INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106
NELSON, Steven Robert, 1948-
THE USE OF MULTIPLE-CRITERION WEIGHTING TECHNIQUES IN EDUCATIONAL DECISION-MAKING.
The Ohio State University, Ph.D., 1975
Education, general

Xerox University Microfilms, Ann Arbor, Michigan 48106
THE USE OF MULTIPLE-CRITERION WEIGHTING TECHNIQUES
IN EDUCATIONAL DECISION-MAKING

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Steve Robert Nelson, A.A., B.A., M.A.

* * * * *

The Ohio State University
1975

Reading Committee:
Robert Lange
Gregory Trzebiatowski
Norval McCaslin

Approved By

Adviser
Faculty of Educational Development
ACKNOWLEDGEMENTS

This study is dedicated to each and all of the individuals who have contributed to the learning environment within which I have grown and will continue to develop. The study is in acknowledgement of those individuals who have dared to improve education through the application of the evaluation methods available, those individuals who have improved education in response to these evaluations and the few individuals dedicated to the improvement of evaluation practices.

Specifically, this dissertation was written in response to the encouragement of Robert Lange, Gregory Trzebiatowski, and Norval McCaslin, members of my graduate committee, the work of Robert Eckenrode of the American Stock Exchange, the advice of Mark Greene of the Northwest Regional Educational Laboratory and Arthur Mittman of the University of Oregon, and the exhortations of my wife, Karen.
VITA

February 18, 1948................Born- Vancouver, Washington

1968............................A. A., Clark College, Vancouver, Washington


1970-1972.......................Research Assistant, Northwest Regional Educational Laboratory, Portland, Oregon

1972............................M. A., The Ohio State University, Columbus, Ohio

1972-1973.......................Research Associate, Center for Vocational and Technical Education, Columbus, Ohio

1973-............................Evaluation Specialist, Northwest Regional Educational Laboratory, Portland, Oregon

FIELDS OF STUDY

Major Field: Educational Evaluation
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Heading</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACKNOWLEDGEMENTS</td>
<td>ii</td>
</tr>
<tr>
<td></td>
<td>VITA</td>
<td>iii</td>
</tr>
<tr>
<td></td>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>I.</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>The Problem</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The Purpose of the Study</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Definition of Terms</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The Study Questions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Limitations of the Study</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overview of the Methodology</td>
<td></td>
</tr>
<tr>
<td>II.</td>
<td>REVIEW OF THE LITERATURE</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Decision Theory</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evaluation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Management Theory</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Psychological Scaling</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Operations Research</td>
<td></td>
</tr>
<tr>
<td>III.</td>
<td>PROCEDURES</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Methodological Approach</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Research Design</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pilot Studies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Selection of Subjects</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Instrumentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Field Procedures</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data Reduction and Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Methodological Assumptions and Limitations</td>
<td></td>
</tr>
<tr>
<td>IV.</td>
<td>RESULTS</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>Case Study One</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Case Study Two</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Case Study Three</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Case Study Four</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Summary of Results</td>
<td></td>
</tr>
</tbody>
</table>
V. CONCLUSIONS AND RECOMMENDATIONS

Summary of the Study
General Conclusions
Recommendations

APPENDIX

A.......................................................... 164

BIBLIOGRAPHY.................................................187
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Research Design</td>
<td>49</td>
</tr>
<tr>
<td>2. Organization of Composite Criterion Weights</td>
<td>66</td>
</tr>
<tr>
<td>3. Results of the Criterion Weighting Method Applications of Study Group One</td>
<td>87</td>
</tr>
<tr>
<td>4. Derived Inter-Method Correlations (r) for Study Group One</td>
<td>89</td>
</tr>
<tr>
<td>5. Method Convergent Reliabilities for Study Group One</td>
<td>90</td>
</tr>
<tr>
<td>6. Elapsed-Time in Minutes for Application of the Criterion Weighting Methods by Study Group One</td>
<td>91</td>
</tr>
<tr>
<td>7. Method Difficulties Perceived by Study Group One</td>
<td>93</td>
</tr>
<tr>
<td>8. Composite Method Difficulty Weights for Study Group One</td>
<td>94</td>
</tr>
<tr>
<td>9. Composite Applicability Weights of the Five Criterion Weighting Methods for Study Group One</td>
<td>95</td>
</tr>
<tr>
<td>10. Summary of Study Group One Findings</td>
<td>97</td>
</tr>
<tr>
<td>11. Results of the Criterion Weighting Method Applications of Study Group Two</td>
<td>99</td>
</tr>
<tr>
<td>12. Derived Inter-Method Correlations (r) for Study Group Two</td>
<td>101</td>
</tr>
<tr>
<td>13. Method Convergent Reliabilities for Study Group Two</td>
<td>102</td>
</tr>
<tr>
<td>14. Elapsed-Time in Minutes for Application of the Criterion Weighting Methods in Study Group Two</td>
<td>104</td>
</tr>
<tr>
<td>15. Method Difficulties Perceived by Study Group Two</td>
<td>106</td>
</tr>
<tr>
<td>16. Composite Method Difficulty Weights for Study Group Two</td>
<td>107</td>
</tr>
<tr>
<td>17. Composite Applicability Weights of the Five Criterion Weighting Methods for Study Group Two</td>
<td>108</td>
</tr>
<tr>
<td>Table</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>18. Summary of Study Group Two Findings</td>
<td>110</td>
</tr>
<tr>
<td>19. Results of the Criterion Weighting Method Applications of Study Group Three</td>
<td>113</td>
</tr>
<tr>
<td>20. Derived Inter-Method Correlations (r) for Study Group Three</td>
<td>115</td>
</tr>
<tr>
<td>21. Method Convergent Reliabilities for Study Group Three</td>
<td>116</td>
</tr>
<tr>
<td>22. Elapsed-Time in Minutes for Application of the Criterion Weighting Methods by Study Group Three</td>
<td>118</td>
</tr>
<tr>
<td>23. Method Difficulties Perceived by Study Group Three</td>
<td>120</td>
</tr>
<tr>
<td>24. Composite Method Difficulty Weights for Study Group Three</td>
<td>121</td>
</tr>
<tr>
<td>25. Composite Applicability Weights of the Five Criterion Weighting Methods for Study Group Three</td>
<td>122</td>
</tr>
<tr>
<td>26. Summary of Study Group Three Findings</td>
<td>124</td>
</tr>
<tr>
<td>27. Results of the Criterion Weighting Method Application of Study Group Four</td>
<td>127</td>
</tr>
<tr>
<td>28. Derived Inter-Method Correlations (r) for Study Group Four</td>
<td>129</td>
</tr>
<tr>
<td>29. Method Convergent Reliabilities for Study Group Four</td>
<td>130</td>
</tr>
<tr>
<td>30. Elapsed-Time in Minutes for Application of the Criterion Weighting Methods by Study Group Four</td>
<td>132</td>
</tr>
<tr>
<td>31. Method Difficulties Perceived by Study Group Four</td>
<td>134</td>
</tr>
<tr>
<td>32. Composite Method Difficulty Weights for Study Group Four</td>
<td>135</td>
</tr>
<tr>
<td>33. Composite Applicability Weights of the Five Criterion Weighting Methods for Study Group Four</td>
<td>136</td>
</tr>
<tr>
<td>34. Summary of Study Group Four Findings</td>
<td>138</td>
</tr>
<tr>
<td>35. Nonparametric Convergent Reliabilities (τ) Observed Across Study Groups</td>
<td>140</td>
</tr>
<tr>
<td>36. Parametric Convergent Reliabilities (r) Observed Across Study Groups</td>
<td>141</td>
</tr>
</tbody>
</table>
Table | Page
---|---
37. Mean Elapsed-Time (In Minutes) Required For Method Applications Across Study Groups | 145
38. Composite Difficulty Weights Observed for Each Method Across Study Groups | 148
39. Composite Applicability Weights Observed for Each Method Across Study Groups | 150
40. Results of the Study Questions and Hypotheses | 152
I. INTRODUCTION

A. The Problem

Decision-making in education is often based upon judgements of both a subjective and qualitative nature. The methods enlisted in education for processing such judgemental information are presently inadequate. Formalized methods have rarely been utilized by educational decision-making bodies for prioritizing alternative needs, solutions or criteria in a systematic manner. Although the systematic structuring of preferential judgements would benefit the decision-making process, little research has been conducted for determining what methods are most applicable for educational decision-making and what characteristics such methods should manifest. If criterion weighting methods are to be applied in educational settings, it would be desirable to collect information which could be utilized for determining which existing methods are appropriate.

Within the field of operations research, a number of methods have been developed for deriving the relative weights of objectives through the use of preferential judgements. Although these methods have been tested and utilized in other disciplines, they have entertained little notoriety in education. These methods would appear to be appropriate for prioritizing alternatives in a variety of educational decision-making settings, such as the prioritization of alternative needs, goals,
objectives, solution strategies, and the criteria for evaluating alternatives. However, little information is available regarding the adequacy of the weighting methods for processing educational decisions. Furthermore, there is a lack of knowledge concerning the comparability of the outcomes generated by the alternative methods. As a result, it is difficult to determine the relative efficiency of these methods.

Assuming that these methods are appropriate for processing preferential judgements in education, it would be beneficial to test the compatibility of these methods in educational settings. Such information could provide a basis for the selection and exclusion of existing criterion weighting techniques for various decision settings encountered in education and for determining which methods may be more appropriate under given circumstances.

B. The Purpose of the Study

The purpose of the present study is to provide an initial source of evidence for determining the applicability of five multiple-criterion weighting methods for educational decision-making. Four characteristics of the weighting methods will be described as a basis for determining their applicability. The first variable, inter-method reliability, will be used to determine the consistency of the outcomes generated by the alternative methods. The second variable, elapsed-time for method application, focuses upon the comparability of the time required to apply the processes utilized by each method, while the final two variables, method difficulty and perceived method applicability, relate to the perceptions of the individuals utilizing each method.
The study focuses upon the determination of the four characteristics for each of the five weighting methods. As an initial source of evidence, the study is based upon the application of these methods by a sample of educational personnel in four simulated decision-making settings. Thus, the study serves as a source of information for determining the applicability of the weighting methods for processing preferential judgements in decision settings encountered in education.

C. Definition of Terms

The variables and weighting methods under consideration in the present study are operationally defined in this section. The four study variables are defined below:

1. Inter-method reliability. The degree of association, as measured by Kendall's rank correlation coefficient and Pearson's product-moment correlation coefficient, between the derived criterion weights of a method and the composite weights of the criterion derived from the balance of the methods.

2. Elapsed-time for method application. The average time required, in minutes, to make comparative judgements among a specified number of criteria, e.g., seven and ten criteria, respectively, for each method used.

3. Method difficulty. The judgement of subjects regarding the method perceived as easiest and most difficult to apply in weighting a specified number of criteria, e.g., seven and ten criteria, respectively, as measured by responses to questions posed of the subjects in the study instruments.
4. **Perceived method applicability.** The preferential judgement of subjects regarding the method perceived as most applicable to the decision-making needs of each subject, as measured by responses to a question posed of the subjects in the study instruments.

Five multiple-criterion weighting methods have been identified as potentially applicable for processing preferential judgements in educational decision-making. The methods under consideration are described below:

1. **Ranking.** Judges assign a numerical rank, from 1 to $k$, to each of the $k$ criteria, where the numerical value of 1 indicates the most important criterion and the numerical value of $k$ represents the least important criterion.

2. **Rating.** Judges assign a numberical value along a continuum, which ranges from 0 to 10, to each of the $k$ criteria, where the numerical value of 10 represents the maximum importance possible and 0 indicates no importance.

3. **Partial-paired comparisons.** Judges indicated in each of $k(k-1)/2$ pairs of criterion comparisons, which criterion is more important in the pair.

4. **Complete-paired comparisons.** Judges indicate in each of $k(k-1)$ pairs of criterion comparisons, which criterion is more important in the pair.

5. **Successive comparisons.** Judges assign initial weights between 0 and 10 to each of the $k$ criterion, where 10 represents maximum importance and 0 represents no importance.
Through an iterative process, the judges successively compare the relative values of the criterion of most importance with the combined values of the balance of the criteria. Adjustments are made to the numerical values of the criteria, based upon these successive comparisons. A normalization process results in a weight assigned to each criterion ranging from 1 to 0, where 1 represents the most important criterion and 0 a criterion having no importance.

D. The Study Questions

Each of the four variables under consideration by the study was selected in response to a specific question regarding the characteristics of the five multiple-criterion weighting methods. The questions posed of the study are:

1. What is the inter-method reliability of each of the five methods?

2. What is the average elapsed-time required to apply each of the five methods when applied to seven and ten criteria, respectively?

3. Which methods are perceived by the subjects as being most difficult and easiest to use?

4. Which method(s) are considered to be most applicable to the decision-making needs of the decision-makers sampled?

Furthermore, hypotheses have been advanced in predicting the outcomes of two study questions. First, it is expected that the inter-method reliabilities for each method will all be significantly high. Thus,
it is not felt that the dimension of reliability will discriminate between the methods under study. Secondly, it is hypothesized that the ranking method will require the least elapsed-time for application and the successive comparisons method will require the greatest elapsed-time for application, regardless of whether seven or ten criteria are being weighted.

The four study questions have been selected as components of the overall dimension of "method applicability." The variables under study are not considered, however, to represent all characteristics relevant to method applicability. Thus, the ultimate question of method applicability to educational decision-making can be addressed, but not necessarily answered, by the present study.

E. Limitations of the Study

The findings of the present study are delimited along two dimensions, e.g., internal validity and generalizability. Along the first dimension, the study of inter-method reliability required the application of the five methods in close succession. Although the sequence of the methods is controlled, the ratings are not totally independent since the effects of prior ratings are not erasable. As a result, it is cautioned that the inter-method reliabilities generated may be liberal estimates of "true reliability."

The second dimension, generalizability, is constrained along four lines. First, the sample of subjects is not generalizable to the universe of all educational decision-makers. Secondly, the decision-settings "sampled" do not necessarily represent all decisions
faced by the field of education. Third, it is assumed that the
number of criteria being weighted will affect the relative applicability
of the methods. Although the number of criteria could potentially
range substantially, only seven and ten criteria were studied.
Thus, the outcomes of the study are limited to these numbers.
Furthermore, although second-order effects may exist between
independent variables, such as an interaction between the number of
criteria and the level of difficulty of the decision setting, an
analysis of such effects are not within the parameters of the
present study. Finally, the study setting involves the application
of all five methods, rather than one method. Since the application
of one method for solution of a problem is the usual strategy
utilized in the field, the study is not necessarily representative
of actual decision-making processes. These limitations are not
considered serious in light of the exploratory purpose of the study.
As an initial source of information, the results of the study
emphasize feasibility at the expense of extensive generalizability.

F. Overview of the Methodology

A field study approach serves the exploratory nature of the
study. The study is based upon the application of five multiple-
criterion weighting methods in four simulated decision settings.
In the first setting, five educational evaluators were randomly
selected from the evaluation staff of a regional educational
laboratory in the State of Oregon. These individuals applied the
methods to seven criteria proposed for assessing the quality of
educational evaluation. The second setting consisted of five educational evaluators randomly selected from the evaluation staff of a metropolitan school district in the State of Oregon. These individuals applied the methods to the same seven criteria utilized in the first setting.

In the third setting, the five Oregon State Department of Education personnel responsible for administering projects funded through Title III of the Elementary and Secondary Education Act (ESEA) were selected. These individuals applied the methods to ten criteria proposed for determining the continuation of exemplary educational projects. The fourth setting consisted of a random sample of five ESEA Title III project directors in Oregon. These individuals applied the methods to the same ten criteria utilized in the third setting.

Although the study is primarily descriptive in nature, and cause and effect relationships are not under investigation, the study has been designed to control for multiple-treatment interference. The design incorporates the random sequencing of the methods to be applied by each of the subjects.

Two instruments were developed and pilot tested which were used for instructing the subjects on the application of the methods, recording the results of the method applications, and querying the subjects of their perceptions of the method's difficulty and applicability.

Data reduction and analysis first involved the transformation of the derived method weights into comparable scales and then reduction
of the subjects' weights, for each method, into a composite score. The analysis, in response to each of the study questions, was conducted separately for each of the four study settings. This analysis is summarized below:

1. **Inter-method reliability.** A parametric and nonparametric measure of association was generated for each method by measuring the correlation between the composite criterion weights derived by a method with the sum of the composite weights derived by the balance of the methods. The resultant measures of association can best be described as convergent reliability and are based upon Kendall's rank and Pearson's product-moment correlation coefficients.

2. **Elapsed-time for application.** The mean and standard deviation of the five subjects' elapsed-time required to apply each of the five methods provides the descriptive base for this dimension. The Friedman rank-sum test was applied to these data to test the hypothesis that the ranking method requires the shortest elapsed-time and the successive comparison method requires the largest elapsed-time.

3. **Method difficulty.** A composite difficulty weight was generated for each method based upon the subjects' responses to the questions regarding ease and difficulty of method use.

4. **Perceived method applicability.** A composite applicability weight was generated for each method based upon the
subjects' responses to a question regarding the method most applicable to the decision-making needs of the subjects.

In summary, the methodological procedures undertaken in the study consist of the following steps:

1. Selection of four study settings.
2. Instrument and field procedure development and pilot testing.
3. Subject selection in each study setting.
4. Method application by the five subjects in accordance with the field procedures in each of the four study settings.
5. Data transformation and reduction.
6. Data analysis and hypothesis testing.
II. REVIEW OF RELATED LITERATURE

The rationale for the present study is supported by at least six strands of related evidence. These strands include empirical and theoretical studies within the general fields of decision theory, evaluation, education, management theory, psychological scaling and operations research. Each of these strands of evidence will be presented in turn, culminating in a delineation of the problem.

A. Decision Theory

Decision making is an underlying process in all problem oriented fields. Thus, a generic field of decision theory has evolved across the boundaries of these disciplines. Some of the key issues of the present study have been addressed by decision theorists.

Braybrooke and Lindblom¹ identified eight major barriers to the utilization of a rational, comprehensive method of decision making: (1) man's limited problem solving capacities, (2) the inadequacy of information available, (3) the costliness of analysis, (4) the failures in constructing a satisfactory evaluative method, (5) the closeness of observed relationships between fact and value, (6) the openness of the systems of variables with which the decision maker must

contend, (7) the analyst's need for strategic sequences of analytical moves, and (8) the diverse forms in which problems actually arise.

Human judgement is not a totally rational process, but it attempts to be a very economical process. Beyond the contentions of Braybrooke and Lindblom, others have pointed out similar difficulties. For example, Simon notes that information costs money, and most managements in organizational decision-making stop searching for alternative courses of action when they locate a "satisficing" option. Likewise, even if it is assumed that all possible alternative courses of action could be found, the capacity of human judgement becomes a barrier. Miller states that, "On the basis of the present evidence, it seems safe to say that we possess a finite and rather small capacity for making such unidimensional judgements and that this capacity does not vary a great deal from one simple sensory attribute to another."  

This points out another aspect of human judgement which confounds the rational decision-making process. Redding noted that "the more complex or ambiguous the stimulus, the more the perception is


3 George A. Miller, "The Magical Number Seven, Plus or Minus Two: Some Limits On Our Capacity for Processing Information," The Psychological Review, Vol. 63, no. 2 (March 1956) p. 86.
determined by what is already 'in' the subject and the less by what is in the stimulus.'"4

It would appear that decision-making processes are largely subjective in nature. The process is limited by human abilities to collect and utilize information. This contention is further substantiated by Simon,5 who has coined the phrase "bounded rationality" in acknowledging the fact that individuals have both perceptual and information processing limits. Whether an individual intends to make a rational decision or not, he can only do so to a limited extent. He must act upon the basis of sufficient rather than complete knowledge, using simple rules and short cuts in searching for a solution to a problem.

Thus, actual decision-making processes are much more complex, but less formalized than one might expect. Decision-making is not necessarily a rational process, by nature, if not by design. Decision theorists are faced with the issue of developing satisfactory models for decision-making which are within the confines of "bounded rationality," while providing satisfactory outcomes.

An alternative to rational decision making is one which Braybrooke and Lindblom label the "naive criteria method." This method holds that by:

---


merely announcing a few general values (the method) supplies enough evaluative machinery to propel descriptive knowledge toward definitive recommendations...

Even if the values are expressed as postulates... the naive criteria method is sure to break down in any serious test. Other criteria press for attention. Should they be treated as if they were irrelevant? Unexpected conflicts develop among the criteria postulated. Are these conflicts to be disregarded? The postulates make no provision for conflicts or for other values. They do not even rank the values they express.6

Accordingly, the naive criteria method does not appear to be an acceptable model for decision-making. Defects in this method may serve as a caution in assuming that all criteria in a decision have equal value. Decision models must provide a systematic process for prioritizing or weighting these decision-making criteria.

Multiple-criterion weighting methods enable the criteria for a decision to be arrayed and prioritized. The weighting process, although systematic, is based upon subjective judgements on the part of decision-makers. Since decision-making is a subjective process in and of itself, the use of multiple-criterion weighting methods is a "nesting" process. That is, the components or steps in decision-making are each addressed by individual decisions subsumed within the "total decision." This process has a number of advantages.

