DOCUMENT RESUME

ED 219 398 TM 820 402

AUTHOR Peterson, Gary W.
TITLE A Meta-Evaluation of a Generic Skills Approach to Evaluating Academic Programs.
PUB DATE 23 Mar 82
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Standards; Case Studies; Educational Development; Educational Planning; Higher Education; *Institutional Evaluation; *Models; Performance; *Program Evaluation; *Self Evaluation (Groups); Skill Development
IDENTIFIERS *Academic Program Evaluation Paradigm; Meta Evaluation

ABSTRACT

The Academic Program Evaluation Paradigm (APEP) is a five-stage process for participating institutions and their faculties to structure inquiry into their academic programs and develop concrete procedures to effect institutional changes. APEP was developed and implemented by 10 member institutions of the American Association of State Colleges and Universities. In the Paradigm, institution faculties define generic skill outcomes of their academic programs; select or develop student outcomes and program portrayal measures; identify desired performance standards; and make judgments about discrepancies, defined as "gaps" between the observed and desired levels of performance. Policies and procedures are then formulated to rectify high priority gaps. The generic skills of communication, analysis, synthesis, quantification and valuing are key components of the Paradigm. Analysis of the two and one-half year project included an institution which completed the Paradigm and six other institutions in which limitations in the implementation of the final stage made results uncertain. Limitations of the Paradigm in its potential goal as a guide for program evaluation include the time factor in completion of all stages, whether population samples are adequate, and the validity of measures of skills. (Author/CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED219398

A Meta - Evaluation

of a

Generic Skills Approach to Evaluating Academic Programs

Gary W. Peterson

Florida State University

Center for Educational Technology

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it

Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G.W. Peterson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

March 23, 1982

Paper to be presented at the Annual Convention of the American Educational Research Association, New York, March 23, 1982.

TM 820 402

Abstract

A five stage paradigm for evaluating academic programs was developed and implemented by ten member institutions of the American Association of State Colleges and Universities. The Paradigm is based on having faculty: define generic skill outcomes of their academic programs; select or develop student outcome and program portrayal measures; identify desired performance standards; and make judgements about any discrepancies (gaps) observed between the observed and desired levels of student and program performance. Then, policies and procedures are formulated to rectify high priority "gaps". In the 2½ year time span allowed for the project, one institution was able to identify performance gaps and to formulate policies and procedures to rectify them. Six other institutions reached the final stage but the extent to which the policies and procedures proposed were based on the derivation of clearly documented performance gaps is uncertain. Some of the limitations in implementing the Paradigm include the amount of time to proceed through all the stages, obtaining adequate population samples, and obtaining or developing valid measures of skills. These and other problems must be solved before the Paradigm reaches its full potential as a guide for structuring program evaluation activities.

A Meta - Evaluation

of a

Generic Skills Approach to Evaluating Academic Programs[1]

Gary W. Peterson[2, 3]

Florida State University

Introduction and Contextual Factors

In the Summer of 1978, a conference was held in Ashville N.C. with over a dozen of the Vice Presidents of the member institutions of the American Association of State Colleges and Universities (AASCU) to formulate the basic parameters of the proposed Academic Program Evaluation Paradigm herein referred to as the Paradigm or APEP. From the proceedings of this conference, broad conceptual descriptions of a five-stage evaluation process subsequently evolved: Stage I: Definitions; Stage II: Establish Levels of Performance; Stage III: Assessment; Stage IV: Evaluation; and Stage V: Policy, Management and Feasibility Issues Related to Program Evaluation (Buhler-Miko, 1979). At this same conference Jonathan "Bud" Warren of the Educational Testing Service also presented the Vice Presidents with a broad conceptual framework of high, medium and low performance levels for each of three designated generic skills, Communication, Analysis and Synthesis (Warren, 1979). Shortly thereafter, with funding support from FIPSE, 17 institutions applied to the Resource Center for Planned Change, AASCU, and 10 were selected to engage in the formal development and implementation of the Paradigm. The following paper presents the outcomes to date regarding the development and implementation of APEP by the ten institutions participating in the project. The paper concludes with a discussion of theoretical issues that undergird the Paradigm, its limitations, and directions for further research.

Central Concepts on Which the Paradigm is Based

In order to use the Paradigm successfully, several pivotal concepts must be thoroughly understood: generic skill, performance level, performance gap, program portrayal, policy development, and procedural development. A series of seven workshops, with supportive materials, was held for

1. The information about the projects is current through March 23, 1982.

2. The author was an external consultant to the project who assisted in the development of the APEP Guidelines (Buhler-Miko, Peterson, and Stakenas, 1982).

3. Gratitude is expressed to Robert Stakenas and Lin Webster for their constructive critical comments on initial drafts of the paper.

the Vice Presidents and their respective faculty teams to help them develop a working knowledge of these concepts. The major source documents provided for the teams included a precise, Developing Generic Skills: A Model for Competency-Based General Education (Woditsh, 1977), an article, "Describing college graduates in 87 phrases or less" (Warren, 1976), excerpts from Florida Competency-Based Articulation Project: Final Report (Peterson and Watkins, 1978), initial drafts of the APEP Guidelines (Buhler-Miko, Peterson and Stakenas, 1982), an occasional paper (Peterson and Stakenas, 1980), and specific quidelines related to test development and selection prepared by the present author. The project staff also provided the teams with annotated bibliographies related to aspects of generic skills and organizational development. For the Paradigm the above terms were defined as follows:

Generic Skill. According to Woditsch (1977), the term, generic, connotes a function or a pattern of activity that is recurrent in a wide series of discrete purposive behaviors. "Generic skills are basic in the sense that they are ubiquitious: they show up again and again as components or instances of successful behavior" (pg.8). The faculty teams were also given an additional set of attributes for generic skills (Peterson and Stakenas, 1980).

