a multi-analysis approach within the contrasted instructed and uninstructed groups, we departed from existing construct validity methods such as factor analysis, convergent and discriminant validity methods. We preferred a pattern analysis approach within two contrasted groups and explored the differences and similarities in patterns.

The strategies employed for validating assessment techniques, and the competence definition and its developmental nature were characterized by a reciprocal process in which preliminary results from validation studies were communicated to Competence Divisions (analogous to Discipline Divisions in their function) who in turn expand or generate the questions that further direct the analysis. Such a field approach provides prompt feedback to faculty for instrument revision and ensures the researchers' sensitivity to the faculty's internal frame of reference.

We wish to emphasize that the purpose of the present study is to create a validation model rather than to report results. Since the two empirical illustrations involve small numbers of students we have chosen to emphasize the method by which we analyzed and interpreted the data rather than the actual outcome or results. If our methods prove effective in validating behavioral data within a competence-based program, we would suggest further tests of our validation strategies with larger samples.

The Instruments

Since the following empirical studies are part of our continuing efforts to develop a validation model applicable across instruments irrespective of the competence or discipline they intend to measure, we selected generic instruments at level 4 for our empirical illustrations. Generic instruments are designed to ensure cross-college credentialing on the same instruments instead of using a variety of instruments for the same purpose.[1]

Briefly, a generic instrument is one which assesses a competence level across content areas instead of using a large variety of instruments, each of which must be validated. In a generic instrument, general criteria can be integrated with criteria specific to the way the ability appears in the

[1] After approximately two years in college, students contract for credentialing at level 4 of a given competence.

discipline or content area in which the instrument is used.

Several Competence Divisions have created generic assessment techniques to assess for their respective competences at level 4. This represents a step toward increased consistency in assessment at the level where general education requirements are certified. It also allows greater comparability across disciplines as we evaluate our assessment techniques, and provides a more uniform data base for comparison with external criterion measures and for longitudinal studies. Validation studies on two of these instruments, the Valuing generic instrument and the Communications generic instrument, were conducted for the purposes of this study.

The Valuing in Decision Making competence focuses on developing the student's ability to use a valuing process with a number of components including discerning value and moral issues and resolving value and moral conflicts.[1] The instrument is designed to elicit value and moral judgments and decision making through written, oral and group decision-making modes.

The Communications competence focuses on the process of effective clarification and involvement between a presenter and an audience. Students performing on the generic Communications instrument after two years in college demonstrate abilities in four components or modes of the Communications competence: Speaking, Writing, Listening and Reading. Their performance provides data across different modes of the same competence.

The Valuing Division was seeking to validate the pilot administration of the Valuing generic instrument. At that juncture they were concerned with a variety of validity issues: How well does the instrument measure

[1] For a description of how Valuing is taught and assessed at Alverno, refer to: M. Earley, M. Mentkowski and J. Schafer, Valuing at Alverno: The Valuing Process in Liberal Education (Milwaukee: Alverno Productions, 1980).

the effects of instruction? How well does it measure competence levels? Does it discriminate between a group of instructed and uninstructed students? Preliminary results from an initial study (Valuing Generic Instrument: Study A) were reported to the Valuing Division who then generated further questions which focused on evaluation of the instrument criteria, and the developmental, sequential and cumulative nature of the Valuing competence. The researchers then proceeded to respond to these newly defined faculty interests through a broader scope of validation strategies. The subsequent analysis explored variability of students' performance, distribution of scores within the separate competence levels, and the sequential, cumulative nature of the competence by way of correlation matrices.

The Communications study benefited from the insights learned from the Valuing study. We also had a better understanding of faculty concerns, probably because communications has always been explicitly taught in college. Although the Communications study began with a more narrow perspective appropriate for purposes of instrument revision and criteria evaluation, the scope and the range of issues again broadened, directed at the attempts to validate the competence model. It is this part of the study, we believe, which contributes the most to validation methodology in competence-based programs in higher education. The multi-analysis approach employed within two contrasted groups provided insight into differences as well as similarities in patterns of student performance on the Communications competence. The analysis techniques selected are more commonly used in other areas of the social sciences. Yet they yielded a greater understanding of the relationship between competence criteria and students' actual performance.

VALUING GENERIC INSTRUMENT: STUDY A

Method

Subjects

The generic instrument assessing the Valuing competence, level 4 was administered in January, 1978 to a group of new students entering Weekend College[1] (WEC) who had no previous instruction at Alverno College (uninstructed group). These 20 students were randomly selected from all new students entering WEC Semester II (n = 60). During Spring, 1978, the same generic instrument was administered to 11 Weekday College students contracted for an assessment of Valuing, level 4 as part of their learning sequence (instructed group). Level 4 is usually achieved at the end of ~~the~~ general educational sequence after two years in college.

Design

We selected students from Weekend College for this comparison in order to control for the effects of maturation. Since the Valuing competence

[1] In Fall, 1977, Alverno College instituted a Weekend College. The Weekend College is an opportunity to earn a four year college degree by going to college every other weekend from late August through May. It was planned for women of all ages who wish to earn a college degree, but are unable to attend weekday and evening classes. Classes involve intensive study, a close working relationship with instructors and fellow students, and maximum opportunity for self-directed study. A semester of Weekend College is equivalent to a semester of Weekday College. The scheduling of courses within a limited time frame and the resulting intensification and concentration of study distinguish Weekend from Weekday College. Bachelor programs are available in the major areas of communications, management, and nursing. All students take courses in liberal arts, which are designed to complement the major and provide a breadth of knowledge. Because of the intensive nature of Weekend College, it is necessary for students to function as self-directed learners. An introductory course designed to provide students with the independent learning skills they need is, therefore, a required part of the curriculum. Currently, approximately 500 women are enrolled in the Weekend College. Median age is 33 years. About 90% of these women currently hold full-time jobs in the Milwaukee area.

can be expected to have a cognitive-developmental component, we felt it wise to select a group of students who could be expected to have developed the Valuing ability to some extent even though they had not had college instruction. (Median age for the WEC is 33; WDC median age is 22 years.) Further, a true pre- and post-instruction comparison is not generally feasible because constant changes in the curriculum and instruments preclude giving the same instrument to the same students before and after instruction (two years would separate the two administrations of the instrument). We were also interested in comparing two groups that are most dissimilar given the type of construct validity we were using.

Procedure

Students who have had instruction "participate" in the validation study as part of the assessment process. How did we motivate new students in the uninstructed group on the second day of their college career to take these instruments? We were concerned that asking them to "take tests" would increase anxiety and influence the results. We tried to resolve this problem by presenting a rationale to the uninstructed group. The Office of Evaluation administered the Learning Style Inventory[1] to the group on a Friday night and provided feedback on Sunday just before they were involved in the validation study. In a talk to the students, the Director of Evaluation labeled the instruments "practice assessments" and called attention to the positive outcomes of participation. She suggested that taking a practice assessment would assist them by answering the following questions usually raised by new students:

[1] David A. Kolb, Learning Style Inventory: Self Scoring Test and Interpretation Booklet (Boston: McBer & Company, 1976).

- What can I learn about myself that will assist me in becoming a better learner (e.g., feedback on the Learning Style Inventory)?
- What is meant by a "competence"?
- What is meant by "demonstrating a competence"?
- What are assessments like here at Alverno?
- What are my initial capabilities?
- Would I be as competent if I hadn't come to college?