As Miller points out, "We have a variety of techniques for increasing the accuracy of our judgements: (a) to make relative rather than absolute judgements, (b) to increase the number of dimensions along which the stimuli can differ, and (c) arrange the

task in such a way that we make a sequence of several absolute judgements in a row." The methodologies under consideration are, by definition, multiple criterion techniques. This provides a means for increasing the accuracy of judgements in each of the three ways which Miller suggests. First, criterion are weighted in relation to each other, rather than in relation to an absolute standard. Secondly, more than one criterion is used, increasing the number of "dimensions" upon which a decision is to be evaluated. Finally, the multiple weighting techniques are used in a sequential decision-making process. The first step involves the selection of criteria, the second step results in the weighting of these criteria, while a subsequent step provides the final outcome of the decision in question. In this respect, multiple-criterion weighting methods are designed as one component of a decision-making model. They do not provide all of the processes necessary for decision-making.

This caution is forwarded in response to concerns cited by Braybrooke and Lindblom of the "naive priorities method of decision making." They state that "In conception, (the naive priorities method) ignores questions of source or relevance, and it ignores the differences of opinion and the need for justifying the simplifications that it claims to effect." Both differences of opinion in multiple-judge settings and the relevance of priorities can be dealt with through the use of these methodologies. However, it is cautioned that these

7 Miller, "The Magical Number Seven, Plus or Minus Two," p. 90.
methods, as well as any method of decision-making, are based upon
the subjective, intuitive judgements of individuals and, as a result, the outcomes are only as appropriate as the decision-makers involved.

In summary, decision theorists do not appear particularly optimistic for the resolution of human decision-making needs. However, the decision-making process can at least be upgraded by the continued development of alternative methodological paradigms.

B. Evaluation

Evaluation may be defined as "the process of delineating, obtaining, and providing useful information for judging decision alternatives." Evaluation has been faced with the task of providing information sufficient for making decisions. However, what is "sufficient information"?

The evaluation profession in education has primarily focused upon providing descriptive information about the potential decision alternatives and their consequences, based upon the judgmental criteria by which the decision is to be made. Such descriptive information is necessary, but is it sufficient for making an effective decision? Considering the complexity of most decision-making settings, the answer is invariably, "no."

---

It appears that descriptive information may be insufficient in serving the decision-making needs of the educator. Beyond descriptive information, there may exist a need for efficient and reliable decision making methodologies. In complex decision settings, the decision maker can not completely depend upon his intuitive capabilities. He requires a means of arraying, prioritizing, and selecting potential decision alternatives.

If evaluation is to serve educational decision-making in the most effective manner, it is not only the responsibility of the evaluator to provide appropriate and timely information for serving these needs, but it is also necessary to provide useful and appropriate decision-making tools which may be included in the evaluator's repertoire. Thus, it is considered a "sin of omission" to deny educational decision-making tools which may be useful for processing information into decision outcomes.

It is likewise the responsibility of the evaluator to ensure that the tools advocated are useful and efficient methodologies which are appropriate for the decision setting. Thus, it is a "sin of commission" to suggest esoteric or inappropriate methodologies which may be inadequate or less efficient than other existing tools suited to the decision-making process. The role of the evaluator must not only deal with the provision of descriptive information, but must also encompass the provision of judgemental methodologies in reducing uncertainty in the decision-making process.
This is clearly articulated by Stake in his description of the role of human judgement in evaluation:

In education, as elsewhere, judgements will continue to rest on incomplete knowledge, imprecise measurements, and inadequate experience. No error-free system is possible, but improvements are within easy reach. The evaluator may lessen the arbitrariness of judging and decision-making by introducing data-gathering methods already developed by other social scientists, economists, political scientists, and historians who routinely study opinions, preferences and values. Many of their methods can be used to measure judgements that shape an educational program.10

Stake goes on to assert that, "we researchers need to find out whether or not specific priorities aid the direction of school programs. It is my belief that excessive attention has been given to precise goal-specification and insufficient attention to statements of priorities."11

Scriven further outlines this judgemental task in defining evaluation as an "activity (which) consists simply in the gathering and combining of performance data with a weighted set of goal scales to yield either comparative or numerical ratings, and in the justification of (a) the data-gathering instruments, (b) the weightings, and (c) the selection of goals."12


11 Ibid., 185.

However, the judgemental process is not a simple task. Values are ingrained into the decision-making process and must be addressed by the evaluator. Suchman points out that, "There can be little question that values play a large role in determining the objectives of public service programs and that any evaluation study of the desirable and undesirable consequences of such programs must take social values, especially conflicting values, into account."¹³

Stake outlines this consideration in yet another way. "Rational judgement in educational evaluation is a decision as to how much to pay attention to the standards of each reference group (point of view) in deciding whether or not to take some administrative action."¹⁴

Thus, human judgement is ingrained both in the process of evaluation and the process of educational decision-making. Human judgement is not invalid in and of itself, but a potential exists for its misuse on the part of decision-makers. The use of subjective judgements for evaluation provides a case in point. Reicken points out that especially when objectives have not been operationally defined, it happens that reliable and valid objective measurement techniques are omitted...(and) reputable and skillful practitioners or subject-matter specialists may be asked to render a judgement about the effectiveness of the program. Such judgements usually fail to make explicit the assumptions, procedures for collecting information and criteria for appraising data that the experts have used.


This remark is not intended to condemn expert judgement but merely to point out that it is difficult to know exactly what was done, to repeat it and to get consistent agreement among judges.\textsuperscript{15}

Stake has identified six major areas in which human judgement enter the evaluative process. These include, (1) objectives as valued goals of education, (2) personal value commitments, (3) priorities assigned to objectives, (4) external standards assigned by authority-figures, (5) outcome data measured in the affective domain, and (6) summative judgements regarding the overall program.\textsuperscript{16}

The literature suggests that the opportunity, if not the need, exists for the application of multiple-criterion weighting methods. These methods have been designed for processing preferential judgements. Such judgements are key elements in a variety of educational decision-making processes. For example, Kaufman considers one of the steps in conducting a needs assessment is to "Place priorities among the discrepancies and select those on which action will be taken" and, goes on to suggest that "representative panels or even a large group ballot might be used to set priorities among the identified and reconciled discrepancies."\textsuperscript{17}


\textsuperscript{16}Stake, "Objectives, Priorities, and Other Judgement Data," p. 182-186.

More generally, Kaufman states that there is a need for a systems approach to educational planning as

a process by which needs are identified, problems selected, requirements for problem solutions are identified, solutions are chosen from alternatives, methods and means are obtained and implemented, results evaluated, and required revisions to all or part of the system are made so that the needs are eliminated.\(^\text{18}\)

However, he does not cite the specific procedures to be undertaken for the prioritization and selection of alternatives.

Metfessel and Michael offer a model for the evaluation of the effectiveness of school programs. In this process a key step was the construction of a cohesive paradigm of broad goals and specific objectives arrayed in a hierarchical order from general to specific outcomes. An underlying step in this process would be a formulation of "judgemental criteria for defining relevant and significant objectives and for setting realistic priorities in terms of societal needs and expectations..."\(^\text{19}\) They did not, however, state exactly how these priorities and judgemental criteria would be determined.

Steele and Moss provide further evidence of the need for such methodologies in their paper on "The Criteria Problem in Program Evaluation."\(^\text{20}\) They propose that the effectiveness of evaluation


rests mainly with the quality of the criteria used. They assert that criteria are not adequately used and understood because (1) there is too narrow a concept of evaluation, (2) there is too great an emphasis on information and program description, (3) there is a wish to avoid decision-making, (4) there is a wish to escape challenge by avoiding subjectivity, and (5) there are poorly defined guides to data interpretation.

As Alkin and Bruno\textsuperscript{21} point out, any general systems approach has five basic elements for systems analysis: (1) the objectives, (2) the alternatives, (3) the costs, (4) the model (or models), and (5) the decision rule. With these five steps, the researcher should be able to specify the objectives, select alternatives, determine the cost of each alternative, develop a representative model to predict the extent to which each alternative will achieve the specified objective, and arrange the alternatives in order of preference. In an equal vein, Coster and Morgan state that the decision-maker is responsible for administrative functions, such as allocating resources and ordering objectives into a hierarchy.\textsuperscript{22}

The point is well taken that these authors assert the need for priority setting and judging between alternatives and/or criteria,


but none suggest methodologies for doing this task. One of the major problems faced by evaluators of social action programs, as indicated by Fortune, is the presence of value conflicts and the absence of procedures for their negotiation.  

Popham states as one of the major evaluation guidelines that the "educational evaluator should clarify value preferences of various groups regarding desired educational goals by having sets of precisely stated objectives rated by the individuals involved, then translating these ratings into composite indicators of each objective's worth." However, in a personal discussion, Popham indicated that one of the greatest problems in educational evaluation is the need for methodologies appropriate for such individual rating procedures.

In an evaluation workshop package developed by the Center for the Study of Evaluation, more detailed procedures are proposed in assigning weights to alternative need statements for conducting needs


assessment. The package addresses both the issue of weighting the rater as well as the needs. The methods, however, are rating techniques not including methods addressed by the present study. The workshop package further suggests using the rating techniques in providing the project director with tools to assist in making planning decisions.

The Sweigert model for needs assessment, in a manner similar to the UCLA model, addresses the use of rating and ranking techniques for prioritizing needs statements. Some alternative methods for dealing with inter-rater differences, such as weighting, pooling or consensus techniques, are also discussed.

Glass has offered a method for establishing the priorities in collecting evaluative data. These priorities will depend upon "(1) the costs of gathering different data, (2) estimates of the prior probabilities that each alternative embodied in the decision will be supported by the data—if they were to be gathered, and (3) the costs of implementing each of the alternatives of the decision."28

In conclusion, the evaluation literature appears to suggest at least four major points: (1) A need exists in the educational


setting for methods of processing judgemental information, 
(2) educational evaluators are aware of this need, (3) educational evaluators are willing to accept a portion of the responsibility for resolving this need, and (4) current solutions offered by evaluators have been few and incomplete.

C. Education

Education, as a social institution, is faced with the responsibility of selecting, from a multiplicity of alternatives, a satisfactory system for the enculturation of individuals. The process of identifying needs, determining solution alternatives, selecting and implementing strategies, and evaluating their performance, has been further complicated by the educational milieu. Educational decision making is currently constrained by five major elements: (1) it is composed of a variety of decision-making bodies; (2) it must confront a vast array of problems of differential severity, (3) it has available an incomprehensible number of potential solution alternatives of varying or unknown quality, (4) in selecting and combining these solutions with needs, it must address a number of both qualitative and quantitative criteria by which the effectiveness of the potential solutions will be assessed, and (5) as a "social science," it has utilized a limited number of methodological tools which may prove ineffective or inappropriate.

The methods suggested by the present study are included under the broad category of operations research algorithms. However a search of the operations research (OR) literature for educational applications of
the methodologies resulted in the following conclusions cited by Hertz and Eddison. "The reader may have observed, and perhaps deplored, the smallness of the number of OR people working in health services, compared to the hordes of analysts inquiring into military and industrial problems. If so, he will deplore more the even smaller group of names associated with research in education."

However, quantitative methods have been applied in the educational realm to some extent. For example, Dyer describes a procedure for identifying educational goal areas to be emphasized in elementary schools. The method produces an index number for each goal area that approximates the incremental increase in the decision-maker's utility function arising from emphasizing it. This utility function is defined on performance measures for a particular group of students. Feichtinger developed a model for optimal educational planning utilizing dynamic programming. Abbott outlined a systems approach which allows quantitative assessment of the causes and consequences of organizational change within the school setting. Sisson considered

the obstacles and potentialities of contributions of OR to education. In his study Sisson asserts that OR has been limited by funds, a lack of theoretical models of the educational process and difficulty in defining the learning process. Armor and Dyer suggest a decision model for evaluating potential change in instructional programs using operations research paradigms, particularly mathematical utility functions. Likewise, Geoffrion, Dyer and Feinberg have developed an interactive mathematical programming approach to multi-criterion optimization, which they have applied in the operation of an academic department.

These findings are descriptive of the relationship between education and operations research in general. Specifically, education has not been totally disregarded by operations researchers and operations research techniques have been installed in educational settings. However, the techniques which have been developed for educational implementation appear to have two characteristics: (1) the paradigms are usually quite sophisticated, based upon linear programming and other highly technical processes, and (2) the techniques have been largely developed in the area of educational management and planning.


Such a finding has a number of implications. First, a barrier does not necessarily exist between education and operations research since models are being developed in educational settings. Secondly, the need expressed in the present study has been largely ignored, regardless of the relative simplicity of the paradigms and the articulated need for a solution to the problem of processing judgements. Finally, it would appear that the specific problem under consideration has not received the attention of operations researchers possibly because a solution to the problem would not serve as a panacea to all of the planning and decision-making needs of educators. More bluntly stated, multiple-criterion weighting methods will not automatize a decision-maker's work for him.

The role of values and human judgement in the educational process has been addressed by a number of researchers in the field of education. Maguire36 studied the value components of teacher judgements of educational objectives. He based his study on the assertion that the value orientation of individuals can be determined by their preference for ranked educational objectives. It was found that the output of a judgemental decision concerning an educational objective could be accurately described using a linear model in which value aspect scores for objectives were combined to predict the decision.

Coughlin\textsuperscript{37} has applied paired-comparison techniques in studying teachers' work values. Although prioritization and value judgements have been investigated by Maguire and others, the studies were conducted to determine the value frameworks rather than as a method of weighting the objectives.

Taylor and Maguire\textsuperscript{38} obtained value ratings of objectives from subject matter experts, teachers and curriculum development specialists. They found considerable agreement among these groups in prioritizing these objectives.

These findings suggest that, regarding the focus of the present study, the researchers had the correct answers but were asking the wrong questions. The literature has shown that, at least to a limited degree, the field of education is at least partially aware of the multiple-criterion weighting methods. However, the methods have not been used to resolve the particular problem under consideration.

Stake\textsuperscript{39} provides an excellent review of methods by which objectives, priorities and other judgement data are utilized in


\textsuperscript{39}Stake, "Objectives, Priorities, and Other Judgement Data," pp. 181-212.
education. It is concluded from this review, however, that the methods suggested by the present study have had little recognition in the educational field. In a later review of the literature, Johnstone summarized the mathematical models developed for use in educational planning. In his conclusion Johnstone states,

Hence, the information (the models) provide must be supplemented by information from other planning techniques and sources. These others could include the Delphi Technique, the Cross-Impact Matrix, PPBS, the use of expert committees, and certainly not to be forgotten, the use of human intuition and judgement by the planners themselves. ⁴⁰

In conclusion, a review of the educational literature has revealed no attempts to determine the applicability of the multiple-criterion weighting methods to education. However, the literature appears to suggest that such information would be of value for selecting such methods for resolving decision-making needs of educators.

D. Management Theory

Individuals in the field of management and administration have developed a wide range of quantitative planning and management techniques. A preliminary review of the literature suggests that theorists in the field have outlined a number of systems approaches

to planning and management functions. For example, Quade identified five basic elements of process analysis. Briefly, these include (1) the objective or objectives, in which the analyst discovers the decision-makers objectives and a method for measuring the extent to which these objectives are attained by various choices, (2) the alternatives, in which the analyst identifies the means by which the objectives will be approached, (3) the costs of the alternatives, (4) a model is generated with which the cost incurred and the probability of success can be estimated for each objective, and (5) a criterion or standard used to rank the alternatives in their order of desirability is established.

Quantitative management techniques and the study of decision processes are abundant within the management field. The utility and prioritization of multiple objectives have been addressed, among others, by Bartee, Brisken, and Souder.

---


Research in the management sciences was found to be quite sophisticated. As indicated, the problems of weighting multiple-criteria have been addressed by the field. The discipline, in fact, has moved beyond the methods suggested by the present study. However, this movement may be reversed, as suggested by Bartee:

The restrictive influence of requiring that an objective function be numerical seems to limit the areas of application so that the relevance of such methodology to the important problems of today is diminishing. Such problems as inventory management seem to get the attention they do because they fit within the restrictions of the problem definition, since items of inventory are more adoptable to the counting and numerical process.46

Thus, management science has not resolved all of the problems relevant to the topic under study. As Cleland and King point out, "the seemingly straightforward establishment of objectives is intrinsically confounded both by the multiple value systems within each individual and by the difficulties in making consensus value judgements on the part of the management group who must set the objectives."47 Though management science is not without its problems, it does not appear hesitant to apply quantitative approaches to decision-making.


E. Psychological Scaling

Four major references provide a theoretical background for the development and application of the multiple-criterion weighting methods. These are Coombs, Torgerson, Guilford, and Shelly and Bryan.

Coombs, in his expose of the theoretical basis of quantitative data, provides a logical framework for the derivation of multiple criterion weighting techniques. He discusses four levels of preferential choice data, of which parallelogram analysis and unfolding analysis are of particular relevance.

Torgerson, in his classification of scaling methods, distinguishes between judgements and responses. The multiple criterion weighting techniques utilize the judgement approach in which the task set for the subject is to evaluate the stimuli with respect to some designated attitude along a defined continuum. Thus the methods are considered to result in at least ordinal measurement. Torgerson further provides a number of dimensions by which the multiple-criterion weighting methods can be described. First, the responses made by the subjects in applying the methods are comparative in nature, e.g., a relation between two or more stimuli regarding the subject's preference.


Secondly, the methods themselves differ procedurally, e.g., ranking, rating, rank ordering, etc. Third, the methods are based upon a stimulis-centered or judgement approach, which Torgerson describes as "systematic variation in the reactions of the subjects to the stimulis... attributed to differences in the stimulis with respect to a designated attribute." Furthermore, relationships between criteria can be interpreted directly in terms of the proportion of times one criterion is judged as having more of the attribute in question than another criterion.

Torgerson also supports the contention's of the present study regarding inter-method reliability. He states that:

There is considerable evidence that the rank order obtained is substantially invariant with respect to the different experimental methods that might be used. For example, we would expect substantially the same rank order to obtain for a given set of stimulis with respect to a designated attribute, whether we used judgements involving paired comparisons, ranking, single-stimulis rating, or sorting into successive intervals to obtain the data.3

The multiple-criterion weighting methods are further differentiated by the judgement methods. Ranking and paired comparison methods are considered to be variability of judgements approaches, "wherein the variability of judgements with respect to each stimulis or stimulis combination is used to derive a unit of measurement,"

---

52 Torgerson, Theory and Methods of Scaling, p. 46.
53 Ibid., p. 53.
54 Ibid., p. 54.
while rating and successive comparison methods are considered to be quantitative-judgement approaches, where "the unit (of measurement) is obtained directly from quantitative judgements of the stimuli with respect to the attribute."\textsuperscript{54}

The quantitative-judgement approaches are based upon a number of theoretical assumptions, which Torgerson has summarized. These include,

1. We take as given a series of stimuli that vary along some discriminable attribute. We assume that the subject is capable of making direct quantitative judgements of the amount of the attribute associated with each stimulus.

2. The subject's judgement is considered to be a direct report of the value of the stimulus on a linear subjective continuum of the attribute of interest.

3. Some variability of judgement with respect to any given stimulus may occur as is true of any measurement procedure. If so, the variability is treated as error, and some averaging procedure is used to give an estimate of the scale value of each stimulus.\textsuperscript{55}

The distinctions between quantitative-judgement and variability of judgement approaches are based upon justifications of the methods as producing interval scale data.

Guilford has described the properties and procedures for application of each of the multiple-criterion weighting methods, which have a historical background in psychological scaling. These citations are summarized by their respective methods below.

\textsuperscript{55}Ibid, p. 62-63.
In the method of pair comparisons, all stimuli to be evaluated on a psychological scale are typically presented to the observer 0 in all possible pairs.

0 judges whether one of the pair is of greater quantity than the other in some defined respect... the stimuli are of a similar nature...the response of 0 is ostensibly a comparative judgement.56

The rank order method was noted as having a greater degree of popularity in psychometric methods, and

bears some superficial resemblance to that of successive categories on the one hand and to that of paired comparisons on the other. It resembles the former in that the stimuli are arranged by 0 along some psychological continuum... The resemblance to pair comparisons is more fundamental in that each stimulus is in essence compared with every other stimulus...One difference here is that in the rank-order method all the stimuli are present for simultaneous observation...Another difference is that pair comparisons give opportunity for deviation from a strictly consistent rank order.57

The successive comparisons method is parallel to the constant sum method developed for scaling interval judgements, as Guilford notes it must name the ratio he thinks exists between two or more stimuli...Assuming that the observations are correct within sampling errors, the stimuli are thus placed at appropriate distances from a zero point. Letting any one stimulus have the value of unity we can readily scale the others.58

56Guilford, Psychometric Methods, p. 154.
57Ibid., p. 178
58Ibid., p. 214
Numerical rating scales, finally, are typically defined as including

a sequence of defined numbers...supplied to the observer. 0 assigns to each stimulus an appropriate number in line with these definitions or descriptions.59

The reader is referred to these aforementioned references for a detailed discussion of the scaling methods' use, errors and statistical issues involved. The reader should note that, for all practical purposes, psychological scaling and multiple-criterion weighting methods are synonymous. Each are concerned with the assignment of criteria to value positions along a continuum.

