Evaluation quality: an analysis of the Emergency School Aid Act evaluation work plans.

Lawrence Bussey
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_1

Recommended Citation

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
EVALUATION QUALITY: AN ANALYSIS OF THE EMERGENCY SCHOOL AID ACT EVALUATION WORK PLANS

A Dissertation Presented
By

LAWRENCE BUSSEY, JR.

Submitted to the Graduate School of the University of Massachusetts in partial fulfillment of the requirements for the degree of DOCTOR OF EDUCATION

December 1975

ADMINISTRATION
EVALUATION QUALITY: AN ANALYSIS OF THE EMERGENCY SCHOOL AID ACT EVALUATION WORK PLANS

A Dissertation

By

Lawrence Bussey, Jr.

Approved as to style and content by:

Dr. David G. Coffing, Chairman of Committee

Dr. Ernest Washington, Member

Dr. Alan I. Gross, Member

Dr. John A. Finger, Jr., Member

Louis Fischer, Acting Dean
School of Education
EVALUATION QUALITY: AN ANALYSIS OF
THE EMERGENCY SCHOOL AID ACT
EVALUATION WORK PLANS
DECEMBER 1975

By

Lawrence Bussey, Jr.

Abstract

This study uses data collected in the fall and spring of 1974 on a panel review of evaluation work plans and an Evaluation Procedure Survey (EPS) in order to determine the relationship between the two scores and the type and stage of development of the project, the size of the project, the type of decision required, and the cost of the evaluation.

The population of interest of this study consists of 54 Emergency School Aid Act school districts and nonprofit organizations in the six New England states.
The procedures are described for collecting and analyzing the data.

The results showed the following:

**Hypothesis I**

The total Evaluation Procedure Survey score is related to panel ratings of evaluation quality. The correlations between panel ratings and each of the background factors type of grant, size of project, evaluation cost and evaluation use were not statistically significant.

**Hypothesis II**

There is a significant difference among the types of grants in terms of panel ratings, EPS score, size of project, evaluation cost, and evaluation use. Of these variables, the most discriminating in describing the four types of grants were panel ratings, size of the project, and EPS Score. However, further analysis showed that these three variables were not statistically significant in distinguishing among the four groups.

**Hypothesis III**

There is a significant relation between the EPS score and panel ratings after the effects of the size of project and evaluation cost were taken into account.
However, only evaluation cost added significantly to the multiple correlation between the three predictor variables and panel ratings.

**Additional Analysis**

The criterion related validity of the EPS scores was reflected in the statistically significant ($p < .03$) correlation with panel ratings. The survey's reliability was measured through the use of the coefficients of reproducibility and scalability, in terms of the six evaluation components. Four components (33%) met the stated criteria of unidimensionality and cumulativeness. The EPS instrument was moderately reliable but sufficient for measuring the gross indicators of this study.

Overall the EPS is a fair predictor of panel ratings. The correlation of .29 accounted for 9% of the variance between EPS scores and panel ratings. However, the multiple correlation between panel ratings and five predictor variables was .49 which accounted for 24% of the variance.

The following findings of this study conflict with the recommendations of the Urban Institute's study on federal evaluation policy. Proportionately, the evaluation costs of a project servicing 5,000 students were similar to one that serviced 100. Secondly, the
study showed that project impact data could be obtained at a lower cost than project strategy information, scores on the EPS tend not to be related to evaluation costs or the number of students to be evaluated.

This study also showed that the credibility of the methodologies of local evaluations would tend not to be related to increased costs for evaluations or to the size or type of grant of the program but to the direct improvement of specific procedures identified in this study.
ACKNOWLEDGMENTS

The author would like to express his thanks and appreciation to the following people for their contribution to the completion of this dissertation.

To Dr. David G. Coffing, for his moral support and technical guidance in the development of this dissertation.

To Dr. Ernest Washington, for his advice and careful reading of the substantive issue of this document.

To Dr. Alan J. Gross, for his assistance and advice in statistical techniques.

To Dr. John A. Finger, Jr., for his support and advice in data preparation and computer programming.

Finally, to my wife, Ellen, for her patience, understanding, and editorial review of the initial documents.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td></td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>I Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Overview</td>
<td>1</td>
</tr>
<tr>
<td>Background of the Study</td>
<td>3</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td>6</td>
</tr>
<tr>
<td>Objectives of Study</td>
<td>9</td>
</tr>
<tr>
<td>Conceptual Background</td>
<td>10</td>
</tr>
<tr>
<td>Basic Assumptions</td>
<td>12</td>
</tr>
<tr>
<td>The Hypotheses of the Study</td>
<td>12</td>
</tr>
<tr>
<td>Procedures of the Study</td>
<td>14</td>
</tr>
<tr>
<td>Organization of Study</td>
<td>15</td>
</tr>
<tr>
<td>II REVIEW OF RELATED LITERATURE</td>
<td>17</td>
</tr>
<tr>
<td>Introduction</td>
<td>17</td>
</tr>
<tr>
<td>Educational Evaluation Defined</td>
<td>18</td>
</tr>
<tr>
<td>Evaluation Models</td>
<td>21</td>
</tr>
<tr>
<td>Evaluation Procedures</td>
<td>29</td>
</tr>
<tr>
<td>Problems with Educational Evaluation</td>
<td>32</td>
</tr>
<tr>
<td>Federal Evaluation Policy</td>
<td>35</td>
</tr>
<tr>
<td>CHAPTER</td>
<td>REVIEW OF RELATED LITERATURE (Continued)</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>II</td>
<td>Concluding Remarks . . . . . . . . . . . 39</td>
</tr>
<tr>
<td>III</td>
<td>STUDY DESIGN AND PROCEDURES . . . . . . 41</td>
</tr>
<tr>
<td></td>
<td>Overview . . . . . . . . . . . . . . . . . 41</td>
</tr>
<tr>
<td></td>
<td>Research Design . . . . . . . . . . . . . 41</td>
</tr>
<tr>
<td></td>
<td>Data Collection . . . . . . . . . . . . . 42</td>
</tr>
<tr>
<td></td>
<td>Sample Selection . . . . . . . . . . . . 45</td>
</tr>
<tr>
<td></td>
<td>Questionnaire Development . . . . . . . 46</td>
</tr>
<tr>
<td></td>
<td>Data Processing and Analysis Procedures . 53</td>
</tr>
<tr>
<td></td>
<td>Limitations . . . . . . . . . . . . . . . 55</td>
</tr>
<tr>
<td>IV</td>
<td>FINDINGS OF STUDY . . . . . . . . . . . . 57</td>
</tr>
<tr>
<td></td>
<td>Overview . . . . . . . . . . . . . . . . . 57</td>
</tr>
<tr>
<td></td>
<td>Major Hypothesis I . . . . . . . . . . . . 57</td>
</tr>
<tr>
<td></td>
<td>Major Hypothesis II . . . . . . . . . . . . 61</td>
</tr>
<tr>
<td></td>
<td>Major Hypothesis III . . . . . . . . . . . . 66</td>
</tr>
<tr>
<td></td>
<td>Additional Analysis . . . . . . . . . . . . 69</td>
</tr>
<tr>
<td>V</td>
<td>SUMMARY AND CONCLUSIONS . . . . . . . . . 100</td>
</tr>
<tr>
<td></td>
<td>Overview . . . . . . . . . . . . . . . . . 100</td>
</tr>
<tr>
<td></td>
<td>Summary . . . . . . . . . . . . . . . . . . 100</td>
</tr>
<tr>
<td></td>
<td>Analysis of Hypothetical Model . . . . . . 105</td>
</tr>
<tr>
<td></td>
<td>Conclusions . . . . . . . . . . . . . . . . 114</td>
</tr>
<tr>
<td>Chapter</td>
<td>Page</td>
</tr>
<tr>
<td>------------------</td>
<td>------</td>
</tr>
<tr>
<td>APPENDIX A</td>
<td>120</td>
</tr>
<tr>
<td>APPENDIX B</td>
<td>126</td>
</tr>
<tr>
<td>APPENDIX C</td>
<td>129</td>
</tr>
<tr>
<td>APPENDIX D</td>
<td>130</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>133</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>SUMMARY OF INDEPENDENT AND DEPENDENT VARIABLES</td>
</tr>
<tr>
<td>II</td>
<td>NUMBER OF BASIC, NPO, PILOT, AND BILINGUAL GRANTS, BY STATE</td>
</tr>
<tr>
<td>III</td>
<td>MEASUREMENT QUALITY</td>
</tr>
<tr>
<td>IV</td>
<td>PRETEST OF EPS ITEMS 12 AND 18 BY FREQUENCY OF RESPONDENTS</td>
</tr>
<tr>
<td>V</td>
<td>CHI SQUARE PANEL RATINGS WITH TYPE OF GRANT, SIZE OF PROJECT, EVALUATION COST, EVALUATION USE, AND EVALUATION PROCEDURE SURVEY SCORE</td>
</tr>
<tr>
<td>VI</td>
<td>NON-PARAMETRIC CORRELATIONS</td>
</tr>
<tr>
<td>VII</td>
<td>DISCRIMINANT ANALYSIS BY FOUR TYPES OF GRANTS</td>
</tr>
<tr>
<td>VIII</td>
<td>MULTIPLE CORRELATION OF PANEL RATING WITH TYPE OF GRANT, SIZE OF PROJECT, EVALUATION COST, EVALUATION USE, AND EVALUATION PROCEDURE SURVEY SCORE</td>
</tr>
<tr>
<td>IX</td>
<td>EVALUATION PROCEDURE SURVEY</td>
</tr>
<tr>
<td>X</td>
<td>STEPWISE REGRESSION PREDICTION OF PANEL RATING FROM 36 PREDICTOR VARIABLES</td>
</tr>
<tr>
<td>XI</td>
<td>SUMMARY GUTTMAN SCALES: OBJSCALE, MEASCALE, QUASCALE, EVSCALE, STASCAL AND ADMSCALE</td>
</tr>
<tr>
<td>XII</td>
<td>TWELVE FACTORS EXTRACTED AFTER FIVE ITERATIONS OF EIGENVALUES AND PERCENT OF VARIANCE</td>
</tr>
<tr>
<td>XIII</td>
<td>PEARSON CORRELATION COEFFICIENTS, PANEL RATING WITH EVALUATION PROCEDURE SURVEY SCORE</td>
</tr>
</tbody>
</table>
### TABLE

<table>
<thead>
<tr>
<th>IVX</th>
<th>CORRELATION COEFFICIENTS FOR PANEL RATING, EPS SCORES, EVALUATION COST AND SIZE OF PROJECT</th>
<th>106</th>
</tr>
</thead>
<tbody>
<tr>
<td>XV</td>
<td>MULTIPLE REGRESSION OF EPS ON SIZE OF PROJECT AND EVALUATION COST: AND PANEL RATING ON EPS, SIZE OF PROJECT, AND EVALUATION COST</td>
<td>109</td>
</tr>
<tr>
<td>XVI</td>
<td>FREQUENCY DISTRIBUTIONS OF PANEL RATING, TYPE OF GRANT, SIZE OF PROJECT, EVALUATION COST, EVALUATION USE, AND EVALUATION PROCEDURE SURVEY SCORE</td>
<td>115</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>PLOTS OF GROUP CENTROIDS AND LOCATIONS</td>
</tr>
<tr>
<td>II</td>
<td>FREQUENCY DISTRIBUTION OF PANEL RATINGS</td>
</tr>
<tr>
<td>III</td>
<td>FREQUENCY OF ITEM RESPONSES FACTORS GOAL SPECIFICATION, MEASUREMENT AND DATA COLLECTION, MEASUREMENT QUALITY, EVALUATION DESIGN, STATISTICAL ANALYSIS, AND ADMINISTRATION OF EVALUATION PROCEDURES</td>
</tr>
<tr>
<td>IV</td>
<td>CORRELATION COEFFICIENTS FOR PANEL RATINGS, EPS SCORES, EVALUATION COST, AND SIZE OF PROJECT</td>
</tr>
<tr>
<td>V</td>
<td>PATH COEFFICIENTS FOR PANEL RATING, EPS SCORES, EVALUATION COST, AND SIZE OF PROJECT</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

Overview

This study focuses on examining the evaluation work plans proposed by local school districts and nonprofit organizations in New England. The evaluation work plans were developed by these organizations to evaluate the impact and effectiveness of their Emergency School Aid Act project and submitted to a technical review panel for a quality rating.

However, the questions of the validity of self evaluation work plans for generating objective data for local and national decision making have been challenged by many scholars in this field. These procedures may have corroded the credibility of evaluation as an instrument of policy in the views of many researchers.

Despite these serious questions, the federal government has continued to require evaluation of each funded program and project. The ESAA program has established

---


2 Ibid.

guidelines for project evaluation and by comparison, these requirements are more rigorous than other educational programs. The lack of expertise, on the local level, to carry out evaluation activities has been recognized by Stufflebeam, Scriven, and others.

This study will look at the quality of evaluation work plans of New England ESAA projects, as measured by panel ratings and scores on an Evaluation Procedure Survey (EPS) in relation to the type and stage of development of the project, the size of the project, the type of decisions required, and the cost of the evaluation.

---


Background of the Study

The quality of local project evaluation findings has been questioned by the public and professional educators. The Office of Education has challenged the findings of the Title I projects, "... because the findings were based upon teacher opinion." It also noted that other educational programs such as the Emergency School Aid Program produced questionable findings, "... because no control schools were visited to determine what changes would have occurred in schools if they had not had the impact of ESAP funds."

The lack of credible information obtained by local projects may be due to a number of factors, including limited resources, authority or the lack of inventiveness of evaluators. Local school administrators often engage in evaluation efforts to fulfill some federal or state requirement for funding. Airasian noted that:

Many teachers, as well as administrators, view formal evaluation and evaluators as being unnecessary, counter-productive, and alien. Formal evaluation studies are often perceived to be threatening, and they are unwelcome, which makes it difficult to

---


7 Ibid, p. 74.
obtain cooperation and assistance from either teachers or administrators.  

Other constraints in designing local evaluation work plans often include the existence of applicable theory, methods or educational models, sufficient time and money to carry out the evaluation, the availability of appropriate comparison or controls groups and the ability to collect the required data.  

Wholey also indicated that the federal government has not established adequate standards for evaluation planning, implementation, and dissemination.

However, Congress has taken steps to require the Office of Education to conduct periodic evaluations of national and local programs for, "... planning for the succeeding year for any such program ..." The passage of

---


10 Ibid.

this evaluation requirement by Congress has signaled the increasing awareness of the importance of formal evaluation of national and local programs and projects. Nonetheless, to date the reports of the American Institutes of Research and the Urban Institute are two of the most significant studies funded to examine the adequacy of local and national evaluation strategies and methods. The present study attempts on a limited scale, to bridge this gap.

A final word is needed on the value of additional research on project evaluation. The author accepts as valid the worth of evaluation activity in regard to improving the quality of education, but recognizes the many limitations. First, some educators have argued that the very imperfect nature of behavioral science data reduces the validity of measures and interpretation to near absurdity. Second, evaluation theory, research, and practice are in their beginning stages. Traditional experimental research techniques may be grossly inappropriate for in situ public educational programs. Third, one could assume that countless final evaluation reports gather dust on some dark shelf of federal and state

officials. Evaluation is often perceived as a pejorative and arcane term for classroom teachers and synonymous with program audit for administrators.