Since the uninstructed group already had feedback on the Learning Style Inventory, we did not feel we had to provide feedback on their performance, a usual college procedure on any assessment. Giving students individual feedback on assessments used for research purposes would overburden the assessment system. We observed that the uninstructed group did seem to take their performance on the instruments seriously.

Instrument

The following paragraphs describe the Valuing generic instrument:

> "In 1977, the Valuing Division followed the suggestion of the Assessment Committee (charged with the college-wide evaluation, revision, and validation of assessment instruments) to develop a "generic" instrument that would examine student performance across curriculum levels of Valuing. The result was an additional instrument—a "generic instrument"—to assess levels 1 through 4 which all students would demonstrate at the end of their general education sequence. Because its content and setting were external to any of the student's course experiences, this instrument was expected to provide a summative assessment of her development in valuing to this point. This generic instrument reiterates every one of the criteria by which the student's several instructors have assessed her developing ability up to level 4. Yet it applies them as part of a tool which is in no way dependent upon the specific assessments or courses she has taken.
>
> Space does not permit more than a brief description of this generic instrument. It consists of four parts that ask the student (1) to infer values from a literary work; (2) to analyze the relationship of values to scientific and technological developments; (3) to participate in a moral dilemma group discussion; and (4) to analyze her own decision-making process.

Various sets of stimuli can be developed for the ins rument, reflecting a range of issues. One such set involves students in the issue of genetic engineering—using a short story, newspaper article and an article from a scientific journal, a moral dilemma, and directions for her response to each. She is first asked to compare the values she infers from the short story to her own value system, and then to that of American society. She then writes an editorial for either the local newspaper or a scientific journal on 'How our decisions regarding scieitific developments influence our value systems, cause value conflict, and raise questions regarding the relationship between private decision-making and public policy.' She next participates in the facilitator-led small group discussion of a moral dilemma, and then analyzes her own decision-making process throughout the experience and writes a letter to a congressman on genetic screening 'stating her case, describing her action plan and relating how her own values motivated her decision.'

The student's performance is measured according to 67 criteria in all:

29 of these repeat the faculty's criteria from levels 1 through 4 on which she has already been wholly or partly credentialed;

21 were developed by the Valuing Division for the student's self-assessment on the moral discussion; and

17 were developed by the Division for instructor assessment of the student's participation in the moral discussion and for tallying the occurrence of her use of the various modes of judgment, her identifying of moral issues and moral orientations categorized in Kohlberg et al., Standard Form Scoring Manual (1978).

The student is credentialed on the 29 'level' criteria, and so these criteria were submitted for validation. The 17 criteria for judging her performance in the discussion also help form a basis for credentialing judgments on some of those 29 criteria, such as 'Recognizes necessity for and utilizes information and knowledge in moral reasoning, judging and deciding' or 'Articulates the point of view of another person or position with empathy and reason.'[1]

What has been created in this 'generic' instrument is an opportunity to elicit and examine the moral reasoning of college students in several situations, to view and analyze their participation in a moral dilemma discussion and to judge the discussion's effectiveness." (Mentkowski, 1980, pp. 42-44).

No cutoff points for credentialing had been specified for this instrument

[1] Thus there is some basis for generalizing from the results on the 29 criteria to the 17 criteria. The 21 criteria by which the student assesses herself were not included in this validation study.

since one purpose of this study was to assist faculty to identify the range of student performance and to provide data for instrument revision. Some of the 29 criteria assess only one level, some all four levels, and some more than one level.

One member of the Valuing Division evaluated both groups to insure consistency in the scoring procedures. While reviewing a student's respons to the stimuli the scorer was looking for evidence in the student's work which would meet each criterion. When such evidence was identified, the criterion was checked. The more criteria checked, the higher the student's score. Thus, by score, we actually mean the number of checked criteria.

Students' checkmarks on each of the criteria were tabulated separately for the instructed and uninstructed group providing frequency of response per criterion within each group, each student's total score, and the total number of level 4 checkmarks per student.

Results

The first analysis asked:

- How does the performance of the instructed group compare to the performance of the uninstructed group on each criterion?

Table 1 shows the frequency of student responses (checked criteria) to each of the 29 criteria. Frequency of response is reported in percentages. If the uninstructed group performed better or equal to the instructed group, the criterion is labeled "non-discriminative." If the instructed group performed better than the uninstructed group, the criterion is labeled "discriminative" and starred (*).

At this preliminary stage, Table 1 was created to provide a descriptive analysis only, rather than a statistical analysis.[1] The extent to which the

[1] In the following Communications study, a statistical analysis was employed to identify discriminative items.

Table 1

Criterion Response Frequency (Percent) for
Instructed and Uninstructed Students

Criteria and Levels Assessed

Groups										
	LEVEL 1									
	1	2*	3*	4	5*	6*	7*	16*		
Uninstructed	100	45	15	100	30	75	55	35		
Instructed	100	100	27	100	81	100	72	63		
	LEVELS 2 & 3									
	17	19*	20	21*	22*	24*				
Uninstructed	15	5	35	0	0	35				
Instructed	9	18	36	9	27	54				
	LEVEL 4									
	8*	9	10*	11*	12*	13*	14*	15*	18	23*
Uninstructed	5	0	25	30	5	25	15	30	80	5
Instructed	18	0	90	54	27	27	54	36	72	9
	LEVELS 3 & 4									
	27*	28*	29*							
Uninstructed	0	25	10							
Instructed	9	63	18							
	LEVELS 1, 2, 3, & 4									
	25	26								
Uninstructed	50	20								
Instructed	0	0								

NOTE: Discriminative criteria are starred.

[illegible]

[illegible] than the main treated group.

[illegible] compared, but the instructed [illegible] significantly better than the noninstructed group.

[illegible]

[illegible] the instructed group was [illegible]

[illegible] Table [illegible] presents the [illegible]

[illegible]

Table [illegible]

[illegible] Means and Standard Deviations [illegible]
[illegible] Instructed Students

Group	[illegible]	Total

[illegible]	M = [illegible]
	SD = [illegible]
		M = [illegible]
		SD = [illegible]

Instructed [illegible]	M = [illegible]
	SD = [illegible]
		M = [illegible]
		SD = 2.9

The instructed group performed significantly better on the [illegible] scores [illegible] and on the total score as well (t [illegible] p < .01).

VALUING GENERIC INSTRUMENT: STUDY B

Results

Study B was conducted in response to questions raised by the Valuing Division when the preliminary findings of criterion frequency response and differences in performance between the two groups, as described in Study A, were presented to the Division.

Some questions raised concerned the future revision of the instrument:

- How do criteria that assess the Valuing process compare with those that assess the content of the discipline?

- Which criteria assess student responses that are overtly elicited by the instrument stimulus and which assess responses that are covertly elicited?

- To what extent should the number of instances of supportive evidence in the student's work affect our decision to credential student performance.

- Should we assign different weights to different criteria? How would this impact our credentialing of students?

Faculty also generated questions that went beyond concerns for the content validity of the instrument, and our study responded to these:

- Are we concerned with 100% mastery at all levels? At level 4? Or rather, are we concerned with reducing variability of student performance (mastery students should fal into a narrow performance range)?

- Are the competence levels sequential? If a student attained level 4 and not levels 2 and 3, what does that tell us? Are level 2 and 3 criteria clear enough for instructional purposes? How are levels 2 and 3 linked to the developmental sequence?

The first analysis asked:

- To what extent does reduction in variability of student performance imply effectiveness of instruction?