Shelly and Bryan60 provide an alternative approach to psychological scaling as it applies to human judgements in decision-making. A major point made in their work is that human judgement is fallible—the difference between a quantitative judgement and its error is referred to as the true value.61 The statistical implications of this fact are discussed within the work cited. A second consideration which they point out is the problem of weights and the conflicts between goals. "In the last analysis...the uncertain and shifting character of over-all evaluative judgements is probably an unavoidable consequence of the inherent instability or indeterminancy of the system of weights associated with the underlying subgoals."62

59 Ibid., p. 263.
60 Shelly and Bryan, Human Judgements and Optimality, 1964.
61 Ibid., p. 57.
62 Ibid., p. 276.
In concluding this section, a number of potential contributions to human decision-making cited by Shelly and Bryan are listed:

1. The development of axiomatic systems of measurement that increase the range of behavior which can be described partly by interval and ratio scales. (and)

2. The development of better techniques for testing the adequacy of formalized models of human and institutional behavior. 63

Thus, although psychological scaling methods have been advanced over the years, they have by no means been perfected for resolving all of the problems encountered in human judgement.

F. Operations Research

Hatry best summarizes the problems to be expected in evaluating public service programs when he states:

multiple criteria are a complicating fact of life. Programs should be evaluated against all relevant criteria. Qualitative measurement is appropriate wherever purely quantitative measurement is insufficient. Each program selection problem is likely to involve its own special considerations, which means that for each problem, the question of appropriate evaluation criteria needs to be raised and criteria selected only after careful deliberation. 64

The combined efforts of psychologists and operations researchers have contributed to the initial development and modification of multiple criterion weighting techniques. The psychological discipline in the early 20th Century began to use multiple scaling techniques extensively.

63 Ibid., p. 276.

Barrett\textsuperscript{65} compared ranking and paired comparison techniques as early as 1914. Buel\textsuperscript{66} suggested a simplified version of the paired comparisons method developed by Hays near the turn of the century. Churchman, Ackoff, and Arnoff\textsuperscript{67} suggest a method of successive comparisons. Even newer and more sophisticated methods have been suggested as recently as 1972 by Goodwin.\textsuperscript{68}

Eckenrode\textsuperscript{69} compared the reliability and efficiency of five methods for weighting multiple criteria. He compared the methods of ranking, rating, three versions of paired comparisons, and a method of successive comparisons. The results of the experimental situation, which was conducted in an air defense setting, indicated that no significant differences existed in the derived weights of sets of criteria as a result of the use of the five methodologies. He further concluded that

\begin{itemize}
\end{itemize}
ranking is the most efficient method and, when compared to the paired comparisons methods, becomes increasingly more efficient as the number of criteria to be judged increases from six to thirty. The present study has been designed, in part, as a replication of this study by Eckenrode.

Eckenrode went on to review the literature on research comparing such studies, including his own research. The judgements upon which the research was based ranged from preferences for cereals to the belief in propositions. However, in none of the studies were judgements made of decisions exemplary of an educational setting.

In a study conducted by Pound a ranking technique, termed the "expected value model", was compared to the intuitive evaluations of four decision-makers. The decision-makers' alternatives for judgement in this case were a number of potential research projects available for potential funding. Pound found that the resulting ranking of the projects agreed with an intuitive evaluation made by the decision-makers. Pound's findings suggested that such models may be useful in

---


the area of research project selection. Baker and Pound identified ten distinct algorithms for the selection of R & D projects.

The qualitative aspects of decision-making have been studied by a number of individuals in the field of operations research. Such studies include those of Sheridan and Carlson, Deutsch, Becker, DeGroot and Marschak, and Fishburn. DiRoccaferrara has conducted a survey of multiple-criteria decision-making methods.

In a second study by Eckenrode, a study of the time efficiency for utilization of six methods of criterion weighting was conducted. The methods compared were ranking, rating, three versions of paired comparisons and successive comparisons. The results indicated that


ranking was consistently the most efficient method for recording preferences regardless of the number of comparisons. This work has been used for founding hypotheses undertaken by the present study.

The work conducted within the general field of operations research has been abundant. However, this body of knowledge has failed to gain impact upon the decision-making community. Baker and Pound point out that before these methodologies are accepted for formal use, R & D people must become familiar with the details of these methods and build up a trust in their value. At present there is a lack of detailed exposure to the mathematical and statistical techniques being proposed, and there has been no clear demonstration that using a formal method is highly advantageous.80

The conclusion cited by Baker and Pound is considered to be a major justification for the present study. The need for such techniques has appeared in a number of areas of educational decision-making. Such methods are particularly relevant to evaluative processes such as needs assessment, the prioritization of objectives and the selection of evaluative criteria. The absence of the methods in the field of education is considered to be significant. In summarizing the findings of Baker and Pound, this absence is an artifact of a lack of familiarity with the procedures and mistrust of a clear demonstration of their relative advantages. Thus, as an initial source of evidence for the validation of these methods, the present study affords an opportunity to improve decision-making processes in education.

80 Baker and Pound, "R & D Project Selection: Where We Stand" p. 130.
III. PROCEDURES

A. Methodological Approach

A field study approach is used as the method of inquiry. A small number of cases are examined as an initial source of evidence for describing specific properties of the criterion weighting methods. Thus, both the method and purpose of the study are exploratory in nature.

The field study focuses upon the application of the five weighting techniques by a sample of educational personnel in four simulated decision-making settings. The evidence derived from these applications is descriptive in nature.

The methodological approach manifests a number of characteristics. As a descriptive study, attention is concentrated upon the systematic collection of detailed information. Causal relationships are not under investigation. Although the study is conducted in the field, the procedures are applied under controlled, simulated conditions. Furthermore, the study utilizes a small number of subjects in exploring the properties of the criterion weighting methods. In this way, exploration of the field of inquiry is initiated in a relatively efficient manner.

Specifically, the study is based upon the application of the five criterion weighting methods in four settings. Each of these settings are described below:
Study One. Five educational evaluators were randomly selected from the evaluation staff of a regional educational laboratory in the State of Oregon. The five weighting methods were applied by each individual to a simulated problem in which seven criteria for assessing the quality of evaluative information were to be weighted.

Study Two. Five educational evaluators were randomly selected from the evaluation staff of a metropolitan school district in the State of Oregon. The five weighting methods were applied by each individual to a simulated problem in which seven criteria for assessing the quality of evaluative information were to be weighted.

Study Three. The five staff members of a state department of education who are responsible for the management and coordination of exemplary projects (Elementary and Secondary Education Act, Title III, as amended). The five weighting methods were applied by each individual to a simulated problem in which ten criteria for assessing the quality of exemplary projects were to be weighted.

Study Four. Five directors of Title III projects were randomly selected from the Title III projects currently operating in the State of Oregon. The five weighting methods were applied by each individual to a simulated problem in which ten criteria for assessing the quality of exemplary projects were to be weighted.
B. Research Design

Five multiple criterion weighting techniques are under consideration in the present study. The study focuses upon the application of these methods to seven and ten criteria, respectively. The five methods are operationally defined below as they would be applied to seven and ten criteria:

1. **Ranking.** Each subject independently weights the seven (or ten) criteria by assigning a number 1 to the criterion which the subject perceives as most important; the number 2 to the criterion considered to be of the next most importance; continuing until the subject has assigned the number 7 (10) to the least important criterion,

2. **Rating.** Each subject independently weights the seven (or ten) criteria by assigning a discrete value between 0 and 10 to each of the criteria, where 10 represents the maximum value of a criterion, while 0 represents no value. Criteria may be assigned similar values in this method.

The application of the method results in the numerical weighting of a value between 10 and 0 for each criterion, where the greater the numerical value, the greater the weight of the criterion.

*The operational procedures for the methods when using ten criteria are defined as a special case by parenthesis.*
3. **Partial-Paired Comparisons.** Each subject independently weights the seven (or ten) criteria by making a preferential judgement of the more important criteria in each of 21 (45) pairs of criteria. Each of the seven (ten) criteria are matched with each of the remaining six (nine) criteria.

The output of this method is a weight represented by the number of times each criteria is designated as preferred over other criteria. Thus, weights between 6 (9) and 0 are possible, where the greater the numerical value, the greater the weight of that criterion.

4. **Complete-Paired Comparisons.** Each subject independently weights the seven (or ten) criteria by making a preferential judgement of the more important criteria in each of 42 (90) pairs of criteria. Each of the seven (ten) criteria are matched twice with each of the remaining six (nine) criteria. The ordering of each pair is different.

The output of this method is a weight based upon the number of times each criteria is designated as preferred over other criteria. Thus, weights between 12 (18) and 0 are possible, where the greater the numerical value, the greater the criterion weight.

5. **Successive Comparisons.** Each subject independently weights the seven (or ten) criteria by applying the following procedure:

   a. The relative value of the criterion are ranked by assigning a number 7 (10) to the most important criterion;
a number 6 (9) to the next most important criterion; continuing until the least important criterion is assigned the value of 1.

b. The criteria are randomly assigned to three groups of two (three) criteria each. The seventh (tenth) criterion is assigned a numerical value of 1.0 and included as the third (fourth) criterion in each of the three groups. The subject tentatively assigns a numerical value between 10 and 0 to each of the remaining two (three) criteria in each of the three groups, where the larger the numerical value, the greater the relative weight.

c. For each group and based upon the numerical value assigned each criterion, the subject compares the most valued criterion to the combined values of the two (three) remaining criteria. The subject then decides whether the most important criterion is of more value, equal value or less value than the combination of the two (three) less important criteria. The subject then adjusts the numerical values of the criteria based upon the results of this decision. (For each group and based upon the numerical value assigned each criterion, the subject compares the next most valued criterion to the combined values of the two remaining criteria. The subject then decides whether the
The next most valued criterion is of more value, equal value or less value than the combination of the two less important criteria. The subject then adjusts the numerical values of the criteria based upon the results of this decision.

d. The subject compares the criterion ranks obtained in Step a with the rank of the computed values derived from Step c. If the ranks differ, the subject decides whether the original ranks are correct or the computed ranks are correct. If the original ranks are considered correct, the subject reapplys Steps b, c and d until consistent results are obtained. If the computed ranks are considered correct or the orders do not differ, the subject proceeds to Step e.

e. The subject normalizes the computed values by deriving the sum of the seven (ten) computed values and then dividing each computed value by this sum.

Application of this method results in a numerical weighting of a value between 1.0 and 0 for each criterion, where the greater the numerical value, the greater the weight of the criterion.

Although the study is primarily descriptive in nature and cause and effect relationships are not under investigation, the study has been designed to control for multiple-treatment interference which often threatens the validity of experimental research.
Without this research design, the external generalizability of the study could be potentially threatened by interference between the sequence of the application of the weighting methods and the results derived from these methods. The design incorporates the random sequencing of method application per subject as displayed in Table 1. Thus, the effects of sequencing are controlled, although the effects of the proximity in applying these methods is not totally erasable. Factors of fatigue are likewise controlled by this design.

**TABLE 1**

**RESEARCH DESIGN**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Sequence of Method Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>B</td>
<td>3 2 1 5 4</td>
</tr>
<tr>
<td>C</td>
<td>5 3 4 1 2</td>
</tr>
<tr>
<td>D</td>
<td>2 5 3 4 1</td>
</tr>
<tr>
<td>E</td>
<td>4 1 5 2 3</td>
</tr>
</tbody>
</table>

Two sets of criteria were selected for application of the five weighting methods. The criteria were selected for the study on the basis of a number of considerations. First, criterion could not be conditional. That is, one criterion could not vary as a function of another criterion. If this were the case, the two criteria would not be mutually exclusive and the assignment of weights to the criteria would be confounded.
However, orthogonality was not applied as a restriction. Although two criteria could be inversely proportional, they could both be weighted independently. Finally, the criterion must be consistent with the decision under study. That is, the criteria should generally be unidimensional.

The first set of seven criteria was utilized for the first and second case studies. The second set, containing ten criteria, was utilized for the third and fourth case studies. These sets of criteria are described below:

**Set One:** The criteria upon which the method weights were based in the first two case studies were derived from a list of eleven criteria for assessing the quality of evaluation which was developed by the Phi Delta Kappa National Study Committee on Evaluation. These studies focus upon the weighting of these criteria by their relative values for assessing the quality of evaluative information. The specific criteria utilized are operationally defined as follows:

- **Internal Validity:** the information is a reliable and accurate representation of reality
- **External Validity:** the information is generalizable to the population it purports to represent
- **Objectivity:** the information is unbiased; replicable among independent observers

---

Relevance: the information is applicable to the decision to be served

Credibility: the source of information is believable

Timeliness: the information is provided at the time which it is needed for decision-making

Efficiency: the information is generated through a reasonable expenditure of time, money and resources

Set Two: The criteria upon which the method weights were based in the final two case studies were derived from a discussion with various individuals involved with the management and development of exemplary projects. The two case studies focused upon the weighting of ten criteria for assessing the quality of exemplary educational projects. The specific criteria utilized are operationally defined as follows:

Transportability: the extent to which the project's products and/or processes can be physically transported to adopter sites

Innovativeness: the degree to which the project's processes resolve a problem through a unique approach and/or improve an existing practice within the target area

Generalizability: the magnitude of an existing or predicted target population for which the project has been designed.
Cost: the amount of resources required for developing, installing and maintaining the project

Significance: the degree to which the project focuses upon an important problem

Evidence: the quality of the information used to support the outcomes of the project

Clarity: the extent to which the project processes have been explicitly defined

Maintenance: the probability that the project processes can be continued through local funding sources

Effectiveness: the extent to which the project satisfactorily impacts upon the problem area

Compatibility: the extent to which the project's goals, processes and outcomes are congruent with the existing values and attitudes of the potential adopter population

As a result of the nature of the criterion, the study may be considered as applicable to generalized decision situations, rather than specific decisions. That is, the criteria utilized are not applied to a particular problem-solving decision regarding a precise situation. The subject's responses to such a general problem may be highly transitory.
Four major variables are addressed in the study. These variables are inter-method reliability, elapsed-time for method application, method difficulty and perceived method applicability.

1. **Inter-method reliability.** Goldberg and Werts, in a study of clinician inferential reliability, stated that the specific components of such reliability include, "generalizability over (a) time, for the same judges using the same data (stability), (b) judges, for the same data from the same occasion (conensus) (c) data sources, administered on the same occasion and interpreted by the same judge (convergence)."\(^2\) Furthermore Campbell and Fisk\(^3\) suggest that it is possible to generate reliability coefficients between methods as well as judges.

The reliability under consideration in the present study, using Goldberg's terms, is convergence. Specifically, the degree of association between the derived method weights and a composite of the weights derived from the balance of the methods. Although of importance, stability and consensus (inter-judge reliability) will not be addressed by the present study.

---


Stability, in the present instance, would be difficult to establish because of the transitory nature of the criterion used. Similarly, inter-judge reliability is a moot question in a situation where no external standard exists for determining a "correct" response.

Two measures of association will be used. These are Kendall's rank correlation coefficient and Pearson's product-moment correlation coefficient, which are further described in the analysis section.

2. Elapsed-time for method application. This variable focuses upon the average time required to make comparative judgments among seven and ten criteria, for each method used. This variable is measured by the elapsed time, in minutes, required to successfully complete comparative judgments among the criteria using a specific weighting method. As Eckenrode points out, "...it is relevant in choosing a method to consider how efficient it is from the standpoint of the time which the judges require to use it."^4

As indicated in the analysis section, the mean and standard deviation of the elapsed time required for the five subjects to complete each weighting task will be used as descriptive indices of elapsed-time.

3. **Method Difficulty.** This variable is based upon the preferential judgment of the five subjects regarding the method perceived as easiest and most difficult to use in weighting the criteria under study. It is felt that perceived ease of use is a contributing factor in the selection of a method for application.

This variable is measured by the responses to two items posed of the subjects upon completion of the five method applications. The questions are, "Which method was easiest to use?" and "Which method was most difficult to use?" Analysis will result in a composite difficulty weight based upon a weighted average of the responses to the ease and difficulty of use questions. Derivation of this weight is further described in the analysis section.

4. **Perceived Method Applicability.** The preferential judgment of the five subjects regarding the method most applicable to the decision-making needs of each subject provides the basis for this variable. It is felt that the method preferred by the judges provides an indication of the extent to which a method will actually be utilized when a need arises and when an awareness of the method exists.

This variable is measured by the subjects' responses to an item which asks, "Which method do you feel is most applicable to your decision-making needs?" The question is posed directly after application of the five methods in the study.
Analysis will result in a composite preference weight through dividing the number of times a method is indicated as being preferred, by five, the number of subjects.

Each of the four variables in the study focuses upon distinct characteristics of the weighting methods under consideration. Thus, the description of each of the methods regarding each variable is in response to a specific question:

1. What is the inter-method reliability of each of the five methods used for weighting the study criteria?
2. What is the elapsed-time required for the application of each of the five methods when weighting seven and ten criteria, respectively?
3. Which methods are perceived by the subjects as being most difficult and easiest to use, when weighting seven and ten criteria?
4. Which methods are considered to be most applicable to the decision-making needs of a limited sample of users?

Furthermore, hypotheses have been established in predicting the outcomes of two of these questions. These hypotheses are based upon the findings of previous research cited in the present study. The hypotheses are:

1. Inter-method reliability. The inter-method reliabilities for each method will be significantly different from chance at the .05 level of probability. This hypothesis is based
upon the findings of Eckenrode⁵ in which the consistency between each of these methods were all found to be high. The hypothesis is further substantiated by Torgerson,⁶ who has indicated that inter-method differences do not tend to be large.

2. **Elapsed-time for method application.** Based upon the findings of Eckenrode,⁷ it is hypothesized that ranking will require the least elapsed-time for application and successive comparisons will require the greatest elapsed-time for application. The difference in elapsed-time for these methods will be significantly different at the .05 level of probability, regardless of whether seven or ten criteria are being considered. A nonparametric test will be utilized.

Suppositions have not been advanced for the two remaining variables regarding method difficulty and perceived method applicability. Previous research has not provided consistent evidence of the superiority of a method along these lines. It is felt that the outcomes of these variables are highly dependent upon the decision setting in question.

---

⁵ Eckenrode, "Weighting Multiple Criteria," p. 186.
⁶ Torgerson, Theory and Methods of Scaling, p. 53.
C. Pilot Studies

The development of the data collection instruments and procedures was undertaken through a three-phase process. First, initial prototypes of the data collection instruments were developed. The prototypes were modified from an earlier instrument devised by Eckenrode. The instruments were then administered to three subjects for the purpose of determining the clarity of directions, the appropriateness of the criteria and the difficulty of the task. Preliminary findings indicated that minor revisions were necessary in further clarifying the directions. It was also found that the subjects could not accomplish the application of the successive comparisons method because of the complexity of directions.

The second phase of pilot testing encompassed a second revision of the instrumentation and an initial trial of the procedures. As a result of this pilot test, both general instructions and specific directions for application of the methods were further modified. Field procedures were developed for conducting the study. The procedures and revised instruments were administered to two subjects in order to determine the clarity of the procedures and directions, the difficulty of the task and the approximate time required for the administration of the instruments to subjects.

The final phase of pilot testing focused upon the completion of minor revisions to the study instruments and procedures. The subjects were able to accomplish all tasks with only minor difficulty. The

---

8Eckenrode, "Weighting Multiple Criteria."
The approximate time required for completion of the two instruments was 45 minutes and one hour, respectively. This phase of the pilot testing also provided an opportunity for an initial trial of the data analysis procedures.

D. Selection of Subjects

As a means of providing an initial sample of educational decision-makers, each of the four case studies involved a sample of a different group of individuals from the education profession. It is again cautioned that the limited sample is neither directly generalizable to all decision settings in education, nor all individuals in the education profession. However, it is felt that a design which incorporates multiple replications of inquiry will enhance the reliability with which the study questions can be answered.

The four groups of individuals participating in the study included:

1. Five subjects randomly selected from a list of the evaluation personnel of the Northwest Regional Educational Laboratory, Portland, Oregon. The list consists of all personnel who are identified as evaluators by title or function.

2. Five subjects randomly selected from a list of the evaluation personnel of Portland Public School District Number 1J, Portland, Oregon. The list consists of all evaluation specialists employed in the district or area offices.

3. The five specialists at the Oregon State Department of Education responsible for coordinating projects funded through Title III of the Elementary and Secondary Education Act.
4. Five subjects randomly selected from a list of the directors of Title III projects currently operating in the State of Oregon.

Where random selection procedures were utilized, randomness was limited to the extent that voluntary participation could be solicited. However, this did not prove to be a constraint in the selection process.

Five subjects were considered to be an adequate number for each case study as an initial source of evidence in describing the selected characteristics of the weighting methods. Guilford substantiates this contention when he points out that

...most gains in reliability come from multiplying raters when initial (inter-rater) reliability is low, and that in adding raters the law of diminishing returns sets in very rapidly. There is usually much to be gained by adding the first two or three raters, but not much after reaching five.  

It should be noted that although inter-rater reliability is not of direct consideration in the present study, the stability of the data derived for the analysis is enhanced by an increase in the number of subjects.