- A generic skill is an ability or capability that possesses its own unique hierarchy of discrete related component skills;
- A generic skill is pervasive and recurs across academic or professional disciplines of study and even across life or job tasks;
- The mastery of a knowledge base underlies the development and demonstration of generic skills;
- The demonstration of generic skills requires the mastery and integration of discrete lower order component skills and knowledge; and
- Individuals who have mastered generic skills are able to apply them in a variety of real life situations or contexts to solve problems encountered in adult roles in society.

The faculty teams were initially presented with four (4) generic skills, Communication, Analysis and Synthesis, and Quantification (Warren, 1979) from which to further develop their unique conceptual and operational definitions. A Valuing skill was added after the inception of the project to make the development of five skills the focus of the evaluation. Each of the faculty teams was encouraged to consider the skills in terms of their attributes (i.e., developing inventories of subskills), their performance levels (Warren, 1979) and in terms of their developmental hierarchies (Gagne, 1968). It was assumed that through these perspectives, the faculty

would gain sufficient understanding of the skills so as to be able to develop and/or select valid measures consistent with their respective missions, goals and curricular offerings.

Performance level. Using examples proposed by Warren (1979), the faculty were to describe each of the five generic skills conceptually in terms of attributes of high, medium and low performance levels. From such conceptual descriptions, faculty could then develop rating scales with which to evaluate student performance on given assessment tasks. Through an understanding of skill definitions and performance levels, it was assumed that faculty could negotiate a "cognitive leap" from conceptual to operational forms of the skills and be able to determine the validity of multiple choice tests available through various commercial testing firms.

Performance gap. The performance gap may be thought of as the "linch pin" of APEP. Basically, the "gap" is the discrepancy between an observed performance level and a desired level of performance of a program element in question (Kaufman, 1972, Kaufman and English, 1979). In APEP, the "gap" refers to not only differences between desired and observed performance levels of generic skill measures but also differences between desired and observed levels of program portrayal dimensions such as number of essays assigned and graded in a given time period in selected courses. The "gap," in effect is the operational definiton of an organizational problem that lays the foundation for subsequent policy and procedural considerations.

Program portrayal. According to Stake (1967), a program can be described in terms of variables related to Antecedents, Transactions, and Outcomes. Within each of these areas, each program element can be analyzed with respect to intents and observations. The former is the program element designated for implementation while the latter is a documentation of actual observations of the ways in which the program element became operational. For example, a Transactional element might be, 'student written productions'. An intent might be the 'the assignment of writing samples' while an observation might be the number of papers, essay tests, and quizzes assigned in a random sample of courses in a program during a given time period. The program portrayal elements selected for observation are logically (and hopefully causally) related to the development of generic skills. The purpose of incorporating the Stake model in APEP is to encourage faculty to accrue information about instructional practices that may account for the observed level of student performances on generic skill measures.

Policy development. Policies may be considered as general statements of plans, principles and priorities that guide decision-making and commit the organization to a set of alternative actions, goals and values (Baldridge et. al., 1978 and Cronbach et. al., 1980). The Paradigm is chiefly concerned with policies related to the structure of the curriculum and to instructional practices. An example of a policy statement stemming from a writing deficiency identified and judged to be significant, might be, "Midwestern State University insists that all graduates are capable of writing eloquent, articulate and grammatically correct prose and that it is

the responsibility of all faculty members to encourage and foster such capability in all undergraduate courses."

Procedural development. Procedures allude to the processes and rules employed to execute and enforce policy (Baldridge, et. al., 1978). In the above example, procedures might include requiring all sophomores to pass a writing proficiency examination, as well as such logistical factors as the persons responsible for developing, administering and scoring the writing test, how often the test will be offered, and the designation of remedial courses to help instruct students who fail, and so on.

The above concepts undergird the process of the Paradigm. As will be discussed later, while they seem to be simple at first glance, these concepts proved to be challenging and complex during implementation.

Research Questions Guiding the Study

Two basic research questions served as the focus for the collection and analysis of the data for the present paper:

- To what degree did institutions implement stages of the Paradigm in the amount of time and resources available to the project?
- What institutional changes were observed as a result of attempting to implement APEP (to date)?

Method

Subjects (i.e., the Institutions). Ten institutions agreed to participate in the development and implementation of the Paradigm. Four institutions had enrollments of less than 5000 students, one enrolled between 5000 and 10,000 and five had enrollments larger than 10,000. Seven were residential and three were commuter colleges. Geographically, three were located in the Northeast, two in the Southeast, four in the Midwest, and one in the Far West. They were all members of the American Association of State Colleges and Universities.

Instrumentation. The principal data sources for the meta-evaluation were case histories written by the project coordinators and their associates at the respective institutions at the close of the project two and a half years after its inception. At the outset of the project, all participants were informed that they were to write a case history describing their processes, accomplishments, problems, difficulties, and outcomes of their respective attempts to implement the Paradigm. Four outlines for the case histories were circulated among the institutions prior to their writing by the project staff. The project teams were also informed that these outlines were meant to be suggestive of ways to structure their histories.

The case histories were between 20 and 56 double spaced pages with four at 20 pages and three more than 40. The styles varied considerably

with one in the form of a scientific journal article, another in the form of the acts of a theatrical production and yet another in the form of a dialogue between a faculty member and an administrator. The others were narratives that described their involvement in each of the stages of the Paradigm. The histories also varied in terms of use of statistics for their analysis of the data. Three used multivariate statistics while the remaining four projects that had completed the collection of data relied on descriptive statistics. In the portrayal of data in Tables 1 through 3, project coordinators at the ten institutions were encouraged to report any inaccuracies or updates up to one week prior to the delivery date of this paper.