We should ask the question, are the benefits derived from program evaluation commensurate with the costs?

Statement of the Problem

Millions of dollars are spent each year on noncomparable self evaluations of local projects, such as the Emergency School Aid Act. All projects do not require the same level of funds for evaluations. The appropriate level of evaluation funding is best expressed as a ratio of the total project costs in relation to the following considerations:

The type of project and its stage of development effect the cost and the credibility of the final results. Pilot or experimental projects tend to be more difficult to evaluate than well established projects that have proven their value and thus require limited evaluation. For this

---

13 It is estimated that over $400,000 was requested in 1974 in the ESAA program in New England for Individual project evaluations. This amount represented approximately 13% of the total funds requested.
study, there are four types of ESAA projects: Basic, Non-Profit, Pilot, and Bilingual.

The number of project participants tends to effect the level of difficulty of evaluation procedures and costs. On a percentage basis, smaller projects tend to cost more to evaluate than larger projects.

The types of decisions that evaluation findings will be used tend to influence evaluation costs. For example, the evaluation costs of programs like Follow Through, Head Start, and Manpower Development and Training Programs tend to be higher than Title I, ESEA-Vocational Education, and various programs funded under the Economic Opportunity Act.14

The cost of evaluation is a prime consideration in examining ESAA work plans. However, the most important factor in defining the potential quality of evaluation results is related to the technical review panel ratings.

Each ESAA application is reviewed in its entirety by panels of experts in school desegregation. Members of the panels included teachers, superintendents, state officials, university professors, civic leaders, and educational

---

consultants. The panel members received training in rating each evaluation work plan, using the following criteria established for the program: 15

1. Objective measurement of project impact.
2. Management procedures for collecting and reviewing evaluation data.
3. Description of the evaluative instruments.
4. Methods for validating these instruments.
5. Use of norms, comparison, or control groups.

Each panelist received and used an evaluation quality rating sheet with points assigned to each of the above criteria. The maximum possible score for each applicant was 24 points.

A review of a sample (52) of scores indicated a mean of nine points which reinforces the assumption of low quality of the evaluation plans, as rated by the panel.

A follow-up study in the spring of 1974 was conducted and additional data was obtained through the Evaluation Procedure Survey, a self rating instrument that assesses the level of specific evaluation procedures.

Data was compared by matching panel ratings and EPS Scores for each applicant. Several questions were raised in examining this data. First, is there a relationship between panel and EPS scores? Are other indicators of evaluation quality, such as cost, size, type of project, or type of decision related to panel ratings? Are measurement errors and unmeasured variables more influential in effecting the panel scores than EPS scores or other indicators of evaluation quality?

The present study will examine the above questions to determine if there is a systematic relation between panel and EPS scores and whether the relationship is statistically significant after the scores of the type of project, size of project, evaluation cost, and evaluation use have been taken into account.

Objectives of Study

The two major objectives of this study are related to examining evaluation procedures in order to determine if one is a good predictor of the other. First, if EPS scores prove to be a good predictor, i.e., significant regression coefficient, of panel ratings, technical assistance could be given to each applicant before being reviewed. The EPS, if validated, could be a handy guide for local projects in developing evaluation work plans.

The second objective focuses on the need to build a better understanding of how evaluation costs, size of project,
type or project, and type of decision factors are related to
evaluation quality, as measured by panel ratings and EPS
scores.

Conceptual Background

The Urban Institute study of evaluation policy pro-
vided the major conceptual background for this study. The
writings of Stufflebeam, Scriven, and Stake were consulted
in the literature review. The Urban Institute's study noted five critical
areas that effect the level of funding and quality of evalua-
tions.

1. The stage of development of the program.
2. The size of the program.
3. The types of decisions required.
4. The probability that the findings will be used.
5. The availability of methodology.

16 Ibid, Wholey
Op. cit., Scriven
R.E. Stake, "Objectives, Priorities and Other
If one were to include these factors within a simplified model of the relationship between panel ratings and EPS scores, an hypothetical model would look thus:

\[ X_1 \rightarrow X_4 \rightarrow X_2 \]
\[ X_3 \rightarrow X_2 \]
\[ X_5 \rightarrow X_2 \]
\[ X_6 \rightarrow X_2 \]

\[ X_1 = \text{type of project; } X_2 = \text{evaluation cost; } X_3 = \text{size of project; } X_4 = \text{evaluation use; } X_5 = \text{EPS score; and } X_6 = \text{panel ratings.} \]

The cost of evaluation is positively affected by the type of project and evaluation use and negatively affected by the size of the project. Evaluation costs are directly or indirectly related to panel scores. Evaluation procedures are directly related to panel ratings.

Intuitively, one could say that a work plan that receives a higher panel and EPS score would cost more to implement than one that received a lower panel score.

There would be \( X_u \) and \( X_w \) variables effecting panel ratings, \( X_6 \) as measurement error and unmeasured variable. The size of variables \( X_u \) and \( X_w \) should be small in comparison to evaluation cost, \( X_2 \) and EPS scores, \( X_5 \).
As stated as this study's second objective, this hypothetical model will be tested with the data obtained in the survey so that inferences could be made about the variables included in this study.

Basic Assumptions

It is assumed that a mean change in panel scores, for a given change in EPS score, is the same as the change that would always occur if all other variables could be controlled.

It is also assumed that all other variables included in this model have been controlled or do not vary. The effects of the confounding variables, type of project, evaluation cost, size of project, and type of use are negligible.

Finally, it is assumed that measurement errors in the Evaluation Procedure Survey are negligible. Whatever, outside factors influencing panel scores are creating variations that are completely unrelated to that produced by EPS scores.

The Hypotheses of the Study

The following section will define the formation of this study's three major hypotheses.
Hypothesis I will test whether panel scores are significantly related to Evaluation Procedure Survey scores. The criterion related validity of EPS is important in establishing the need to further analyze the data. The second hypothesis will be tested to determine whether the four subgroups of ESAA applicants require separate analysis as a result of major discriminating variables. The last hypothesis will be tested to determine whether evaluation procedures (EPS) contributes significantly to the multiple correlation of panel ratings with the other four predictor variables.

Hypothesis I  The total Evaluation Procedure Survey score is significantly related to the panel ratings of evaluation quality.

Hypothesis II  There is a significant difference among the types of grants in terms of panel ratings, EPS score, size of project, evaluation cost, and evaluation use.

Hypothesis III  There is a significant relationship between the total Evaluation Procedure Survey score and panel ratings after the effects of evaluation costs, size of project, type of grant, and evaluation use factors have been taken into account.
Procedures of the Study

The entire ESAA population of 52 applicants were rated by the technical review panel in February 1974. In the spring of 1974 additional data was obtained through the use of an Evaluation Procedure Survey, a self rating questionnaire.

Panel ratings for each applicant were matched to the EPS scores and on the other descriptor variables, type of project, cost of evaluation, size of project, and evaluation use.

The EPS consisted of 32 self descriptive statements which described a desired level of evaluation activity. Each response ranged from 1-4; 1=none; 2=some; 3=many; and 4=all of the objectives. Four is the highest level of response. Within the 32 items in the EPS there were six subscales, each measuring different components of the evaluation work plan.

Research Design  The research design consists of a cross sectional analysis of the scores of 52 ESAA applicants, as measured by the Evaluation Procedure Survey. Although the EPS scores were obtained after the panel ratings, no temporal sequence is inferred. Additional indicators of evaluation quality were included in the analysis to determine
if the correlation between panel ratings and EPS scores is spurious.

**Statistical Analysis** The statistical methods used in this study included Pearson's product moment and Spearman's nonparametric correlations, multiple regression, factor and discriminant analyses. The p-values for all correlation coefficients and F-ratios were set at p=0.05 level of significance.

One dependent variable, panel ratings was examined in relationship to five independent variables, EPS score, type of project, evaluation cost, size of project, and type of decision. An additional analysis of the criterion related validity and the internal reliability of the EPS was conducted. Descriptive statistics on panel ratings and EPS scores were also reported.

**Organization of Study**

The remaining chapters include a review of related literature, research procedures, findings, and conclusions.

Chapter II includes a discussion of current literature on educational evaluation goals, purposes and methodologies to familiarize the reader with substantive issues on this topic.
Chapter III covers sample selection, research procedures and design, instrument development and statistical analysis. This chapter also provides extensive discussion of the reliability and validity of the EPS.

Chapter IV presents the major findings in terms of the variables included in this study and a review of reliability and validity factors.

Chapter V includes a summary, discussion, and recommendation for future research and an overall review of the objectives, procedures, and findings of this study.
CHAPTER II
REVIEW OF RELATED LITERATURE

Introduction

The goals of this chapter are twofold. First, the review of literature on educational evaluation focuses on defining some characteristics of evaluation quality, i.e., some plausible models. Second, the review identifies some factors related to the feasibility of conducting evaluation studies. The review is not intended to be a comprehensive critique of current evaluation theory, research, and procedure, but a report on current applications.

The purposes of this chapter are to report on the substantive issues related to educational evaluation, as opposed to a historical review, and to identify common themes in the literature. The major topics discussed are the definitions and problems of educational evaluation and federal evaluation policies and procedures.

Educational Evaluation Defined

The attention of scholars towards educational evaluation theory, research and procedures has grown over the past 15 years. Much of the growth has not reduced the controversy among evaluation researchers nor adequately defined functional areas.
In the past, evaluation was equated with test and measurement. Later, evaluation was made synonymous with judgments about the quality of schooling.

These narrow definitions were challenged by pressing public demands to improve education and to provide equal opportunities for disadvantaged children. Evaluation's focus changed from evaluating students to assessing the factors related to school success.

Merwin observed that:

"Concepts of evaluation have changed over the years. They have changed in relationship to such issues as who is to be evaluated, what is to be evaluated, and how evaluations are to be made."

---


More recently, attention has been focused on the goals of educational evaluation as an element of the decision making process. Unfortunately, superfluous or inappropriate information is collected without the establishment of adequate criteria for decision making.

The definition of evaluation's goals and purposes have come to be accepted by some researchers as a means to make decisions about program improvement based upon an assessment on ongoing program activity. Stufflebeam has defined evaluation as, "... the provision of information through formal means, such as criteria, measurement, and statistics, to provide rational bases for making judgments which are inherent in decision situations." 22

Stake has expanded this definition of evaluation to include a recognition of the role that values and value systems play in the decision making process. 23

The last characteristic needed to define educational evaluation is related to the methods used to plan, collect,


and analyze the information identified as important to the consumer, i.e., local, state, or federal agency. Wholey noted that evaluation, "... relies on the principles of research design to distinguish a program's effects from those of other forces working in a situation ..."\(^{24}\)

We have defined educational evaluation theory in terms of goals, purposes, and methods.

Evaluation goals are defined along several dimensions, i.e., short term or long range objectives, aims or directions. According to Stake, the goal of evaluation is to assign value or cost to the outcomes of an educational endeavor for its improvement, modification, or termination\(^ {25}\). Evaluation goals are for educational decision making.

Evaluation purposes are related to assessing the merits or costs of various educational strategies to make decisions about the context, input, process, or outcome of these efforts.\(^ {26}\) Evaluation's prime function is to assess


\(^ {25}\)Stake, p. 182.

\(^ {26}\)Stufflebeam, p. 53.
worth. According to Wholey, evaluation methods are defined as systematic procedures for assessing the value or cost of some educational endeavor for the purpose of decision making.27

The following section will review specific evaluation methods that are current in theory and practice.

Evaluation Models

As previously stated, evaluation is conducted to assess the value or cost of various educational strategies aimed at decision making.

It is a form of scientific inquiry which is significantly different from classical research in several ways. Pure research is based on hypothetico-deductive methods; it is founded in theory, aimed at theory testing and explanation.28 Evaluation is more often decision-oriented and founded in value systems and judgments about

27 Wholey, p. 23.

desired ends. Evaluative studies that have attempted to use traditional research models have often produced uninterpretable and costly information.

Types of Evaluation Models Evaluation models can be differentiated by their content and goals. For example, curriculum evaluation's purpose is to assess the effectiveness of its design and content. Evaluation of institutions is focused on determining effects of various strategies on some educational outcome.

Guba and Stufflebeam have identified four types of goals for decision making in their CIPP model.29

Context evaluation serves decision making for the planning of an educational program. It is similar to needs assessment strategies and discrepancy evaluation. Input evaluation is focused on making program goals operational. It is an assessment of human and material resources and the feasibility of achieving the stated goals. Process evaluation serves to determine the quality of the ongoing program. Process evaluation, as defined in the CIPP model is similar to Bloom's formative model.30 Product evaluation serves to


measure the impact and effectiveness of the project on achieving some stated objective. Again, Bloom has defined product evaluation in terms of an assessment of an ongoing program in achieving its stated objectives. Both product and summative evaluations are aimed at providing information towards those who set policy or fund programs. Another model, Curriculum evaluation varies in terms of assessments of student learning in developed courses of study. Cronbach reported on new curriculum efforts and their effects on student outcomes in comparison to established programs. 31 Scriven focused his attention on clarifying the role of the evaluator to be sensitive to relative outcomes across many dimensions. He also stressed that evaluators cannot avoid making "goal free" judgments about a program's worth. 32 Stake theorized that judgments about a program's worth must consider the context of the program and students:

31 Lee J. Cronbach, "Course Improvement Through Evaluation." Teaching College Record, 1963, 64, pp. 672-683.

serves to measure the impact and effectiveness of the project on achieving some stated objective. Again, Bloom has defined product evaluation in terms of an assessment of an ongoing program in achieving its stated objectives. Both product and summative evaluations are aimed at providing information towards those who set policy or fund programs. Another model, Curriculum evaluation varies in terms of assessments of student learning in developed courses of study. Cronbach reported on new curriculum efforts and their effects on student outcomes in comparison to established programs. Scriven focused his attention on clarifying the role of the evaluator to be sensitive to relative outcomes across many dimensions. He also stressed that evaluators cannot avoid making "goal free" judgments about a program's worth. Stake theorized that judgments about a program's worth must consider the context of the program and students:

31 Lee J. Cronbach, "Course Improvement Through Evaluation." Teaching College Record, 1963, 64, pp. 672-683.

"Evaluation requires judgment. Decision-making requires judgment. Both are judgmental in themselves, but also depend on judgments previously made. A school and a curriculum are where they are because of judgments from within and from without. Judgments are made early, and late, and in between times. To understand what a school is doing requires an understanding of what a school is expected to do."\(^3_3\)

Curriculum evaluation theories have focused equal attention on the process of implementing a new program as well as an assessment of learner outcomes. Many well developed curriculi become failures because teachers and students deliberately sabotaged them.

Provis reported that evaluations should focus on the discrepancies between program plans and actual program operations.\(^3_4\) Discrepancy evaluation uses descriptive and case study methods as well as experimental. Several interpretations of summative evaluations have ignored the fact

\(^{33}\)Stake, p. 181.

that planned program activities were never implemented. In this regard, discrepancy evaluation studies are similar to program evaluation review techniques (PERT) and management-by-objectives. Discrepancy evaluation may also be defined as determining the difference between pre- and post-tests, needs assessment, and goal discrepancies.  

Discrepancy evaluation has been generally applied to studies of comparisons between performance and planned design. It is akin to process monitoring of program activities. Suchman's level of effort criteria for evaluation is similar to discrepancy evaluation in terms of process monitoring.  