E. Instrumentation

As previously indicated, the instruments utilized for data collection were modified from an earlier version developed by Eckenrode.  

A final version of each instrument was developed after

---

9 Guilford, Psychometric Methods, p. 397.

10 Eckenrode, "Weighting Multiple Criteria."
pilot testing. Each instrument consists of three major sections. The first section contains an introduction and general instructions, the second section contains the directions and response sheets for application of each of the five weighting methods, while the final section consists of a brief questionnaire regarding the subjects' perceptions of the appropriateness of the alternative methods.

The two instruments differed in only three respects. First, the criteria under consideration differed, the first instrument addressing criteria for assessing the quality of evaluative information, while the second instrument focused upon criteria for determining the value of innovative educational projects. Secondly, the first instrument contained seven criteria, while the second instrument contained ten criteria. Finally, an additional methodological step was required for applying the successive comparisons technique to ten criteria.

The order of the methods was randomly assigned a different sequence for each subject. Thus, 20 different instruments were used which contained the same elements, but in differing orders. As indicated in the discussion of the research design, this random ordering reduces the likelihood of sequencing effects appearing in the resultant data. Copies of exemplary instruments have been attached in Appendix A.

F. Field Procedures

The field procedure utilized the distribution of a questionnaire. The mailed packet contained the study instrument, a letter of transmittal and a self-addressed, stamped return envelope. The letter of transmittal addressed the following key points:
1. **Purpose of the instrument.** "Enclosed is a criterion weighting study instrument. You have been selected to complete this instrument as part of a doctoral study. Your anonymity is assured."

2. **Procedural Considerations.** "In completing the instrument, the following important points should be noted:

   - Read through **all** of the directions before beginning.
   - When completing the five sections of the instrument, **record the time required to accomplish each section, in minutes and seconds.** Spaces have been provided at the end of each section for noting this elapsed time.
   - Complete the evaluation on the final page.
   - Return the instrument at your earliest convenience in the attached self-addressed envelope."

The name of the subject was noted on each instrument. The specific contents of the study instruments have been previously discussed. Refer to Appendix A for samples of these instruments.

Two weeks after the questionnaire distribution, a follow-up letter was mailed to each non-respondent. The follow-up letter contained the message below:

Would you please complete the criterion-weighting study instrument which I have mailed to you? Your prompt reply would be sincerely appreciated.

If you have already forwarded the instrument, please disregard this message. Do not hesitate to call me should you have any questions.

Thank you for your cooperation.
Four weeks after the initial questionnaire distribution, a follow-up by phone was made to each non-respondent. An appeal was made for the return of the instrument. Additional copies of the instrument were provided upon request.

Upon the return of each instrument, the responses were reviewed for completeness. Where omissions appeared to be accidental, the subject was contacted by telephone and asked to complete the particular item in question. Where omissions were noted by the subject as intentional, the responses were maintained as received.

Although a laboratory procedure may well be better suited for instructing and observing each subject, and for controlling the conditions under which the methodologies are applied, such a controlled laboratory procedure would be essentially contrived. Decisions are not usually performed under optimal conditions. To this end, the use of laboratory procedures would detract from the representativeness of the method applications.

G. Data Reduction and Analysis

The resultant data are collapsed across sets of subjects for analysis. Initial analysis was conducted separately for each of the four sets of subjects. This initial analysis was designed to address the study questions and hypothesis for each study group. The initial analysis procedures are described below.

Composite weights are generated for each of the criteria from the raw data computed by each subject. The reduction process, which
has been suggested by Eckenrode, results in comparable scales for each of the five weighting methods. The reduction procedures employed for each method are described below.

1. **Ranking.** Raw ranks are reversed in order. The ranks for each criterion are then summed across judges in the following manner:

\[
(a) \quad R_c = \sum_{j=1}^{5} R_{aj} \quad \text{where,}
\]

\[
R_c = \text{sum of converted ranks across judges for criterion } c
\]

\[
R_{aj} = \text{converted ranks assigned by judge } j \text{ to criterion } c.
\]

These sums of converted ranks are then modified to composite weights:

\[
(b) \quad W_c = \frac{R_c}{\sum_{\alpha=1}^{m} R_{\alpha}} \quad \text{where,}
\]

\[
W_c = \text{composite weight of criterion } c \text{ across all judges}
\]

\[
m = \text{number of criteria; in the present instance, seven or ten.}
\]

2. **Rating.** Raw ratings are interpreted to the nearest tenth along the scale from 0 to 10. An initial weight for each criterion for each of the five judges is derived as indicated on the following page.

---

11Eckenrode, "Weighting Multiple Criteria."
(c) \( W_{cj} = \frac{pc_j}{\sum_{c=1}^{m} pc_j} \) where,

\[ \]

\( W_{cj} = \) weight computed for criterion \( c \) from the rating given by judge \( j \)

\( pc_j = \) rating given by judge \( j \) to criterion \( c \).

Composite weights are then derived across judges:

(d) \[ W_c = \frac{\sum_{j=1}^{5} W_{cj}}{\sum_{j=1}^{5} \sum_{j=1}^{m} W_{cj}} \]

3. Partial-Paired Comparisons. The number of times each criterion is chosen over each other criterion is tabulated for each judge, and the number of times each criterion is selected over all other criteria is derived as follows:

(e) \[ f_{cj} = \frac{\sum_{c=1}^{m-1} f(c/c')}j \]

where,

\( f_{cj} = \) frequency of choice of judge \( j \) of criterion \( c \) over all other criteria

\( f(c/c') = \) frequency of choice of criterion \( c \) over a given criterion \( c' \).

These frequencies are then normalized:

(f) \[ W_{cj} = \frac{f_{cj}}{J} \]

where,

\( J = \) total number of judgments being made; \( m(m-1)/2 \).

Composite weights are then derived using equation (d) above.

4. Complete-Paired Comparisons. The same procedures are used as in the partial-paired comparisons method. However, equation (f) is adjusted to account for the larger number of pairwise
judgements, $J$, being made:

$$W_{cj} = f_{cj}/J$$

where,

$J =$ total number of judgements being made; $m(m-1)$.

5. Successive Comparisons. The raw weights produced by this method are comparable to the raw scale of the rating method. Composite weights are generated by applying the same procedures listed under rating, i.e., equations (c) and (d).

The data reduction procedures result in the generation of a composite weight for each criterion by each method. Table 2 portrays an example of the data organization when seven criteria are being used.

**TABLE 2**

**ORGANIZATION OF COMPOSITE CRITERION WEIGHTS**

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Method</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ranking</td>
<td>$W_{11}$</td>
<td>$W_{12}$</td>
<td>$W_{13}$</td>
<td>$W_{14}$</td>
<td>$W_{15}$</td>
<td>$W_{16}$</td>
<td>$W_{17}$</td>
</tr>
<tr>
<td></td>
<td>Rating</td>
<td>$W_{21}$</td>
<td>$W_{22}$</td>
<td>$W_{23}$</td>
<td>$W_{24}$</td>
<td>$W_{25}$</td>
<td>$W_{26}$</td>
<td>$W_{27}$</td>
</tr>
<tr>
<td></td>
<td>Partial-Paired Comparisons</td>
<td>$W_{31}$</td>
<td>$W_{32}$</td>
<td>$W_{33}$</td>
<td>$W_{34}$</td>
<td>$W_{35}$</td>
<td>$W_{36}$</td>
<td>$W_{37}$</td>
</tr>
<tr>
<td></td>
<td>Complete-Paired Comparisons</td>
<td>$W_{41}$</td>
<td>$W_{42}$</td>
<td>$W_{43}$</td>
<td>$W_{44}$</td>
<td>$W_{45}$</td>
<td>$W_{46}$</td>
<td>$W_{47}$</td>
</tr>
<tr>
<td></td>
<td>Successive Comparisons</td>
<td>$W_{51}$</td>
<td>$W_{52}$</td>
<td>$W_{53}$</td>
<td>$W_{54}$</td>
<td>$W_{55}$</td>
<td>$W_{56}$</td>
<td>$W_{57}$</td>
</tr>
</tbody>
</table>

Table 2 indicates that a single method-criterion weight has been generated as a composite of the five judges application of the methods.
A data analysis procedure has been designed to address each of the study questions. These analyses utilize the composite criterion weights and/or other information derived from the completed instruments. The analysis procedures are specifically described in the sections which follow.

1. **Method Reliability.** The inter-method reliability for each weighting method is established by both nonparametric and parametric procedures. The nonparametric procedures are described as follows:

   (a) Kendall's coefficient of concordance (\(\omega\)) is applied to the composite weights across all criterion and all methods in order to determine the overall relationship of the five methods. The general equation is described by Siegal\(^{12}\) as

   \[
   \omega = \frac{S}{1/12 \ k^2(N^3-N)} \quad \text{where,}
   \]

   \[S = \text{sum of squares of the observed deviations from the mean of } R^j\]

   \[k = \text{number of sets of rankings}\]

   \[N = \text{number of entities ranked}\]

   \[1/12 \ k^2(N^3-N) = \text{maximum possible sum of the squared deviations.}\]

   (b) The hypothesis is tested that the five sets of rankings are independent by comparing the derived \(S\) to the critical value of the \(S\) distribution, where \(p<.05\).

---

(c) Siegal notes that the "pooled ordering (of criteria) may serve as a 'standard,' especially when there is no relevant external criterion for ordering the objects."\(^{13}\) He further points out that "Kendall suggests that the best estimate of the 'true' ranking of the N objects is provided, when \( \omega \) is significant, by the order of the various sums of ranks, \( R_j. \)^{14}\n
Since the criteria under consideration lack an external standard for comparison, the sum of \( R_j \), i.e., the sum of the ranks of the composite weights for each criterion across all methods, might provide a best estimate of the "true" weight of each criterion. However, a correlation between the \( \Sigma R_j \) and the ranks derived by a method would be spuriously high, since the \( \Sigma R_j \) is in part a function of that method rank. As an alternative, for each method, \( j \), the \( \Sigma R_j \) will be computed as follows:

\[
\Sigma R_j = \Sigma R_j^{4} \quad \text{where the } R_j \text{ of the method to be ranked is deleted from the } \Sigma R_j.
\]

This procedure is analogous to that of removing the influence of an item from the total score of a test when estimating internal consistency through item-remainder (item-test) correlations.\(^{15}\) The procedure, however, is not entirely

\(^{13}\) Siegal, *Nonparametric Statistics*, p. 238.

\(^{14}\) Ibid.

satisfactory since the corrected $\Sigma R_j$'s will vary slightly from method to method.

Kendall's rank correlation coefficient ($\tau$) is applied to each of the method ranks and the corrected $\Sigma R_j$. The equation, as described by Siegal,\(^{16}\) for this coefficient is

$$\tau = \frac{S}{1/2N(N-1)}$$

where,

$N$ = the number of objects ranked

$S$ = the sum of the +1 and -1 scores for all pairs, when a +1 is assigned to natural orderings of the pairs.

Application of this method results in a single inter-method reliability coefficient for each of the weighting methods.

(d) The null hypothesis is tested that each method is independent of the $\Sigma R_j$, by a comparison of the derived $\tau$ to the critical value of the $\tau$ distribution, where $p<.05$.

The second procedure to be employed for further substantiating the inter-method reliabilities of the weighting methods is parametric in nature. The Pearson product-moment correlation coefficient ($r$) is used to determine the relationship between the composite weights derived by each of the methods. The general equation for the Pearson $r$ formula,

\(^{16}\)Siegal, *Nonparametric Statistics*, p. 216.
as described by Games and Klare,\textsuperscript{17} is portrayed below:

\[ r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2} \sum n\sum Y^2 - (\sum Y)^2}} \quad \text{where}, \]

\( X \) and \( Y = \) the scores of the variables under consideration
\( n = \) the number of pairs of measures.

In a manner parallel to the nonparametric approach, the composite weights of the individual methods are compared to the sum of the method weights, less the weights of the particular method in question. Thus, a single inter-method reliability coefficient is generated for each of the methods through the parametric procedure as well.

A final step in this procedure is a test of the null hypothesis that the reliability coefficient is greater than 0. The reliability coefficient must exceed the 95 percent confidence limits, based on the given \( n \), in order to be considered tenable.

The product-moment correlations observed between each pair of methods will also be reported. Although there is no hypothesis directly connected to this analysis, the inter-method correlations will be reported for descriptive purposes.


\textsuperscript{18} Games and Klare, \textit{Elementary Statistics}, p. 142.
2. **Elapsed-Time for Method Application.** The arithmetic mean and standard deviation of the elapsed-time required, in minutes and seconds, for the five subjects to complete each weighting task will be determined for descriptive purposes. The computational formula suggested by Games and Klare\textsuperscript{18} will be used for deriving the standard deviation. The formula is:

\[
\text{sd} = \frac{\sqrt{\sum (X-X)^2}}{n-1} \quad \text{where,}
\]

\(n=\) the number of subjects

\(X=\) the individual elapsed-time per subject per method.

In response to the hypothesis that ranking will require the least elapsed-time, successive comparisons will require the greatest elapsed-time, and the difference between these elapsed-times will be statistically significant, the differences between these methods will be tested. In addition, the differences between all methods will be tested in order to determine whether or not other unpredicted differences exist. The null hypothesis is established that the elapsed-time for applying each method is equal, i.e., all observations of elapsed-time are derived from the same population.

The Friedman rank sums analysis will be applied to these data. This test has been designed to detect differences in

\textsuperscript{18}Games and Klare, *Elementary Statistics*, p. 142.
location among multiple samples when the treatments are blocked by a second variable. In the present instance, the second (blocking) variable will be the subject, since each individual will apply each of the five methods. The analysis could best be described as a nonparametric counterpart to the two-way (treatments-by-block) analysis of variance model. The equation for this test, as described by Wolfe and Hollander,\(^1\) is:

\[
S = \frac{12n}{k(k+1)} \sum_{j=1}^{k} (R_j - R_{..})^2
\]

where,

- \(k\) = the number of treatments, in this case five methods
- \(n\) = the number of subjects, in this case five
- \(R_j = \sum_{i=1}^{n} r_{ij}\), the sum of the ranks of observations across treatment \(j\)
- \(R_{..} = \frac{R_j}{n}\), the average rank for treatment \(j\)
- \(R_{..} = \frac{k+1}{2}\).

Assuming that the null hypothesis is rejected, i.e., the derived \(S\) exceeds the critical value of the upper-tail probabilities for the distribution of \(S\) when \(p < .05\), a distribution-free multiple comparison will be applied. The multiple comparisons procedure is parallel to the experiment-wise multiple comparison methods used in parametric analyses. The procedure is based upon a comparison of the absolute differences of pairs of methods against a critical value.

established for that particular sample size, number of treatments, and experimentwise error rate. The experimentwise error rate established for the present study is a .05 level of confidence.

3. **Method Difficulty.** The analysis of the method difficulties will result in a composite difficulty weight for each method, based upon the following equation:

\[
D_m = \frac{\Sigma_{em} - \Sigma_{dm}}{n + \Sigma_{dm}}
\]

where,

- \(\Sigma_{em}\) = the number of times the method, \(m\), is nominated as being "the easiest to use"
- \(\Sigma_{dm}\) = the number of times the method, \(m\), is nominated as being "the most difficult to use"
- \(n\) = the number of subjects

\(D_m\) can potentially range from +1, in which the method is considered the easiest to use, to -.5, in which the method is unanimously considered to be the most difficult to utilize.

4. **Perceived Method Applicability.** The analysis of the perceived method applicabilities will result in a composite applicability weight for each method, based upon the following equation:

\[
A_m = \frac{\Sigma_{am}}{n}
\]

where,

- \(\Sigma_{am}\) = the number of times the method, \(m\), is nominated as being "the most applicable."

---

Am may range from +1, in which the method was rated as most applicable by all subjects, to 0, in which the method was not nominated as most applicable by any of the subjects.

These analysis procedures will be individually applied to each of the four study groups. The results of these analyses will be reported separately. Finally, the results will be aggregated across the study groups for comparative purposes. While no particular statistical procedures are suggested for such an overall comparison, an analysis of the results across study groups is of value in determining the extent to which the results are replicable.

The analysis procedures which have been described focus upon the provision of descriptive information. Accordingly, the results should provide an empirical basis for selecting between methods. However, it is reiterated that the four study variables under consideration in the present research do not represent all dimensions which could be used for determining method applicability. It is hoped that as an initial exploratory study, the present research will provide a groundwork for future analysis of these methods, both on a broader scale and in greater depth. Ultimately, substantial information can be provided for making a rational decision regarding the applicability of these methods to particular decision-making situations within education.
H. Methodological Assumptions and Limitations

Assumptions regarding the processes and outcomes of the study are along three dimensions, i.e., procedural, statistical and methodological. Procedural assumptions focus upon the application of the research design, statistical assumptions are directed toward the analysis, while methodological assumptions refer to the operational parameters of the weighting methodologies. The assumptions and limitations are individually addressed for each of these dimensions in the paragraphs which follow.

1. Procedural Assumptions and Limitations. It should be stressed that the present study is not experimental to the extent that variables are not manipulated. However, the control of extraneous variables remains a concern in the current design.

Stanley and Campbell\textsuperscript{21} have identified 12 potential threats to the validity of an experimental design. These 12 extraneous factors are history, maturation, testing, instrumentation, statistical regression, selection, experimental mortality, selection-maturation interaction, interaction effects of selection, interaction effects of pretesting, reactive experimental effects and multiple treatment interference. A number of these factors are considered to be of consequence to the study design. Each of these relevant factors are discussed

in terms of their relationship to procedural assumptions and limitations.

(a) The effects of history are remote in the present study design. All weighting methods under consideration are applied within a relatively short interval. The effects of differential intervening history would, however, pose a serious threat to a study of the stability dimension of reliability. Such a "test-retest" situation, particularly with judgements as transitory as in the present instance, would be particularly vulnerable to this factor.

(b) The effects of fatigue (maturation) are controlled through the random sequencing of the weighting methods for each subject. It is assumed that by averaging the five subjects' responses into a composite value, the effects of fatigue on responses will be controlled, i.e., evenly distributed across methods.

(c) The effects of testing pose a threat to the present study. Each weighting method is essentially based upon the more general methods of either ranking or rating. As a result, application of one method may alter a subject's responses when applying subsequent methods. The design has utilized two strategies in order to limit this effect. First, the sequence of methods is different for each subject. Secondly, the subjects are asked not to refer back to the responses given through a particular methodology once it has been completed. It is assumed that by averaging the five subjects' responses
into a composite weight, the effects of method sequence will be controlled. However, a limitation which has not been totally controlled for is the lack of independence between method applications. The effects of the "proximity" in applying these methods is not totally erasable.

(d) The effects of instrument decay have been accounted for through the use of a single instrument in each study setting. The measurement of the study variables does not call for a situation in which assessments are repeated over time.

(e) Statistical regression is not considered to be a relevant factor in the present study since repeated measurements are not called for. Furthermore, subject selection is not based upon any of the measured study variables.

(f) The effect of selection upon the internal validity of the present study is absent, since the design does not incorporate the use of an experimental-control group paradigm.

(g) Experimental mortality, as in selection, does not effect the present design since comparison groups are not called for by the analyses.

(h) Interaction between selection and maturation does not pose a threat since these effects occur in (experimental) treatment-control group comparisons.

(i) The interaction effects of selection poses a threat to the external generalizability of the study. Both random samples and complete census are used in the study. It is assumed that
the study groups are representative of their respective populations and that nonrespondent error can be minimized by the utilization of extensive follow-up procedures.

A limitation of the study, however, is that the results can not be generalized to the universe of all educational personnel. Furthermore, findings are not necessarily generalizable across time, since approximately a three-week interval is "sampled" by the study. These limitations are not considered serious in light of the exploratory purpose of the effort. Furthermore, the approach does utilize multiple replications for each study question, which should lend credence to the results ultimately obtained. As an initial source of evidence, the procedures utilized in the study emphasize feasibility at the expense of universal generalizability.

It is assumed that the effects of prior method knowledge or experiences are equally controlled for by the selection procedures and analysis methods. Prior knowledge is considered to be a random factor which subjects may or may not possess. (j) It is again reiterated that since a "pre-post" design is not utilized by the present study, interaction effects of pretesting is not considered to be a relevant factor. (k) Reactive effects of the study may be considered a limitation. Since the subjects are aware that they are participating in a study, the responses to the study variables
may be biased by socially acceptable responses or other reactive behaviors. However, the subjects are not aware that the study is concerned with inter-method reliability. The probability of a conscious bias of this variable is remote.

It is assumed that the results of the elapsed-time for application of the methods will probably represent the lower limit of the time spent in applying such methods under "normal" conditions. The study situation is to some extent speeded, since the subjects are asked to work as quickly as possible. However, it is likewise assumed that the order of the elapsed-time required across methods will not be affected by such a speeded situation, e. g., ranking will take less time than successive comparisons, regardless of the respective times required for their application.

(1) Multiple-treatment interference is not totally controlled for since five methods (treatments) are simultaneously under consideration. This effect, in the present study, is parallel to that previously described as a testing effect. Similarly, the design used for partially controlling for this effect has previously been discussed.