Analysis. The content analysis of the case histories was structured along the lines of the Paradigm itself in order to explore the variety of ways and the extent to which each of the five stages was implemented. The ten institutions were also grouped according to common purposes by the present author so that the reader may observe the of ways in which the respective institutions implemented the Paradigm to achieve common project objectives. Thus, a two-dimensional matrix of Purpose X Stage was created to highlight commonalities and differences. For the present analysis, an attempt was made to use only information recorded in the case histories and to temper the use of impressions derived from other contexts. At times, however, it was difficult to separate these two sources of information and to completely exclude the latter.

Results of the Analysis

Several of the key components of the Paradigm are highlighted for the analysis: definitions of generic skills, the measures selected for student outcomes and program portrayal, evaluation designs, results of the respective inquiries, and subsequent policies and procedures adopted as a result of the investigation. These elements then provided a step by step overview of the ways in which institutions implemented APEP. As will be seen, no two institutions implemented the Paradigm in exactly the same way.

Definitions of generic skills. Using Travers (1980) discussion on taxonomies and classifications of educational objectives, the degree to which institutions were able to explore definitions was analyzed in terms of the following hierarchy of classification schemas moving from elementary to advanced levels of exploration: 1) conceptual descriptions; 2) inventories; 3) classifications within inventories; 4) hierarchial classifications; and 5) relationships among categories that ultimately relate to a higher order synthesis of all skills. As is portrayed in Table 1 on the next page, two institutions did not progress beyond the first level of broad conceptual descriptions. Five developed inventories of subskills within each of the generic skill areas (second level) while three were able to establish classification schemas within generic skill categories. None of the institutions reached levels four, or five although, in two of the case histories, one referred to the need to develop skill hierarchies, and the other suggested that the skills may actually be subordinate to an overarching program solving process. Without achieving the

Table 1: Extent of Implementation of Academic Program Evaluation Project: Stages I and II

Extent of Implementation / Purpose	Type Institutions (see key below)	Definitions Level of Accomplishment (see key below)	Student Outcome Measures Used	Program Protrayal Measures Used
I Improve curriculum				
a) PSC	a) Small, Residential	a) Classifications within categories	a) ETS Gen Ed -Local essay (Val) -Local M-C (Anal/ Syn, Quant)	a) Local student questionnaire -Local faculty questionnaire
b) WSC	b) Small Residential	b) Inventories within categories	b) Classroom tests (Comm, Anal, Syn, Quan)	b) none
II Initiate Formal Evaluation Procedures				
a) NASC	a) Small, residential	a) Inventories within categories	a) ETS Gen Ed, parallel forms	a) Local 2-item rating scale for students
b) SIU-E	b) Large commuter	b) Inventories within categories	b) ETS Gen Ed -Local Quant -Local Val	b) Pace, College Experiences Questionnaire for Students - Local faculty questionnaire
III Exploratory Pulse Reading of Gen Ed				
a) BSU	a) Large, residential	a) Classifications within inventories	a) ETS Gen ED	a) none
b) RC	b) Small commuter	b) Inventories within categories	b) Watson Gleaser Critical Thinking -STEP Math -Local Comm	b) none
c) UNO	c) Large commuter	c) Conceptual descriptions	c) ETS Gen Ed -Local Comm. -Rest Defining issues	c) Interviews with students and faculty
d) WCU	d) Medium residential	d) Inventories within categories	d) ETS Gen Ed -Neslon-Denny Read -Local Problem-Solving, Communication Analysis (M-C and essay), and Quantification -Huey-Johnson List.	d) local faculty questionnaire

Continued

Table 1: (Continued)

Extent of Implementation / Purpose	Type Institutions (see key below)	Definitions Level of Accomplishment (see key below)	Student Outcome Measures Used	Program Protrayal Measures Used
e) WKU	e) Large residential	e) Classifications within categories	e) ETS Gen Ed -ACT/COMP -Cornell test of Critical Thinking -Local M-C Synthesis	e) none
IV Enhance on-going Program Evaluation a)CS-C	a) Large, residential	a) Conceptual descriptions	a) Local Comm. -COOP English -NAEP Math -ETS Gen Ed -McBer TAT -Rokeach Dog. -CLEP,AVLSV	a) none
KEY	Small < 5,000 Medium 5,000-10,000 Large > 10,000	5. Dynamic relationships 4. Hierarchies 3. Classifications within categories 2. Inventories within classifications 5. Conceptual descriptions		

fourth and fifth levels of definition and classification, evaluation teams could only be left with a bewildering array of as many as 50 to 75 separate skill statements on which to select or develop measures. The question is raised concerning whether, in frustration, a number of teams reached for tests that, on the basis of "title" and face validity, appeared to measure at least some of the subskills they had identified.

Selection of student outcome measures. Eight of ten institutions selected the ETS Measures of General Education (Warren, 1980) as valid indicators of their skills. With respect to attempts to develop local tests (affectionality known as "home growns"), four institutions developed essay tests to assess Communication, and two used essay tests to assess Valuing. Two institutions developed mathematics tests and one developed a Problem Solving test. Other tests that were administered by only one institution included Watson Gleaser Test of Critical Thinking, Nelson-Denny Reading Test, COOP English Test, STEP Math Test, the Rest Defining Issues Test, NAEP Math Test, Rokeach Dogmatism Scale, Alport, Vernon, Linsey Study of Values Inventory, the McBer Thematic Analysis Test, ACT/COMP Communication Test, and the Cornell Critical Thinking Test. One institution used only existing classroom tests, quizzes, and term papers on which to observe generic skill performance. Three institutions attempted to develop their own multiple choice analysis and synthesis tests. The preponderance of student outcome testing involved the use of commercially prepared multiple choice tests. Possibly the teams, even though they were encouraged to develop their own tests, lacked either the time, technical assistance or self-confidence to engage in much experimentation with their own measures.