It was decided to further investigate variation in total score performance of the instructed group compared to the uninstructed group. The uninstructed group responses formed a normal distribution whereas the instructed group formed a negatively skewed distribution where most of the students fell at the mean range or above (Figure 1). The omega squared statistic[1] (Hays, 1973), which estimates the amount of statistical association implied by the obtained difference between means ($w^2 = .22$), suggests that 22% of the instructed group total score variation (levels 1 through 4) was due to instruction.

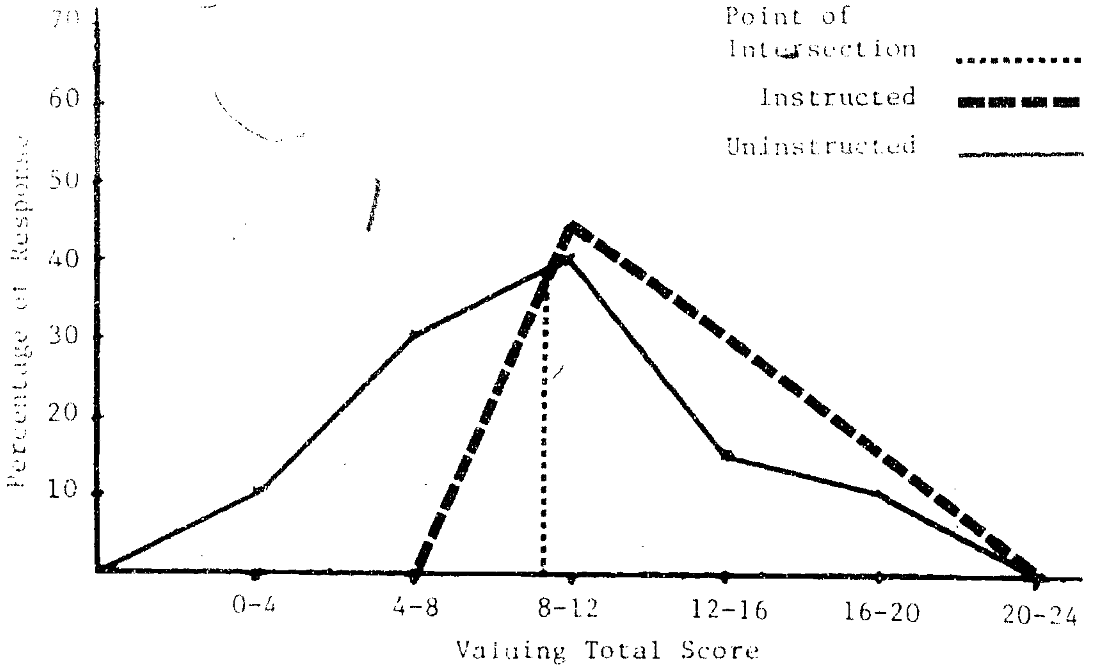

Figure 1. Frequency polygons of Valuing total scores for instructed and uninstructed students.

Note: Point of intersection can ~~be~~ considered as an optimal cutoff point for an acceptable level of performance (Berk, 1976).

The level 4 score distribution was then examined. Figure 2 shows that the uninstructed group displayed a positively skewed distribution whereas the instructed group formed a normal distribution. The small amount of overlap between the distributions of the two groups (Figure 2)

[1] $$w^2 = \frac{t^2 - 1}{t^2 + N_1 + N_2 - 1}$$

shows that instructional intervention is effective in differentiating between the two groups, and is more effective at level 4 than in all four levels combined.

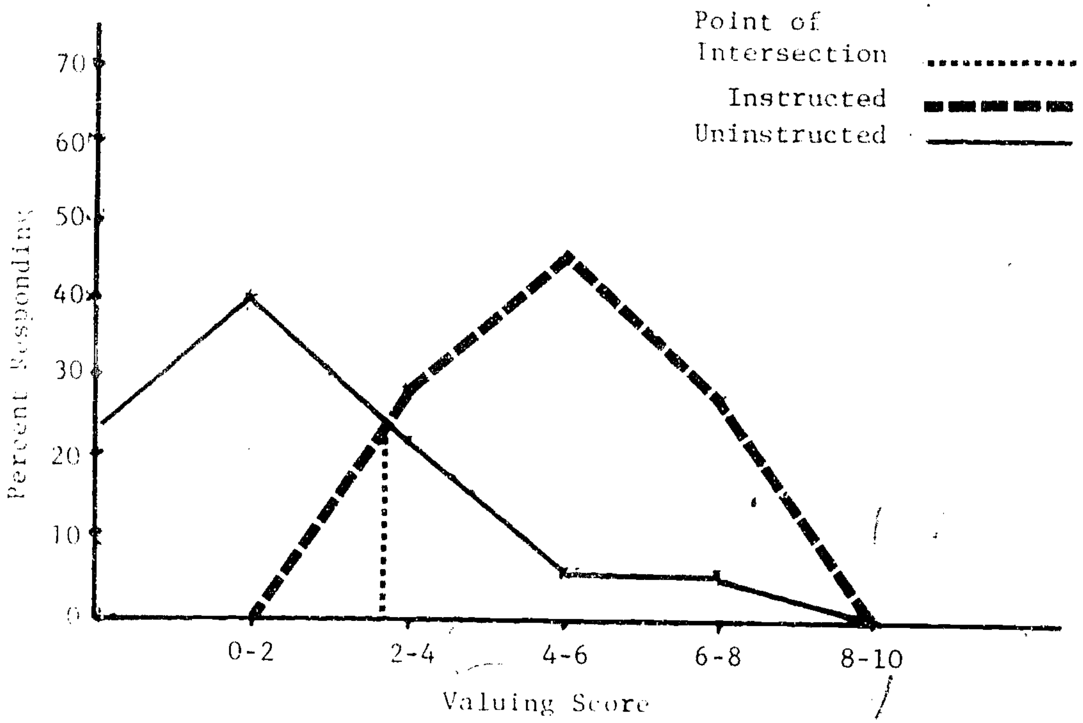

Figure 2. Frequency polygons of Valuing Level 4 scores for instructed and uninstructed students.

Note: Point of intersection can be considered as an optimal cutoff point for an acceptable level of performance (Berk, 1976).

Omega squared statistic computed on level 4 scores ($w^2 = .37$) estimated that 37% of the variation in the instructed group may be accounted for by the instructional treatment. Variability within the instructed group at level 4 (SD = 1.6) was smaller compared to the uninstructed group (SD = 1.9). When the mean scores from levels 1 through 3 were examined, no significant differences were obtained between the two groups.

The second analysis asked:

- Are the competence levels sequential, developmental and cumulative?

One correlation matrix was generated from the responses of the instructed group and another from the uninstructed group to assist in investigating the following questions:

- To what extent are the 4 levels of the Valuing competence, assessed by the generic instrument, sequential? Our assumption is that criteria at one competence level will intercorrelate more highly with each other than with criteria at different competence levels.
- To what extent do the criteria reflect a cumulative, developmental sequence? A particular competence level should correlate more highly with preceding levels than with the next level in the sequence. For example, there should be a higher correlation between level 3 and levels 1 and 2 and a lower correlation between level 3 and level 4. We assume that a student who has reached level 3 has also mastered the first two levels.

In the instructed group, the correlation matrix did not show clusters of higher intercorrelations among the criteria at each level.[1] This matrix did not support the sequentiality of levels 1 to 4. However, there is some evidence for the cumulative nature of the levels, because criteria at the higher levels tend to form clusters of intercorrelations with lower level criteria but not with upper level criteria. For example, students who responded to level 4 criteria tended to respond to levels 1, 2 and 3 as well.