The limitations of method interaction are considered to be less severe than the effects which would be encountered when applying the alternative procedure of distributing the method applications across time to reduce the effects of prior responses. It would be difficult to determine how long the
intervening interval should be in order to erase the effects of memory, while assuring that the derived reliabilities were not deflated by intervening experiences. Again it is noted that such a procedure would be highly susceptible to variations in responses due to the transitory nature of the criterion under study.

2. **Statistical Assumptions and Limitations.** The analysis procedures utilize a number of statistical tools, many of which require assumptions about the underlying characteristics of the variables in question. The nonparametric statistical methods of Kendall's rank correlation coefficient, Kendall's coefficient of concordance and Friedman's rank sum statistic assume that the observations are ordinal in nature, i.e., that the data can be ranked in their order of magnitude. Considering the nature of the weighting methods, this assumption appears quite tenable.

Application of the coefficient of concordance and rank correlation also requires the assumptions that the bivariate observations are mutually independent and come from the same continuous bivariate population. The Friedman rank sums procedure has three underlying assumptions. First, the analysis of variance model $X_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$ is applied, where $\mu$ is the unknown overall mean, $\beta_i$ is the block effect of subject $i$, $\tau_j$ is the unknown effect of treatment $j$, and $\epsilon_{ij}$ is the error term associated with each observation. The model
assumes that the sum of $\beta_i$ and the sum of $\tau_j$ are both zero, the error associated with each observation is independent and comes from the same continuous population. It is assumed that these qualifications are satisfied by the data under analysis.

It is also assumed, based upon the findings of Siegal,\textsuperscript{22} that Kendall's coefficient of concordance ($\omega$) bears a linear relationship to the average of the rank correlations taken over all methods. Furthermore, it is assumed that the sum of the ranks computed in deriving $\omega$ is the best estimate of the "true" or consensual orderings of the criterion and by removing the direct effects of a particular method, i.e., the sum of ranks, from this grand sum, an unbiased measure of reliability can be obtained.

More stringent constraints are placed upon the data when utilizing the parametric analyses. The Pearson product-moment correlation coefficient requires interval data of two variables which have underlying normal population distributions. It is assumed, therefore, that the data under study are of an interval nature. As Eckenrode points out, "if the assumption made here is tenable, i.e., that an interval scale may legitimately be derived from the judgments recorded by the methods used, then a correlation statistic at the interval

\textsuperscript{22} Siegal, \textit{Nonparametric Statistics}, p. 238.
level of measurement is appropriate."^23

It is further assumed that the data are normally
distributed. This contention is supported by Dixon and Massey,
who state "Since a mean is usually more nearly normally
distributed than individual measurements, it is sometimes
convenient to draw a series of samples and record the means for
use in the analysis in place of the single observation."^24

In this case, a modified arithmetic mean is computed as the
composite weights derived from each method for each criterion
across subjects.

3. Methodological Assumptions. In applying the multiple-criterion
weighting methods, Churchman, Ackoff and Arnoff^25 note that
there are a number of operational assumptions. These
methodological assumptions include:

(a) Given a range of real numbers, an individual can assign
values to each criterion along this scale which provides
information about the relative value of each criterion.

(b) It is assumed that this assigned numerical value represents
the true value of the criterion.

^23Eckenrode, "Weighting Multiple Criteria," p. 186.

^24Wilfred J. Dixon and Frank J. Massey, Introduction to Statistical

(c) In the case of the successive comparisons method, the combined numerical values of two or more criteria is equal to their combined importance.

(d) The criteria under consideration are mutually exclusive.

(e) The criteria are unidimensional.

An assumption is also made regarding the outputs of these methodologies in preparation for the analysis. It is assumed that the between-subject differences in weighted values, if existent, will be evenly distributed across the criteria and will, on the average, be similar for each method. That is, the value preferences of the judges will not effect the inter-method reliability.

In a parallel manner, it is assumed that the study variable of inter-method reliability is independent of the balance of the study variables. Under such an assumption, if the inter-method reliabilities are found to be of equal or similar value, then the variables of elapsed-time for application, method difficulty and perceived method applicability become key to the consideration of method applicability. However, the question of reliability must take the forefront of the study. If the methods do not exhibit satisfactory reliabilities, then there is no means by which the "truly" valid method can be identified. Thus, all subsequent considerations would be conducted on methods of questionable efficacy.

An additional methodological consideration which should be
made, particularly when considering the time required for applying the methods, is the immediate utility of the output of each method. For example, rating, ranking and successive comparisons methods generate values which can be directly utilized for decision-making. On the other hand, the partial- and complete-paired comparisons methods require an intermediate step for transforming the observations into usable weights.

The assumptions can be summarized as being three-fold, i.e., procedural, statistical and methodological. Procedural assumptions focus upon the vigor of the research design, statistical assumptions are concerned with the conditions under which the analysis procedures may be applied, while the methodological assumptions address the constraints under which the weighting techniques are to be employed.

A limitation of the study design is the threat of the effects of method interaction upon the computed reliability scores. The reliability coefficients may be unjustifiably large. However, this was considered to be a trade-off in lieu of the alternative design, in which test-retest reliability would be determined. Such an alternative design introduces a time factor in which the transitory nature of the judgements can not be controlled. Regarding the statistical assumptions, Siegal notes that "All decisions arrived at by the use of any statistical test must carry with them this qualification: 'If the model used was correct, and if the measurement requirement was satisfied, then..."  

\[\text{Siegal, Nonparametric Statistics, pp. 18-19.}\]
Again, it is stressed that the generalizability of the study is limited in scope. Although the sample sizes are somewhat limited, replicability has been utilized as an alternative for broadening the generalizability of the findings. As an initial source of evidence for determining the applicability of these weighting methods, generalizability is not considered to be the primary emphasis of the study.
IV. RESULTS

The results derived from each of the four study groups are reported separately in the sections which follow. At the conclusion of these individual sections, a fifth section is presented in which the results of the four study groups are aggregated. This final section provides an overall comparison of the study outcomes.

A. Study One

The data collection instrument was administered to the five subjects in accordance with the established protocol. The subjects, randomly selected from the evaluation staff of the Northwest Regional Educational Laboratory, ranged in age from 30 to 54. All held advanced degrees, two having Ph.D.'s, while three having Ed.D.'s. Positions included Division Director, Research and Evaluation Coordinator of a major program, Program Evaluation Coordinator, Project Evaluation Specialist, and Research and Development Specialist of a major program. Research and evaluation experience ranged from five to fourteen years.

All subjects successfully completed the instrument. The data derived from the subjects' application of the five methods were transformed and reduced across subjects in accordance with the analysis procedures. The resultant outcome data are displayed in Table 3.
TABLE 3

RESULTS OF THE CRITERION WEIGHTING METHOD APPLICATIONS
OF STUDY GROUP ONE.

<table>
<thead>
<tr>
<th>CRITERIA</th>
<th>Weight</th>
<th>Rank</th>
<th>Partial-Paired Comparisons</th>
<th>Complete-Paired Comparisons</th>
<th>Successive Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>weight</td>
<td></td>
<td>weight</td>
<td>rank</td>
<td>weight</td>
</tr>
<tr>
<td>Internal Validity</td>
<td>.181</td>
<td>6</td>
<td>.175</td>
<td>6</td>
<td>.181</td>
</tr>
<tr>
<td>External Validity</td>
<td>.095</td>
<td>2</td>
<td>.120</td>
<td>2</td>
<td>.095</td>
</tr>
<tr>
<td>Objectivity</td>
<td>.162</td>
<td>5</td>
<td>.139</td>
<td>4</td>
<td>.133</td>
</tr>
<tr>
<td>Relevance</td>
<td>.238</td>
<td>7</td>
<td>.177</td>
<td>7</td>
<td>.238</td>
</tr>
<tr>
<td>Credibility</td>
<td>.076</td>
<td>1</td>
<td>.138</td>
<td>3</td>
<td>.076</td>
</tr>
<tr>
<td>Timeliness</td>
<td>.143</td>
<td>4</td>
<td>.141</td>
<td>5</td>
<td>.181</td>
</tr>
<tr>
<td>Efficiency</td>
<td>.105</td>
<td>3</td>
<td>.110</td>
<td>1</td>
<td>.095</td>
</tr>
</tbody>
</table>

Mean Weight = .143
Standard Deviation = .056
Table 3 provides a summary of the composite normalized weights and the rank of these derived weights for each of the seven criteria across the five methods. The Table reveals slight differences in the weights assigned to the criteria when different methods are utilized. For example, the internal validity criterion was given a weight of .181 through the ranking method, while the weight of .175 was generated by the rating method. However, when the criteria were ranked, the internal validity criterion received a ranking of 6 from both methods.

Since the weights have been normalized, the mean weight for each method is equal, i.e., .143. However, the variation in the criterion weights differed from method to method. For example, it can be seen that the standard deviation of the weights ranged from .025 for the rating method to .059 for the partial-paired comparisons method.

On the basis of the findings summarized in Table 3, along with the balance of information derived through the application of the data collection instrument, the following sections individually address the research questions and hypotheses posed by the study.

1. Inter-Method Reliability

As an initial source of evidence, product-moment correlations were derived between methods for descriptive purposes. These inter-method correlations are reported in Table 4.


A review of Table 4 reveals that methods ranged in association from .82 to .99. Thus, relationships appear to exist between the data derived by application of the various methods.

As outlined by the methodology, both parametric and nonparametric measures of association were applied to the data. Comparisons were made between the outcomes of each method and the sum of the outcomes of the balance of methods. In addition, an overall correlation of the five methods was computed. The results of the application of these correlational approaches are summarized in Table 5.
As noted in Table 5, all methods differ from chance at or beyond the .05 level of probability. The reliability values derived through nonparametric methods (Kendall \( \tau \)) range from .81 for rating to .71 for both ranking and successive comparisons methods. On the other hand, the reliability values which were found through parametric procedures (Pearson \( r \)) range from .97 for successive comparisons to .88 for rating.

The overall inter-method reliability was found to be .91, when the nonparametric coefficient of concordance (\( \omega \)) was applied.

The results substantiate the hypothesis that relationships do exist between the various criterion weighting methods. However, the measures of association generated by the alternative statistical

<table>
<thead>
<tr>
<th>Method</th>
<th>Kendall ( \tau )</th>
<th>Pearson ( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.71*</td>
<td>.96*</td>
</tr>
<tr>
<td>Rating</td>
<td>.81*</td>
<td>.88*</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>.73*</td>
<td>.95*</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>.73*</td>
<td>.94*</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>.71*</td>
<td>.97*</td>
</tr>
<tr>
<td>Total (( \omega ))</td>
<td>.91*</td>
<td></td>
</tr>
</tbody>
</table>

*\( p < .05 \)
procedures do not consistently identify a method as having the greatest reliability, but rather are diametrically opposed.

Secondly, the nonparametric correlations for each of the five weighting methods range in value by .10, while the parametric correlations range in value by .09. Thus, substantial differences in the reliabilities of the methods are not readily apparent.

2. Elapsed-Time for Method Application

The elapsed-time for method application is based upon the time required for each subject to accomplish the tasks described for each of the weighting methods. The range of elapsed-times required for the application of each of the five methods are reported in Table 6.

<table>
<thead>
<tr>
<th>Method</th>
<th>Observed Elapsed Time in Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>shortest</td>
</tr>
<tr>
<td>Ranking</td>
<td>.75</td>
</tr>
<tr>
<td>Rating</td>
<td>1.00</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>1.00</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>3.00</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>10.00</td>
</tr>
</tbody>
</table>
As predicted by one of the study hypothesis, the ranking method generally required the least time for application, while the successive comparisons method required the greatest time for application. The Friedman rank-sum test was applied to the ranks of the elapsed-time required for each method. The results of the test were statistically significant, in which the derived $S$ of 16.8 exceeded the critical value of the upper-tail probabilities of $S (n=5, k=5, \alpha=.049)$. The critical value in this instance was shown to be 8.96.

The results of this test indicates that not all alternative methods require an equal elapsed-time. However, the test does not directly reveal which methods are significantly different. Therefore distribution-free multiple comparisons were made based upon the Friedman rank sums derived for each method. Only the comparisons made between ranking and successive comparisons, and rating and successive comparisons were found to be significant. These elapsed-time differences exceeded the critical value for the comparisons ($\alpha=.04, k=5, n=5$) which in this instance was 14.

Thus, the hypothesis is substantiated that a significant difference exists between the elapsed-time required for applying the ranking and the successive comparisons methods.

Further analysis of Table 6 reveals that the differences in elapsed-time for application of the ranking and rating methods were slight, each requiring about one to two minutes for application. The partial-paired comparisons method required slightly more time, about three minutes, while the complete-paired comparisons method required
PLEASE NOTE:

This page not included in material received from the Graduate School. Filmed as received.

UNIVERSITY MICROFILMS
TABLE 8

COMPOSITE METHOD DIFFICULTY WEIGHTS
FOR STUDY GROUP ONE

(possible range = 1.0 to -.5)

<table>
<thead>
<tr>
<th>Method</th>
<th>Composite Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>+.8</td>
</tr>
<tr>
<td>Rating</td>
<td>0</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>0</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>+.2</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>-.5</td>
</tr>
</tbody>
</table>

Thus, the major differences in perceived method difficulties lies between the ranking and successive comparisons methods.

If one can assume that the results derived from the application of two alternative weighting methods are essentially the same and that the time required for their application does not differ significantly, then a major criterion for selecting between the two methods would be the perceived difficulty of applying the methods. The results of the study indicate major differences between the ranking and successive comparisons methods along this line.

Differences between the relative difficulties of the rating, partial- and complete-paired comparisons methods are not substantially apparent. Thus, the results appear to indicate which method is the
most difficult and which method is considered to be the least difficult, but little evidence of the relative difficulties of the balance of methods.

4. Perceived Method Applicability

The five subjects responded to a question regarding the method which they perceived as being most applicable to their decision-making needs. Three of the five subjects indicated that the ranking method was most applicable to their decision-making needs, while the remaining two subjects indicated a preference for the rating method as being most applicable to their decision-making needs. The results of an application of the composite applicability equation described in the methodology section is summarized in Table 9.

<table>
<thead>
<tr>
<th>Method</th>
<th>Applicability Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.60</td>
</tr>
<tr>
<td>Rating</td>
<td>.40</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>0</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>0</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>0</td>
</tr>
</tbody>
</table>
The results shown in Table 9 reveal a preference for the applicability of ranking and rating methods. On the other hand, the paired comparisons and successive comparison methods were not preferred by any of the subjects. Thus, it appears that the ranking and rating methods experience a preferential advantage.

5. Summary of Study One Results

A comparison of the results of the Study One application of the five methods across the four study variables is summarized in Table 10. Ranking was found to have the second largest parametric correlation (+.96) with the criterion, while requiring the least time for application and was perceived to be the least difficult and most applicable. Rating was found to have the largest nonparametric correlation (+.81) with the criterion, but the smallest parametric correlation (+.88) with that criterion. However, it required only slightly more time for application than ranking, and received an applicability weight only slightly less than that received by the ranking method. The partial-paired comparisons method received reliability values through the nonparametric and parametric methods of .73 and .95, respectively. The method required about twice the time the ranking method required for application. The method was not perceived as being particularly easy or difficult. It did not receive nominations of perceived applicability.

The complete-paired comparison method received reliability values very similar to the partial-paired comparison method. It required about four times the elapsed-time required for applying the
<table>
<thead>
<tr>
<th>Method</th>
<th>Convergent Reliability</th>
<th>Mean Elapsed-Time</th>
<th>Difficulty Weight</th>
<th>Applicability Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>τ</td>
<td>r</td>
<td>In Minutes</td>
<td></td>
</tr>
<tr>
<td>Ranking</td>
<td>.71</td>
<td>.96</td>
<td>1.15</td>
<td>+.8</td>
</tr>
<tr>
<td>Rating</td>
<td>.81</td>
<td>.88</td>
<td>1.62</td>
<td>0</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>.73</td>
<td>.95</td>
<td>2.73</td>
<td>0</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>.73</td>
<td>.94</td>
<td>4.85</td>
<td>+.2</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>.71</td>
<td>.97</td>
<td>15.20</td>
<td>-.5</td>
</tr>
</tbody>
</table>
The ranking method. It was indicated as being the easiest method by one of the five judges, but did not receive nominations of applicability. The successive comparisons method received the highest parametric reliability coefficient (.97), but was tied with the ranking method as having the lowest nonparametrically derived correlation (.71) with the criterion. The successive comparisons method required by far the greatest time for application, on the average about thirteen times longer than the ranking method. Accordingly, the method was unanimously rated as being the most difficult, and was not perceived as being particularly applicable by the subjects in Study Group One.

B. Study Two

The data collection instrument was administered to the five subjects in accordance with the established protocol. The subjects, randomly selected from the evaluation staff of the Portland Public School District, ranged in age from 33 to 55. All held graduate degrees. One individual had a Master of Science, one individual held an Ed. D., while three subjects held Ph. D.'s. Positions included two district-level Evaluation Specialists, two area-level Research and Evaluation Coordinators, and an area-level Planning and Evaluation Specialist. Research and evaluation experience ranged from three to twenty years.

All subjects successfully completed the instrument without the need for follow-up activities. The data derived from the subject's application of the five methods were transformed and reduced across subjects in accordance with the established analysis procedures. The resultant outcome data are displayed in Table 11.
<table>
<thead>
<tr>
<th>CRITERIA</th>
<th>Ranking</th>
<th>Rating</th>
<th>Partial-Paired Comparisons</th>
<th>Complete-Paired Comparisons</th>
<th>Successive Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>weight</td>
<td>rank</td>
<td>weight</td>
<td>rank</td>
<td>weight</td>
</tr>
<tr>
<td>Internal Validity</td>
<td>.221</td>
<td>7</td>
<td>.200</td>
<td>7</td>
<td>.248</td>
</tr>
<tr>
<td>External Validity</td>
<td>.129</td>
<td>3</td>
<td>.143</td>
<td>3</td>
<td>.133</td>
</tr>
<tr>
<td>Objectivity</td>
<td>.157</td>
<td>4.5</td>
<td>.146</td>
<td>4</td>
<td>.162</td>
</tr>
<tr>
<td>Relevance</td>
<td>.186</td>
<td>6</td>
<td>.170</td>
<td>6</td>
<td>.190</td>
</tr>
<tr>
<td>Credibility</td>
<td>.093</td>
<td>2</td>
<td>.107</td>
<td>2</td>
<td>.086</td>
</tr>
<tr>
<td>Timeliness</td>
<td>.157</td>
<td>4.5</td>
<td>.147</td>
<td>5</td>
<td>.152</td>
</tr>
<tr>
<td>Efficiency</td>
<td>.057</td>
<td>1</td>
<td>.086</td>
<td>1</td>
<td>.029</td>
</tr>
<tr>
<td>Mean Weight=</td>
<td>.143</td>
<td>1</td>
<td>.143</td>
<td>1</td>
<td>.143</td>
</tr>
<tr>
<td>Standard Deviation=</td>
<td>.055</td>
<td>.038</td>
<td>.071</td>
<td>.072</td>
<td>.066</td>
</tr>
</tbody>
</table>
Table 11 provides a summary of the composite normalized weights and the rank of these derived weights for each of the seven criteria across the five methods. The Table reveals slight differences in the weights assigned to the criteria when different methods are utilized. For example, the internal validity criterion was given a weight of .221 through the ranking method, while a weight of .200 was produced by the rating method. However, when these criterion weights are ranked, the internal validity criteria received a ranking of seven for both methods.

Since the weights have been normalized, i.e., each has been computed as a proportion of the value of one, the mean weight for each method is equal. However, the variation in the criterion weights differed dramatically from method to method. The standard deviation of the criterion weights ranged from .072, for the complete-paired comparison method, to .038 for the rating method. These variations may be a function of the differential scale ranges associated with the methods, e.g., ranking is based upon a scale of one to seven, while rating is based upon a scale from zero to ten.

On the basis of the findings summarized in Table 11, along with the balance of information derived through the application of the weighting instrument, the following sections individually address the research questions and hypotheses posed by the study.
1. **Inter-Method Reliability**

As an initial descriptive statistic, product-moment correlation coefficients were derived between method weights. These inter-method correlations are reported in Table 12.

<table>
<thead>
<tr>
<th></th>
<th>Partial-Paired Rating</th>
<th>Complete-Paired Comparisons</th>
<th>Successive Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranking</td>
<td>.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Rating</td>
<td>.99</td>
<td>.98</td>
<td>.96</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>.99</td>
<td>.96</td>
<td>.96</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>.96</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A review of Table 12 reveals that methods ranged in association from .96 to 1.00. Thus, substantial relationships appear to exist between the weights derived through the application of the various methods. Table 12 reveals that the successive comparisons method maintained the lowest correlations with the other methods, ranging from .98 to .96.

As outlined in the analysis section, both parametric and non-parametric measures of association are to be applied to these data. Comparisons were made between the weights derived through each method and the sum of the weights of the balance of the methods. In addition,
an overall correlation of the five methods was computed. The results of the application of the aforementioned correlational analyses are summarized in Table 13.