Program portrayal measures. Two institutions developed student questionnaires, three developed faculty questionnaires, and one institution administered the Pace College Student Experiences Questionnaire. Six institutions did not administer program portrayal measures, particularly those interested in obtaining only a general reading of student skill achievement (Purposes III and IV).

Evaluation designs. Nine institutions (see Table 2 on the next page) used some form of nonequivalent comparison group, posttest only designs (Campbell and Stanley, 1963) to assess the "value-added" contribution of either time in school (such as comparing freshmen and seniors) or kinds of courses (e.g., structured vs. unstructured general education programs of study). The major reason why these designs were classified as nonequivalent group designs is that the groups were not randomly drawn from the same population, thus introducing potential bias due to selection, mortality, and history. Two institutions used a pretest-posttest only design, one using a 4-month time span and the other a 7-month time span. Two institutions used co-relational designs (Tuckman, 1978) employing regression analyses to determine the amount of variance in generic skill performance attributed to either courses or length of time in school--again, hoping to determine the extent of the value-added benefit of educational experience. There was one time-series design planned as part of a four-year longitudinal study. Because the project period was only 2½

Table 2: Extent of Implementation of Academic Program Evaluation Project: Stage III

PURPOSE / Extent Implementation	Evaluation Designs	Student Outcome Samples	Program Portrayal Samples
I Improve Curriculum			
a) PSC	a) Non-equivalent comparison group, posttest only	a) Volunteers from stratified samples Academic area 8 x year (4) (n=177) -Freshmen non-volunteers (n=960)	a) Student questionnaire (n=177) Faculty Questionnaire (n = not reported)
b) WSC	b) Non-equivalent comparison group, posttest only	b) Faculty volunteered student tests (n=868) Freshmen, Sophomores, Junior,	b) None
II Initiate Formal Evaluation Procedures			
a) NASC	a) Pretest-Posttest (4 mos) -Non-equivalent comparison group	a) Non-volunteer, Freshman, Sophomore, Junior, Senior (n=482,338)	a) Student questionnaire (n = not reported)
b) SIU-E	b) Non-equivalent comparison group posttest only	b) Volunteers, Students Fresh(n=42, Sen(n=29); Non-volunteers (n=248).	b) Student questionnaire (n = 152) Faculty questionnaire (n=170, 32% return)
III Exploratory Pulse Reading of Gen Ed			
a) BSU	a) Pretest-Posttest (7 mos) -Non-equivalent comparison group, posttest only	a) -Volunteer Fresh (n=375,91) -Random sen...s (n=260) -"Distinction" plus honors seniors (n=39) -Distinction only seniors (n=44)	a) None
b) RC	b) Correlational, Course credits X Skills	b) Student volunteers (n=572)	b) None
c) UND	c) Non-equivalent comparison group, posttest only	c) Freshman volunteers (n=20) Senior volunteers (n=124)	c) None

Continued

Table 2: (Continued)

PURPOSE	Extent Implementation	Evaluation Designs	Student Outcome Samples	Program Portrayal Samples
	d) WCU	d) Non-equivalent comparison group, posttest only -Corelational, Year X Skills	d) Non-volunteers, (Psychology class), Fresh (n=62) Soph (44 native, 22 transfers)	d) Faculty questionnaire (n=181, 70% return)
	e) WKU	e) Non-equivalent comparison group, posttest only	e) Volunteer, from stratified random samples (n=56 fresh, 22 seniors)	e) None
IV	Enhance on-going Program Evaluation a)CS-C	a) -Pretest-Posttest (4 ys) -Non-equivalent comparison group, posttest only -Time series	a) Random native seniors 1980, n=30 Random native seniors 1983, n=30 Random freshmen, 1980 n=100 Random senior transfers, 1980 n=30 Random senior transfers, 1983 n=30	a) None

years, the actual amount of time available for testing was less than a year which restricted the use of more rigorous evaluation designs.

Samples. With respect to student sampling, seven institutions used volunteer samples (three of these paid a cash honorarium, while two others used "perks" such as meals or passes to plays or recreational events). Two teams used non-volunteers (by testing students during regular classtime). At one institution, faculty volunteers submitted their student's final exams, papers, and quizzes for external review. Three institutions attempted to use stratified random sampling or matrix sampling but found that the number of subjects in some of the cells was too small for analysis, and thus collapsed the sample into a single volunteer sample. Regarding the collection of data related to program portrayal, three institutions used faculty volunteers to complete questionnaires about their instructional practices and attitudes. As will be discussed later, obtaining representative samples of student cohorts proved to be a major difficulty in implementing the Paradigm.

Results of the analysis of the data. Four institutions found that generic skill performance on ETS General Education measures was related to length of time in school (i.e., seniors earned higher scores than juniors, who in turn earned higher scores than sophomores, and so on) and one found ETS tests did not differentiate between curricula or class membership. One institution, with a majority of students that could be called "adult learners", found that year in school was not related to generic skill performance on ETS General Education measures. Using factor analysis, this institution identified two factors - a multiple choice test factor and a performance test factor with grade point average loading on the multiple choice factor. Another institution found that, for freshmen, length of time in schools, not differences in number or kinds of courses, was related to generic skill performance on ETS measures. Two regression analyses revealed that once academic aptitude (eg. SAT) or academic performance (GPA) are included in an equation, little additional variance in generic skill performance is explained by the accumulation of credit hours. One institution identified a performance "gap" between observed and "expected" levels of performance in the area of writing skills using a locally developed composition test. At the time of this writing, two institutions either had not yet reported their findings or had decided not to release them. (See Table 3 on the next page.)