Recent discussions of the accountability in public education have drawn attention to cost analysis as an evaluation consideration. Cost analysis theories are based upon system theories in that the benefits of programs are determined by the investments made. In the traditional use of cost analysis in educational evaluation, Bowles related the production functions in terms of inputs and


outputs. The difficulty in Bowles' theory is in determining benefits of educational outputs, such as reading achievement on high school graduation. Bowles' theories of educational production function are grounded in economic theory of the firm. When this theory is applied to educational evaluation, it is necessary to identify the commodities and to indicate the extent to which the educational enterprise is efficient. Opportunity costs of an educational output must always be weighted against the options available. Productivity criteria for evaluation is the ratio of output to input or the increase in output per unit increase in input.

Many evaluators have tried to include costs as an important variable in program evaluation. However, the theories are generally fairly simple. Cost and systems


analysis may provide the kinds of gross indicators of quality that public decision makers are interested in.

A fifth model included in this review is generally termed case study method. Case study models are almost always made synonymous evaluative studies for accreditation. Methodology for inquiry for this model may include testing of students, classroom observations, content analyses of records, and interviews with participants. The strength of this model of institutional evaluation is in that it allows for informed judgment on hard to quantify variables, such as morale, institutional commitment, and environmental context. In many instances, the case study model is used for pre- or post-evaluation efforts. As a preliminary investigative tool, this method could identify key variables for further study. As a follow-up to completed evaluative studies, it could throw light on hard to interpret findings. Dyer reported on theories of informed judgments on educational programs.

---


The final evaluation model included in this review has been called Instructional Evaluation. This model is mainly grounded in the use of various instruments for the observation of classroom instruction. In this regard, this model is similar to the case study method since the observer plays a key role in processing data. Evaluation of instructional methods are non-experimental in that the focus is upon direct observation of classroom interactions. It is similar in goals to Bloom's formative and Provus' discrepancy evaluation. The goals for evaluation of instruction are for making recommendations for changes in an instructional program.

Flanders (1970) and Gage (1968) have reported on observational instruments for noting classroom interaction.

Researchers are beginning to recognize the full potential of observational methods in the evaluation and study of instruction. There are three prime problem areas of this model: (1) need for greater differences in alternative

---


educational strategies, (2) improved observational instruments, and (3) additional research on the relationship between instructional activities and student outcomes.

The various models for evaluation are difficult to summarize or categorize in terms of content and goals. For example, they could be divided into areas of investigation, i.e., curriculum, instructional or institutional or by approaches such as discrepancy and cost analysis. The models could be classified by research methods, i.e., observational/statistical or experimental/quasi-experimental. Possibly, Bloom's formative, summative schema or Stufflebeam's CIPP typology may provide a way of classifying various evaluation models.

Evaluation Procedures

Whatever model of evaluation used, each must be organized along basic procedures for implementation: specification of goals, information requirements, and quality, evaluation design, data analysis, and dissemination.

Stufflebeam reported six basic areas for planning evaluations:

1. Focusing the Evaluation
2. Collection of Information
3. Organization of Information
4. Analysis of Information
5. Reporting of Information
6. Administration of the Evaluation.\textsuperscript{43}

Scriven noted thirteen areas for evaluation procedures:

1. Need Justification
2. Market
3. Performance—True Field Trials
4. Performance—Time Consumer
5. Performance—Crucial Comparisons
6. Performance—Long Term
7. Performance-Side Effects
8. Performance—Process
9. Performance—Causation
10. Performance—Statistical Significance
11. Performance—Educational Significance
12. Costs and Cost-Effectiveness
13. Extended Support\textsuperscript{44}

\textsuperscript{43} Stufflebeam, p. 70.

Note that Scriven has included nine factors related to measurement, evaluation design, and statistical analysis.

Suchman lists six steps as essential for evaluation:

1. Identification of goals to be evaluated.
2. Analysis of the problems with which the activity must agree.
3. Description and standardization of the activity.
4. Measurement of the degree of change that takes place.
5. Determination of whether the observed change is due to the activity or to some other cause.
6. Some indication of the durability of the effects.\textsuperscript{45}

The most common features of these procedures are:

1. Goal Specification
2. Measurement Specification
3. Measurement Quality
4. Evaluation Design
5. Statistical Analysis
6. Management and Dissemination

\textsuperscript{45}Suchman, p. 31.
These procedures go beyond the design of evaluation to include pre and post design concerns. This would include the recognition of the importance of values and judgment factors to program's goals. It also recognizes the importance of dissemination of evaluation findings to the consumer, i.e., funding agency, parents, community, etc.

The characteristics of evaluation procedures are substantially different from other kinds of research because it takes place in action settings. Program concerns take precedent over the need for rigorous research.

Problems With Educational Evaluation

There are four basic problems identified in the literature for program evaluation. They are conceptual, methodological, political, and organizational.

Cohen reported that there is frequently poor consensus on the causation of educational problems. This lack of agreement tends to lead to post hoc arguments about the worth of any evaluation. 46

---

There are many problems in evaluation design and methodology. The lack of consensus among researchers on evaluation, theory, research, and practice tend to minimize the credibility of evaluation findings. The traditional experimental design employed in educational research cannot adequately assess the relationship between treatment conditions (the program) and the effects (outcomes).

Many experts agree with Stanley that "... there is a definite though by no means unlimited place in evaluation for controlled, variable manipulating, comparative experimentation." However, evaluators are directed to the self critical use of quasi-experimental designs by Stanley.

The political problems associated with evaluation make it difficult to find and carry out adequate plans. Suchman noted six major political constraints to evaluation:

---

47 Stufflebeam.

1. the use of evaluation for "eye-wash," i.e., selecting only those aspects of a program that appear to be successful
2. the use of evaluation for "white-wash," covering up the program failure by securing "testimonials" from partisans
3. the use of evaluation findings to destroy a program regardless of its worth, because of power interests.
4. the use of evaluation to make gestures towards "objective evaluation" only to promote a favorable image.
5. the use of evaluation to delay decision-making with the hope that the concern will dissipate over time.
6. the use of evaluation to shift attention from one part of a program that has failed to another part that has succeeded.49

The Teacher Corps and Head Start are examples of programs difficult to evaluate because economic and political considerations were more important than educational concerns.

49 Suchman, p. 143.
Finally, the difficulty to manage the evaluation effort has been caused by many organizational constraints. Confidentiality of data, the training and experience of field staff, the support and cooperation with local school district, the use of control groups, and evaluation monitoring have been identified as some organizational problems.

Federal Evaluation Policy

Introduction The General Education Provisions Act requires that the Secretary of HEW transmit to Congress "a report evaluating the results and effectiveness of programs and projects assisted thereunder during the preceding fiscal year." However, in another report on evaluation studies of the Office of Education's programs indicated that, "it will take a number of years before all Office of Education programs have been subjected to systematic formal evaluation.

50 C. Weiss, Organizational Constraints on Evaluative Research, New York: Columbia University, June 1971

51 Ibid.


The report revealed the following:

1. There is a lack of goal specification in the enabling legislation of the programs.
2. The definition of program goals are not adequately completed at the administrative levels of government.
3. There are more frequently immediate objectives specified than long-term goals.
4. Program outcomes, if specified, are not usually related to program costs.
5. State and local government capacity to evaluate programs needs to be developed. Almost no programs exist to improve evaluation capacity of local government.
6. Office of Education is one of few agencies that earmark program funds for evaluation.
7. The General Accounting Office and the Office of Management and Budget should be more involved in substantive evaluation studies.  

Over the past eight years, Congress has appropriated some 75 million for program research and evaluation. The

\[54\] Ibid.

Elementary and Secondary Education Act brought evaluation into national prominence. However, those who have reviewed these evaluations have been disappointed to note that a great many of them reported on the number of students participating, expenditures for equipment, and material on "testimonials" from people administering the programs. The consequences of ESEA for evaluation may mean a rethinking about program aims and criteria for program success.\(^{56}\)

The Follow-Through program was recommended as an experimental strategy and evaluation design. However, the evaluation work plan has run to insurmountable problems. Since non-Follow Through schools are (1) reluctant to serve as control groups, (2) many Follow Through schools also participate in other federal programs thereby compounding the effects of the programs with others, (3) classrooms are used as the unit of analysis rather than students, which further compounds problems of sample size and interclass mobility.\(^{57}\)


\(^{57}\) Ibid.
The Follow Through evaluation experience may provide important information to evaluators on what not to do.

Another evaluation study to assess the impact of the Emergency School Aid Program in desegregated school districts found that the vast majority of respondents felt that the racial climate had changed for the better or had remained the same.\footnote{DHEW, Southern Schools: An Evaluation of the Effects of the ESAP and of School Desegregation, Washington, D.C.: OPBE, December 6, 1973.} The findings were based upon a random sample of 600 schools in 103 southern school districts receiving ESAP grants in 1971-72. A racial preference questionnaire was administered along with interviews with school officials. A standardized achievement test was administered to over 32,000 fifth and tenth grade students. This study used an experimental design of program effects by randomly selecting matched pairs of schools, one selected at random, to receive ESAP funds, and the other to be a control school.\footnote{Ibid.} The effects of ESAP aid were measured

\begin{footnotesize}
\item[59] Ibid.
\end{footnotesize}
by comparing the two groups of schools which differed only in whether or not they were receiving ESAP aid.

However, the experimental evaluation design may be grossly inappropriate for desegregation programs. Suchman noted the ambiguous interactions among social stratification, power in the community, public opinion, prejudice, and personality factors may make traditional experimental research uninterpretable.  

Concluding Remarks

The feasibility of evaluating federal educational programs depends on several factors: First, the type of evaluation desired, i.e., whether context, input, process or product should be of prime concern. Second, the existence of available evaluation theory, procedures or models, i.e., discrepancy, cost analysis, or case studies. Third, whether there are sufficient inputs in terms of personnel, money, and authority related to the evaluation effect (follow through). Fourth, the availability of appropriate

---

comparison groups (ESAP). Fifth, the ability of evaluation producers to collect the required data.

Agency level staffs should analyze existing programs for evaluation feasibility based upon the above factors.
CHAPTER III
STUDY DESIGN AND PROCEDURES

Overview

This chapter focuses on the overall research design, data collection, sample selection, questionnaire development, and data processing and analysis procedures.

Research Design

The research methodology utilized in this study consists of a cross sectional survey design in which panel scores are compared to predictor variables, i.e., the Evaluational Procedure Survey, score, type of grant, size of project, evaluation costs, and evaluation use.

The panel ratings and EPS scores were matched for each case. Although there is a three month difference between the collection of panel ratings and EPS data, there is no claim made that one variable is the cause of the other.

The procedures used in this study consisted of a survey questionnaire that was mailed to each ESAA applicant. The EPS questionnaire was precoded and accompanied by a cover letter of explanation and a return envelope. Each questionnaire was precoded with a three digit number to facilitate follow-up on nonrespondents and to match EPS scores with panel ratings. Approximately ten (20%) questionnaires were
completed by phone with four (8%) applicants not replying to the follow-up request.

Many questions, concepts, or statements contained in the EPS were discussed at several general sessions for all prospective applicants. Overall, every effort was made to insure an adequate response rate and to prepare each applicant concerning the content in the Survey.

Data Collection

This section will define the criteria used to isolate the variables and describe how the criteria are measured.

In this study evaluation quality is defined by the rating of a panel of educators on five components of evaluation work plans. These components are:

1. The use of quantifiable objectives.
2. The use of standardized tests.
3. The validation of the reliability and validity of the measures used.
4. The use of control, comparison groups, or external standards.
5. The use of ongoing monitoring procedures to revise the program.

Panel ratings are an interval level measurement.
Evaluation procedures are defined by measures on the EPS. The EPS is composed of six factors:

1. Goal Specification (3 items)
2. Measurement/Data Collection (7 items)
3. Measurement Quality (6 items)
4. Evaluation Design (5 items)
5. Statistical Methods (6 items)
6. Administration of Evaluation Procedures (5 items)

Evaluation Procedure Survey scores are interval level measurements. 

Descriptor or categorical data were collected from each ESAA applicant in order to determine the relation between panel ratings and EPS scores. The four categorical measures consisted of two nominal level variables, Type of Grant and Evaluation Use. The two ordinal level measures are size of project and evaluation cost. Table I summarizes the five independent and one dependent variables.

---

61 See Appendix A for the 32 items grouped by the six factors.
Table I
SUMMARY OF INDEPENDENT AND DEPENDENT VARIABLES

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Dependent Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁ Type of grant</td>
<td>Y₁ Panel ratings</td>
</tr>
<tr>
<td>X₂ Size of project</td>
<td></td>
</tr>
<tr>
<td>X₃ Evaluation cost</td>
<td></td>
</tr>
<tr>
<td>X₄ Evaluation use</td>
<td></td>
</tr>
<tr>
<td>X₅ Evaluation Procedure Survey score</td>
<td></td>
</tr>
</tbody>
</table>
Sample Selection

The procedures used to select the sample of this study are discussed in this section.

The sample was drawn from a population of 54 ESAA school districts and community organizations that were involved in desegregation and related educational programs.

In 1974 there were some 829 school districts in the six New England states. These districts enrolled some 2,479,206 students, including some 178,099 (7%) minority group children.\(^{62}\) Thirty school districts in New England were involved in some type of desegregation on a voluntary or court ordered basis. An additional 17 applicants proposed to support these districts in the implementation of their desegregation plan and programs. Seven school districts had also applied for funding for bilingual or innovative pilot projects. These districts enrolled substantial numbers of Hispanic and other minority group children.

A master list of the 54 ESAA applicants was developed by the regional Office of Education for 1974. This list established the population frame for this study.\(^{63}\)


\(^{63}\)See Appendix B.
Table II shows the number of ESAA applicants by type of grant and state for 1974.

Ninety-two percent (50) of the ESAA applicants responded to the Evaluation Procedure Survey questionnaire. Four applicants did not respond to the EPS which included one Basic and three Non Profit applicants. Ninety-seven percent of the Basic applicants were represented in this survey and 82% Non Profit and all of the Pilot and Bilingual grants were represented.

Questionnaire Development

This section focuses on the conceptualization of the EPS questionnaire items, instrument format, design, and procedures.

The EPS was based upon studies conducted by Wholey (1971) Stufflebeam (1969) and Scriven (1974). Chapter II provides an extensive review of these studies.

The study's definition of evaluation is that evaluation: (1) is a method of work to determine a program's worth

---

64 Wholey, Federal Evaluation Policy
Scriven, Evaluation in Education.
Table II

NUMBER OF BASIC, NPO, PILOT, & BILINGUAL GRANTS, BY STATE

<table>
<thead>
<tr>
<th>State</th>
<th>BASIC</th>
<th>NPO</th>
<th>PILOT</th>
<th>BILINGUAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connecticut</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Maine</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>17</td>
<td>8</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>3</td>
<td>2</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Vermont</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>30</strong></td>
<td><strong>17</strong></td>
<td><strong>3</strong></td>
<td><strong>4</strong></td>
</tr>
</tbody>
</table>
in comparison to a stated goal or external criteria, (2) is based upon scientific principles of research and (3) is aimed at providing objective information to decision and policy makers.