**TABLE 13**

**METHOD CONVERGENT RELIABILITY FOR STUDY GROUP TWO**

<table>
<thead>
<tr>
<th>Method</th>
<th>τ</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.95*</td>
<td>1.00*</td>
</tr>
<tr>
<td>Rating</td>
<td>1.00*</td>
<td>.99*</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>.90*</td>
<td>.99*</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>1.00*</td>
<td>.99*</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>.86*</td>
<td>.97*</td>
</tr>
<tr>
<td>Total</td>
<td>( \omega = .98* )</td>
<td></td>
</tr>
</tbody>
</table>

*\( p < .05 \)

As noted in Table 13, all methods differ from chance at the .05 level of probability. The reliability values derived through non-parametric methods (Kendall \( \tau \)) range from .86, for successive comparisons, to 1.00 for rating and complete-paired comparisons methods. In a quite parallel manner, the reliability values which were computed through parametric procedures (Pearson \( r \)) range from .97, for successive comparisons, to 1.00 for the ranking method. The overall inter-method reliability, as derived through the application of the coefficient of concordance \( (\omega) \), was found to be .98.
The results substantiate the hypothesis that significant relationships do exist between the various criterion weighting methods. The parametric and nonparametric measures of association are fairly similar. Both analyses identified the successive comparisons method as having the lowest reliability.

The nonparametric correlations range in value from .86 to 1.00, a variation of .14 across the five methods, while the parametric correlations range in value from .97 to 1.00, a variation of only .03. Thus, the nonparametric approach revealed the greatest range of association. However, substantial differences in the reliabilities of the methods are not readily apparent. That is, the observed differences would probably not warrant a decision to exclude methods on the basis of reliability.

2. Elapsed-Time for Method Application

The elapsed-time for method application is based upon the time required for each subject to accomplish the tasks described for each of the weighting methods. The range of elapsed-time required for the application of each of the five methods is reported in Table 14,
TABLE 14
ELAPSED-TIME IN MINUTES FOR APPLICATION
OF THE CRITERION WEIGHTING METHODS
BY STUDY GROUP TWO

<table>
<thead>
<tr>
<th>Method</th>
<th>shortest</th>
<th>mean</th>
<th>longest</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.17</td>
<td>.84</td>
<td>2.25</td>
<td>.846</td>
</tr>
<tr>
<td>Rating</td>
<td>.42</td>
<td>1.23</td>
<td>2.33</td>
<td>.885</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>1.00</td>
<td>1.61</td>
<td>3.32</td>
<td>.961</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>1.33</td>
<td>3.31</td>
<td>7.33</td>
<td>2.341</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>5.33</td>
<td>15.63</td>
<td>25.55</td>
<td>7.793</td>
</tr>
</tbody>
</table>

As predicted by one of the study hypotheses, the ranking method generally required the least time for application, while the successive comparisons method required the greatest time for application. The balance of the methods required elapsed-times which ranged within these two extremes. The Friedman rank sum statistic was applied to the ranks of the elapsed-time required by each subject in applying each method. The result of the test was statistically significant, in which the derived $S$ of 17.2 exceeded the critical value of the upper-tail probabilities of $S(n=5, k=5, \alpha=.049)$. The critical value in this instance was shown to be 8.96.

The result of this test indicates that not all alternative methods require an equal elapsed-time for application. Since the test only reveals general differences in location, the test does not
directly identify which methods are significantly different. Therefore distribution-free multiple comparisons were made based upon the Friedman rank sums derived for each method. Only the comparisons made between ranking and successive comparisons, and rating and successive comparisons were found to be significant. These elapsed-time differences exceeded the critical value for the comparisons \((\alpha=.04, k=5, n=5)\) which in this instance was 14.

Thus, the hypothesis is substantiated that a significant difference exists between the elapsed-time required for applying the ranking and the successive comparisons methods. In addition, the unpredicted significant difference observed between the rating and successive comparison methods' time for application provides additional information for determining the relative amounts of time required for the alternative methods.

Further analysis of Table 14 reveals that the elapsed-time for application of the ranking method averaged about .84 minutes, when seven criteria were under consideration. Rating and partial-paired comparisons methods required slightly more time, 1.23 and 1.61 minutes, respectively. The complete-paired comparisons method required 3.31 minutes, on the average, while the successive comparisons method required an average of 15.63 minutes and in one case required more than 25 minutes to apply. Thus, the successive comparisons method required, on the average, eighteen times longer to apply than the ranking method. Such findings further substantiate the contentions of the elapsed-time study hypothesis.
3. **Method Difficulty**

The five subjects indicated the methods which they perceived as being easiest and most difficult to apply. These perceptions were measured by responses to two questions. First, which method was easiest to use, and secondly, which method was most difficult to use. The responses to these questions are summarized in Table 15.

**TABLE 15**

**METHOD DIFFICULTY PERCEIVED BY STUDY GROUP TWO**

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Which method was easiest to use?</td>
<td>ranking</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>rating</td>
<td>1</td>
</tr>
<tr>
<td>2. Which method was most difficult to use?</td>
<td>successive comparisons</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>complete-paired comparisons</td>
<td>1</td>
</tr>
</tbody>
</table>

As revealed by the Table, the majority of the subjects indicated that the ranking method was easiest to use, while the successive comparisons method was nominated by the majority of subjects as the most difficult procedure to use. An application of the composite difficulty equation described in the analysis section reveals the differential perceived difficulty weights noted in Table 16.
As can be seen in Table 16, the perceived difficulty weights range from +.8, for the ranking method, to -.44, for the successive comparisons method. Although differences in perceived difficulty are apparent for each method, based upon such a small sample of subjects the difference observed between ranking and successive comparisons would be the only difference which could be considered as substantial. Thus, if perceived method difficulty is an important consideration in selecting and/or rejecting weighting methods, the results would appear to provide a basis for such a decision regarding ranking and successive comparisons.

Differences between the relative difficulties of the rating, partial- and complete-paired comparisons methods are not substantial. Thus, the results appear to indicate which method is the most difficult and which method is considered to be the least difficult,
based upon consensus. On the basis of such a small number of observations, little evidence is provided regarding the relative difficulties of the balance of the methods.

4. Perceived Method Applicability

The five subjects responded to a question regarding the method which they perceived as being most applicable to their decision-making needs. Four of the five subjects indicated that the ranking method was most applicable to their decision-making needs, while the remaining subject indicated a preference for the rating method. The results of an application of the composite applicability equation described in the analysis section is summarized in Table 17.

TABLE 17

COMPOSITE APPLICABILITY WEIGHTS OF THE FIVE CRITERION WEIGHTING METHODS FOR STUDY GROUP TWO

(possible range = 1.0 to 0)

<table>
<thead>
<tr>
<th>Method</th>
<th>Applicability Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.80</td>
</tr>
<tr>
<td>Rating</td>
<td>.20</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>0</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>0</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>0</td>
</tr>
</tbody>
</table>
The results depicted in Table 17 reveal a preference for the ranking and rating methods. While the ranking method was preferred by the majority of the subjects, the paired comparisons and successive comparison methods were not preferred by any of the subjects as being the most applicable method for their decision-making needs. Thus, it appears that the ranking method experiences a preferential advantage, with the rating method receiving a mild preference.

5. Summary of Study Two Results

A comparison of the results of the Study Two application of the five methods across the four study variables is summarized in Table 18. Ranking was found to have the second largest nonparametric correlation (.95) and the largest parametric correlation (1.00) with the criterion. The method required the least time for application and was considered to be both the easiest to use and most applicable method for the subjects' needs.

Rating was found to have the largest nonparametric correlation (1.00) and second largest parametric correlation (.99) with the criterion. The rating method required slightly more time to apply than the ranking method, averaging about 1.23 minutes for processing seven criteria. The method received both a mildly positive difficulty weight (+.2) and a mildly positive applicability weight (.2).

The partial-paired comparisons method received a moderate nonparametric reliability of .90 and a high parametric reliability of .99. The method required an average of 1.61 minutes to apply in processing seven criteria. This was about twice the time required
<table>
<thead>
<tr>
<th>Method</th>
<th>Convergent Reliability</th>
<th>Mean Elapsed-Time</th>
<th>Difficulty Weight</th>
<th>Applicability Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\tau$</td>
<td>$r$</td>
<td>In Minutes</td>
<td></td>
</tr>
<tr>
<td>Ranking</td>
<td>.95</td>
<td>1.00</td>
<td>.84</td>
<td>+.8</td>
</tr>
<tr>
<td>Rating</td>
<td>1.00</td>
<td>.99</td>
<td>1.23</td>
<td>+.2</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>.90</td>
<td>.99</td>
<td>1.61</td>
<td>0</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>1.00</td>
<td>.99</td>
<td>3.31</td>
<td>-.16</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>.86</td>
<td>.97</td>
<td>15.63</td>
<td>-.44</td>
</tr>
</tbody>
</table>
for the ranking method. The method was not perceived as being particularly easy or difficult. It did not receive any applicability ratings by the subjects.

The complete-paired comparison method received a high nonparametric correlation of 1.00 with the criterion. The parametric correlational value was nearly as substantial at .99. The method required about four times the elapsed-time required for applying the ranking method. As would be predicted, since the complete-paired comparisons method contains twice as many criterion pairs as the partial-paired comparisons method, the elapsed-time of 3.31 was almost exactly twice as long as the elapsed-time required for the partial-paired comparisons procedure. The complete-paired comparisons method received a mildly negative difficulty weight of -.16. The method did not receive nominations of applicability.

The successive comparisons method received the lowest observed reliability values. The nonparametric correlation was .86, while the parametric measure of association was .97. The successive comparisons method required by far the longest time to apply, on the average 15.63 minutes for processing seven criteria. This average elapsed-time was about eighteen times longer than the time required for ranking. Four of the five subjects nominated the method as being the most difficult, providing a strongly negative composite difficulty weight of -.44. Finally, as indicated in Table 18, the method was not perceived as being applicable by any of the subjects in Study Group Two.
C. Study Three

The data collection instrument was administered to the five subjects in accordance with the established protocol. The subjects, representing the Title III administrative staff of the Oregon State Department of Education, ranged in age from 45 to 54. All subjects held Masters degrees. Positions included Director of Exemplary Programs, Program Grants Operation and Review Specialist, Program Dissemination Specialist, Program Evaluation Specialist and Coordinator of Planning and Evaluation. Program management experience ranged from 10 to 15 years.

All subjects successfully completed the instrument. Two subjects required follow-up briefings to clarify the successive comparisons procedure. The data derived from the subjects' application of the five methods were transformed and reduced across subjects with the established data reduction procedures described in the analysis section. The resultant outcome data are displayed in Table 19.
<table>
<thead>
<tr>
<th>CRITERIA</th>
<th>Weight</th>
<th>Rank</th>
<th>Weight</th>
<th>Rank</th>
<th>Weight</th>
<th>Rank</th>
<th>Weight</th>
<th>Rank</th>
<th>Weight</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transportability</td>
<td>.084</td>
<td>3</td>
<td>.094</td>
<td>5</td>
<td>.071</td>
<td>3</td>
<td>.067</td>
<td>3</td>
<td>.090</td>
<td>4</td>
</tr>
<tr>
<td>Innovativeness</td>
<td>.116</td>
<td>8</td>
<td>.122</td>
<td>8</td>
<td>.107</td>
<td>7</td>
<td>.098</td>
<td>5.5</td>
<td>.135</td>
<td>8</td>
</tr>
<tr>
<td>Cost</td>
<td>.091</td>
<td>5</td>
<td>.087</td>
<td>3.5</td>
<td>.102</td>
<td>6</td>
<td>.102</td>
<td>8</td>
<td>.106</td>
<td>6</td>
</tr>
<tr>
<td>Evidence</td>
<td>.113</td>
<td>7</td>
<td>.118</td>
<td>7</td>
<td>.133</td>
<td>8</td>
<td>.133</td>
<td>7</td>
<td>.054</td>
<td>3</td>
</tr>
<tr>
<td>Clarity</td>
<td>.087</td>
<td>4</td>
<td>.087</td>
<td>3.5</td>
<td>.093</td>
<td>5</td>
<td>.098</td>
<td>5.5</td>
<td>.107</td>
<td>7</td>
</tr>
<tr>
<td>Maintenance</td>
<td>.040</td>
<td>1</td>
<td>.046</td>
<td>1</td>
<td>.031</td>
<td>1</td>
<td>.038</td>
<td>1</td>
<td>.029</td>
<td>1</td>
</tr>
<tr>
<td>Effectiveness</td>
<td>.153</td>
<td>9.5</td>
<td>.151</td>
<td>10</td>
<td>.178</td>
<td>10</td>
<td>.176</td>
<td>10</td>
<td>.196</td>
<td>10</td>
</tr>
<tr>
<td>Compatibility</td>
<td>.055</td>
<td>2</td>
<td>.052</td>
<td>2</td>
<td>.049</td>
<td>2</td>
<td>.044</td>
<td>2</td>
<td>.047</td>
<td>2</td>
</tr>
<tr>
<td>Mean Weight</td>
<td>.10</td>
<td></td>
<td>.10</td>
<td></td>
<td>.10</td>
<td></td>
<td>.10</td>
<td></td>
<td>.10</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>.037</td>
<td></td>
<td>.036</td>
<td></td>
<td>.045</td>
<td></td>
<td>.045</td>
<td></td>
<td>.050</td>
<td></td>
</tr>
</tbody>
</table>
Table 19 provides a summary of the composite normalized weights and the rank of these derived weights for each of the ten criteria across the five methods. The Table reveals differences in both the weights assigned to the criteria, as well as the ranking of these weights. For example, the criterion of transportability was given a weight of .084 through the ranking method and .094 through the rating method. When these weights were ordered, they ranked 3 and 5, respectively. Thus, some differences were apparent in the weightings provided by the alternative methods.

Since the weights have been normalized, the mean weight for each method is equal, a value of .10. However, the variation in the criterion weights differed from method to method. The standard deviation of the criterion weights ranged from .036, for the rating method, to .050 for the successive comparisons method.

On the basis of the findings summarized in Table 19, as well as the balance of the information derived from the data collection instrument, the following sections individually address the research questions and hypotheses posed by the study.

1. **Inter-Method Reliability**

As an initial source of descriptive information, product-moment correlation coefficients were derived between method weights. These inter-method correlations are reported in Table 20.
An analysis of Table 20 reveals that the inter-method correlations ranged in strength from .79 to .99. Thus, the degree of association between methods varies considerably. The Table further indicates that the ranking method maintained the highest correlations with other methods, ranging from .98 to .84, while the successive comparisons method exhibited the weakest associations with the other methods, ranging from .84 to .79.

As noted in the analysis plan, both parametric and nonparametric measures of association were to be applied to the method weights. Comparisons were made between the weights derived through each method and the sum of the weights of the balance of the methods. In addition, a measure of the overall association of the five methods was computed. The results of the application of the aforementioned correlational analyses are summarized in Table 21.
Table 21 indicates that the reliabilities of all methods differ from chance at the .05 level of probability. The convergent reliability values derived through nonparametric methods (Kendall $\tau$) range from .98, for ranking, to .69 for the successive comparisons method. When a parametric measure of association (Pearson $r$) was applied, similar results were observed. In this instance, the convergent reliabilities ranged from .97, for the ranking method, to .83 for the successive comparisons method. The overall inter-method reliability, as derived through the application of the coefficient of concordance ($\omega$), was found to be .87.

The results substantiate the hypothesis that significant relationships do exist between the various criterion weighting methods.

The results of the parametric and nonparametric measures of association...
are fairly parallel. Both analyses identify the ranking method as having the strongest degree of association with the criterion variable and the successive comparisons method as having the weakest degree of association with the criterion variable.

A review of Table 21 reveals that the nonparametric measures of association range from .98 to .69, a variation of .29 across the five methods, while the parametric correlations range in value from .97 to .83, a variation of .14. Thus, the nonparametric measures exhibit the greatest range in values. These differences are quite large. That is, the observed differences, particularly between ranking and successive comparisons methods, might warrant a decision to exclude the latter method on the basis of convergent reliability, given the acceptance of the composite weights as a valid estimate of the "true score." All things being equal, successive comparisons might be considered less appropriate than ranking on this basis.

2. Elapsed-Time for Method Application

The elapsed-time for method application is based upon the time required for each subject to accomplish the tasks described for each of the weighting methods. The range of elapsed-time required for the application of each of the five methods is reported in Table 22.
TABLE 22
ELAPSED-TIME IN MINUTES FOR APPLICATION
OF THE CRITERION WEIGHTING METHODS
BY STUDY GROUP THREE

<table>
<thead>
<tr>
<th>Method</th>
<th>shortest</th>
<th>mean</th>
<th>longest</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.50</td>
<td>1.72</td>
<td>4.25</td>
<td>1.467</td>
</tr>
<tr>
<td>Rating</td>
<td>1.00</td>
<td>1.50</td>
<td>2.25</td>
<td>.468</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>2.25</td>
<td>4.08</td>
<td>8.00</td>
<td>2.258</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>2.00</td>
<td>6.75</td>
<td>13.50</td>
<td>4.709</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>15.50</td>
<td>25.17</td>
<td>33.00</td>
<td>8.048</td>
</tr>
</tbody>
</table>

Contrary to the results predicted by one of the study hypotheses, rating rather than ranking required the least (average) time for application. However, the balance of the original hypothesis is supported by the time required for the successive comparisons method, which was the largest elapsed-time. The Friedman rank sum statistic was applied to the ranks of the elapsed-time required by each subject in applying each method to the ten criteria. The result of the test was statistically significant, in that the derived $S$ of 18.6 exceeded the criterial value of the upper-tail probabilities for $S$ ($n=5$, $k=5$, $\alpha=.049$) of 8.96.

The result of this test indicates that not all alternative methods require an equal elapsed-time for application. Since the test only reveals overall differences in location, it does not directly
identify which methods are significantly different. Therefore, distribution-free multiple comparisons were made on the basis of the Friedman rank sums. The comparisons made between ranking and successive comparisons, and rating and successive comparisons were both found to be significant. These rank sum differences in elapsed-time exceeded the critical value for such multiple comparisons \((a=.04, k=5, n=5)\), which in this instance was 14.

Thus, although the ranking method did not require the least average elapsed-time, the hypothesis is substantiated that a significant difference exists between the elapsed-time required for applying the ranking and successive comparisons methods. In addition, the unpredicted significant difference between rating and successive comparisons provides additional information for determining the relative amounts of time required for the alternative methods when they are applied to ten criteria.

Further analysis of Table 22 reveals that the elapsed-time for application of the ranking method averaged 1.72 minutes, when ten criteria were under consideration. Rating required slightly less time, averaging 1.50 minutes. Partial- and complete-paired comparisons required, on the average, 4.08 and 6.75 minutes, respectively. The successive comparisons method required an average of 25.17 minutes and in one case required 33 minutes to apply. Thus, the successive comparisons method required, on the average, nearly fifteen times longer to apply than the ranking method. Such findings further substantiate the contentions of the elapsed-time hypothesis.
3. **Method Difficulty**

The five subjects indicated the methods which they perceived as being easiest and most difficult to apply. These perceptions were measured by responses to two questions in the study instrument. First, the subjects were asked to identify which method was easiest to use, and secondly, which method was most difficult to use. The responses to these questions are summarized in Table 23.

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Which method was easiest to</td>
<td>ranking</td>
<td>3</td>
</tr>
<tr>
<td>use?</td>
<td>rating</td>
<td>2</td>
</tr>
<tr>
<td>2. Which method was most difficult to use?</td>
<td>successive comparisons</td>
<td>5</td>
</tr>
</tbody>
</table>

The Table reveals that the subjects' perceptions were divided regarding the method easiest to use. Three subjects indicated ranking as easiest, while two subjects noted rating as the easiest method. The successive comparisons method was unanimously nominated by the subjects as the most difficult procedure to utilize. An application of the composite difficulty equation described in the analysis section reveals the differential perceived difficulty weights noted in Table 24.
TABLE 24
COMPOSITE METHOD DIFFICULTY WEIGHTS
FOR STUDY GROUP THREE
(possible range = 1.0 to -0.5)

<table>
<thead>
<tr>
<th>Method</th>
<th>Composite Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>+0.6</td>
</tr>
<tr>
<td>Rating</td>
<td>+0.4</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>0</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>0</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>-0.5</td>
</tr>
</tbody>
</table>

As can be seen in Table 24, the perceived difficulty weights range from +0.6, for the ranking method, to -0.5 for the successive comparisons method. Rating received a moderately positive difficulty weight, while the paired comparisons methods received neutral weightings.

The differences observed between the rating and ranking methods, on the one hand, and the successive comparisons method, on the other hand, may be large enough to warrant attention. However, on the basis of the small number of observations, evidence regarding the relative difficulties of the methods is not particularly substantial.

4. Perceived Method Applicability

The five subjects responded to a question regarding the method which they perceived as being most applicable to their decision-making needs. Two of the subjects indicated that the ranking method
was most applicable, two of the subjects regarded rating as being most applicable, while the last subject indicated a preference for the partial-paired comparisons method. The results of the application of the composite applicability equation described in the analysis section are summarized in Table 25.