These "early returns", while certainly inconclusive, suggest the following: 1) ETS General Education measures which most institutions used may assess fundamental intellectual abilities or academic aptitude more than generic skills[3]; 2) ETS General Education measures may be highly sensitive to maturation during late adolescence - however, this effect may be influenced by experimental mortality or selection bias inherent in the evaluation designs; and 3) the general lack of adequate controls in the respec-

3. See Catell's (1971) and Horn's (1968) discussions of Fluid and Crystallized abilities.

Table 3: Extent of Implementation of Academic Program Evaluation: Stages IV - V

PURPOSE / Extent of Implementation	Results of the Analysis of Data	Outcomes: Policy Alternatives/ Implications	Outcomes: Procedural Recommendations
I Improve Curriculum			
(a) PSU	a) -Few statistical differences among curricula or classes -ETS was related to GPA and SAT scores -Faculty stress communication -Analysis and Synthesis more than Valuing -Faculty teach skills using primarily the content of their courses.	a) None yet reported	a) None
(b) WSC	b) Performance rating (1= not effective, 5=highly effective) by grade level, academic area, and general education for each skill -No conclusions made	b) Generic Skills should be nurtured in all courses and programs (implied)	b) Recommendations by VPAA 1. Rewrite course syllabi to reflect skill development 2. Develop "capstone" courses for interpretation of skills 3. Form ad-hoc committee to review skill development in general education 4. Evaluate course by course contribution to skills in general education 5. All programs articulate new skills engendered and measured
II Initiate Formal Evaluation Procedures			
(a) NASC	a) -Performance on tests was related to years in school -Posttest scores lower than pretest scores	a) Revise preamble to Gen Ed curriculum	a) Criteria for inclusion of courses in Gen Ed will include strategies for generic skill development

Continued

Table 3: (Continued)

PURPOSE / Extent of Implementation	Results of the Analysis of Data	Outcomes: Policy Alternatives/ Implications	Outcomes: Procedural Recommendations
(b) SIU-E	b) Seniors earned higher scores than freshmen on ETS Analysis/Synthesis and local Quantitative test.	b) None yet reported	b) None
III Exploratory Pulse Reading of Gen Ed			
(a) BSU	a) -Time, not courses, associated with increment in skill -No differences between honors & non-honors seniors in generic skills	a) -Continue to assess generic skills in freshmen and seniors -Revise Gen Ed program to include more structure -Enhance graduate research in undergrad instruction and curriculum development -Generic skills should become part of courses	a) Recommendation by APEP committee 1. Establish assessment Center 2. Include generic skills achievement in course objectives 3. Conduct faculty development workshops
(b) RC	b) -Positive correlation between credits and generic skills -GPA greatest predictor of generic skills -Credits account for little variance in regression analysis	b) Generic skills should become part of all courses (implied)	b) Recommendations by APEP committee 1. Each course should address critical thinking and communication 2. Conduct follow-up testing
(c) UNO	c) "Gap" identified in writing proficiency at senior level -ETS did not differentiate freshmen from seniors (adult learners).	c) Written communication should be emphasized across curriculum	c) Recommendations by APEP committee 1. Consider upper division writing proficiency requirement 2. Conduct follow-up writing test

Table 3: (Continued)

PURPOSE / Extent of Implementation	Results of the Analysis of Data	Outcomes: Policy Alternatives/ Implications	Outcomes: Procedural Recommendations
(d) WCU	d) -Sophomores earned higher scores than freshmen -3 semesters accounted for 0 - 7.5% of variance in generic skills. -Instructors exceeded "ideal" in portrayal dimensions.	d) -Policies should not be made on test evidence alone -Continue process of data gathering -Generic skills should become part of gen ed -Writing across the curriculum should be encouraged	d) Recommendations by APEP committee 1. Focus on skill identification and validity of measures 2. Investigate valuing dimension 3. Clarity performance standards 4. Create gen ed monitoring committee (APEP influenced) 5. Gen ed courses should address generic skill development
(e) WKU	e) None reported	e) None yet reported	e) None
IV Enhance On-going Program Evaluation (a) CS-C	a) None released (policy decision)	a) Generic skills are part of policy on goals of Gen Ed	a) Procedures adopted 1. Faculty must address generic skills in course syllabi in Gen Ed -Procedures recommended by Advisory committee 1. Junior level writing test 2. Information day for student testing

DR. HERBERT LUBS
REGIONAL GENETICS PROGRAM DIRECTOR
MAILMAN CENTER
1601 N.W. 12TH AVENUE
MIAMI, FL 33136

DR. JAIME FRIAS
DEPARTMENT OF PEDIATRICS
COLLEGE OF MEDICINE
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32601

DR. THOMAS A. TEDESCO
REGIONAL GENETICS PROGRAM
UNIVERSITY OF SOUTH FLORIDA
DEPARTMENT OF PEDIATRICS
BOX 15
TAMPA, FLORIDA 33620

DR. BORIS KOUSEF
REGIONAL GENETICS PROGRAM DIRECTOR
UNIVERSITY OF SOUTH FLORIDA
DEPARTMENT OF PEDIATRICS
BOX 15
TAMPA, FLORIDA 33620

tive methologies left the interpretations of data very tentative for all institutions. It may well be that greater care is required in developing and selecting measures and in formulating designs before there can be much confidence in the results.