The correlation matrix from the uninstructed group did not show consistent patterns of correlations that would support the sequential or cumulative nature of the competence.

Discussion

Presenting the preliminary criteria evaluation and the results supporting the overall validity of the instrument in measuring instruction to the Valuing Division in a simple descriptive manner proved to be effective in stimulating a group of interdisciplinary liberal arts faculty

[1] Correlation matrices are available from the authors.

to become involved with instrument validation. They generated questions which directed additional analyses. The descriptive analysis of the criteria's internal consistency was effective in identifying criteria with potential "problems." When faculty started to explore alternative reasons for why these criteria were non-discriminative, the broad nature of the competence itself was discussed, and faculty raised issues related to construct validity.

The graphic presentation of the frequency distribution of scores dramatically demonstrated that the instrument was effective in discriminating between the instructed and uninstructed groups (presumably measuring the effects of instruction), providing a powerful motivator for faculty to continue instrument validation.

The analysis indicated that Weekend students (median age, 33) enter college with similar Valuing abilities compared to Weekday students (median age, 22) at levels 1 through 3 (no significant differences were found when levels 1 to 3 scores were compared between the two groups). In contrast, the instructed group made a successful leap to level 4. Instruction appears to be effective in reducing instructed students' variation at level 4 as compared to the uninstructed group and brings instructed students closer together on the mastery continuum. The magnitude of the instructional effect accounted for 37% of the variance at level 4 within the instructed group.

The correlation matrices did not support the sequence of the competence levels. But a correlation matrix based on 11 students is hardly a basis for generating conclusions. The cumulative nature of the competence levels was more apparent in the performance of the instructed group, than the uninstructed group, and so receives some support.

The very fact that older, more experienced women (WEC) performed lower on level 4 criteria illustrates that instruction was effective for the younger Weekday students and not due entirely to maturation. The lack of significant differences between the instructed and uninstructed groups at levels 1, 2 and 3 provided a stimulus to the Valuing Division members who then rewrote and clarified criteria at levels 1 through 4. Members are currently meeting with faculty who teach Valuing, to introduce them to the revised criteria. Further, Division members concluded that older, more experienced women were not as likely to need in-depth instruction on the awareness of their values (level 1), but that there is a "leap" in the Valuing ability that is enabled by college instruction. Valuing Division members currently question the extent to which levels 2 and 3 are actually sequential in nature as they are currently defined.

The small sample size did not allow us to form definite conclusions in response to the questions raised. The study did allow us to test out a process for validating instruments incorporating faculty questions and feedback, and to try out various methods. The study outcomes stimulated further testing of validation strategies which helped us to build a conceptual framework for validating assessment techniques applicable across the various competences and disciplines. A similar study on a larger group of students will provide a basis for possible cutoff points for accepted levels of performance and also yield more information about entry or baseline performance.

COMMUNICATIONS GENERIC INSTRUMENT

Method

Subjects

The generic instrument assessing the Communications competence, level 4, was administered in January, 1978 to a group of 20 new students entering Weekend College (WEC)[1] who had no previous instruction at Alverno College (uninstructed group). During Spring, 1978, the same generic instrument was administered to 20 Weekday College students (WDC) contracted for an assessment of Communications, level 4 as part of their learning sequence (instructed group). Level 4 is usually achieved at the end of the general education sequence after two years in college.

A group of uninstructed students was selected from all WEC entering students (n = 60) who had some previous course work in a content area because the Communications instrument at level 4 demands that performance of the competence be integrated with content. The most frequent common content base was a previous course or two in psychology. Twenty students were then randomly selected from the group who had some psychology to take the Communications instruments (uninstructed group). The instruments were administered to those in the instructed group with a comparable psychology background.

[1] Weekend College students are generally older experienced women who hold a full-time job during the weekdays and come to the college full time in a weekend time frame that allows achievement of a degree in four years. The Weekday College students are generally younger full-time students, most of whom enter college after graduating from high school.

Design

We selected students from Weekend College for this comparison, since the Communications competence can be expected to develop to some extent due to life experience. We felt it wise to select a group of students who could be expected to have developed the Communications ability without formal instruction after high school.

Further, a true pre- and post-instruction comparison is not generally feasible because constant changes in the curriculum and instruments preclude giving the same instrument before (beginning of the first year) and after (end of the second year) instruction. For the type of construct validity we were employing, we were interested in comparing two groups that are dissimilar.

Instrument

The generic instrument designed by the Communications Division to assess effective Communications as an outcome of general education integrates several modes of communication. It assesses Writing, Speaking, use of Media and analytic Reading and Listening from college-entry level to the summative performance level which represents the completion of general education.

This instrument involves content—though not necessarily a specific course—because it assesses the communication of concepts related to an academic discipline or a comparable area of study. In the form of the instrument that is administered to instructed and uninstructed groups in this study, the content is psychology. However, the format and criteria are sufficiently generic to permit substitution of different content with relative ease. In effect, the instrument provides a criterion measure that is external to courses. The Communications Division set criteria and provide assessors. The criteria in the generic instrument are the same as those by which the student is assessed for Communications in all of her courses.

Specifically the instrument consists of four parts: (1) Directions to prepare and actually give a speech (including use of a visual); (2) Directions to write a letter; (3) An article to read; and (4) A taped lecture to listen to. A letter provides the initial stimulus and establishes the setting and context. In addition to answering a series of open-ended questions to analyze an article and a lecture, the student is required to take new information from these two sources and integrate it into her present understanding of the concept involved. (In the form of the instrument used for this study, the concept is "the influence of an infant's environment on human development.") The student is also required to assess herself in each of the Communications modes involved.

The student's performance is measured by a total of 64 criteria:

19 assess Writing performance

27 assess Speaking, including Media performance

9 assess Reading performance

9 assess Listening performance

The Communications battery is a generic assessment technique used to credential students at level 4 of the Communications competence. Since Alverno faculty view the Communications competence as a developmental, pedagogical sequence, competence levels 1 through 4 are cumulative and sequential. Students who wish to be credentialed at level 4 must again demonstrate satisfactory performance on the three preceding levels for which they have already been credentialed. Thus each of the four exercises is divided into four hierarchical levels. The generic assessment technique is criterion referenced since each competence level states explicitly the behavioral criteria for satisfactory performance. The student does not respond directly to the instrument criteria, but reacts to stimuli designed for each of the four communication modes and her performance is judged by faculty who evaluate her demonstrated behavior against the defined criteria. Once the assessor finds evidence in the student behavior which meets the criterion, the criterion is checked. The student has mastered the competence level if all criteria specified for that level are checked. She will be credentialed for level 4 if the preceding three levels and all the criteria at level 4 are checked. Each exercise provides a score for each of the four competence levels, a total score for the exercise mode, and a combined total score for performance on the Communications competence.

Results

Based on our experience analyzing the Valuing Instrument, we [illegible] the same set of faculty-generated questions to begin analysis of the Communications data. The first question guiding the analysis was:

- How does the performance of the instructed group compare to the performance of the uninstructed group on each competence level within each Communications mode (Speaking, Writing, Listening, and Reading)?

The mean and standard deviation per level within each Communications mode is presented in Table 1. Univariate ANOVA was employed to investigate significant differences between group performances. The univariate analysis shows significantly higher performance by the instructed group in Speaking, levels 2, 3, and 4; Writing, levels 2 and 3; Listening, levels 2 and 4; and Reading, levels 2 and 3.