**TABLE 25**

COMPOSITE APPLICABILITY WEIGHTS OF THE FIVE CRITERION WEIGHTING METHODS FOR STUDY GROUP THREE

(possible range = 1.0 to 0)

<table>
<thead>
<tr>
<th>Method</th>
<th>Applicability Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.40</td>
</tr>
<tr>
<td>Rating</td>
<td>.40</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>.20</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>0</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>0</td>
</tr>
</tbody>
</table>

The results depicted in Table 25 reveal a preference for the ranking, rating and partial-paired comparisons. The complete-paired comparisons and successive comparisons methods were not preferred by any of the subjects as being the most applicable to their decision-making needs. Thus, on the basis of the present study group, ranking, rating and partial-paired comparisons methods would appear to exhibit at least a slight preferential advantage.
5. Summary of Study Three Results

A comparison of the results of the Study Group Three application of the five methods is summarized across the four study variables in Table 26. As noted in the Table, ranking was found to have both the largest parametrically derived reliability (.97) and nonparametrically calculated reliability (.98). The method required an average of 1.72 minutes to apply to the criteria. This elapsed-time was only slightly larger than the time required for rating. The method was considered to be the easiest to use and was considered to be applicable to the decision-making needs of two of the five subjects.

Rating was found to have the second highest parametrically derived reliability (.96), but a lower nonparametric reliability of .80. The rating method required the least time to apply; on the average 1.50 minutes were needed to apply the method to the ten criteria. The rating method received a moderately positive difficulty weight and was considered to be applicable to the decision-making needs of two of the five subjects.

The partial-paired comparisons method received the relatively high reliability values of .91 and .96 through the respective nonparametric and parametric analyses. The method required an average elapsed-time of 4.08 minutes to be applied to the ten criteria. Although the method was not indicated as being particularly difficult or easy, it received a mild applicability weight of .2.
<table>
<thead>
<tr>
<th>Method</th>
<th>Convergent $\tau$</th>
<th>Reliability $r$</th>
<th>Mean Elapsed-Time in Minutes</th>
<th>Difficulty Weight</th>
<th>Applicability Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.98</td>
<td>.97</td>
<td>1.72</td>
<td>+.6</td>
<td>.4</td>
</tr>
<tr>
<td>Rating</td>
<td>.80</td>
<td>.96</td>
<td>1.50</td>
<td>+.4</td>
<td>.4</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>.91</td>
<td>.96</td>
<td>4.08</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>.80</td>
<td>.95</td>
<td>6.75</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>.69</td>
<td>.83</td>
<td>25.17</td>
<td>-.5</td>
<td>0</td>
</tr>
</tbody>
</table>
The complete-paired comparisons method received a nonparametric correlation with the criterion of .80, while a parametric correlation with the same criterion was .95. The method required an average of 6.75 minutes to apply, an interval which was slightly less than expected. Since this method contains twice the number of criterion pairs, it would be anticipated that the complete-paired comparisons method would require about twice the time needed for applying the partial-paired comparisons. The method was neither nominated as being particularly easy or difficult, nor most applicable to the decision-making needs of the subjects.

The successive comparisons method received the lowest reliability values, .69 and .83, respectively, for the nonparametric and parametric analyses. The elapsed-time required for applying the method to the ten criteria ranged from 15.50 to 33 minutes. The average elapsed time of 25.17 minutes was found to be more than 14 times the average time required for ranking the same criteria. The method was unanimously nominated as the most difficult to apply and subsequently was not indicated as being applicable to the decision-making needs of the subjects in Study Group Three.

D. **Study Four**

The data collection instrument was administered to the five subjects in accordance with the established protocol. The subjects were randomly selected from a list of project directors of E.S.E.A. Title III projects currently operating in the State of Oregon.
The projects were distributed throughout the State and included sites in Eastern, Southern, Mid-Willamette, and the Portland Metropolitan areas. The sites were located at three school districts and two intermediate education districts.

The subjects ranged in age from 32 to 46. Educational backgrounds included one Bachelor of Arts, three Masters of Science, and one Doctor of Philosophy. Program management experience ranged from three to 15 years.

All subjects successfully completed the instrument. Four of the subjects required follow-up briefings to clarify the successive comparisons procedure. The data derived from the subjects' application of the five methods to the ten criteria were transformed and reduced across subjects with the previously described data reduction procedures. The resultant criterion weights are displayed in Table 27.

Table 27 provides a summary of the composite normalized weights and the rank of these derived weights for each of the ten criteria across the five methods. The Table reveals differences in both the weights assigned to the criteria, as well as the rankings of these weights, across the five methods. For example, the criterion transportability was given a weight of .131 when the ranking method was applied and a weight of .128 when the rating method was utilized. When these weights were ordered across the ten criteria, they ranked 9 and 10, respectively. Even more dramatic differences can be observed when other criterion weights are considered, such as the weights
TABLE 27

RESULTS OF THE CRITERION WEIGHTING METHOD APPLICATIONS OF STUDY GROUP FOUR

<table>
<thead>
<tr>
<th>CRITERIA</th>
<th>Ranking</th>
<th>Rating</th>
<th>Partial-Paired Comparisons</th>
<th>Complete-Paired Comparisons</th>
<th>Successive Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>weight</td>
<td>rank</td>
<td>weight rank</td>
<td>weight rank</td>
<td>weight rank</td>
</tr>
<tr>
<td>Transportability</td>
<td>0.131</td>
<td>9</td>
<td>0.128 10</td>
<td>0.107 8</td>
<td>0.125 8</td>
</tr>
<tr>
<td>Innovativeness</td>
<td>0.116</td>
<td>7</td>
<td>0.094 6</td>
<td>0.085 6</td>
<td>0.062 1</td>
</tr>
<tr>
<td>Generalizability</td>
<td>0.080</td>
<td>3.5</td>
<td>0.086 4</td>
<td>0.080 4</td>
<td>0.093 7</td>
</tr>
<tr>
<td>Cost</td>
<td>0.087</td>
<td>5</td>
<td>0.090 5</td>
<td>0.093 7</td>
<td>0.082 5</td>
</tr>
<tr>
<td>Significance</td>
<td>0.128</td>
<td>8</td>
<td>0.126 9</td>
<td>0.169 9.5</td>
<td>0.160 9</td>
</tr>
<tr>
<td>Evidence</td>
<td>0.091</td>
<td>6</td>
<td>0.085 3</td>
<td>0.075 3</td>
<td>0.069 2.5</td>
</tr>
<tr>
<td>Clarity</td>
<td>0.065</td>
<td>1</td>
<td>0.081 2</td>
<td>0.071 2</td>
<td>0.080 4</td>
</tr>
<tr>
<td>Maintenance</td>
<td>0.073</td>
<td>2</td>
<td>0.078 1</td>
<td>0.067 1</td>
<td>0.069 2.5</td>
</tr>
<tr>
<td>Effectiveness</td>
<td>0.149</td>
<td>10</td>
<td>0.124 8</td>
<td>0.169 9.5</td>
<td>0.171 10</td>
</tr>
<tr>
<td>Compatibility</td>
<td>0.080</td>
<td>3.5</td>
<td>0.107 7</td>
<td>0.084 5</td>
<td>0.089 6</td>
</tr>
</tbody>
</table>

Mean Weight = 0.10

Standard Deviation = 0.029
derived for the innovativeness criterion for the partial- and complete-paired comparisons methods. Thus, initial differences were apparent in the weightings provided by the alternative methods.

Since the weights have been normalized, the mean weight for each method is equal, a value of .10. However, the variation in the criterion weights differed from method to method. The standard deviation of the criterion weights ranged from .020, for the rating method, to .043 for the successive comparisons method.

On the basis of the findings summarized in Table 27, as well as the balance of information derived from the data collection instrument, the following sections individually address the research questions and hypotheses upon which the study is focused.

1. Inter-Method Reliability

As an initial source of descriptive information, product-moment correlation coefficients were derived between method weights for each pair of methods. These inter-method correlations are reported in Table 28.
A perusal of Table 28 reveals that the inter-method correlations ranged in magnitude from .77 to .95. Thus, the degree of association exhibited between methods varies considerably. The Table further indicates that the rating method maintained the strongest correlations with the other methods, ranging from .83 to .93, while the ranking method exhibited the weakest associations when all other methods are considered, ranging from .77 to .85.

As outlined in the analysis plan, both parametric and nonparametric measures of association were to be applied to the method weights. Comparisons were made between the weights derived through each method and the sum of the weights derived through the balance of the methods. In addition, a measure of the overall association between the five methods was computed. The results of the application of the aforementioned correlational analyses are summarized in Table 29.
An examination of Table 29 indicates that all reliabilities differ from zero at the .05 level of probability. The convergent reliabilities derived through nonparametric analyses (Kendall \( \tau \)) range from .76, for the partial-paired comparisons method, to .53 for the ranking method. When a parametric measure of association (Pearson \( r \)) was applied, the correlation coefficients ranged from .94, for the complete-paired comparisons method, to .85 for the ranking method. The overall inter-method reliability, as derived through the application of the coefficient of concordance (\( \omega \)), was found to be .78.

The results substantiate the hypothesis that the reliabilities of the methods would all significantly differ from zero. The results of the parametric and nonparametric measures differ slightly. While both analyses identify the ranking method as having the weakest degree
of association with the criterion variable, the nonparametric measure indicates the partial-paired comparisons method as having the strongest convergent reliability (.76) and the parametric measures identifies the complete-paired comparisons method as having the strongest convergent reliability (.94).

Further examination of Table 29 reveals that the nonparametric measures of association range from .53 to .76, a difference of .23, while the parametric coefficients range from .85 to .94, a range of .09. This would suggest that the nonparametric method is affected more by differences in the ranks of the method weights. None of the present nonparametric coefficients are particularly substantial, while the parametric coefficients are of a more "respectable" magnitude. Because of the limited range of the reliabilities observed through the parametric approach, it would be difficult to justify a decision to exclude methods on the basis of the relative magnitudes of these reliabilities.

2. Elapsed-Time for Method Application

The elapsed-time for method application is based upon the time required for each subject to accomplish the tasks described for each of the weighting methods. The range of elapsed-time for the application of each of the five methods is reported in Table 30.
**TABLE 30**

ELAPSED-TIME IN MINUTES FOR APPLICATION
OF THE CRITERION WEIGHTING METHODS
BY STUDY GROUP FOUR

<table>
<thead>
<tr>
<th>Method</th>
<th>shortest</th>
<th>mean</th>
<th>longest</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.75</td>
<td>2.73</td>
<td>6.58</td>
<td>2.253</td>
</tr>
<tr>
<td>Rating</td>
<td>.33</td>
<td>2.48</td>
<td>4.17</td>
<td>1.447</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>3.17</td>
<td>8.17</td>
<td>17.83</td>
<td>5.815</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>4.33</td>
<td>14.72</td>
<td>40.42</td>
<td>14.953</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>28.00</td>
<td>38.20</td>
<td>46.00</td>
<td>6.916</td>
</tr>
</tbody>
</table>

Contrary to the results predicted by one of the study hypotheses, rating rather than ranking required the least (average) time to apply to the ten criteria. However, the balance of the original hypothesis is directly supported by the time required for applying the successive comparisons method, which was by far the longest elapsed-time observed for the methods. The Friedman rank sum statistic was applied to the ranks of the elapsed-time required by each subject in applying each method to the ten criteria. The result of the test was statistically significant, in that the derived $S$ of 19.04 exceeded the critical value of the upper-tail probabilities for $S$ ($n=5$, $k=5$, $a=.049$) of 8.96.

The result of this test indicates that not all alternative methods require an equal elapsed-time for application. Since the
test only reveals overall differences in location, it does not directly identify which methods are significantly different. Therefore, distribution-free multiple comparisons were made using the Friedman rank sums derived for each method. The comparisons made between ranking and successive comparisons, and rating and successive comparisons were both found to be statistically significant. These rank sum differences in elapsed-time exceeded the critical value for such multiple comparisons \((a=.04, k=5, n=5)\) which was 14.

Thus, although the rating method required, on the average, less time to apply than the ranking method, the difference between these methods is not statistically significant. Furthermore, since a significant difference does exist between the elapsed-times for ranking and successive comparisons, the original hypothesis is substantiated. In addition, the unpredicted significant difference between the elapsed-times of rating and successive comparisons lends further information for determining the relative time intervals required for the alternative methods.

Further study of Table 30 shows that the elapsed-time for application of the ranking method averaged 2.73 minutes, when ten criteria were under consideration. Rating required slightly less time, averaging 2.48 minutes. Partial- and complete-paired comparisons required an average of 8.17 and 14.72 minutes, respectively. The successive comparisons method required an average of 38.20 minutes.
3. **Method Difficulty**

The five subjects indicated the methods which they perceived as being easiest and most difficult by responding to two questions at the conclusion of the study instrument. First, the subjects were asked to identify which method was easiest to use, and secondly, which method was most difficult to use. The responses to these questions are summarized in Table 31.

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Which method was easiest to use?</td>
<td>ranking</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>rating</td>
<td>2</td>
</tr>
<tr>
<td>2. Which method was most difficult to use?</td>
<td>successive comparisons</td>
<td>5</td>
</tr>
</tbody>
</table>

The Table would appear to indicate that the subjects' perceptions were divided regarding which method was easiest to use. Three subjects regarded ranking as the easiest to use, while two subjects noted that the rating method was easiest to use. The successive comparisons method was unanimously nominated by the subjects as the most difficult procedure to utilize. An application of the composite difficulty equation described in the analysis section reveals the differential perceived difficulty weights summarized in Table 32.
TABLE 32

COMPOSITE METHOD DIFFICULTY WEIGHTS
FOR STUDY GROUP FOUR

(possible range = 1.0 to -.5)

<table>
<thead>
<tr>
<th>Method</th>
<th>Composite Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>+.6</td>
</tr>
<tr>
<td>Rating</td>
<td>+.4</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>0</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>0</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>-.5</td>
</tr>
</tbody>
</table>

As can be seen in Table 32, the perceived difficulty weights range from +.6, for the ranking method, to -.5 for the successive comparisons method. Rating received a moderately positive difficulty weight, while the paired comparisons methods received neutral weightings.

The differences observed between the rating and ranking methods versus the successive comparisons method may be sufficiently large to warrant attention. However, on the basis of the small number of observations, evidence regarding the relative difficulties of the methods is not particularly substantial.

4. Perceived Method Applicability

The five subjects responded to a question regarding the method which they perceived as being most applicable to their decision-making needs. Two of the subjects indicated that the ranking method
was most applicable, two subjects regarded rating as being most appropriate, while the last subject indicated a preference for the partial-paired comparisons method. The results of the composite applicability equation described in the analysis section are summarized in Table 33.

**TABLE 33**

**COMPOSITE APPLICABILITY WEIGHTS OF THE FIVE CRITERION WEIGHTING METHODS FOR STUDY GROUP FOUR**

(possible range = 1.0 to 0)

<table>
<thead>
<tr>
<th>Method</th>
<th>Applicability Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.40</td>
</tr>
<tr>
<td>Rating</td>
<td>.40</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>.20</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>0</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>0</td>
</tr>
</tbody>
</table>

The composite weights depicted in Table 33 indicate a preference for the ranking, rating and partial-paired comparisons methods. The complete-paired comparisons and successive comparisons methods were not preferred by any subjects as being the most applicable to their decision-making needs. Thus, on the basis of the present study group, ranking, rating and partial-paired comparisons methods would appear to experience at least a slight preferential advantage over the other two methods.
5. Summary of Study Four Results

A comparison of the results of the Study Group Four application of the five methods is summarized across the four study variables in Table 34. The Table reveals that the ranking method received the lowest reliabilities of any of the methods, .53 for the nonparametric procedure and .85 for the parametric procedure. The method required an average of 2.73 minutes to apply to the ten criteria. This elapsed-time was only slightly longer than the time required for rating. The method was considered to be the easiest to use and was nominated by two of the subjects as being most applicable to their decision-making needs.

Rating was found to have moderate convergent reliabilities on the basis of the nonparametric (.73) and parametric (.92) coefficients. The rating method required the least time to apply, an average of 2.48 minutes, when ten criteria were under consideration. The rating method received a moderately positive difficulty weight and was considered to be applicable to the decision-making needs of two of the five subjects.

The partial-paired comparisons method received the largest nonparametrically derived reliability (.76) and a moderate parametric reliability of .92. The method required an average elapsed-time of 8.17 minutes to be applied to the ten criteria. Although the method was not noted as being particularly easy or difficult to use, it was considered to be most appropriate for the decision-making needs of one of the subjects.
TABLE 34
SUMMARY OF STUDY GROUP FOUR FINDINGS

<table>
<thead>
<tr>
<th>Method</th>
<th>Convergent ( \tau )</th>
<th>Reliability ( r )</th>
<th>Mean Elapsed-Time in Minutes</th>
<th>Difficulty Weight</th>
<th>Applicability Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.53</td>
<td>.85</td>
<td>2.73</td>
<td>+.6</td>
<td>.4</td>
</tr>
<tr>
<td>Rating</td>
<td>.73</td>
<td>.92</td>
<td>2.48</td>
<td>+.4</td>
<td>.4</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>.76</td>
<td>.92</td>
<td>8.17</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>.62</td>
<td>.94</td>
<td>14.72</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>.69</td>
<td>.92</td>
<td>38.20</td>
<td>-.5</td>
<td>0</td>
</tr>
</tbody>
</table>
The complete-paired comparisons method received a moderate nonparametrically derived reliability coefficient of .62 and the largest parametrically derived reliability coefficient (.94). The method required an average of 14.72 minutes to apply. The method was neither considered to be particularly difficult or easy, nor regarded as being applicable to the decision-making needs of the subjects.

The successive comparisons method received moderate convergent reliabilities on the basis of the nonparametric (.69) and parametric (.92) coefficients. The method required by far the longest time to apply, averaging 38.2 minutes. This elapsed-time was about 14 times the average elapsed-time required for ranking the same ten criteria. The method was unanimously nominated as being the most difficult to apply and subsequently was not considered to be applicable to the decision-making needs of the subjects in Study Group Four.

E. Summary Results and Discussion

The results observed in each of the study group applications have been individually reported in the four preceding sections. In the present section the overall results observed across study groups are summarized and discussed. This concluding summary will address the four study questions and related hypotheses in the paragraphs which follow.
1. Inter-Method Reliability

The convergent reliabilities derived through nonparametric analysis (Kendall $\tau$) for each of the five methods are summarized in Table 35.

**TABLE 35**

NONPARAMETRIC CONVERGENT RELIABILITIES ($\tau$) OBSERVED ACROSS STUDY GROUPS

<table>
<thead>
<tr>
<th>Method</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.71</td>
<td>.95</td>
<td>.98</td>
<td>.53</td>
</tr>
<tr>
<td>Rating</td>
<td>.81</td>
<td>1.00</td>
<td>.80</td>
<td>.73</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>.73</td>
<td>.90</td>
<td>.91</td>
<td>.76</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>.73</td>
<td>1.00</td>
<td>.80</td>
<td>.62</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>.71</td>
<td>.86</td>
<td>.69</td>
<td>.69</td>
</tr>
</tbody>
</table>

In reviewing Table 35 it can be seen that the rating method maintained the strongest reliability coefficients, ranging from 1.00 to .73. This method was followed by partial-paired comparisons, ranking, complete-paired comparisons and successive comparisons methods, in that order. The successive comparisons method received the lowest reliabilities, when all four study groups are considered, ranging from .86 to .69.

It is further observed that the reliabilities obtained varied from study to study. If an average of the study group reliabilities is estimated, the average convergent reliabilities range from .97,
in the second study, to .67 in the fourth study. Thus, consistent reliability values were not obtained across groups. In each case, however, the reliability values were significantly different from zero at the .05 level of probability. Such a finding is consistent with the study hypothesis.

The convergent reliabilities derived through parametric analysis (Pearson r) for each of the five methods are summarized in Table 36.

**TABLE 36**

**PARAMETRIC CONVERGENT RELIABILITIES (r) OBSERVED ACROSS STUDY GROUPS**

<table>
<thead>
<tr>
<th>Method</th>
<th>Study Group 1</th>
<th>Study Group 2</th>
<th>Study Group 3</th>
<th>Study Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>.96</td>
<td>1.00</td>
<td>.97</td>
<td>.85</td>
</tr>
<tr>
<td>Rating</td>
<td>.88</td>
<td>.99</td>
<td>.96</td>
<td>.92</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>.95</td>
<td>.99</td>
<td>.96</td>
<td>.92</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>.94</td>
<td>.99</td>
<td>.95</td>
<td>.94</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>.97</td>
<td>.97</td>
<td>.83</td>
<td>.92</td>
</tr>
</tbody>
</table>

In reviewing Table 36 it can be seen that the paired comparisons methods maintained the strongest reliability coefficients, ranging from .92 to .99. These methods were followed by rating, ranking and successive comparisons methods, in that order. The successive comparisons method received the lowest reliabilities, when all four study groups are taken into consideration, ranging from .83 to .97.
The Table further reveals that the reliabilities varied slightly from study to study. If an average of the study group coefficients is computed, the average convergent reliabilities range from .99, in the second study, to .91 in the final study. Thus, reliability values varied from group to group. The complete-paired comparisons method varied only .07 points in reliability across studies, while the successive comparisons coefficients ranged as much as .14 across studies. Thus, the reliability values obtained were not entirely consistent from group to group. However, in each case the reliability values were significantly different from zero well beyond the .05 level of probability. Such a finding is consistent with the study hypothesis.