Policy outcomes. At the time of this writing, six institutions have moved, or are planning to move, toward the adoption of statements alluding to generic skill development as goals of general education or other college programs. (See Table 3 on the previous page.) One institution is considering the adoption of policies regarding research efforts in course and curriculum development, continued assessment of generic skills and greater structure to is general education curriculum (even though the results of the analysis of their data found that generic skill development was independent of the degree of curricular structure). Two institutions are considering the adoption of a policy regarding emphasis on writing across all courses. Finally, one asserts that policies should not be made on the basis of test information alone.

Procedural outcomes. Six institutions are considering procedures requiring or encouraging faculty to include instructional objectives or strategies in their course syllabi. Four institutions are recommending procedures for further testing of students either in courses or programs. Only one institution is considering the adoption of procedures alluding to faculty development. Two institutions are considering the implementation of a writing proficiency requirement for passage to upper division. One is mapping out plans for further investigation into generic skill identification and measurement.

Discussion and Conclusions

Amid the data presented in Tables 1 through 3, several issues became paramount concerning the Paradigm as a set of procedures to structure the process of institutional inquiry leading to orderly and effective change. Among these are: Was the Paradigm implemented to such a degree so as to provide an indication of its utility? How valid and useful are the concepts which underlie the Paradigm? What are conceptual and operational limitations of the Paradigm for the variety of purposes for which it was employed? If the Paradigm provides a mechanism for observing and evaluating institutional performance, are there directions for further investigation that may contribute its utility and validity? Such questions structure the ensuing discussion.

First, Was there an APEP event?

Let us assume that in order to qualify as an "APEP event", an institution must have completed three tasks: 1) developed a set of generic skill definitions; 2) determined whether a program "performance gap" exists, and 3) if gaps were evident, formulated policies and procedures to rectify them. According to these criteria, one institution was able to closely approach an "APEP event" in the time alloted for the project. While insti-

tutions dealt with the definitions to some degree, only two were able to identify a performance standard with which to compare observed levels of performance. Furthermore, while six institutions have adopted or are considering adopting certain policies, a question remains concerning the extent to which the adoption of these policies was owed to having engaged in the first four stages of the Paradigm. Perhaps one year from now several more of the institutions, proceeding through the stages carefully and meticulously and who now have just completed stage III, or looping back to retrace steps, will eventually realize an "APEP event". Nevertheless, all institutions implemented parts of the Paradigm. Therefore, by "piecing together" the collective experiences of the 10 institutions, some inferences may be advanced about the utility and validity of the Paradigm.

The Paradigm: Theoretical Foundations. A formal theory may be considered as consisting of a set of assumptions, definitions, and operations which can be used for observing, describing, explaining, prescribing, or predicting phenomena. (See Wolman, B.B, 1973.) Taking first assumptions, at the present time, the developers and implementers (of which I am one) have not yet declared a set of assumptions on which the Paradigm is based. In this regard, what is assumed about the nature of outcomes of the higher educational experience? Are there, or could there be, a set of common "trans-disciplinary" outcomes which can serve as referents with which to compare student achievement across programs within institutions or between institutions? What philosophical propositions are made about the nature of the individual, programs and institutions of higher learning, and society to which common modes of thought are related? What is the relationship between generic skills and human performance? Does the Paradigm assume a completely rational, data-based approach to organizational decision-making?

With respect to definitions, two concepts may be considered vital to understanding and implementing the Paradigm: generic skill and performance gap. Could it be that describing the essential learning outcomes of baccalaureate education in terms of Communication, Analysis, Synthesis, Quantification, and Valuing today may be at the same stage of development in the evolution of classifications and taxonomies (according to Travers, 1980) as in midieval times when chemists classified all of matter in terms of earth, fire, air and water or oils, flowers and butters? The state of the art in defining educational outcomes may still be a far cry from today's atomic chart in Chemistry. In this vein, if the implementers had more time to deliver more deeply into their definitions and to try out their own measures of them, would the kinds of tests that were selected and implemented have been different? Would the implementers have relied so heavily on the use of tests prepared by commercial firms? Unfortunately, the project came to a close before such challenging questions could be deliberated and resolved.

Finally, the evaluation procedures set forth in the APEP Guidelines (Buhler-Miko,Peterson, and Stakenas, 1982) may yet undergo refinement after the assumptions and definitions on which it is based stabilize and become sharper and clearer. For example, it may be well to have faculty committees first develop direct measures (Stiggins, 1981; Sachse, 1981) of these

skills by actually observing designated cognitive processes under controlled conditions and then to have faculty identify, select, and validate indirect measures such as published multiple choice tests which consistently predict high and low performers on direct measures. By employing such a strategy, it can be documented that multiple choice tests are valid measures of generic skills and not primarily measures of academic aptitude or Spearman's g.

Regarding the concept of performance gap, setting performance standards a-priori to the administration of the tests appeared to be troubling as evidenced by the fact that no institution identified a gap between a desired level and observed by level of performance. One institution set "expected" (not desired) performance standards for student performance measures and another set "ideal" performance standards for program portrayal dimensions. As will be discussed later, many unresolved issues remain regarding the process of standard setting for generic skills and instructional practices.

Limitations on the Utility of the Paradigm

Some of the major limitations and constraints, in addition to the theoretical and conceptual difficulties discussed earlier, appear to be: 1) time and resources to conduct the inquiry and to develop policies and procedures; 2) the procurring of adequate population samples from which to draw inferences; and, 3) the drawing of logical conclusions from the analysis of the data on which to propose policies and procedures.