According to the reviews conducted by Stufflebeam and Metfessel and Micheals, there are six components for developing evaluation work plans. They are:

1. Program goals, objectives, or criteria
2. Measurement procedures
3. Measurement reliability and validity
4. Evaluation Design
5. Data Analysis
6. Management of evaluation activities

The paradigms of Stufflebeam and Metfessel/Micheals are presented below.  

---

Components of Evaluation Work Plans

<table>
<thead>
<tr>
<th>Stufflebeam</th>
<th>Metfessel-Micheals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Focusing the evaluation.</td>
<td>Involvement of total community.</td>
</tr>
<tr>
<td></td>
<td>Construction of broad goals and specific objectives.</td>
</tr>
<tr>
<td>2. Information collection</td>
<td>Development of instrumentation.</td>
</tr>
<tr>
<td>3. Information organization</td>
<td>Conduct periodic observations.</td>
</tr>
<tr>
<td>4. Information analysis</td>
<td>Analyze the data</td>
</tr>
<tr>
<td>5. Information reporting</td>
<td>Interpret data</td>
</tr>
<tr>
<td>6. Administration of evaluation.</td>
<td>Formulate recommendations</td>
</tr>
</tbody>
</table>

There is a great deal of similarity between the two models and the model used in this study for the formulation of the evaluation items. However, the six components for this study were further refined into a total of 32 questionnaire items. The items within each component were ordered by their degree of difficulty.

Finally, the criteria for the selection of each item within the six components were related to the following three areas:

1. Each item within a component should be measuring movement towards or away from the same thing.
2. The components themselves should be relatively independent of each other.

3. The items within each component should be reliable in the sense of yielding internally consistent response patterns.

The six components noted by Stufflebeam and Metfessel and Micheals are included in the EPS. They are listed below along with the number of items that compose the components:

Six Factors in the Evaluation Procedure Survey

1. Goal Specification (3 items)
2. Measurement and Data Collection (7 items)
3. Measurement Quality (6 items)
4. Evaluation Design (5 items)
5. Statistical Methods (6 items)
6. Administration of Evaluation Procedures (5 items)

Some components of the EPS were more difficult to develop than others. Questionnaire items related to Measurement Quality were perhaps the most difficult to conceptualize and develop. Table III lists the items of this component. It is expected that items 1 to 6 are ordered in difficulty. The literature review tends to support the notion that evaluators and school administrators often include other student information in the interpretation of test results.
Table III
MEASUREMENT QUALITY

The Evaluation Work Plan:

1. Specifies how other measures, in addition to standardized tests, will be used to determine the achievement of the objectives, such as motivation, behavior, or performance factors.

2. Indicates procedures that monitor the testing of students to validate the consistency of test instruction, pupil characteristics, and test situations.

3. Specifies how test-taking skills, anxiety, motivation, speed, guessing, and test instructions will be examined in the analysis of test results.

4. Specifies statistical procedures that examine subgroup scores on alternative forms of a test to validate its stability.

5. Indicates the frequency of which separate norms and reliability coefficients will be established for various subgroups participating in the program.

6. Indicates how race, language, SES, or personality will be examined to determine the degree to which these factors account for current achievement test scores.
Standardized tests such as the Woodcock Reading Test and the California Achievement Test, 1970 edition, include information on how race, language dominance, SES and personality factors account for various achievement levels. Item 6 assesses the degree that cultural bias is addressed in the examination of measurement reliability and validity. The EPS questionnaire items were also developed in conformance to prevailing ESAA regulations and guidelines concerning project evaluation. Finally, the 32 items in the questionnaire were critically reviewed by two Office of Education officials as for clarity and relevance.

A pre-test was conducted in March 1974 with a random sample of ten ESAA applicants to determine whether the a priori items were ordered correctly by their level of difficulty. Items 12 and 18 were selected as representative questions. Item 12 states: "The specification of criteria that defines the attainment of attitudinal, interest, appreciation, or self-concept objectives." Item 18 states: "The specification of a set of criteria that defines the

---


attainment of affective and cognitive objectives." Item 18 is expected to have fewer "yes" respondents than item 12 because it is a more difficult procedure. Table IV summarizes the results of the pre-test. More respondents tended to answer "yes" on the less difficult item (12) than the more difficult item (18). The results showed that item 18 received more "no" responses than item 12, seven to five. However, the differences do not appear to be significant. It could be expected that the other items in the EPS questionnaire are moderately ordered by difficulty. Final tests on the Guttman scale and factor analysis will test for statistical significance at .05 level. Further evaluation and comments of the Office of Education officials showed that the pilot study had several questions that were poorly worded or confusing. These items were revised or deleted from the final questionnaire.

Data Processing and Analysis Procedures

The SPSS program, 2nd edition, was used to process the Evaluation Procedure Survey data. The responses to the items were key punched in a fixed column format. There was one card for each case in the file. The data was processed in batch and printouts were produced to describe the output of the programs used.

Several programs in SPSS were utilized to analyze the data. Simple correlations coefficients were produced from the Pearson and Nonparametric programs. Discriminant analysis
Table IV
PRETEST OF EPS ITEMS 12 AND 18 BY FREQUENCY OF RESPONDENTS

<table>
<thead>
<tr>
<th>Item 12 (Less Difficult)</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>The specification of criteria that defines the attainment of attitudinal, interest, appreciation, or self-concept objectives</td>
<td>5 (.63)</td>
<td>5 (.42)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item 18 (Difficult)</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>The specification of a set of criteria that defines the attainment of affective and cognitive objectives.</td>
<td>3 (.38)</td>
<td>7 (.58)</td>
</tr>
</tbody>
</table>

YULES'S $Q = -.40$ correlation of coefficient.
program was used to determine whether there were significant differences among the types of grants. The multiple regression program (step-wise) was used to test the third hypothesis. Additional analysis was made through the use of a factor analysis program. Some calculations were performed by pocket calculator to verify the general output of the various programs.

Limitations

There are some methodological limitations in the design and implementation of this study.

First, there was only one administration of the EPS questionnaire. It would have been preferred that the EPS instrument was administered two times over a six month period to determine the stability of the questionnaire.

Secondly, there were only 50 cases in the sample. A larger sample of 150 to 200 cases would have been preferred so that each type of grant would have larger cell frequencies and be near equity. A larger sample would insure a more normal distribution in the criterion variable, panel ratings.

Lastly, the nominal level variables, type of grant, and evaluation use were treated as ordinal or interval level measures. These variables should have been translated into "dummy" variables, i.e., each subgroup should be treated as separate variables in order to meet the requirements of
higher level statistics, such as multiple regression.

Limited sample size, single administration of the EPS questionnaire and nominal level variables greatly reduced the generalizability of this study's findings. These problems should be addressed in subsequent studies.
CHAPTER IV
FINDINGS OF STUDY

Overview

The major purpose of this study is to discover what are the relationships between evaluation procedures and panel ratings of evaluation work plans. Information on evaluation costs, size of project, type of grant, and types of decisions for the use of evaluation findings were collected through the use of an Evaluation Procedure Survey in the spring of 1974.

The remainder of this chapter will focus on the findings related to the three major hypotheses, the examination of the reliability and validity of the EPS instrument and additional descriptive statistics.

Major Hypothesis I

Hypothesis I. The total Evaluation Procedure Survey score is significantly related to panel ratings of evaluation quality.

Null Hypothesis I. The relationship between the total EPS score and panel ratings of evaluation quality is not statistically significant.
Test Hypothesis I. The correlation between the total EPS score and panel ratings was examined to determine whether the relationship is statistically significant. The correlation coefficient for the two variables was $r = -0.29$ and significant at $p < 0.03$. However, the EPS score accounted for only 9% of the variance of panel ratings. Approximately, 91% of the variance in panel ratings is determined by variables other than EPS score or measurement errors. The null hypothesis is rejected, therefore, confirming the initial research hypothesis.

Further analysis was conducted by examining chi-square statistics on each independent variable and panel ratings. Table V summarizes the raw chi-square panel ratings with each independent variable. Only the chi-square for panel ratings and total EPS score was found to be statistically significant at $p < 0.00$. Chi-square values for the other pairs were not statistically significant.

Spearman's correlation was used to verify the relationship between type of grant, evaluation use variables, and panel ratings. Table VI presents the relationships between panel ratings and each nominal level independent variable. The initial values of type of grant and evaluation use were replaced by ordinal rankings to meet the assumptions of Spearman's nonparametric correlations. The correlations
Table V

CHI SQUARE PANEL RATINGS (EVSC)

WITH

TYPE OF GRANT, SIZE OF PROJECT, EVALUATION COST,
EVALUATION USE, AND EVALUATION PROCEDURE SURVEY SCORE

<table>
<thead>
<tr>
<th>Variables</th>
<th>Raw Chi Square</th>
<th>df</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel Rating/Type of Project</td>
<td>47.36</td>
<td>42</td>
<td>.26</td>
</tr>
<tr>
<td>Panel Rating/Size of Project</td>
<td>53.26</td>
<td>56</td>
<td>.57</td>
</tr>
<tr>
<td>Panel Rating/Evaluation Cost</td>
<td>42.92</td>
<td>52</td>
<td>.81</td>
</tr>
<tr>
<td>Panel Rating/Evaluation Use</td>
<td>15.96</td>
<td>24</td>
<td>.88</td>
</tr>
<tr>
<td>Panel Rating/EPS</td>
<td>478.97</td>
<td>364</td>
<td>.00</td>
</tr>
</tbody>
</table>
Table VI
NON-PARAMETRIC CORRELATIONS

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. PANEL RATINGS</td>
<td>.02</td>
<td>.30*</td>
<td>- .23</td>
<td>.24*</td>
<td>.28**</td>
<td></td>
</tr>
<tr>
<td>2. TYPE OF GRANT</td>
<td>.11</td>
<td>.16</td>
<td>-.07</td>
<td>.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. SIZE OF PROJECT</td>
<td>.07</td>
<td>.08</td>
<td>-.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. EVALUATION COST</td>
<td></td>
<td>-.34*</td>
<td>.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. EVALUATION USE</td>
<td></td>
<td></td>
<td>.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. EVALUATION PROCEDURE SURVEY SCORE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Significant at p .05
**Significant at p .01
between panel ratings and type of grant was negligible. However, the correlations for evaluation use, size of project, EPS score variables, and panel ratings were significant at \( p < .05 \). The correlation between panel ratings and evaluation cost was negative, -.23, which further supports the findings of the pilot project.\(^67\)

Overall, the relationship between panel ratings and EPS scores were significant for Pearson and Spearman correlations and chi-square values. The size of project and evaluation use variables were found to be sizable for Spearman correlations. Nonetheless, the correlation between panel ratings and EPS scores were found to be weak when compared to expected criteria of .85, as is the criteria for most achievement test criterion-related validity.

Major Hypothesis II

**Hypothesis II.** There is a significant difference among the types of grants in terms of panel ratings, EPS score, size of project, evaluation cost, and evaluation use.

\(^{67}\)See Appendix C.
**Null Hypothesis II.** There are no significant differences among the types of grants in terms of major discriminating variables.

**Test of Hypothesis II.** The effects of the major differences among the types of grants by distinguishing variables is determined by discriminant analysis procedures. The results of this analysis is summarized in Table VII.

An examination of the five variables in Table VII shows that three variables contribute the most in differentiating among the types of grants. They are panel ratings, evaluation costs, and EPS scores. Discriminating variables are listed and one analysis is performed on all variables. The values for Wilks Lambda were obtained through a stepwise program. The F ratios and canonical correlations denote the ability of each function to separate the four groups. Only three variables (functions) were used in the discriminant analysis (panel ratings, evaluation cost, EPS score):

<table>
<thead>
<tr>
<th>Function</th>
<th>Eigenvalue</th>
<th>Wilks Lambda</th>
<th>Chi-Sq.</th>
<th>df</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.27</td>
<td>.68</td>
<td>17.07</td>
<td>15</td>
<td>.31</td>
</tr>
<tr>
<td>1</td>
<td>.15</td>
<td>.86</td>
<td>6.30</td>
<td>8</td>
<td>.61</td>
</tr>
<tr>
<td>2</td>
<td>.00</td>
<td>.99</td>
<td>.01</td>
<td>3</td>
<td>.99</td>
</tr>
</tbody>
</table>

The last function contributes little towards the discrimination among the four groups. Wilks Lambda and chi-square tests were not significant for any of the three functions.
### Table VII
DISCRIMINANT ANALYSIS BY FOUR TYPES OF GRANTS

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables Entered/Removed</th>
<th>F Ratio</th>
<th>Wilks Lambda</th>
<th>Raos</th>
<th>Change in Raos</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>COST</td>
<td>3.08</td>
<td>.82*</td>
<td>9.26</td>
<td>9.26</td>
</tr>
<tr>
<td>2</td>
<td>NOPART</td>
<td>1.36</td>
<td>.75</td>
<td>13.47</td>
<td>4.21</td>
</tr>
<tr>
<td>3</td>
<td>TEPS</td>
<td>.63</td>
<td>.72</td>
<td>15.79</td>
<td>2.31</td>
</tr>
<tr>
<td>4</td>
<td>EVUSE</td>
<td>.71</td>
<td>.69</td>
<td>18.36</td>
<td>2.56</td>
</tr>
<tr>
<td>5</td>
<td>EVSC</td>
<td>.19</td>
<td>.68</td>
<td>19.17</td>
<td>.80</td>
</tr>
</tbody>
</table>

*Significant at p < .05

<table>
<thead>
<tr>
<th>Number Removed</th>
<th>Eigenvalue</th>
<th>Canonical Correlation</th>
<th>Wilkes Lambda</th>
<th>Chi-Sq</th>
<th>df</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.273</td>
<td>.463</td>
<td>.681</td>
<td>17.07</td>
<td>15</td>
<td>.31</td>
</tr>
<tr>
<td>1</td>
<td>.151</td>
<td>.362</td>
<td>.867</td>
<td>6.30</td>
<td>8</td>
<td>.61</td>
</tr>
<tr>
<td>2</td>
<td>.000</td>
<td>.019</td>
<td>.999</td>
<td>.01</td>
<td>3</td>
<td>.99</td>
</tr>
</tbody>
</table>
Before any functions were removed, lambda was .68132 which indicates modest discriminating power. However, chi-square is not significant. The second function removes yet more discriminating power, a nonsignificant lambda is still found. One conclusion could be drawn that neither functions significantly discriminate between groups. This conclusion is substantiated by the fact that the centroids for the four groups are fairly homogeneous except for Group 3 on function 1. The standardized discriminant function coefficients supports this conclusion. **Evaluation cost** coefficient on function 1 is quite large, .78 and represents .61% of the variance of this function. Both groups 1 and 2, **Basic**, and **Non Profit**, are close on functions 1, 2, and 3, i.e., **evaluation cost**, **size of project**, and **EPS**. An examination of plots of group centroids and locations, groups 1 and 2 are very close to each other. However, there is a great deal of overlap among all groups.

Of the three functions entered into the discriminant analysis, none were statistically significant in terms of their distinguishing powers. Functions 1 and 2 were at .61 and .99 levels respectively. Of the valid cases, only 32.7%

---

68 See Figure I
Figure I
were correctly classified. Chi-square, at 1.531, was not significant. Null Hypothesis II is accepted. There are no statistically significant differences between the types of grants on major distinguishing variables. Further analysis and classification by groups would be meaningless.