Table 1

Means (M), Standard Deviations (SD), and F Ratios for Instructed and Uninstructed Groups Per Level within Modes of Communications

Communications Mode/ Competence Level	Group Uninstructed (n = 20)	Instructed (n = 20)	F Ratio[a]
Speaking			

Level 1	M = .8
	SD = .4
		.9
		.3
			.76

Level 2	M = 11.2
	SD = 2.2
		13.8
		1.8
			17.56**

Level 3	M = 3.3
	SD = 2.2
		5.5
		1.6
			12.33**

Level 4	M = 1.1
	SD = .9
		2.5
		1.3
			15.05**

Writing			
Level 1	M .8 SD .4	.7 .4	.14

Level 2	M = 7.8
	SD = 3.3
		.7
		2.5
			2.98*

Level 3	M = 2.3
	SD = 1.3
		3.9
		1.4
			12.47**

Level 4	M = .9
	SD = .6
		1.2
		.8
			1.80

Listening			

Level 1	M = .9
	SD = .3
		.8
		.4
			.22

Level 2	M = 2.7
	SD = .7
		3.0
		.0
			2.43*

Level 3	M = 1.0
	SD = .9
		1.2
		.6
			.67

Level 4	M = .7
	SD = .8
		1.4
		1.2
			5.19**

Reading			

Level 1	M = .8
	SD = .4
		.8
		.4
			.00

Level 2	M = 2.1
	SD = 1.0
		2.8
		.4
			8.96**

Level 3	M = .2
	SD = .4
		.8
		.7
			13.56**

Level 4	M = .6
	SD = .9
		.8
		1.0
			.43

[a] df (1, 38)
* p < .05
** p < .001

The next questions to guide the analysis were:

- To what extent does the instructional treatment produce differences in variability in student performance and what are the directions of such differences?

- To what extent is reduction of variability in the instructed group an indication of the validity of the instrument for measuring the effectiveness of instruction?

Table 3 shows a pattern of reduced variability in the instructed group compared with the uninstructed group within Speaking, Levels 1, 2, and 3; Writing, Level 2; Listening, Levels 2 and 3; and Reading, Level 2. Although the instructed group demonstrates a higher mean performance in Level 4 of Speaking, Writing, Listening and Reading, the variability within this group was higher than in the uninstructed group instead of lower, as expected.

To further investigate variation in student performance, the distribution of the students' combined total scores in the Communications competence were examined (Figure 3).

The uninstructed group displays a positively skewed distribution. Most students fall in the lower score range with a few in the higher score range. The instructed group displays a negatively skewed distribution. Most the students fall in the higher score range with a few in the lower. The amount of overlap between the two distributions may indicate the extent to which the instructional treatment (as measured by the instruments) discriminates between the two groups. Lack of overlap, or a small amount of overlap may indicate the discriminative power of the instrument and its validity in measuring effectiveness of instruction. The lined area in Figure 3 represents the amount of variation in the instructed group which may be due to the instructional treatment. How much variation could be attributed to the instructional treatment and how much to individual

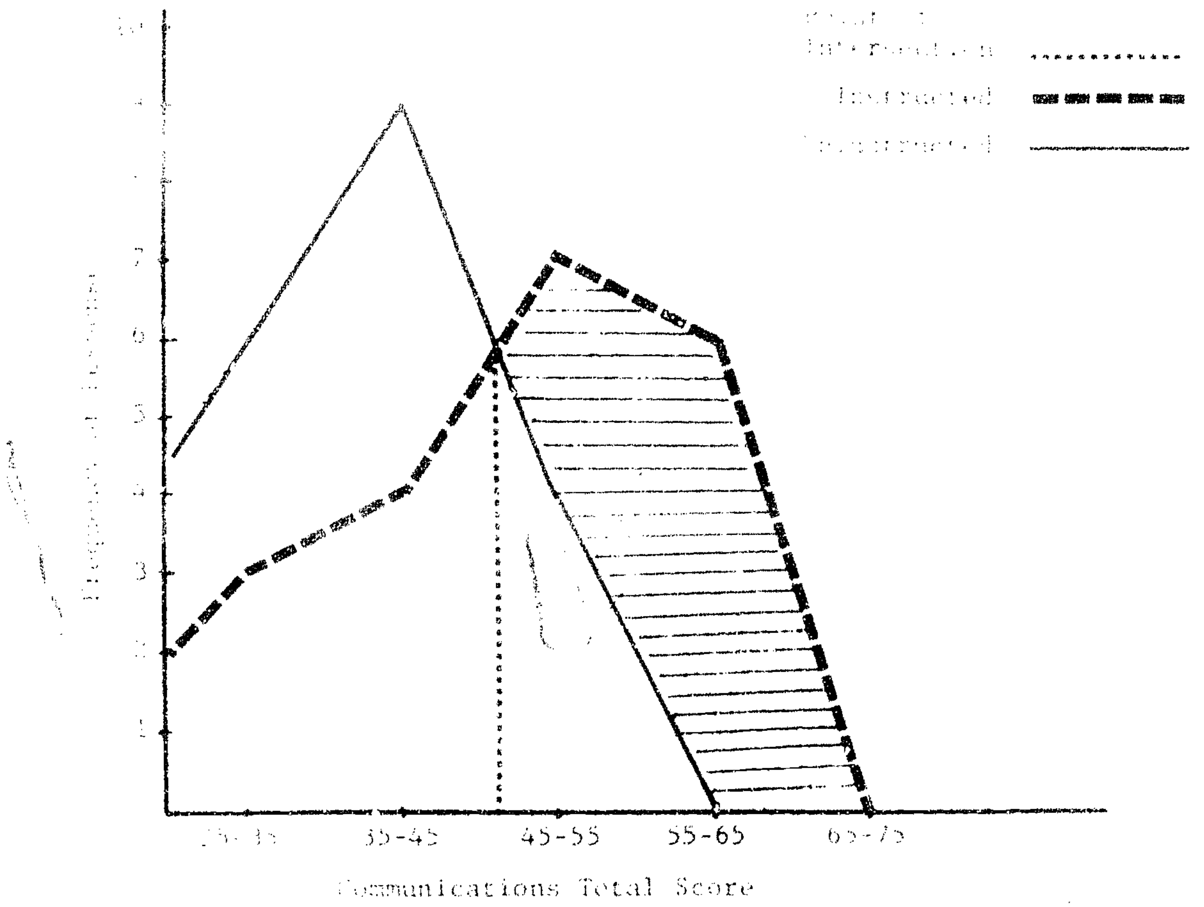

Figure 3. Frequency polygons of Communications combined total scores for instructed and uninstructed groups.

Note: Point of intersection may indicate cutoff point for an acceptable level of performance (Berk, 1976).

differences within the instructed group? A t test computed for the Communications total score for each group (instructed M = 49.5, SD = 9.8) (uninstructed M = 37.4, SD = 9.8) resulted in a significant value of t(38) = 3.84, p < .01). The instructed group performs significantly better overall.

Omega squared statistic was computed to estimate how much of the variation in student performance can be attributed to the effects of instruction. Twenty-six percent of the variations (w^2 = .26) in the instructed group is attributable to the instructional effects as measured by the instrument.

Since the ultimate goal of the [illegible] program is to facilitate student mastery of [illegible] competency levels, we then asked the following question:

- To what extent did the instructed group master[1] each competence level within each communications mode compared to the uninstructed group?