First, this author believes there is much more to APEP than simply purchasing tests on the basis of "title" or face validity, administering them to groups of freshmen and seniors and observing what happens from there. It is far more demanding than this. The proper implementation of the Paradigm requires that faculty devote time and effort to understand the nature of generic skills as outcome criteria and to relate them to the mission of the institution and to on-going instructional activities within courses and programs. APEP calls for faculty to be able to develop or select valid measures of these skills and to be able to come to some agreement in terms of desired performance standards. Faculty must be able to formulate a defensible evaluation design, to analyze the data appropriately, and to present the results of their inquiry in a meaningful and cogent manner. Then in the evaluation and action phase (Stages IV and V), faculty members and administrators must be able to work together to initiate and carry out policies and procedures to effect change while withstanding the stress of such "human" factors as suspicion, territorial imperatives and general resistance to change. Such accomplishments take both time and commitment, more than the 2½ years allotted to the present investigation. The fact that institutions were not able to reach higher levels of skill definitions, that only five used portrayal measures, and that only two identified an a-priori standard gives testimony to an insufficient amount of time to thoroughly work through the stages of the Paradigm. As a writer of one case history put it, "We ought to take four years with adequate resources and do it right."

Secondly, the procurring adequate population samples proved to be a major stumbling block. There appeared to be two successful strategies for soliciting students to take tests: either pay subjects on an hourly basis, (honoraria of $3 to $5 per hour were used), or use a mechanism for "capturing" students such as using regular class time, making testing a course requirement, or employing freshman orientation proceedings. Even the offering of "perks" such as tickets to plays or sports events or meals in the cafeteria proved to be unproductive. The least effective method was to appeal to students' "good will". This motive consistently resulted in less than a 20% response rate.

Finally, moving from empirical data about student performance to policy and procedural considerations appeared to be a difficult transition in the implementation of the Paradigm. This is perhaps not all that uncommon of a problem in evaluation which might be owed to the fact that policy considerations involve not only logical analysis, but social, historical, and political analyses as well (Baldridge, 1978, and Lindblom and Cohen, 1979). The complexities of this leap from data to policy were reflected in curious anomalies within the project itself. For example, one institution came to no conclusion regarding the performance of its students on generic skills, but nevertheless proposed a rather elaborate set of policies and procedures related to the fostering of generic skill development in courses and programs. Another found that student achievement on generic skills was unrelated to the degree of structure an individual's program of study in general education. Nevertheless, policies and procedures were proposed to impose greater structure on the distribution of kinds of general education courses students may take to fulfill their general education requirement. Could a "gap(s)" have been inferred so as to compel change? It appears that in order for the empirical data to have any relationship to or bearing on subsequent policy and procedural considerations, no matter how tentatively, speculations about the potential outcomes of the project should be discussed early, ostensibly in the clarification of project purposes. At this time, an institution may consider not only Why and What to evaluate, but also potential implications for change that might be reflected in the eventual adoption of policies and procedures. "Futures" scenarios are often effective in helping to identify potential project outcomes.

Future directions and unresolved issues

In the course of the conduct of the project a number of issues were raised by participants, members of the project staff and consultants. Many of the more fundamental questions related to the utility of the Paradigm concern the nature of generic skills, their measurement, and the concept of performance gap.

First, the nature of generic skills and their properties at the operational level appears to warrant further investigation. What is the relationship between the mastery of content and the demonstration of generic skills? Perhaps in order to be a generic thinker, one must first possess a mastery of a broad range of knowledge. How are generic skills different

from other intellectual abilities? Are they more or less subject to the accepted principles of learning (such as forgetting, extinction, operant and classical conditioning) than other intellectual abilities? Are there similarities between the development of psychomotor skills and the development of generic skills? To what extent do they conform to developmental phenomena such as staging or critical periods? What is the relationship between short term and long term memory and the demonstration of generic skills? Should generic skills be assessed using content that has already become part of long term memory (as in the ETS tests) or by supplying content and using short term memory (as in the case of ACT/COMP tests)? To what extent do intelligence factors and academic aptitude factors contribute to the demonstration of generic skills? Can generic skills be thought of in terms of the use of content in the service of intelligence? Are generic skills more than the idiosyncratic fusion of subject matter content and fundamental intellectual factors such as proportionate logic, controlling variables, syllogistic reasoning, and analogies? In the process of addressing such issues we may begin to understand more fully the relationship between instructional events in higher education, the development of "thinking skills," and the kinds of measures more suited to assess them.

There are also avenues of inquiry to be explored in the area of the measurement of generic skills. Can individual generic skills, as described in terms of Communication, Analysis Synthesis, Valuing etc., be validly assessed using a multiple choice test? Ostensibly, each multiple choice test item may be viewed as a problem solving task in its own right requiring the use of all generic skills in the identification of a correct response. (See Sternberg, 1980.) Each item requires that an examinee read the stimulus (Communication) and understand the requirements of the task (Analysis), consider alternative solutions (Synthesis), test each alternative against the conditions of the task and arrive at a best-fit solution (Valuing). Perhaps this is why one faculty group using factor analysis found that the all generic skills tests loaded on two factors: a performance test factor or a multiple choice test factor. The question is raised: How can an Analysis item, for instance, not also measure Communication, Synthesis, and Valuing at the same time? Must each item, say on an Analysis test, demonstrate that it discriminates between high and low Analyzers but not high and low Communicators, Synthesizers, and Valuers (to coin a term)? A test composed of a majority of such items would indeed be a challenge to develop. Nevertheless, if institutions desire the convenience and low cost of indirect, objectively scored measures, and if the Paradigm is to achieve a high degree of dissemination and usage, the difficult task of developing valid indirect measures that capture the "value-added" variance attributed to classroom instruction assumes high importance.