Major Hypothesis III

Hypothesis III. There is a significant relationship between the total Evaluation Procedure Survey score and panel ratings after the effects of evaluation cost, size of project, type of grant, and evaluation use factors have been taken into account.

Null Hypothesis III. The relationship between the total Evaluation Procedure Survey score and panel ratings is not significant after the effects of evaluation cost, size of project, type of grant, and evaluation use factors have been taken into account.

Test of Hypothesis III. A stepwise multiple regression program was used to determine the relationship between panel ratings and EPS score, controlling for the effects of type of grant, size of project, evaluation cost, and use. The results of this analysis is presented in Table VIII. The overall multiple correlation between panel ratings and the five
Table VIII
MULTIPLE CORRELATION OF PANEL RATING WITH
TYPE OF GRANT, SIZE OF PROJECT, EVALUATION COST, EVALUATION
USE AND EVALUATION PROCEDURE SURVEY SCORE

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables Entered/Removed</th>
<th>F Ratio</th>
<th>MR</th>
<th>R sq.</th>
<th>R sq. Change</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Size of Project</td>
<td>64.94</td>
<td>.29</td>
<td>.08</td>
<td>.08</td>
<td>.74</td>
</tr>
<tr>
<td>2</td>
<td>Evaluation Proc. Survey</td>
<td>3.40</td>
<td>.38</td>
<td>.14</td>
<td>.05</td>
<td>.74</td>
</tr>
<tr>
<td>3</td>
<td>Evaluation Cost</td>
<td>6.08*</td>
<td>.49</td>
<td>.24</td>
<td>.09</td>
<td>.51</td>
</tr>
<tr>
<td>4</td>
<td>Type of Project</td>
<td>.09</td>
<td>.49</td>
<td>.24</td>
<td>.00</td>
<td>.64</td>
</tr>
<tr>
<td>5</td>
<td>Evaluation Use</td>
<td>.01</td>
<td>.49</td>
<td>.24</td>
<td>.00</td>
<td>.67</td>
</tr>
</tbody>
</table>

* Significant at p  .05
independent variables was significant at p .00. The multiple correlation was .49. These predictor variables accounted for 24% of the variance of panel ratings.

An analysis of the individual contributions of each independent variable towards the multiple correlation was examined in the stepwise solution. The correlation between size of project and EPS scores was .74. However, the size of project was entered first into the regression equation and EPS was entered second. The multiple correlation for EPS was .39. This variable added .05 to the change in the multiple correlation. The F-ratio for EPS was not significant. However, evaluation cost added .09 to the change in the multiple correlation but evaluation cost was significant at p < .05. The type of grant and evaluation use variables contributed little towards the explanation of the multiple correlation. The null hypothesis is accepted. The contribution of EPS scores to the multiple correlation was not significant. This variable added 5% increase to the variance of panel ratings. Evaluation cost contributed 9% increase in the variance of panel ratings.

These findings should be considered with caution since EPS scores are correlated with panel ratings, more so than the other independent variables. However, the addition of EPS scores in establishing a linear relationship with panel ratings was not significant.
Additional Analysis

What are the distributional characteristics of the panel ratings? Are the scores normally distributed?

The integer frequencies for panel ratings on Figure II showed the following results:

<table>
<thead>
<tr>
<th>Score</th>
<th>f</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>23</td>
<td>44%</td>
</tr>
<tr>
<td>5-9</td>
<td>2</td>
<td>4%</td>
</tr>
<tr>
<td>10-14</td>
<td>8</td>
<td>16%</td>
</tr>
<tr>
<td>15-19</td>
<td>10</td>
<td>19%</td>
</tr>
<tr>
<td>20-24</td>
<td>9</td>
<td>17%</td>
</tr>
</tbody>
</table>

Panel ratings were not normally distributed. Although the mean score was 9, the standard deviation was 8.9. Most of the scores were clustered to the left of the mean, in the lower end of the scale. The kurtosis of -1.561 indicated a flat distribution in comparison to a normal curve. Twenty-three cases (44%) had panel scores between 0-4. Several of these cases had zero panel scores and had to be assigned a value which would not greatly alter the characteristics of the distribution.
Figure II

FREQUENCY DISTRIBUTION OF PANEL RATINGS

Percentage of All Panel Scores

Panel Score

45 |
40 |
35 |
30 |
25 |
20 |
15 |
10 |
0 |

4 9 14 19 24
The mean panel score by type of grant was:

<table>
<thead>
<tr>
<th>Type</th>
<th>N</th>
<th>Basic</th>
<th>NPO</th>
<th>Pilot</th>
<th>Bilingual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>29</td>
<td>14</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>9.2</td>
<td>8.1</td>
<td>10.3</td>
<td>11.6</td>
</tr>
<tr>
<td>Adjusted Mean</td>
<td></td>
<td>8.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The mean panel ratings of the 50 ESAA applications on evaluation work plans were relatively low, 9 out of 24 points.

**Description of Independent Variables.** Type of grant is measured on a nominal scale. Fifty cases were divided among these four categories:

<table>
<thead>
<tr>
<th>Type</th>
<th>N</th>
<th>Basic</th>
<th>Non Profit</th>
<th>Pilot</th>
<th>Bilingual</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>29</td>
<td>14 (27%)</td>
<td>3 (6%)</td>
<td>4 (8%)</td>
<td>2 (3%)</td>
</tr>
</tbody>
</table>

The size of the project is measured on an interval scale. The size of project variable has a fairly normal distribution with a mean of 3.0 and a standard deviation of 1.2. The majority of the cases planned to service between 500 to 5,000 students as compared to 10% of the projects to service over 5,000, and 15% for less than 100 students.

**Evaluation cost** is measured on an interval scale. This variable was not normally distributed. The mean score was 2.2 and the standard deviation 1.3. The evaluation costs
of the proposed projects were relatively modest with over 42% spending less than 1% for evaluation. Unfortunately, this descriptor was not refined enough to distinguish the frequency of zero funding for evaluations. Ten percent of the projects planned to expend over 11% on program evaluation.

The evaluation use variable is measured on a nominal scale 1 to 4 for each category. Seventy-five percent of the projects planned to use their evaluation results to "assess the overall impact of the program." Only 11% of the cases planned to use evaluation results to decide on the "best" program strategy and 14% to "redefine project goals."

The frequency distributions for items 1 to 32 on the EPS varied considerably. However, the overwhelming number of cases fell within the (1) none or (2) some categories. The initial assumption of poor evaluation work plans was upheld in the histogram for each independent and variable and their respective statistics. See Table IX.

Multiple Regression Analysis. Although Null Hypothesis III was accepted, the reanalysis of the regression of panel ratings with variables, type of grant, size of project, evaluation cost, evaluation use, and the items on the EPS produced significant results related to EPS scores. A step-wise multiple regression program was used to produce multiple R for dependent variable panel rating with regression forced
### Table IX

**EVALUATION PROCEDURE SURVEY**

Please check one of the following:

1. **Type of Grant:**
   - (29) Basic 58%
   - (14) Non Profit 28%
   - (3) Pilot 6%
   - (4) Bilingual 8%

2. **Number of project participants**
   - (8) less than 100 15%
   - (8) 101-500 15%
   - (14) 501-1,000 27%
   - (17) 1,001-5,000 33%
   - (5) 5,001-plus 10%

3. **Cost of the Evaluation**
   - (21) 0-1% 42%
   - (11) 2-3% 22%
   - (9) 4-6% 18%
   - (4) 7-10% 8%
   - (5) 11-plus % 10%

4. **The Evaluation results will be used to:**
   - (5) Redefine project goals 14%
   - (0) Determine project input
   - (4) Decide on best program strategy 11%
   - (27) Determine impact of project 75%

---

1. The specification of a schedule to monitor all of the evaluation activities to determine adherence to established time-table and procedures. (26) (14) (9) (3)

2. The description of statistical methods that compares changes in means, ranks, standard deviations, percentages, or signs to determine the effects of the program (45) (2) (4) (1)
Table IX (Cont.)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3. The specification of an evaluation design that systematically describes and analyzes a single program based upon the observation of project staff.</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>4. The specification of how other measures, in addition to standardized tests, will be used to determine the achievement of the objectives, such as motivation, behavior, or performance.</td>
<td>(41)</td>
<td>(7)</td>
<td>(4)</td>
</tr>
<tr>
<td>5. The identification of standardized achievement or ability tests to measure the attainment of cognitive objectives.</td>
<td>(25)</td>
<td>(14)</td>
<td>(10)</td>
</tr>
<tr>
<td>6. The specification of criteria that defines the attainment of cognitive objectives, such as knowledge, comprehension, understanding, skills or applications</td>
<td>(26)</td>
<td>(1)</td>
<td>(20)</td>
</tr>
<tr>
<td>7. The description of the frequency of the training of the staff to determine the quality of the evaluation work plan.</td>
<td>(42)</td>
<td>(8)</td>
<td>(2)</td>
</tr>
<tr>
<td>8. The description of statistical methods that examines categories of project participants, such as race, sex, or performance levels, to account for patterns of correlation.</td>
<td>(44)</td>
<td>(3)</td>
<td>(5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>9. The specification of an evaluation design that investigates the effects of a program on its participants by using a pre and post test</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>10. The indication of evaluation procedures that monitors the testing of students to validate the consistency of test instruction, pupil characteristics and testing situations.</td>
<td>(17)</td>
<td>(10)</td>
<td>(14)</td>
</tr>
<tr>
<td>11. The identification of standardized tests to measure the attainment of cognitive objectives.</td>
<td>(47)</td>
<td>(1)</td>
<td>(3)</td>
</tr>
<tr>
<td>12. The specification of criteria that defines the attainment of attitudinal interest, appreciation, or self-concept objectives.</td>
<td>(26)</td>
<td>(2)</td>
<td>(17)</td>
</tr>
<tr>
<td>13. The specification of the frequency that evaluation results will be reported to various interest groups</td>
<td>(19)</td>
<td>(12)</td>
<td>(19)</td>
</tr>
<tr>
<td>14. The indication of an evaluation design that investigates the extent to which variations of scores on one test corresponds to variations on another</td>
<td>(25)</td>
<td>(15)</td>
<td>(10)</td>
</tr>
</tbody>
</table>
Table IX (Cont.)

15. The specification of methods that measures differences between the categories of actual and predicted frequencies, such as race, sex, or performance levels to test their statistical significance. (46) (3) (3) (0)

16. The specification of how test-taking skills, anxiety, motivation, speed, guessing, and testing instructions will be examined in the analysis of test results. (50) (1) (0) (1)

17. The indication of the use of frequency counts of absences, lateness, discipline referrals, attendance or suspensions, as indirect measures of student behavior (28) (6) (17) (1)

18. The specification of a set of criteria that defines the attainment of affective and cognitive objectives. (38) (6) (7) (1)

19. The specification of the frequency of the monitoring of the administration and scoring of tests by project staff to validate the adherence to established procedures. (49) (1) (1) (1)
20. The description of evaluation methods that will examine how current test scores are accounted for by predictive variables, such as aptitude, program attendance, or individual background factors.

21. The specification of methods that compares the relationship between two scores while holding a third score constant and allowing the others to vary.

22. The specification of statistical procedures that examines sub-group scores on alternative forms of a test to validate its stability.

23. The specification of teacher-made tests to be related with standardized tests results.

24. The specification of the use of frequency counts of teacher conferences, school visits or attendance at special school events by parents of project participants, to assess their attitude or understanding of the program.
25. The description of evaluation methods that assigns students to treatment or control groups in order to experimentally examine the program's effects. 

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.</td>
<td>(47)</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

26. The specification of statistical methods that tests whether there is a difference between the "between" and "within" group variance to determine the effects of the program on the treatment and control groups. 

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.</td>
<td>(51)</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

27. The indication of the frequency of which separate norms and reliability coefficients will be established for various subgroups participating in the program. 

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.</td>
<td>(50)</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

28. The specification of the use of rating scales or checklists for observing teacher and student classroom behavior or performance to measure intra program effects

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>28.</td>
<td>(31)</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

29. The indication of statistical methods that will test whether the results between two or more independent measures are statistically significant.

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>29.</td>
<td>(45)</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Table IX (Cont.)

30. The indication of the frequency of use of anecdotal records, teacher logs or case studies of student performance during the program

31. The indication of how race, language, SES or personality factors will be examined to determine the degree to which these factors account for current achievement test scores

32. The specification of the use of self-rating reports, scales, checklists or inventories in which students compare perceived levels of achievement with future educational plans.
through origin. Independent variables that did not add substantially to the prediction of panel rating, once other independent variables were included, were deleted.

The overall test that the multiple Rs of panel ratings are zero for a majority of variables in the equation showed that the sample was drawn from a population in which the multiple correlations were more than zero. The overall F-ratio for the 22 independent variables in the equation was 69.48, significant at the .000 level. The null hypothesis that the multiple regression of panel rating on the 36 variables is zero was rejected. The comparison of $R^2$ change for categorical and continuous variables showed that size of project and evaluation cost contributed to 27% of the variance of panel rating as compared to the 53% accounted for by EPS. The prediction of the panel ratings by size of project, evaluation cost, and selected EPS was significant at the pr .01 level.