Although the univariate ANOVA shows significantly higher mean performance by the instructed group at the various levels of the Communications modes, this does not imply that the in[illegible]ted group was made up of more "mastery" students per level within each mode. A Chi-square was computed to investigate the extent of the relationship between mastery and instruction.

[illegible] analysis [illegible] Table 2 indicates [illegible] significant [illegible] instructed students are [illegible]

[illegible]

- [illegible]

[illegible] the [illegible]

Table 2

Chi-Square Values Comparing Mastery Students in Instructed and Uninstructed Groups

	Mastery Students		
Mode	Instructed (n = 20)	Uninstructed (n = 20)	Chi-Square
Speaking: Level			
1	18	16	.2
2	12	2	8.9*
3	10	3	4.1*
4	8	0	7.6*
Writing: Level			
1	15	16	.0
2	12	8	.9
3	10	2	6.8*
4	8	3	2.0
Listening: Level			
1	17	18	.0
2	20	17	1.4
3	0	0	---
4	6	0	4.9*
Reading: Level			
1	17	17	.2
2	17	11	2.9
3	3	0	1.4
4	2	0	.62

*p < .05

Table 5

Criterion Response Frequency within Four Modes of Communications

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
SPEAKING Criteria																											
	Level 1	Level 2															Level 3							Level 4			
Uninstructed	16	18	17	14	16	15	16	7	20	16	20	8	8	20	17	14	9	7	15	5	7	11	13	3	6	3	11
Instructed	18	19	19	20	19	19	20	14	18	18	20	17	18	20	20	17	16	17	17	13	13	17	17	9	12	14	17
Chi-square	.19	.00	.27	6.2 **	.91	1.7	2.5	3.6 *	.52	.00	---	6.8 **	8.9 **	---	1.4	.57	3.8 *	8.4 **	.15	4.9 *	3.6 *	7.9	1.2	7.9	7.5	10.2 **	1.8

Group,	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
WRITING Criteria																			
	Level 1	Level 2											Level 3					Level 4	
Uninstructed	16	18	15	15	10	15	10	13	16	11	18	16	11	10	4	9	14	3	15
Instructed	15	20	19	17	16	17	17	16	18	18	15	16	17	16	14	15	16	9	15
Chi-square	.00	.52	1.7	.15	2.7	.15	4.1 *	.50	.19	4.5 *	.69	.15	5.3 *	2.7	8.1 **	2.6	.13	2.9	.13

*p < .05
**p < .001

Note: Starred criteria were identified as discriminative items.

Group	47	48	49	50	51	52	53	54	55
LISTENING Criteria									
	Level 1	Level 2			Level 3		Level 4		
Uninstructed	18	18	19	18	10	12	16	0	10
Instructed	17	20	20	20	7	18	8	7	14
Chi-square	.00	.52	.00	.52	.40	4.5 *	1.0	6.2 **	.93

Group	56	57	58	59	60	61	62	63	64
READING Criteria									
	Level 1	Level 2			Level 3		Level 4		
Uninstructed	17	14	17	12	0	4	0	0	10
Instructed	17	20	20	17	7	10	3	15	9
Chi-square	.19	4.9 *	1.4	2.0	6.2 **	2.7	1.4	3.6 *	1.0

Speaking criterion 1

		CRITERIA	
		Responded	Did not respond
GROUPS	Instructed n = 20	16	4
	Uninstructed n = 20	18	2

The Chi-square analysis identified discriminative criteria reflecting the effectiveness of instruction. The significance level of each Chi-square value as indicated in Table 3 for each criterion shows a strong association between group and criterion. Instructed students tend to perform better on the starred criteria (see Table 3).

The following analysis was performed to validate the conceptual framework of the competence model:

- Can we identify clusters of response to criteria in the instructed group that are different from the uninstructed group? Where do clusters occur?

Cluster analysis was employed[1] on all 64 criteria within each mode of the Communications competence. Twenty percent error level was chosen as the level for comparison. This generated 10 clusters for each group.

A number of similar clusters was formed in both groups. The first common cluster for both groups is formed within the Speaking mode. Students who are able to make inferences tend also to make relationships to other sources of information or among the parts of their speech. This ability appears to describe both groups, independent of instruction. A second common cluster indicates a relationship between Speaking and

[1] See Donald J. Heldman, Fortran Programming for the Behavioral Sciences (Chicago: Holt, Rinehart, & Winston, 1967). This program is based on an article by J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," American Statistical Association Journal, 1963, 58, 236-244. Additional subroutines were developed by Larry W. Claflien and Fred Ostapik, University of Wisconsin-Milwaukee.

Listening. Students who support their statements with examples in the Speaking exercise, speak on their feet, create a positive image and articulate clearly using appropriate sentence structure tend also to identify the central idea of another speaker, state examples used by the speaker and distinguish a speaker's facts from his or her opinions. Since this cluster is formed in a similar manner in both groups, the assumption is made that such an ability is formed independent of instruction.

The third common cluster indicates a similar characteristic within the Writing mode. Students who distinguish among sources of information tend also to internalize ideas and structure their writings appropriately. This ability also appears to form independent of instruction.

The fourth common cluster indicates that students who are able to recognize their own strengths and weaknesses in Writing and Speaking tend to do so regardless of the number of self-assessment criteria involved.

The instructed group demonstrates a number of clusters which differ from the uninstructed group. Some of the clusters formed within the instructed group raise the following questions regarding the competence:

- Does the attainment of one mode of Communications facilitate attainment of other modes?
- Is growth toward the mastery of the Communications competence a function of the student's initial competence level upon entering college?

The clusters formed within the instructed group presumably reflect the instructional treatment, and show a definite improved association between Speaking and Writing abilities. This suggests that skills learned in one mode may generalize to the other. Such a cluster is not evident in the uninstructed group, and suggests weak associations between Speaking and Writing abilities.

Since both groups formed a cluster of Writing abilities independent of instruction, one may conclude that students entering college with certain writing skills are more likely to further develop their writing ability and to acquire Speaking skills as an instructional outcome. Students who lack such basic Writing skills may not develop all the required Writing and Speaking skills within a period of two years.

The patterns of student response were further investigated employing correlation matrices:

- To what extent can we attribute differences between the instructed and uninstructed groups to the cumulative nature of the competence levels?

The correlation matrices indicate clearly that the instructed group demonstrates high intercorrelations among the criteria for levels 2, 3, and 4 within the Speaking mode, reflecting a cumulative pattern of student response. Students who mastered levels 3 and 4 tend to master level 2 also. This pattern is not evident in the uninstructed group. (The previous analysis clearly indicated that instructed students master Speaking levels 2, 3, and 4 significantly better than the uninstructed group.)

The correlation patterns support the cluster analysis results. Similar patterns of high intercorrelations within the Writing mode for the two groups imply a Writing capability independent of instruction. (Within the Writing mode, the instructed group is more likely to achieve mastery than the uninstructed group only at level 3.)

Finally the question was raised:

- To what extent can we attribute differences between the two groups to the sequential nature of the Communications competence?