The concept of "performance gap" also demands further investigation. Tuscher (1971) found that the relationship between costs and educational achievement is not a linear relationship, but more in the form of an "S" curve. At certain ranges the investment of additional resources may result in the familiar economic principle of "diminishing returns". The

implication is that while some "gaps" may require few additional resources to produce appreciable gains, others may require much more depending on the relationship between resources and change in a particular skill. Without such knowledge of the relationship between performance and resources, implementers of the Paradigm may be left with attempting escalator strategies of trying the least costly intervention first, followed by the second, third, and so on.

An element in the "gap" mentioned above requires the identification of a desired standard of performance. What procedures might be employed to assist faculty in deriving a standard? Should faculty look outside the institution to ascertain the level required for entry jobs typically acquired by the graduates? Should a faculty look at the academic aptitude level of its student population and arrive at an estimate of a level that is feasible? Should faculty teams employ techniques such as Ebel's (1972) or Nedelsky's (1954) methods for setting standards when multiple choice tests are used? Or should faculty teams observe how other institutions perform on similar measures and then set standards by employing a "keeping-up-with-the-Joneses" ethos. One institution adopted the rule, the lowest senior should score no lower than the average freshman. Evidently this institution is pretty satisfied with its average freshman.

Conclusion

When this author was given the proposal to this project and asked to participate, a first-blush response was one of, "My gosh, another ambitious, well meaning, but too short and too underfunded FIPSE project." The challenge faced by this project is one we all face as researchers, administrators, and faculty in higher education. The Paradigm offers institutions and their faculty a procedure to structure inquiry into their academic programs, and to develop concrete steps to effect institutional change. The process is demanding and troubling questions are inevitably raised about the very purposes of higher education in contemporary society. The Paradigm compels faculty to contemplate the very mission of their institutions and the ways they intend to influence the growth of students. Even if clear and precise definitions have not yet been achieved, the measures questionably administered and the results ignored or misinterpreted, by merely providing a logical structure for faculty and for administrators to emerge from their departmental enclaves and daily routines to contemplate the broad questions of higher educaiton in new ways with new concepts, may well make the attempt to implement the Paradigm worthwhile. For the salutory benefit of the Paradigm may be not in products at demands, but in the process it compels, and likewise, not in the answers to the questions it addresses, but more in the questions it raises. Let us not forget an old adage that people are energized far more by a good question than by being given the correct answer.

BIBLIOGRAPHY

Baldridge, J.V., et.al. Policy making and effective leadership. San Francisco: Jossey Bass, 1978.

Buhler-Miko, M. (Ed.) Alternatives, Resource Center for Planned Change, American Association of State Colleges and Universities, 1979, 10-11.

Buhler-Miko, M., peterson, G.W., & Stakenas, R.G. APEP Guidelines. Resource Center for Planned Change, American Association of State Colleges and Universities (In print).

Campbell, D.T., & Stanley, J.C. Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963, 171-246.

Cattell, R.B. Abilities: Their structure, growth, action. Boston: Houghton-Mifflen, 1971.

Cronbach, L.J. Toward reform of program evaluation. San Francisco: Jossey Bass, 1980.

Ebel, R.L. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice Hall, 1972.

Gagne, R.M. Learning hierarchies. Educational Psychologist, 1968, 6, 1-9.

Horn, J.L. Organization of abilities and the development of intelligence. Psychological Review, 1968, 75, 242-259.

Kaufman, R. Educational system planning. Englewood Cliffs, N.J.: Prentice Hall, 1972.

Kaufman, R., & English, F. Needs assessment: Concept and application. Englewood Cliffs, N.J.: Educational Technology Publications, 1979.

Lindblom, C.E., & Cohen, D.K. Useable knowledge: Social science and social problem solving. New Haven: Yale University Press, 1979.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Peterson, G.W., & Watkins, K. Florida competency-based articulation project: Final report. Florida Competency-based Articulation Project. Resources in Education (ERIC), ED169839, September, 1979.

Peterson, G.W., & Stakenas, R.G. On moving from defining generic skills to designing assessments for them. Occasional paper, Resource Center for Planned Change, American Association of State Colleges and Universities, 1980.

Sachse, T.P. The role of performance assessment in tests of problem solving ability. Clearinghouse for Applied Performance Testing: Northwest Regional Laboratories, Portland, Oregand, April 1981.

Stake, R.E. The countenance of educational evaluation. Teachers College Record, 1967, 68, 523-540.

Sternberg, R.J. Factor theories of intelligence are all right almost. Educational Researcher, 1980, 9(8), 6-13.

Stiggins, R.J. An analysis of direct and indirect writing assessment procedures. Clearinghouse for Applied Performance Testing: Northwest Regional laboratories. Portland, Oregon, April 1981.

Travers, R.M.W. Taxonomies of educational objectives and theories of classification. Educational Evaluation and Policy Analysis, 1980, 2(2), 5-23.

Tuckman, B.W. Conducting educational research. New York: Harcourt Brace Jovanovich, 1978.

Tuscher, L.J. Cost-utility analysis: An evaluation and resource-allocation model for educational programs. Dissertation, Florida State University, 1971.

Warren, J.R. Describing college graduates in 87 phrases or less. Findings, 1976, 3(2), 5-8.

Warren, J.R. Alternatives. Resource Center for Planned Change, M. Buhler-Miko (Ed.), American Association of State Colleges and Universities, 1979, 20.

Warren, J.R. ETS Measures of General Education, Berkeley, CA: Educational Testing Service, 1980.

Woditsch, G. Developing generic skills: A model for competency-based general education. CUE Project, Bowling Green State University, Bowling Green, Ohio, May 1977.

Wolman, B.B. Concerning psychology and the philosophy of science. In Handbook of General Psychology, B.B. Wolman (Ed.). Englewood Cliffs, N.J.: Prentice Hall, 1978, 22-48.