Those items that were significantly (p .10) correlated (multiple) with the criterion were:

Variables:

001 (administration of evaluation procedures)
017 (measurement and data collection)
004 (reliability and validity of data)
006 (objectives)
013 (administration and evaluation procedures)
024 (administration and evaluation procedures)
018 (objectives)
The analysis of the above indicates that all of the a priori factors have some items that panel rating scores regress on. Approximately 53% of the EPS items were significantly correlated with the criterion. The results from above indicate that the background variables, size of project, and evaluation cost account for less of the variance of panel ratings than selected EPS scores. These items should be used in subsequent studies to validate the prediction of the criterion.
### Table X

**STEPWISE REGRESSION PREDICTION OF PANEL RATING FROM 36 PREDICTOR VARIABLES**

<table>
<thead>
<tr>
<th>Step</th>
<th>Variable</th>
<th>R</th>
<th>R Sq.</th>
<th>R Sq. Increase</th>
<th>F</th>
<th>In/Out</th>
<th>Variable Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Size of Project</td>
<td>.48</td>
<td>.23</td>
<td>.09</td>
<td>41.5</td>
<td></td>
<td>Size of Project</td>
</tr>
<tr>
<td>3</td>
<td>Eval. Cost</td>
<td>.58</td>
<td>.34</td>
<td>.11</td>
<td>34.1</td>
<td></td>
<td>Evaluation Cost</td>
</tr>
<tr>
<td>4</td>
<td>Item 17</td>
<td>.59</td>
<td>.35</td>
<td>.01</td>
<td>25.6</td>
<td></td>
<td>Measurement Coll.</td>
</tr>
<tr>
<td>5</td>
<td>Item 4</td>
<td>.61</td>
<td>.37</td>
<td>.02</td>
<td>21.1</td>
<td></td>
<td>Measurement Quality</td>
</tr>
<tr>
<td>6</td>
<td>Item 6</td>
<td>.63</td>
<td>.40</td>
<td>.03</td>
<td>18.3</td>
<td></td>
<td>Goal Specification</td>
</tr>
<tr>
<td>7</td>
<td>Item 13</td>
<td>.65</td>
<td>.42</td>
<td>.02</td>
<td>16.1</td>
<td></td>
<td>Adm. Eval. Proc.</td>
</tr>
<tr>
<td>8</td>
<td>Item 24</td>
<td>.66</td>
<td>.44</td>
<td>.02</td>
<td>14.5</td>
<td></td>
<td>Adm. Eval. Proc.</td>
</tr>
<tr>
<td>9</td>
<td>Item 18</td>
<td>.68</td>
<td>.46</td>
<td>.02</td>
<td>13.2</td>
<td></td>
<td>Goal Specification</td>
</tr>
<tr>
<td>10</td>
<td>Item 32</td>
<td>.68</td>
<td>.47</td>
<td>.01</td>
<td>11.9</td>
<td></td>
<td>Measurement Coll.</td>
</tr>
<tr>
<td>11</td>
<td>Item 31</td>
<td>.69</td>
<td>.48</td>
<td>.01</td>
<td>10.9</td>
<td></td>
<td>Measurement Coll.</td>
</tr>
<tr>
<td>12</td>
<td>Item 5</td>
<td>.70</td>
<td>.50</td>
<td>.02</td>
<td>10.1</td>
<td></td>
<td>Measurement Coll.</td>
</tr>
<tr>
<td>13</td>
<td>Item 3</td>
<td>.72</td>
<td>.51</td>
<td>.01</td>
<td>9.5</td>
<td></td>
<td>Evaluation Design</td>
</tr>
<tr>
<td>14</td>
<td>Item 23</td>
<td>.73</td>
<td>.53</td>
<td>.02</td>
<td>9.0</td>
<td></td>
<td>Measurement Coll.</td>
</tr>
<tr>
<td>15</td>
<td>Item 30</td>
<td>.73</td>
<td>.54</td>
<td>.01</td>
<td>8.4</td>
<td></td>
<td>Measurement Coll.</td>
</tr>
<tr>
<td>16</td>
<td>Item 29</td>
<td>.74</td>
<td>.55</td>
<td>.01</td>
<td>7.9</td>
<td></td>
<td>Statistical Method</td>
</tr>
<tr>
<td>17</td>
<td>Item 20</td>
<td>.75</td>
<td>.56</td>
<td>.01</td>
<td>7.5</td>
<td></td>
<td>Evaluation Design</td>
</tr>
<tr>
<td>18</td>
<td>Item 21</td>
<td>.76</td>
<td>.58</td>
<td>.02</td>
<td>7.2</td>
<td></td>
<td>Statistical Method</td>
</tr>
<tr>
<td>19</td>
<td>Eval. Use</td>
<td>.77</td>
<td>.59</td>
<td>.01</td>
<td>6.8</td>
<td></td>
<td>Evaluation Use</td>
</tr>
<tr>
<td>20</td>
<td>Item 22</td>
<td>.77</td>
<td>.60</td>
<td>.01</td>
<td>6.4</td>
<td></td>
<td>Measurement Quality</td>
</tr>
<tr>
<td>21</td>
<td>Item 28</td>
<td>.78</td>
<td>.62</td>
<td>.02</td>
<td>6.2</td>
<td></td>
<td>Measurement Coll.</td>
</tr>
<tr>
<td>22</td>
<td>Item 15</td>
<td>.79</td>
<td>.62</td>
<td>.00</td>
<td>5.9</td>
<td></td>
<td>Statistical Methods</td>
</tr>
</tbody>
</table>
Technical Considerations: Reliability and Validity of Evaluation Procedure Survey

Internal Reliability. Were the various scales within the EPS relatively reliable in terms of measuring a single dimension? Were the obtained item scores predictable in terms of scale patterns? Did the EPS measure what it purported to measure?

The reliability of EPS will have to be determined by the repeated use and assessment of the instrument over several test periods and with a larger number of cases. However, the relationships between panel ratings and EPS scores would be spurious if the survey instrument was unreliable. The types of data that might be given as evidence of reliability are Guttman scales, intercorrelations among items and standard error measurements. Table XI displays the results of the evaluation of the undimensionality and cumulativeness of the EPS items.

The Guttman or cumulative-type scales consist of a relatively small cluster of homogeneous items that should be measuring the same attribute. Item scores were accumulated over the total scale score for individuals or cases. The EPS items were ordered in difficulty from low to high so that a correct response on the most difficult item implies a correct response on all preceding items. The coefficient of
Table XI

SUMMARY GUTTMAN SCALES: OBJSCALE MEASCALE QUASCALE EVSCALE, STASCALE AND ADMSCALE

<table>
<thead>
<tr>
<th>EPS SCALES</th>
<th>COEFFICIENT OF REPROD.</th>
<th>COEFFICIENT OF SCALABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Objscale</td>
<td>.82</td>
<td>.40</td>
</tr>
<tr>
<td>2 Meascale</td>
<td>.87</td>
<td>.55</td>
</tr>
<tr>
<td>3 Quascale</td>
<td>.98</td>
<td>.77</td>
</tr>
<tr>
<td>4 Evascale</td>
<td>.93</td>
<td>.52</td>
</tr>
<tr>
<td>5 Stascale</td>
<td>.95</td>
<td>.30</td>
</tr>
<tr>
<td>6 Admscale</td>
<td>.87</td>
<td>.17</td>
</tr>
</tbody>
</table>
reproducibility, a measure of the extent to which a respondent's scale score is a predictor of one's response pattern, should be greater than .90. The coefficient of scalability, a measure of the unidimensionality and cumulativeness of a scale, should be greater than .60.

An evaluation of the six scales, in terms of the two criteria showed:

<table>
<thead>
<tr>
<th>EPS Scales</th>
<th>Coefficient of Reproducibility</th>
<th>Coefficient of Scalability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Objscale</td>
<td>.82</td>
<td>.4043</td>
</tr>
<tr>
<td>2 Meascale</td>
<td>.87</td>
<td>.5534</td>
</tr>
<tr>
<td>3 Quascale</td>
<td>.98</td>
<td>.7778*</td>
</tr>
<tr>
<td>4 Evascale</td>
<td>.93</td>
<td>.5263</td>
</tr>
<tr>
<td>5 Statscale</td>
<td>.95</td>
<td>.3000</td>
</tr>
<tr>
<td>6 Admscale</td>
<td>.87</td>
<td>.1795</td>
</tr>
</tbody>
</table>

*Attained both criteria

Quascale attained both criteria, in terms of unidimensionality and cumulativeness. However, there is no overall test for evaluating the EPS scales relative to the two criteria. Scales 3, 4, and 5 clearly met the criteria for predicting one's response patterns. The difficulty or value-loading of these items were correctly ordered.

Objscale on item 018, (high difficulty) had 85% of the cases failing this item in comparison to 65% failing item 006 and 60% for item 012 (low difficulty). The
coefficient of reproducibility was equal to .82, less than criteria. The coefficient of scalability represented the largest value that the percentage improvement may attain. In this case, the coefficient was .40. However, the coefficient of marginal (minimum) reproducibility, which is the lowest value that could have occurred for the Objscale, was .69, a 12% improvement. It is not clear whether changes in the cutting points or the order of the items would improve the Objscale. The inter-item correlations suggests that it would not:

<table>
<thead>
<tr>
<th></th>
<th>Item 006</th>
<th>Item 012</th>
<th>Item 018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 006</td>
<td>1.0000</td>
<td>.1275</td>
<td>.3636</td>
</tr>
<tr>
<td>Item 012</td>
<td>.1275</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Item 018</td>
<td>.3636</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Scale item</td>
<td>.1360</td>
<td>.4255</td>
<td>.6848</td>
</tr>
</tbody>
</table>

Items 012 and 018 are highly correlated at perfect Rs of 1.0000. However, item 006 has a weak correlation with the other two items and its biserial correlation was .1360, a negligible association with the sum of all other items.

Meascale consisted of seven items with three division points each. Coefficient of reproducibility equalled .87, scalability equalled .55, slightly less than criteria. However, many of the inter-item correlations were quite large. Correlations between items 005 and 011 was .99 and items
028 and 023 was .91. Bi-serial correlations for the majority of the seven items was above .61.

Quascale consisted of six items with three division points each. Coefficients of reproducibility and scalability met the stated criteria. Except for items 004 and 010 biserial and inter-item correlations were high.

Evascale consisted of five items with three division points each. The total number of errors were 18. Ninety six percent of the cases failed item 009 (scored 1 or 2), the least difficult question. The coefficient of reproducibility was .93, slightly above the criteria. The coefficient of scalability equalled .52, below criteria. An examination of the inter-item correlations showed that item 003 had little relationship to the other items in this scale. The correlations between this item and the others were almost perfectly negative.

Statscale consisted of six items measuring various dimensions of statistical analysis procedures. The total errors for the six items was 14 cases. Item 021 had 100% of the cases failing in comparison to 88% for item 029, the most difficult item. The other four items' percent failures ranged between these two extremes. The percent improvement between the coefficients of reproducibility and minimum marginal reproducibility equalled .0992, which met criteria.
However, the coefficient of scalability was .30, well below criteria of .60. The inter-item correlations were high, ranging from .69 to 1.0000.

Admscale was composed of five items and had a total number of errors of 32. Items 007 and 013 had negative inter-item correlations of -1.0000 which would indicate that they are measuring different dimensions. Except for item 019, which had a bi-serial correlation .92, all other items had correlations below .65. Neither coefficients met criteria. See Figure III.

The final evaluation of the six Guttman scales in meeting the stated criteria would tend to indicate that the scales adequately measure the extent to which a respondent's scale score is a predictor of one's response pattern. All coefficients of reproducibility were greater than .82, with three scales attaining over .93. In terms of the ordering of the scales by the degree of difficulty, only Qualscal attained criteria. Admscal and Statscal were well below criteria in this regard with the remaining three scales above .4043 for the coefficient of scalability.

It is clear from Figure III that a few questionnaire items will have to be eliminated in this and subsequent studies. For the time being, the stepwise multiple regression program will automatically delete those items that have little relationship with the criterion. The possibility of building
Figure III

Frequency of Responses (3) and (4)

FACTORS:

Goal Specification
  006
  012
  018

Measurement & Data Collection
  005
  011
  017
  023
  028
  030
  032

Measurement Quality
  004
  010
  016
  022
  027
  031

Evaluation Design
  003
  009
  014
  020
  025
Figure III (Cont.)

Statistical Analysis

| 002 | 008 | 015 | 021 | 026 | 029 |

Administration of Evaluation Procedures

| 001 | 007 | 013 | 019 | 024 |
a scale with two division points instead of three may greatly enhance the internal consistency of the EPS instrument. It would seem that the scale is sufficiently consistent for this initial study.

Another method of presenting evidence of the internal consistency of the Evaluation Procedure Survey is based upon the intercorrelations among the 32 item scores. Again, these intercorrelations were based upon 52 cases.

Intercorrelations between variables or items indicate the extent to which obtained item scores measure some common component or factor. Conversely, the lack of correlation may be used as evidence that the items are measuring something unique or different. In interpreting these intercorrelations it is important that EPS items form a pattern of factors that are orthogonal or independent of each other. In analyzing these correlations, it should be noted that relationships between the scores are weakened by standard measurement errors of each score.

Relatively high correlations among item scores are anticipated to form discrete factors. Some extraneous factors may form large portion of the correlations, such as the technical assistance received by an applicant in preparing its evaluation work plan or experience with previous evaluations. These factors effect the "evenness" of responses to the items in the questionnaire.
The factorial structure of the EPS items was analyzed using the intercorrelation matrix of the 36 items. Principal component analyses were made for the 50 cases. After the principal factor solution was obtained, eight factors were rotated to varimax solution. Table XII shows the results of this analysis. There were 12 factors with eigenvalues of more than 1.08834, which accounts for 3% of the variance. Since six factors are of interest in this study, only that number will be examined.

The estimated commonalities were substantially improved after five iterations using varimax solution. Variables panel rating were increased from .387 to .569, 014 from .623 to .768 and variable 022 from .700 to .920.

The eigenvalues for the eight factors were substantially improved after five iterations. The percent of variation for factor 1 increased from 24.2 to 33.0 and factor 5 from 5.6 to 6.7. The improvement in the percent of variation was substantial for the first five factors. Of the first six factors with significant eigenvalues (1.745) the following variables had significant factor loadings:

<table>
<thead>
<tr>
<th>Factor I</th>
<th>Factor II</th>
</tr>
</thead>
<tbody>
<tr>
<td>014 (Evaluation design)</td>
<td>005 (Measurement/data coll.)</td>
</tr>
<tr>
<td>022 (Reliability/Val.)</td>
<td>009 (Evaluation/design)</td>
</tr>
<tr>
<td>010 (Reliability/Val.)</td>
<td>011 (Measurement/data coll.)</td>
</tr>
<tr>
<td>019 (Reliability/Val.)</td>
<td></td>
</tr>
<tr>
<td>016 (Reliability/Val.)</td>
<td></td>
</tr>
</tbody>
</table>
Table XII
TWELVE FACTORS EXTRACTED AFTER FIVE
ITERATIONS BY EIGENVALUES AND PERCENT OF VARIANCE

<table>
<thead>
<tr>
<th>Factor</th>
<th>Eigenvalue</th>
<th>Percent of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.53</td>
<td>33.0</td>
</tr>
<tr>
<td>2</td>
<td>3.49</td>
<td>13.5</td>
</tr>
<tr>
<td>3</td>
<td>2.92</td>
<td>11.3</td>
</tr>
<tr>
<td>4</td>
<td>1.80</td>
<td>7.0</td>
</tr>
<tr>
<td>5</td>
<td>1.74</td>
<td>6.7</td>
</tr>
<tr>
<td>6</td>
<td>1.52</td>
<td>5.9</td>
</tr>
<tr>
<td>7</td>
<td>1.39</td>
<td>5.4</td>
</tr>
<tr>
<td>8</td>
<td>1.11</td>
<td>4.3</td>
</tr>
<tr>
<td>9</td>
<td>1.04</td>
<td>4.0</td>
</tr>
<tr>
<td>10</td>
<td>.81</td>
<td>3.2</td>
</tr>
<tr>
<td>11</td>
<td>.77</td>
<td>3.0</td>
</tr>
<tr>
<td>12</td>
<td>.68</td>
<td>2.7</td>
</tr>
</tbody>
</table>
The six most significant factors in terms of eigenvalues could be named:

Factor I: Reliability and Validity of Data
Factor II: Measurement and Data Collection
Factor III: Objectives
Factor IV: Reliability and Validity of Data
Factor V: Reliability and Validity of Data
Factor VI: Statistical Analysis
Factors I and IV tend to cluster around items that measure the extent that projects plan to check the reliability and validity of the various instruments, tests and procedures. The four other factors clustered around items classified as "objectives," "measurement and data collection," and "statistical analysis." The significant factor loadings (eigenvalue: 8.53257) on Factor I, "Reliability and Validity of Data," supports the substantial coefficients of reproducibility and scalability of the Guttman scale. However, those items contained in the functions "administration" and "evaluation design" do not appear to be highly interrelated with the other items in order to form meaningful components. This might be due to the wording of the questionnaire items rather from an overlap in evaluation functions.

Whether these factors are the results of basic evaluation functions is ambiguous. Although the items within the factors are highly intercorrelated, the correlations among the factors were moderate to substantial. Factor I's correlation with Factors II and III was .56 and .41. In this regard, Factor II accounts for .31 percent of the variance of Factor I and Factor III for 16 percent. An examination of the transformation matrix shows that the correlations between the factors were small (less than .65). This would give evidence that these factors are measuring different things.
Validity. The review of the validity of the EPS is focused on presenting the basis for subjective judgments concerning whether it actually measures what it purports to measure.