Guttman Scalogram analysis (Hambleton, 1979) was employed to explore the theoretical framework of the pedagogical developmental sequence of the levels within the four modes. The order of the levels within each mode as generated by the Guttman Scalogram is presented in Table 4. The analysis

Table 4

Sequential Order of Levels Demonstrated by Instructed and Uninstructed Groups Within Each Mode of Communications as Generated by Guttman Scalogram Analysis

	Order of Levels			
Group	Speaking	Writing	Listening	Reading
Expected Sequence	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
Instructed Group	1 2 3 4	1 2 3 4	2 1 4 3	2 1 3 4

	COR[a] = 1.000
	MMR[b] = .650
	COS[c] = 1.000
		COR - .950
		MMR = .612
		COS = .871
			COR = 1.000
			MMR = .88
			COS = 1.000
				COR = .92
				MMR = .86
				COS = .45

Uninstructed Group	1 3 2 4	1 2 4 3	1 2 4 3	1 2 4 3

	COR = .975
	MMR = .885
	COS = .778
		COR = 1.000
		MMR = .787
		COS = 1.000
			COR = .950
			MMR = .937
			COS = .200
				COR = 1.000
				MMR = .85
				COS = 1.000

[a] Coefficient of reproducibility.
[b] Minimum marginal reproducibility.
[c] Coefficient of scalability.

indicates that the instructed group demonstrates a sequence identical to that of the expected order in the Writing and Speaking modes. The Speaking levels show perfect scalability and the Writing levels show high scalability. The uninstructed group do not follow the expected developmental sequence in Speaking and Writing. They do demonstrate perfect scalability in an order clearly their own.

The results may indicate that Alverno is successful in teaching its students to follow a pedagogically specified developmental sequence throughout their learning experience and that uninstructed students adopt a different sequence of performance throughout their previous experience.

In the Reading exercise, the instructed group demonstrates level 2 before they do level 1, and the overall performance is not entirely coherent within the group (COS = .45). The uninstructed group demonstrates level 4 before they do level 3, but the overall performance is coherent within the group (COS = 1.000).

Both groups demonstrate level 4 of the Listening exercise before they do level 3. However, the instructed group demonstrates perfect scalability (COS = 1.000); the coefficient of scalability is low (.20) in the uninstructed group.

Discussion

The Communications study results were first presented to the chairperson of the Communications Division. Another meeting was arranged during which all Communications Division members discussed the study, contributed to the interpretation of the data and its implication for instrument revision.

Each research question (listed below) is discussed in light of the statistical findings and the discussion with Communications Division members.

The first question is:

- How does the performance of the instructed group compare to the performance of the uninstructed group on each competence level within each Communications mode (Speaking, Writing, Listening, and Reading)?

The first competence level in each Communications mode consists of one self-assessment criterion which was intentionally designed to be an easy one for the incoming student. Consequently both the instructed and uninstructed students performed equally well on level 1 criteria across four modes of Communications. Within the remaining 2 to 4 levels (a total of 12) the instructed group performed significantly higher on 9 out of the 12 competence levels.

The instructed group does not demonstrate higher performance in level 4 Writing, level 3 Listening, and level 4 Reading. The questions raised by faculty that will guide instrument criteria revision on these three levels are:

1. To what extent are the criteria clearly defined?
2. Are the criteria a sensitive measure of what was learned?
3. Does the stimulus elicit the expected response?
4. Is the assessor's interpretation of student performance consistent with the intended meaning of the criterion?

The second question is:

- To what extent is reduction of variability in the instructed group an indication of the validity of the instrument for measuring the effectiveness of instruction?

We believe that reduction in variation of student performance is one indication of instrument validity in our outcome-centered curriculum, where cutoff points are not always specified. However, the data analysis indicates that absolute statements on reduction in variability cannot be made for the entire instrument. Performance variability changed with the competence levels. The instructed group demonstrated increased variability at level 4 of all Communications modes, whereas at the lower levels there is some indication of a decreased pattern of variability within the instructed group compared to the uninstructed group. These findings suggest that individual differences play a greater role in the variation of performance at the higher level of Communications. The difference in variability at level 4 as compared to the preceding levels may also indicate the leap students are making in integrating competence and content. At the lower levels students follow explicit criteria which structure their learning. Level 4 criteria, however, call for internalization of the competence with more emphasis on the content analysis. It is assumed that such an internalization process requires additional skills, bringing individual differences to the fore. Figure 3 indicates that 45% of the instructed students are below the cutoff point (the point where the instructed and uninstructed curves intersect). Thus, instruction was effective for 55% of the instructed students. When performance of individual students is followed across the four modes of Communications, it is apparent that the same group of students excel at the higher levels, producing a more dispersed score distribution. When we attempt to measure change, individual differences account not only for attaining a competence level but also for the rate of attainment. The distribution of scores may indicate that either instruction was not as effective for 45% of the students, or these students need more time to develop the required Communications skills. A final statement about the validity of the instrument with regard to reduction in instructed students' variability should integrate data pertaining to the rate of competence attainment, i.e., number of attempts to master a competence level, duration of learning experience prior to the mastery attempt, and consistency of performance across ~~the~~ four modes of Communications.

The third question is:

- To what extent did the instructed group master each competence level within each Communications mode compared to the uninstructed group?

Comparison of mastery performance between the two groups indicates a significantly higher performance of the instructed group at levels 2, 3, and 4 of the Speaking mode, level 4 of the Listening mode, and level 3 of the Writing mode. The instructed group does not indicate high mastery performance on the Reading and Listening modes whereas, in the Writing mode, the lack of significant differences in mastery performance are due to the fact that the uninstructed group performed as well as the instructed group (see Table 2).

The results of the present study clearly distinguish between the Speaking and Writing components as one aspect of the Communications competence, and Reading and Listening as a somewhat different aspect. Alverno faculty were aware of such differences and identified the Reading exercise as "Analysis of Written Verbal Construct." The Listening exercise is referred to as "Analysis of Oral Verbal Construct." In these modes of Communications students are required to demonstrate a greater degree of analytical abilities as well as Communications skills. Instructed students also seem to view Listening and Reading as preparatory exercises for Speaking and Writing. This may diminish the seriousness with which they approach the task.

The statistical analysis directs faculty attention to two competences (rather than one) measured by the instrument: Analysis and Communications. The Communications competence is defined as the process of effective clarification and involvement between a presenter and an audience (which represents the Communications aspect of the instrument). Since the instrument is also measuring analytical skills how might we best describe the Communications aspect of Reading and Listening? Is it the way the student articulates her analytical skills? Is the analytical part inseparable from her ability to effectively clarify and articulate her message? Are Communications and Analysis two competences or one? If analytical skills are necessary for effective communications, the instrument is considered unidimensional even though it measures an additional analytical dimension.

The fourth question is:

- Is the instructed group performing better on each of the criteria?

Sixteen criteria out of 64 were found to discriminate between the instructed and uninstructed students. However, the question is whether the statistical analysis employed (Chi-square) is a suitable strategy for identifying such criteria. Should criteria be considered in terms of amount of effort and time invested in the teaching process rather than being evaluated by the frequency of student response (see Table 3)? Faculty decided to study the aspects that contribute to the discriminant power of a criterion?

The fifth question is:

- Can we identify clusters of responses in the instructed group which are different from the uninstructed group? Where do they occur?

Cluster analysis demonstrates common abilities within either the Speaking or Writing modes independent of instruction, suggesting that the uninstructed group enters college with already acquired communications skills in Speaking and Writing. However, instruction produces significantly greater performance by instructed students in the Speaking mode but not in the Writing mode. The cluster analysis also demonstrates a learned pattern or improved associations

[illegible]

[illegible]

[illegible] further develop their [illegible]

[illegible]

[illegible] developed their communica[illegible] [illegible] by the definition of the [illegible]

[illegible] developing students' communication skills in the specified sequence within the Writing and Speaking modes. The instructed students do not follow the expected sequence within the Listening and Reading modes. The uninstructed group follows a different sequence in performing all four modes of communications.

[illegible]

[illegible]

1. [illegible] demonstrate [illegible] different [illegible] disability [illegible]

2. [illegible] alternative developmental [illegible]

3. [illegible]

4. [illegible] related development [illegible]

[illegible]

[illegible] for the [illegible] understanding which abilities are expected to be [illegible] learning experiences. Abilities which [illegible] to be developed [illegible] will then receive greater emphasis in the learning process.

CONCLUSIONS

As more colleges [illegible] competence-based curricula call for different measurement techniques, especially in higher education. Educational researchers, discipline experts and experts in the development of generic abilities such as Valuing and Communications are necessary to create and validate assessment techniques. These techniques must assess student performance for credentialing and also provide adequate diagnosis of student performance to enable prescriptive instruction and structured feedback to students.

[illegible] multidisciplinary [illegible] to instrument construction and validation is a process of instrument development and revision that is not separate from the process of evaluation and validation. Faculty profit from empirical feedback during instrument design as much as they profit during later instrument revision. More important, their input is critical to study design and interpretation.

The validation studies reported in this paper stimulated faculty involvement in redesigning and revising the assessment techniques. The very learning that resulted from the interpretation of the empirical data appeared to be an intrinsic motivator for grappling with old issues or generating new ones, for confirming or reconceptualizing a competence definition, for reinforcing or revising the competence assessment criteria and for supporting or rethinking the instructional objectives.

The two empirical illustrations demonstrate the extent to which a variety of specific issues raised by the studies identify new issues, all directed toward creating a link between student performance measured by the instrument under study and the effectiveness of instruction. Once such a link is demonstrated across all generic instruments, the internal validity of the curriculum is strengthened.

For us, developing a conceptual framework for the validation of assessment techniques in an outcome-centered curriculum demanded that it be derived from the definition of the competences and the characteristics of assessment techniques as well as from faculty questions and concerns. The empirical studies provided the framework for the integration of all the above sources and for a study of competences defined as generic, developmental and holistic.

Generic instruments were employed, allowing comparability between entering students and students at the midpoint of their college career. Thus, we were able to explore the extent to which instructed students demonstrate a given ability across a variety of settings and competence components. When Alverno faculty break open a competence into its separate components, they also spell out a pedagogically developmental sequence of acquiring these abilities that they believe will best facilitate the learning of the ability. This sequence, while it makes sense pedagogically, may not match the developmental sequence of the ability as it is acquired without the benefit of formal instruction. Students who have extensive life experience may develop some aspects of some abilities on their own. Results from this study stimulated questions about the extent to which each of the four competence levels are truly cumulative and sequential. The holistic nature of the competences is explored in this study by examining the extent to which the ability measured is unidimensional—especially at the upper levels.

The assessment system and techniques employed at Alverno are designed to assess mastery learning for credentialing purposes. The present study explored mastery learning without specific reference to cutoff points. Instead, we focused on the reduction in variability within groups since a major problem in measuring change and growth in a competence-based program is how individual differences account for a large proportion of the variance in program outcomes (Fischer, 1978). Reduction in variation of student performance is not consistent across competence levels. Individual differences as well as instruction impact variability in student performance. How can we separate the two? Where is variation in student performance due mostly to individual differences? Instructional methods can then focus on other areas of performance, and take these individual differences into account. In a mastery learning program where individual differences are viewed diagnostically, the rate of competence attainment should also be incorporated into validity statements. Identifying the areas where instruction reduces variation in student performance may facilitate students' upward movement within the learning process.

We would like to conduct additional studies (pre- and post-instruction) on the same students, which will assist us in further exploring the issue of individual differences. Such pre- and post-instruction studies will allow us to explore the following questions which are central to demonstrating program effectiveness: Does the program provide similar opportunity for each student? Is the program more effective with the better student? Does it stimulate enough growth for the excelling student?

In sum, the empirical illustrations demonstrated the importance of recognizing the need for construct validation in outcome-centered programs in higher education. Studies compared competence definitions against empirical data and provided a broader scope for understanding the competence constructs and generated implications for future instruction. A link was created between validation studies and instrument revision with the full cooperation of the faculty. Finally, the studies support recognition of the importance of individual differences and their impact on competence definition. We are not suggesting that competence criteria be adjusted to group "norms." Rather, we suggest that an understanding of the patterns of students' performance provides greater insight into the meaning of competence assessment criteria. Our future studies will incorporate the present findings and will open new ways and strategies for better understanding outcome-center programs, and how the validity of their assessment techniques can be established.

Our awareness of the framework of the curriculum from which instruments are developed assisted us to be more sensitive to the "authentic" meaning of the assessment techniques and to relate to issues which lie at the heart of the curriculum. We support attempts to establish the construct validity of the competences as a major way to establish instrument validity:

> "A significant development is the recognition of the need for construct validation studies. . . . The size of the test development project will influence the scope and the number of construct validation studies, but clearly, more work is needed in this area. . . . The limit of these studies will be the level of creativity and ingenuity of the researchers involved." (Hambleton, 1978)

We agree with Hambleton that we are limited only by our own level of creativity and ingenuity.

REFERENCES

Buck, R. A. Determination of optimal cutting scores in criterion referenced measurement. _Journal of Experimental Education_, 1976, _45_, 4-9.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), _Educational measurement_ (2nd ed.). Washington: American Council on Education, 1971.

Fincher, C. Program monitoring in higher education. _Monitoring Ongoing Programs_ (3rd ed.), 1978.

Gamsan, Z. Assuring survival by transforming a troubled program: Grand Valley State Colleges. In G. Grant & Associates (Eds.), _On competence_. Washington: Jossey-Bass, 1979.

Hambleton, R. K., & Eignor, D. R. _Criterion referenced test development and validation methods._ AERA training program materials. University of Massachusetts, Amherst, 1979.

Hambleton, R. K., & Swaminathan, H. Criterion-referenced testing and measurement: A review of technical issues and developments. _Review of Educational Research_, 1978, _48_, 1-47.

Hays, W. L. _Statistics for the Social Sciences_ (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

King, E. S. Assessment of competence: Technical problems and publications. In G. Grant & Associates (Eds.), _On competence_. Washington: Jossey-Bass, 1979.

Kohlberg, L., Colby, A., Gibbs, J., & Speicher-Dubin, B. _Standard Form Scoring Manual_. Unpublished manuscript, Harvard University, 1978.

Mentkowski, M. Creating a "mindset" for evaluating a liberal arts curriculum where "valuing" is a major outcome. In L. Kuhmerker, M. Mentkowski, and E. Ericksen (Eds.), Evaluating moral development: And evaluating educational programs that have a value dimension. Schenectady, N.Y.: Character Research Press, 1980.

Mentkowski, M., & Doherty, A. _Careering after college: Establishing the validity of abilities learned in college for later success._ Proposal funded by the National Institute of Education, September 1977.

Mentkowski, M., & Doherty, A. _Careering after college: Establishing the validity of abilities learned in college for later careering and professional performance_. Final report to the National Institute of Education. Milwaukee, WI: Alverno Productions, 1983.

Messick, S. The standard problem: Meani and values in measurement and evaluation. _American Psychologist_, 1974, _30_, 955-966.

Payne, J. _Principles of social science measurement_. College Station, Tex.: Lytton Publishing Company, 1975.

Alverno College

3401 South 39th Street / Milwaukee, WI 53215