The validity of EPS will have to ultimately be based on how the ESAA applicants use and implement the evaluation work plans and the quality of the evaluation findings, in terms of relevancy and objectivity. However, there is a need to discuss the thinking that formed the basis for the development of the EPS.

The type of statistical data that might be considered as evidence of EPS validity are correlations with panel evaluation scores, coefficients of reproducibility and scalability, the correlations between rankings on panel ratings, and EPS score and the predictive validity of EPS items. These measures reflect on the validity of the EPS, but do not substantiate it. These tests do not prove that EPS is a bonafide instrument or that the survey measures what it purports to measure. High coefficients of reproducibility and scalability may show that EPS is measuring some unified factor consistently but may not demonstrate its validity. The correlation of panel ratings with EPS is evidence of the validity of EPS but it does not show whether panel ratings is itself a valid measure. The quality of panel ratings will have to be based upon their predictive validity and reliability.
The two most important questions in evaluating the validity of the EPS are:

1. Do the EPS items substantially define what evaluation work plans should contain?
2. Are those work plans required of each applicant those that are important for evaluation?

The inclusion of items on the EPS were based on current federal evaluation guidelines, major issues of evaluation and equal educational opportunity and recent review of the literature.

ESAA evaluation guidelines require:

a. The use of experimental or quasi-experimental evaluation designs.
b. The use or check for reliable and valid program measures.
c. The specification of the process for ongoing evaluation.
d. The use of quantifiable goals.
e. The description of procedures for data collection and analysis.

These criteria were used by the panels that rated each ESAA applicant. The correlation between panel ratings and EPS should be significant. The zero order correlation between the panel ratings and EPS = .29, significant at p = .05 level. Also, nonparametric correlations between panel
ratings and \( \text{EPS} = 0.2849 \), significant at \( p < 0.021 \) level.

One could conclude that the Evaluation Procedure Survey is a modest predictor of panel ratings, only slightly better than Evaluation use variable. EPS scores can account for approximately 9% of the variance in panel ratings.
Table XIII

PEARSON CORRELATION COEFFICIENTS

PANEL RATINGS WITH EVALUATION PROCEDURE SURVEY SCORE

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel rating</td>
<td>51</td>
<td>9.23</td>
<td>9.0</td>
</tr>
<tr>
<td>EPS Score</td>
<td>52</td>
<td>46.69</td>
<td>11.48</td>
</tr>
</tbody>
</table>

Zero order correlation = .29 significant at p .03
CHAPTER V
SUMMARY AND CONCLUSIONS

Overview

This chapter summarizes the nature of the problem, focus, and background of this study. The hypothetical model, research procedures used, and results of these procedures are covered in this section.

Summary

Nature of Problem. The inadequacy of local level evaluation for local, state, and federal decision making has been linked to invalid evaluation procedures, a lack of genuine experimentation, and few valid models of social behaviors. The report entitled *ESEA Title I: A Reanalysis and Synthesis of Evaluation Data from Fiscal Year 1965 Through 1970* (Wargo, et. al.) noted that an:

"Analysis of all possibly relevant data sources immediately indicated that nationally representative and valid impact data are simply not available and that some data relating to participation and expenditures also suffer from severe limitation."\(^{69}\)

\(^{69}\)Wargo, p. 32.
Focus of Study. The focus of this study is on examining evaluation procedures proposed by local ESAA projects and its relationship to panel ratings of evaluation quality, the type of grant applied for, the size of the project, evaluation costs, and plans for evaluation use.

The continued problems of local evaluation adequacy are exacerbated by three factors: federal requirements for project evaluation; the focus of funding agencies on national evaluation efforts; and competing evaluation models. In general, there is a lack of consensus about evaluation procedures as they are linked to the production of credible program information. Issues of the feasibility of implementing evaluation work plans have been noted in Federal Evaluation Policy, (Whooley, et. al.). These issues include level of funding for evaluation, the availability of adequate methodology, the size of the project, the type of project, and the possibility that evaluation findings will be used for decision making. It will be difficult to adequately address these issues until evaluators have a better understanding of the interaction among these factors.

70 Whooley, Federal Evaluation Policy, p. 81
Procedures. The research design for this study consisted of a cross sectional survey of 54 ESAA applicants for funding in 1974. The procedures used for this study included mailing out a questionnaire to obtain information about proposed evaluation procedures from the ESAA sample and matching this data with panel ratings of evaluation quality. Additional background data was collected through the use of an Evaluation Procedure Survey. The data was processed and analyzed by batch, using the SPSS program. A stepwise multiple regression, correlation, and discriminant analysis programs were used. The EPS was checked for reliability and validity, using the Guttman scale and factor analysis programs.

Results Obtained. The total Evaluation Procedure Survey score is significantly related to panel ratings of evaluation quality. There was a modest correlation between panel and EPS scores. This correlation was significant at .03 level. However, the sample size of 50 cases may tend to inflate the correlation between these variables because of measurement error.

Further analysis was conducted by examining each pair of relationships between the predictor and criterion variables. The raw chi square of 478.97 for panel and EPS scores was significant at p .01 level. This may indicate
that there is a systematic relationship between these two variables. The cell frequencies of panel and EPS scores showed that the marginals exceeded predicted values. Gamma statistics showed that positive concordant pairs predominated at .24, a low positive correlation. The raw chi squares for the other variables, type of project, size of project, evaluation cost, and evaluation use were not significant.

Nonparametric statistics were also used since the distributions of some of the predictor and criterion variables were not normal. The correlations for size of project, evaluation use, and EPS scores were significant at p .05 level. The first null hypothesis was rejected.

Hypothesis II was accepted. There is a significant difference among the types of grants in terms of panel ratings, EPS score, size of project, evaluation cost, and evaluation use. The initial analysis showed that three variables or functions contributed most towards the differentiation among the group. The F-ratios for variables, evaluation cost, size of project were significant at p .05. EPS was not significant in approximate F-ratio or the change in Raos V. The initial lambda was .68 which
indicated modest discriminating power among the functions. This value of lambda was not significantly different than that predicted nor were the second and third functions. The three functions did not significantly discriminate among groups. Both the Basic and Non Profit groups were close on scores on variables EPS, Evaluation Use, Size of Project, while Pilot and Bilingual’s mean scores on Evaluation Use were identical. Further analysis and classification by type of grant would be meaningless.

The null hypothesis for III was accepted. The relationship between the total Evaluation Procedure Survey score and panel ratings is not significant after the effects of evaluation cost, size of project, type of grant, and evaluation use factors have been taken into account.

The linear regression of panel ratings on the five independent variables was .49. Although the overall multiple correlation was significant at p < .00, EPS scores did not add much to the equation. EPS scores added .05% to the squared variance of panel ratings in comparison to .90% by evaluation cost. The type of project, size of project, and evaluation use variables did not add significantly to the prediction of panel ratings. The small increment in the multiple correlation accounted for by EPS scores may not warrant the expense of collecting the data.
Analysis of Hypothetical Model

The mathematical model used for examining the relationships between panel ratings and the five predictor variables was based upon the works of Blalock and Kerlinger.\textsuperscript{71} This mathematical language has frequently been called path analysis or path coefficients.

The hypothetical model for examining the relationship between panel ratings and the five predictor variables was outlined in Chapter I.

Since one basic assumption of path analysis methods is that all variables are continuous, only size of project evaluation cost, EPS and panel scores are included in this model.

Table IVX shows the correlation matrix for the four variables of interest. Figure 4 shows the paths of the correlations for the four variables.

\textsuperscript{71}Herbert M. Blalock (ed.) Causal Models in the Social Sciences, Chicago: Aldine, 1970

Table IVX

CORRELATION COEFFICIENTS FOR

PANEL RATINGS, EPS SCORES, EVALUATION COST, AND SIZE OF PROJECT

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Panel Ratings</td>
<td>.29</td>
<td>-.23</td>
<td>.31</td>
<td></td>
</tr>
<tr>
<td>2. EPS</td>
<td></td>
<td>.10</td>
<td>-.00</td>
<td></td>
</tr>
<tr>
<td>3. Evaluation Cost</td>
<td></td>
<td></td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>4. Size of Project</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 4

CORRELATION COEFFICIENTS FOR PANEL RATINGS, EPS SCORES, EVALUATION COST, AND SIZE OF PROJECT
According to this model, \textit{EPS} is a dependent variable for \textit{size of project} and \textit{evaluation cost}. \textit{Panel ratings} is a dependent variable for \textit{EPS} and \textit{evaluation cost}.

We might first examine the regression of \textit{EPS} on \textit{size of project} and \textit{evaluation cost} in order to determine the amount variance accounted for by the two independent variables.

First, the overall multiple correlation of .10 for \textit{EPS}, \textit{size of project} and \textit{evaluation cost} was not statistically significant. The \textit{size of the project} and \textit{evaluation cost} accounted for 1\% of the variance in \textit{EPS} scores. 99\% of the variance in \textit{EPS} scores could be attributed to an unmeasured variable.

Second, the overall multiple correlation of .50 for \textit{panel rating}; \textit{EPS}, \textit{size of project} and \textit{evaluation cost} was significant at .002. \textit{EPS} scores, evaluation costs, and the \textit{size of project} accounted for 25\% of the variance in panel ratings. We shall focus on this second analysis since one might wish to start from the cause closest to the dependent variable and trace backwards to the more distant causes. See Table XV for summary statistics.

The overall percent variance accounted for by the three predictor variables is 25\%. When \textit{evaluation cost} and \textit{size of project} are entered first they jointly account for 15\% of the variance. The increment due to \textit{EPS} is about 10\%. 
<table>
<thead>
<tr>
<th>(A) Step I Variables Entered/Removed</th>
<th>F To Enter</th>
<th>Significance</th>
<th>Cost</th>
<th>Size of Project</th>
<th>R Squ. Change</th>
<th>R Squ. Change Significance</th>
<th>R</th>
<th>F</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5.07</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Overall F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.77</td>
<td>.253</td>
<td></td>
</tr>
<tr>
<td>R Sq. Change MR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R Sq. Change MR Significance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.10</td>
<td>-.00</td>
<td></td>
</tr>
<tr>
<td>R Sq. Change MR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.01</td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>R Sq. Change MR Significance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.01</td>
<td>.01</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(B) Step II Variables Entered/Removed</th>
<th>F To Enter</th>
<th>Significance</th>
<th>Cost</th>
<th>Size of Project</th>
<th>R Squ. Change</th>
<th>R Squ. Change Significance</th>
<th>R</th>
<th>F</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.74</td>
<td>.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Overall F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.73</td>
<td>5.57</td>
<td></td>
</tr>
<tr>
<td>R Sq. Change MR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.23</td>
<td>.31</td>
<td></td>
</tr>
<tr>
<td>R Sq. Change MR Significance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.05</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td>R Sq. Change MR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.23</td>
<td>.29</td>
<td></td>
</tr>
<tr>
<td>R Sq. Change MR Significance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.25</td>
<td>.01</td>
<td></td>
</tr>
</tbody>
</table>

Table XV
MULTIPLE REGRESSION OF EPS ON SIZE OF PROJECT AND EVALUATION COST AND PANEL RATINGS ON EPS, SIZE OF PROJECT AND EVALUATION COST
As shown in Figure 5, the numbers in parentheses indicate zero-order correlations. For example, correlation between panel score and size of project equals .31.

Only the regression of panel ratings on EPS, Evaluation cost, and size of project was significant at $p = .03$, indicating that the multiple correlation was due to direct effects. The correlation between size of project, evaluation cost and EPS was not significant and leads to the conclusion that the present model can be trimmed. A more parsimonious model is presented in Figure 6.

The three predictor variables are orthogonal and each one's multiple correlations can be added together. Figure 6 shows that EPS, evaluation cost, and size of project have separate direct effects upon the variance in panel ratings.

Limitations of this Study

The limitations of the design and implementation of this study are related to sample size, instrumentation, manipulation of antecedent variables, and statistical analysis.

A larger sample of 150 to 200 cases would have been preferred to the 50 cases of this study. Larger sample size could insure a normal distribution for each array of panel scores. For example, 54% of the cases had panel scores in the lower range of 0 to 6 points in comparison to 11% in the
Figure 5
PATH COEFFICIENTS FOR PANEL RATING, EPS SCORES, EVALUATION COST, AND SIZE OF PROJECT

Evaluation Cost
  .01
  Evaluation Cost
  .10 (.10)
  EPS
  .39 (-.00)
  Size of Project
  .23 (-.23)

Panel Rating

EPS
  .50 (29)

Figure 6

Panel Rating

EPS
  .11

Size of Project
  .16

Evaluation Cost
  .23
higher range of 21 to 24. To obtain a larger sample, a
stratified random cluster sample would have to be drawn from
other regions. Equal numbers of Basic, Non Profit, Pilot,
and Bilingual applicants would have to be included in the study.

Secondly, the EPS had several questionnaire items
with weak biserial correlations. These items were identified
in one of the multiple regression programs and in the Guttman
Scale analysis. The inclusion of these weak items in the
computation of the EPS score greatly reduced the possible
effect between this variable and panel ratings. This study
served as an item tryout, so to speak, to strengthen the
construction of the Evaluation Procedure Survey. However,
this study did not focus on determining the item biserial
correlations with the total EPS score rather than the
relationship between each item and the panel score. The
biserial correlation for items in EPS should be examined in
future studies. Technical considerations concerning the
internal consistency of the EPS showed that the instrument
could be substantially improved. Only one factor, "Measurement
Quality," met the criteria of unidimensionality and
scalability. Three scales met the criteria of measuring the
same thing or dimension. Factor analysis of the 32 items in
the survey showed that there were nine distinct components
or clusters as opposed to the expected six. In most cases, items formed subscales within scales as opposed to combining with other items of different scales. Consideration should be given to reducing the number of weak items in each factor so that the components could become more meaningful.

The matching of panel ratings with EPS scores through the use of an *ex post facto* design did not permit the manipulation of antecedent variables, type of grant, etc. It would have been preferable to match all cases by type of grant, size of project and evaluation use and vary evaluation costs. All cases could have been matched on antecedent variables but randomly assigned to one group for special training or another for general proposal development. Many of the problems concerning external validity, particularly the reactive effects of responding to mail out questionnaires.

Lastly, each antecedent factor should be treated as separate variables. For example, type of grant should be four distinct variables as opposed to one. The interpretation of the results from the multiple regression program could have been more intuitive, i.e., the separate effect of each type of grant on the prediction of panel ratings could be judged. These limitations should be addressed in future studies.
Conclusions

The ESAA panel ratings and EPS scores were extremely low. Sixty-five percent received panel ratings below 14 points and 87% of the applicants received EPS scores below 60 points. (See Table XVI). Evaluation costs for the ESAA applicants were negatively related to panel ratings. Ninety-five percent (95%) of those applicants that received high panel ratings had relatively low evaluation costs in comparison to 5% with costs of 11% or more.

Larger projects received higher panel scores than smaller projects. Over 60% of the applicants with high panel scores had more than 5,000 project participants.

Seventy-five percent (75%) of the applicants with high panel scores planned to use their evaluation results to determine the impact of their project in comparison to 25% for context and process evaluation.

Recommendations. Prior to the submission, each organization should check its evaluation work plan to determine whether its score exceeds the median EPS score.

A maximum cost for evaluations should be established rather than a minimum cost. Those applicants that planned to spend 1% of the total project cost on evaluation received higher panel scores.

Projects with limited numbers of participants should be required to devote more attention to developing sound evaluation plans.
Table XVI
FREQUENCY DISTRIBUTIONS OF PANEL RATINGS, TYPE OF GRANT, SIZE OF PROJECT, EVALUATION COST, EVALUATION USE, AND EVALUATION PROCEDURE SURVEY SCORE

<table>
<thead>
<tr>
<th>Panel Rating/Type of Grant</th>
<th>Basic</th>
<th>NPO</th>
<th>Pilot</th>
<th>Bilingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-6</td>
<td>15</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7-13</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>14-20</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>21-24</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Chi square = 47.36  42 df significant = .26

<table>
<thead>
<tr>
<th>Panel Rating/Size of Project</th>
<th>(number of students)</th>
<th>100</th>
<th>500</th>
<th>1,000</th>
<th>5,000</th>
<th>5,000+</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-6</td>
<td>8</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>7-13</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>14-20</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>21-24</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Chi square = 53.26  56 df significant = .57
Table XVI (Cont.)

<table>
<thead>
<tr>
<th>Panel Rating/Eval. Cost</th>
<th>0-1%</th>
<th>2-3%</th>
<th>4-6%</th>
<th>7-10%</th>
<th>11+%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-6</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>7-13</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14-20</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>21-24</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>11</td>
<td>9</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Chi square 42.92  52 df significant = .81

<table>
<thead>
<tr>
<th>Panel Rating/Eval. Use</th>
<th>Context</th>
<th>Process</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-6</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>7-13</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>14-20</td>
<td>2</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>21-24</td>
<td>0</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4</td>
<td>27</td>
</tr>
</tbody>
</table>

Chi square = 15.96  24 df significant = .88

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0-6</td>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>7-13</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>14-20</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>21-24</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>7</td>
</tr>
</tbody>
</table>

Chi square = 478.97  364 df significant = .00
Concluding Remarks. Over 8% of our gross national product is spent on funding educational programs on a federal level. However, many problems still prevail.

Educational inequality is perhaps the most critical of our nation's problems. School desegregation, compensatory and bilingual education and programs for the handicapped are still unresolved issues. We need to develop a national agenda and test solutions for these problems.

Current evaluation procedures, as a scientific method, are not adequate to test solutions to these problems. This fact may be largely due to the complexity of behavioral science.

First, long-term effects of programs are difficult to evaluate because of the action setting and variability of educational environments. Scriven noted that evaluation's focus must be "goal free" to account for factors outside of the school or immediate program.

Second, the linear effects of research and development cannot be uniformly applied to educational programming as is the case with the biological and physical sciences. Products developed for the classroom vary in its effectiveness according to use, training of teachers, the characteristics of the learning, and other historical contexts.

Third, the goals of public education and programs originate from judgments and priorities previously made by Congress, funding agencies and local or state administrators.
Title I and Follow Through are examples of the conflict between deep-seated commitment to program goals and negative evaluation evidence to continue, modify, or terminate these programs. Judgments previously made by sponsors may not be sufficiently different from traditional values concerning program alternatives.

Fourth, research and evaluation are generally perceived as pejorative and arcane terms by local and state administrators. Often more attention is paid to program planning than to evaluation.

Lastly, evaluation theories, practices, and models have tended to be dominated by research of psychologists such as Skinner, Cronbach, and Thorndike. The theories of sociologists, political scientists, and economists need to be included in developing evaluation models. The Office of Education has recognized the importance of these theories and initiated the use of social, economic, and educational indicators in the report *Condition of Education*.

There are some signs that local, state, and national administrators are becoming more concerned about accountability, assessment, and evaluation. However, we need an agenda for planning, using, and disseminating educational evaluations. We need to focus attention on the area of educational
inequality. This task is not an easy one since there are many constraints related to local autonomy, rights of privacy, parent and community involvement. Perhaps most important, we need to focus on the validity of current evaluation procedures and the credibility of the information obtained for decision making.
I Goal Specification

1. The specification of criteria that defines the attainment of cognitive objectives such as knowledge, comprehension, understanding, skills, or applications.

2. The specification of criteria that defines the attainment of attitudinal, appreciation, interest, or self-concept objectives.

3. The specification of a set of criteria that defines the attainment of affective and cognitive objectives.

II Measurement and Data Collection

1. The identification of standardized achievement or ability tests to measure the attainment of cognitive objectives.

2. The identification of standardized test to measure the attainment of affective objectives.
3. The indication of the use of frequency counts of absences, lateness, discipline referrals, attendance or suspensions, as an indirect measure of student behavior.

4. The specification of teacher-made tests to be related with standardized test results.

5. The specification of the use of rating scales or checklists for observing teacher and student classroom behavior or performance to measure intra-program effects.

6. The indication of the frequency of use of anecdotal records, teacher logs, or case studies of students' performance during the program.

7. The specification of the use of self-rating reports, scales, check lists, or inventories in which students compare perceived levels of achievement with future educational plans.

III Measurement Quality

1. The specification of how other measures, in addition to standardized tests, will be used to determine the achievement of the objectives, such as motivation, behavior, or performance.
2. The indication of evaluation procedures that monitor the testing of students to validate the consistency of test instruction, pupil characteristics, and test situations.

3. The specification of how test-taking skills, anxiety, motivation, speed, guessing, and test instructions will be examined in the analysis of test results.

4. The specification of statistical procedures that examine subgroup scores on alternative forms of a test to validate its stability.

5. The indication of the frequency of which separate norms and reliability coefficients will be established for various subgroups participating in the program.

6. The indication of how race, language, SES, or personality will be examined to determine the degree to which these factors account for current achievement test scores.

IV Evaluation Design

1. The specification of an evaluation design that systematically describes and analyzes a single program based upon the observations of project staff.
2. The specification of an evaluation design that investigates the effects of a program on its participants by using a pre and post test.

3. The indication of an evaluation design that investigates the extent to which variations of scores in one test correspond to variations in another.

4. The description of evaluation methods that will examine how current test scores are accounted for by predictive variables, such as aptitude, program attendance, or background factors.

5. The description of evaluation methods that randomly assigns students to treatment or control groups in order to experimentally examine the effects of the program.

V Statistical Methods

1. The description of statistical methods that compare changes in means, ranks, standard deviations, percentages, or signs to determine the effects of the program.
2. The description of statistical methods that examines categories of project participants, such as race, sex, or performance levels, to account for patterns of correlations.

3. The specification of methods that measure differences between the categories of actual and predicted frequencies, such as race, sex, and performance levels, to test their statistical significance.

4. The specification of methods that compare the relationship between two scores while holding a third score constant and allowing the others to vary.

5. The specification of statistical methods that test whether there is a difference between the "between" and the "within" group variance, to determine the effects of the program on the treatment and the control groups.

6. The indication of statistical methods that will test whether the results between two or more measures are statistically significant.
VI Administration of Evaluation Procedures

1. The specification of a schedule to monitor all of the evaluation activities to determine adherence to established time table and procedures.

2. The description of the frequency of the training of the staff to determine the quality of the evaluation work plan.

3. The specification of the frequency that evaluation results will be reported to various interest groups.

4. The specification of the frequency of the monitoring of the administration and scoring of tests by project staff to validate the adherence to established procedures.

5. The specification of the use of frequency counts of teacher conferences, school visits, or attendance at special school events, by parents of project participants, to assess their attitude or understanding of the program.
<table>
<thead>
<tr>
<th>ESAA Population 1974 Region I</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Hartford Board of Education</strong></td>
</tr>
<tr>
<td><strong>2. Charles River Academy</strong></td>
</tr>
<tr>
<td><strong>3. Interchange, Inc.</strong></td>
</tr>
<tr>
<td><strong>4. Holyoke Public School</strong></td>
</tr>
<tr>
<td><strong>5. New Bedford Public School</strong></td>
</tr>
<tr>
<td><strong>6. METCO Boston</strong></td>
</tr>
<tr>
<td><strong>7. Lincoln Public Schools</strong></td>
</tr>
<tr>
<td><strong>8. Lawrence Public Schools</strong></td>
</tr>
<tr>
<td><strong>9. Medford Public Schools</strong></td>
</tr>
<tr>
<td><strong>10. New Haven Board of Education</strong></td>
</tr>
<tr>
<td><strong>11. Newport (MLK)</strong></td>
</tr>
<tr>
<td><strong>12. ABCD Boston</strong></td>
</tr>
<tr>
<td><strong>13. Stamford Board of Education</strong></td>
</tr>
<tr>
<td><strong>14. West Hartford Board of Education</strong></td>
</tr>
<tr>
<td><strong>15. Middletown Board of Education</strong></td>
</tr>
<tr>
<td><strong>16. Farmington Board of Education</strong></td>
</tr>
<tr>
<td><strong>17. Waterbury Board of Education</strong></td>
</tr>
<tr>
<td><strong>18. New Haven Board of Education</strong></td>
</tr>
<tr>
<td><strong>19. New Haven Board of Education</strong></td>
</tr>
<tr>
<td><strong>20. Urban League Stamford</strong></td>
</tr>
<tr>
<td><strong>21. SPHERE Hartford</strong></td>
</tr>
<tr>
<td><strong>22. Hampshire College</strong></td>
</tr>
<tr>
<td><strong>23. Needham Board of Education</strong></td>
</tr>
</tbody>
</table>
24. SPHERE

NPO

25. Cambridge Board of Education

Basic

26. Glastonbury Board of Education

Basic

27. Lincoln-Sudbury Board of Education

Basic

28. Pawtucket Board of Education

Basic

29. New Haven, U.L.

NPO

30. Waltham Board of Education

Basic

31. Lexington Board of Education

Basic

32. Reading Board of Education

Basic

33. Watertown Board of Education

Basic

34. Urban Ethics, Glastonbury

NPO

35. Model Cities, Springfield

NPO

36. Norwalk Board of Education

Bil

37. Fall River Board of Education

Bilingual

38. Newport Board of Education

Basic

39. Providence Board of Education

Basic

40. Hamden Board of Education

Basic

41. Canton Board of Education

Basic

42. Medford Board of Education

Basic

43. Bridgeport Board of Education

Pilot

44. Brookline Board of Education

Basic

45. Springfield Board of Education

Basic

46. Newton Board of Education

Basic

47. Providence Corp.

NPO

48. PYCO, Fall River

NPO
49. Urban League, Springfield
50. Fall River Board of Education
51. Bridgeport PRYO
52. Providence School Department
53. University of Maine
54. Adm. Dist. Maine

NPO
Basic
NPO
Bil
NPO
Basic
APPENDIX C

Pilot Study of Correlation Between Percent Evaluation Cost and Panel Ratings on Sample of 20 ESAA Applicants

<table>
<thead>
<tr>
<th>N</th>
<th>Mean Percent Evaluation Cost</th>
<th>Mean Panel Rating</th>
<th>r</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>4%</td>
<td>11.6</td>
<td>-0.37</td>
<td>p 0.03</td>
</tr>
</tbody>
</table>

The correlation coefficient between the two variables percent evaluation cost and panel scores was -0.37. The relationship is negative and accounts for 13% of the variance between the two variables. Eighty-seven percent (87%) of the variance in panel scores can be accounted for by some other unknown or unmeasured variables.

It would appear that a sample correlation between panel scores and a crude index of evaluation quality would be spurious unless we could control for other variables.
APPENDIX D

REVISED EVALUATION PROCEDURE SURVEY

**Evaluation Procedure Survey**

Name of Organization:

Contact Person:

Phone:

Mark /X/ One

<table>
<thead>
<tr>
<th>I</th>
<th>Type of Grant</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basic: ( )New ( )Continuation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pilot: ( )New ( )Continuation</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>II</th>
<th>Size of Evaluation Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) 50-100</td>
</tr>
<tr>
<td></td>
<td>(2) 101-300</td>
</tr>
<tr>
<td></td>
<td>(3) 301-600</td>
</tr>
<tr>
<td></td>
<td>(4) 601-900</td>
</tr>
<tr>
<td></td>
<td>(5) 901-Up</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>III</th>
<th>% Evaluation Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>0</td>
</tr>
<tr>
<td>(2)</td>
<td>1-2%</td>
</tr>
<tr>
<td>(3)</td>
<td>3-4%</td>
</tr>
<tr>
<td>(4)</td>
<td>5-6%</td>
</tr>
<tr>
<td>(5)</td>
<td>7-Up%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IV</th>
<th>Intended Use of Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Formative</td>
</tr>
<tr>
<td></td>
<td>Summative</td>
</tr>
<tr>
<td>CRITERIA</td>
<td>RATING</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>For each objective, the applicant has:</td>
<td>( )None ( )Some ( )Many ( )All ( ) N.A.</td>
</tr>
<tr>
<td>1. Specified cognitive objectives in measurable terms</td>
<td>( ) ( ) ( ) ( ) ( )</td>
</tr>
<tr>
<td>2. Specified a set of affective and cognitive objectives in measurable terms</td>
<td>( ) ( ) ( ) ( ) ( )</td>
</tr>
<tr>
<td>3. Identified objective instruments to measure cognitive objectives</td>
<td>( ) ( ) ( ) ( ) ( )</td>
</tr>
<tr>
<td>4. Proposed the use of nonobstrusive indicators as an indirect measure of cognitive or affective objectives</td>
<td>( ) ( ) ( ) ( ) ( )</td>
</tr>
<tr>
<td>5. Specified the use of instruments to determine participants' attitudes towards the project</td>
<td>( ) ( ) ( ) ( ) ( )</td>
</tr>
<tr>
<td>6. Specified the use of nonobstrusive measures, in addition to objective tests, to validate student performance</td>
<td>( ) ( ) ( ) ( ) ( )</td>
</tr>
<tr>
<td>7. Specified how student characteristics will be examined to account for variance in objective test scores</td>
<td>( ) ( ) ( ) ( ) ( )</td>
</tr>
<tr>
<td>8. Specified how student characteristics will be examined to account for variance in objective test scores</td>
<td>( ) ( ) ( ) ( ) ( )</td>
</tr>
<tr>
<td>CRITERIA</td>
<td>RATING</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>9. Indicated statistical methods that describe data in summary form, such as means, averages, percentages, percentiles, etc.</td>
<td></td>
</tr>
<tr>
<td>10. Indicated methods to test whether the results between two or more independent measures are statistically significant</td>
<td></td>
</tr>
<tr>
<td>11. Specified methods to determine whether differences among subgroups or independent measures are educationally significant.</td>
<td></td>
</tr>
<tr>
<td>12. Specified a formal schedule of evaluation activities</td>
<td></td>
</tr>
<tr>
<td>13. Specified how evaluation results will be disseminated to interested groups and local decision makers</td>
<td></td>
</tr>
<tr>
<td>14. Specified the formal goals, purposes, and aims of the evaluation work plan</td>
<td></td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY


Blalock, H.M., Causal Models in the Social Sciences; Chicago: Aldine, 1970.


Gage, N.L., et. al., Explorations of the Teacher's Effectiveness; Stanford, Calif.: Stanford Center for Research and Development in Teaching, Stanford, University, 1968.