The overall results substantiate the hypothesis that relationships do exist between the various criterion weighting methods. However, the measures of association generated by the alternative statistical procedures do not consistently identify a method as having the greatest reliability.

Secondly, the variation observed in the reliabilities from study to study make interpretation difficult. The "noise" associated with the between-group differences in reliability conceals the existence of any consistent overall convergent reliability, if one exists. However, convergent reliability is not a measure of the weighting methods alone, but also of the circumstance and the criterion or standard of comparison. Obviously these two considerations
change from study to study. Therefore, the search for an overall reliability may be futile.

Based upon the sample size, the correlations can be shown to be statistically significant from chance. Beyond this, the strength of these relationships becomes a matter of speculation. Considering these outcomes, the results of the present inquiry appear to indicate that methods cannot be readily distinguished based upon their respective reliabilities. In effect, variation in method reliabilities has not been shown to be sufficiently large to justify selective decisions on that basis. This is consistent with a study by Eckenrode, in which the consistency between each of these methods were all found to be high. Torgerson further substantiates this finding by indicating that inter-method differences do not tend to be large.

The results of the reliability portion of the study could suggest at least three possible alternative interpretations. First, the reliabilities of the respective methods are sufficiently large to enable application of any of the methods. Second, the sample size of the study does not provide sufficient evidence for selecting a method on the basis of reliability. Third, the reliabilities of the respective methods do not differ to a degree which warrants selecting a method on the basis of reliability. In any case, the present study

---


has not provided evidence which clearly indicates the superiority of one criterion weighting method over others on the basis of convergent reliability.

It should be noted that convergent reliability is, in effect, a form of validity. As Tiffin and McCormick point out,

Quantitative evidence of the validity of ratings usually is difficult, if not impossible, to obtain. In fact, merit ratings sometimes are used because there are no other criteria of employee merit available. Under such circumstances the validity of merit ratings may have to be inferred from their reliability. This does not necessarily imply that a low inter-method reliability indicates a method has a low validity. Both the validity of the method and the validity of the criterion variable must be considered. Thorndike and Hagen note that,

Validity, insofar as we can appraise it, is the crucial test of a measurement procedure. Reliability is important only as a necessary condition for a measure to have validity. The ceiling for the possible validity of test is set by its reliability.

Other things being equal, the more reliable procedure is preferred if a selection has to be made between procedures. However, other things are usually not equal. Thus, it is suggested that other variables be considered before a selective decision is made.


2. Elapsed-Time for Method Application

The average elapsed-times required for the subjects in each study group to apply the five weighting methods are summarized in Table 37.

**TABLE 37**

**MEAN ELAPSED-TIME (IN MINUTES)**

**REQUIRED FOR METHOD APPLICATIONS ACROSS STUDY GROUPS**

<table>
<thead>
<tr>
<th>Method</th>
<th>1 n=7</th>
<th>2</th>
<th>Study Group</th>
<th>3 n=10</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>1.15</td>
<td>.84</td>
<td>1.72</td>
<td>2.73</td>
<td></td>
</tr>
<tr>
<td>Rating</td>
<td>1.62</td>
<td>1.23</td>
<td>1.50</td>
<td>2.48</td>
<td></td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>2.73</td>
<td>1.61</td>
<td>4.08</td>
<td>8.17</td>
<td></td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>4.85</td>
<td>3.31</td>
<td>6.75</td>
<td>14.72</td>
<td></td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>15.20</td>
<td>15.63</td>
<td>25.17</td>
<td>38.20</td>
<td></td>
</tr>
</tbody>
</table>

Table 37 reveals that the ranking method required the least time to apply, when seven criteria were being processed, while the rating method required the least time to apply when ten criteria were being processed. In either case, the successive comparisons method required the longest time. The time required to apply the various methods could be compared by combining the means in Study Groups One and Two, and Three and Four. The reader is cautioned that such a comparison would be highly speculative, since such a comparison is confounded by both different subjects and different criteria.
Therefore, such a comparison will not be made in the present inquiry. However, the findings of previous research conducted by Eckenrode regarding this relationship is noted below:

The results...indicate that for simple preference judgements using ranking and paired comparisons methods, the time required to make these judgements is a linear function of the number of judgements required. The fact that the number of judgements required in using partial paired comparisons is an exponential function of the absolute number of items on which preferences are expressed makes paired comparisons much less efficient than ranking, where efficiency is measured as the time which a judge requires to complete a set of preference judgements on n items.  

It would be expected that differences in efficiency would become more apparent as the number of criteria increases. Although caution regarding comparisons across groups is reiterated, a re-examination of Table 37 reveals a number of points. First, the differences in the times required for rating and ranking become less apparent when applied to ten criteria. On the otherhand, differences between the partial-paired comparisons, complete-paired comparisons and successive comparisons methods become more apparent when applied to ten criteria. A peculiar observation is the relationship between partial- and complete-paired comparisons. It would be expected that complete-paired comparisons would require twice the time needed for applying the partial-paired comparisons method. When ten criteria were under consideration, this expected difference appeared to diminish slightly.

---

In the first study, only the observed differences in elapsed-time between the ranking and successive comparisons methods were found to be statistically significant. In the three subsequent studies, the differences between the elapsed-time for rating and successive comparisons methods were found to be statistically significant, as well as the ranking and successive comparisons methods. In each case, the hypothesis of the inquiry was confirmed. Thus, the results of the studies focus upon apparent differences between rating and ranking, on the one hand, and successive comparisons, on the other hand.

Depending upon the degree to which elapsed-time for application is considered to be a critical variable, the study findings may provide a basis for not selecting the successive comparisons method for criterion weighting problems. This posture is supported by a statement by Eckenrode that,

The method of successive comparisons was so difficult to train judges to use and took such a long time for them to use in the specific system experiment..., while yielding weights similar to the other methods, that it was dropped from succeeding experiments.  

Thus, differences are observed between the various methods when the time required for their application to a selected number of criteria is considered. However, these differences may change when the number and difficulty of the criteria to be judged are modified.

---

6 Eckenrode, "Weighting Multiple Criteria," p. 186.
Eckenrode points out, "the more difficult the judgement situation the longer it takes to decide upon a preference."^7

3. Method Difficulty

The composite difficulty weights derived for each of the methods in each of the four study settings are summarized in Table 38.

<table>
<thead>
<tr>
<th>Method</th>
<th>Study Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Ranking</td>
<td>+.80</td>
</tr>
<tr>
<td>Rating</td>
<td>0</td>
</tr>
<tr>
<td>Partial-Paired Comparisons</td>
<td>0</td>
</tr>
<tr>
<td>Complete-Paired Comparisons</td>
<td>+.20</td>
</tr>
<tr>
<td>Successive Comparisons</td>
<td>-.50</td>
</tr>
</tbody>
</table>

In reviewing Table 38 at least three major points are apparent. First, ranking appears to maintain a high "ease-of-use" rating across study groups. Secondly, the successive comparisons method maintains a very low "ease-of-use" rating across study groups. Third, rating appears to be considered the second easiest method to use when the ratings across groups are observed.

If it can be assumed for the moment that the difficulty weights can be combined across study groups, more substantial differences become apparent along this dimension. The order of the methods, from easiest to most difficult, would appear to be ranking, rating, partial-paired comparisons, complete-paired comparisons and successive comparisons. Major differences are apparent between the ranking and rating methods, on the one hand, and the successive comparisons method, on the other hand.

The results of the present inquiry provide substantial evidence for discriminating between methods along the method difficulty dimension. Other things being equal, a selective decision between ranking and successive comparisons, or rating and successive comparisons, would appear to be justifiable along this dimension.

4. Perceived Method Applicability

The composite applicability weights derived for each of the methods in each of the four study settings are summarized in Table 39.
An examination of Table 39 reveals at least three major points. First, ranking appears to maintain a degree of perceived utility or applicability across subjects. Secondly, in a parallel manner, rating appears to maintain a degree of perceived utility across subjects. Third, the complete-paired comparisons and successive comparisons method are not perceived as being most applicable to the decision-making needs of any of the subjects. The partial-paired comparisons method was perceived as being appropriate for the decision-making needs of one individual in Study Group Three and one individual in Study Group Four.

Again it appears that the ranking and rating methods experience the advantage along this dimension. Other things being equal, a decision to select the ranking and rating methods, on the one hand,
over the complete-paired and successive comparisons methods, on the other hand, would appear to be justifiable along this dimension.

5. Summary Results

The combined evidence suggest a number of conclusions along two lines: First, conclusions regarding the outcome of the study questions and hypotheses; and, secondly, the consequences of the study regarding the five weighting methods. The outcome of the study regarding these two lines is summarized in Table 40.

Finally, the results are further interpreted in terms of the consequences which the findings have upon the various criterion weighting methods in educational decision-making:

1. The ranking and rating methods have normally received the greatest attention in educational decision-making. The results of the present study do not dispute this preference.

2. The successive comparisons method may provide results which are essentially the same as the ranking and rating methods, but will require a greater amount of time and will be perceived as more difficult and less appropriate than the ranking and rating methods.

3. The paired comparisons methods demonstrated acceptable reliabilities and slightly positive perceptions on the part of the users.

A number of limitations should be considered when examining the aforementioned general interpretations. First, the generalizability
### TABLE 40
RESULTS OF THE STUDY QUESTIONS AND HYPOTHESES

<table>
<thead>
<tr>
<th>Question or Hypothesis</th>
<th>Results</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Inter-Method Reliability</td>
<td></td>
<td>Although substantial reliabilities exist, methods can not be distinguished on the basis of reliability.</td>
</tr>
<tr>
<td>(a) What are the reliabilities?</td>
<td>(See Tables 35 and 36)</td>
<td></td>
</tr>
<tr>
<td>(b) Do they differ significantly from zero?</td>
<td>Yes, beyond .05 level of significance.</td>
<td>Rating and ranking require significantly less time than the successive comparisons method.</td>
</tr>
<tr>
<td>2. Elapsed-Time for Application</td>
<td></td>
<td>Rating and ranking are considered to be easiest, while successive comparisons is considered most difficult.</td>
</tr>
<tr>
<td>(a) What are the elapsed-times?</td>
<td>(See Table 37)</td>
<td>Rating and ranking require significantly less time than the successive comparisons method.</td>
</tr>
<tr>
<td>(b) Do rating and successive comparisons methods differ significantly?</td>
<td>Yes, beyond .05 level of significance.</td>
<td>Rating and ranking are considered to be easiest, while successive comparisons is considered most difficult.</td>
</tr>
<tr>
<td>3. Method Difficulty</td>
<td></td>
<td>Rating and ranking are considered to be easiest, while successive comparisons is considered most difficult.</td>
</tr>
<tr>
<td>(a) Which method is easiest to use?</td>
<td>(See Table 38)</td>
<td>Rating and ranking are considered to be easiest, while successive comparisons is considered most difficult.</td>
</tr>
<tr>
<td>(b) Which method is most difficult to use?</td>
<td>(See Table 38)</td>
<td>Rating and ranking are considered to be easiest, while successive comparisons is considered most difficult.</td>
</tr>
<tr>
<td>4. Perceived Method Applicability</td>
<td></td>
<td>Rating and ranking are considered to be most applicable to the subjects' decision-making needs.</td>
</tr>
<tr>
<td>(a) Which methods are considered to be most applicable to the subjects' decision-making needs.</td>
<td>(See Table 39)</td>
<td>Rating and ranking are considered to be most applicable to the subjects' decision-making needs.</td>
</tr>
</tbody>
</table>
of the study is limited by the number of subjects and the number and kind of criteria under consideration. Secondly, the study was conducted under simulated rather than an actual decision setting. Third, the measures of association upon which the convergent reliabilities are based were shown to be significantly different from zero. They were not shown, however, to differ significantly from each other. The magnitude of a reliability necessary for application has long been a matter of discussion. However, it should be noted that the convergent reliability may be considered to be a measure of not only reliability, but also validity, since the weights derived by each method were compared to a "best estimate" of the true value of the criterion weights.
V. CONCLUSIONS AND RECOMMENDATIONS

A. Summary of the Study

The purpose of the study is to provide an initial source of evidence for determining the applicability of five existing multiple-criterion weighting methods for educational decision-making. The five methods under consideration are ranking, rating, partial-paired comparisons, complete-paired comparisons and successive comparisons. The variables selected for describing the applicability of these methods include inter-method (convergent) reliability, elapsed-time for application, perceived difficulty and perceived method applicability.

The study is based upon four major questions:

1. What is the inter-method reliability associated with each of the five methods?

2. What is the average elapsed-time required to apply each of the five methods to seven and ten criteria of given difficulties?

3. Which methods are perceived by the subjects as being easiest and most difficult to use?

4. Which methods are considered to be most applicable to the decision-making needs of the subjects?
Furthermore, hypotheses were advanced in a number of directions. First, the reliabilities of the methods would all differ significantly from zero. Secondly, the ranking method would require the least average elapsed-time to apply to the criteria. Third, the successive comparisons method would require the greatest average elapsed-time to apply to the criteria. Finally, the differences in the average elapsed-times observed for the ranking and rating methods would be statistically significant.

A review of related research revealed that the methods have entertained little notoriety in education. However, in other disciplines, particularly operations research, psychology and management science, a relatively large amount of research has been conducted in the development, validation and application of these weighting methods. Thus, the literature substantiates the need for evidence which can be utilized for determining the applicability of the various weighting methods for educational decision-making.

The study was undertaken in a field setting under simulated conditions. The focus of the study is descriptive in nature. The study was replicated in four educational decision-making settings. The first setting consisted of five subjects randomly selected from the evaluation staff of a regional educational laboratory. The second setting involved five subjects randomly selected from the evaluation staff of a metropolitan school district. The third setting focused upon the five staff members responsible for coordinating exemplary program efforts in a state department of education. The fourth setting consisted of five project
directors randomly selected from the exemplary (ESEA Title III) projects currently operating in the State of Oregon.

Two sets of instruments were utilized. The first instrument, consisting of seven criteria for assessing the quality of evaluative information, was administered to the subjects in the first and second study settings. The second instrument, which focused upon ten criteria for assessing the quality of exemplary educational projects, was administered to the subjects in the third and fourth study settings.

The instruments consisted of brief definitions of the criteria to be weighted, and directions and recording procedures for applying each of the five weighting methods. At the conclusion of the instrument a brief questionnaire was attached which focused upon the perceived difficulty and applicability of the methods. The body of the instrument consisted of five sections. These five sections, representing the five weighting methods, were randomly ordered for each subject as a means of controlling for the effects of ordering and interaction between methods.

The analysis was conducted separately for each of the four study groups. The analysis, which was based upon the subjects' applications of the methods, utilized both parametric and nonparametric measures of association. The reliability estimates were based upon the relationship between the derived method weights and a composite of the weights derived from the balance of the methods.

The elapsed-time for method application utilized the number of minutes and seconds required by each subject to complete a given method for either seven or ten criteria. Comparisons of the elapsed-times
associated with the various methods were based upon the Friedman (two-way analysis) rank sum statistic.

The results of the analyses revealed that all reliability coefficients were significantly different from zero at the .05 level of probability, supporting the first hypothesis. It was found that although all reliability values observed could be considered acceptable, no substantial differences in the reliabilities were apparent between the methods. Therefore, it was concluded that the convergent reliability variable did not consistently discriminate between methods and further analysis was warranted in other directions. Since the reliabilities were all relatively high, any of these methods would produce essentially similar orderings of the criteria. Therefore, the basis for selecting a weighting method can go beyond considerations of reliability.

Major differences were observed in the elapsed-time required for applying the various weighting methods. Specifically, significant differences were observed in the elapsed-times required for ranking and successive comparisons methods in each study setting. Furthermore, in three of the four settings, significant differences were observed in the time intervals required for rating and successive comparisons methods. Thus, the second hypothesis was likewise confirmed.

When seven criteria were used, the ranking method required the least time to apply, while the successive comparisons method required the longest interval. When ten criteria were under consideration,
either ranking or rating required the least time, while the successive comparisons method required the most time to apply.

Ranking, followed by rating, was indicated as the easiest method to apply. The successive comparisons method was indicated as being the most difficult to apply by the majority of the subjects. Ranking and rating methods were both perceived as being particularly applicable to the decision-making needs of the subjects under investigation.

B. General Conclusions

As an initial source of evidence for determining the applicability of the five alternative multiple-criterion weighting methods for educational decision-making, the study suggests a number of general conclusions:

1. All methods have moderately high convergent reliabilities which are significantly different from zero.
2. The successive comparisons method requires by far the longest amount of time to apply.
3. The successive comparisons method is considered to be the most difficult method to apply.
4. The ranking and rating methods are generally considered to be the easiest to apply.
5. The ranking and rating methods are generally considered to be most applicable to the decision-making needs of the subjects.

When this evidence is combined with the qualifications of the study, the major outcomes of the investigation can be summarized as follows:

1. The inter-method (convergent) reliabilities all differ
significantly from zero, suggesting that essentially the same results will be produced by the alternative methods.

2. Based upon seven and ten criteria of a given difficulty, the successive comparisons method required significantly more time to apply than the ranking method and in three of the four instances, also required significantly more time to apply than the rating method.

3. Upon applying the five methods, the subjects indicated that the successive comparisons method was the most difficult to use. The majority of the subjects indicated that the ranking method was the easiest to apply, while slightly fewer subjects indicated that the rating method was the easiest to apply.

4. Upon applying the five methods, the majority of the subjects indicated that the ranking method was most applicable for their decision-making needs, while slightly fewer subjects indicated that the rating method was the most applicable method for their decision-making needs.

5. The study has not provided evidence which refutes the current use of ranking and rating as the major weighting techniques in education. In a parallel manner, the study has not provided evidence which could be utilized for advocating the adoption of the successive comparisons method for these criterion weighting purposes.

However, these conclusions are restrictive in nature. That is, they focus upon the maintenance of methods currently in use in education and
the rejection of proposed methods which appear to be unacceptable. Further conclusions can be drawn, on the basis of the study findings, which focus upon the potential adoption of alternative methods. The data suggest that the paired comparisons methods may be equally acceptable. The reliabilities derived for these methods were found to be particularly strong, while the time required for their applications were only slightly longer than the ranking and rating methods. The partial-paired comparisons method was considered to be appropriate by some of the subjects, while subjects varied in their perceptions regarding the difficulty of the complete-paired comparisons method. Thus, these methods could be considered as contenders when selecting weighting methods in education.

Such a consideration becomes even more tenable when considering the methods currently utilized. Eckenrode notes that "a decision maker presented with carefully weighted alternatives has much more information than one for whom the alternatives have only been ranked."^1 Thus, rating would be the only alternative when ordinal information is desired. The partial-paired comparisons method was found to require only slightly more time to apply than the rating method, and was perceived as being slightly less applicable than the rating method.

Since the inter-method reliabilities observed between the partial- and complete-paired comparisons methods were all quite high, and the complete-paired comparisons method required a relatively greater amount

---

of time to apply, it is the author's opinion that the partial-paired comparisons method would be the most viable alternative approach to adopt.

Although the data are far from conclusive, the results also appear to indicate a decrease in inter-method reliabilities as the number of criteria increases. This point is more an unsubstantiated trend than an empirically based conclusion. However, if this point is valid, it draws serious light on the utility of these methods when the number of alternatives to be weighted becomes large. A case in point is the successive comparisons method. The author has observed that the application of this method becomes extremely difficult when the number of criteria exceeds ten. When the number of criteria do become quite large, the relative utility of the rating and partial-paired comparisons method might be predicted. Since rating requires the decision maker to place the criteria along a scale, it would be expected that rating (as well as ranking) would become more difficult. On the otherhand, the partial-paired comparisons method requires the decision maker to record preferences in pairs of criteria. Thus, although the partial-paired comparisons method may require more time to apply than the ranking or rating method, the author would expect that the results derived from this method would be more satisfactory. This implication provides further justification for the adoption of the partial-paired comparisons method.

---

Conclusions can be drawn from the standard deviations observed for the weights generated by each method. The weights derived through the rating process consistently demonstrated the smallest standard deviation. The ranking method produced weights which had only slightly greater variation, while the paired and successive comparisons methods produced weights having the largest standard deviations. Such a finding may relate to the sensitivity with which the methods produce criterion weightings. Furthermore, the data suggest that as the number of criteria increases, the standard deviations of the weightings decrease. Thus, the methods may have differing sensitivities which are affected by the number of criteria being processed.

The results hold implications in another direction. Although far from conclusive, the problems encountered by the subjects in applying the various methods could be considered as a source of information. Two subjects in study group three and four subjects in study group four encountered difficulty in applying the successive comparisons method. However, the obverse point is of more general interest. Since subjects did not indicate that they experienced difficulty in using the ranking, rating and paired comparisons methods, regardless of whether seven or ten criteria were being considered, then these four methods would appear to require less training and/or expertise for their application. It is possible to suggest, therefore, that these methods might be applied to more broad-ranged situations where such activities as school- or community-based needs assessments are being accomplished by individuals who would require training in the use of more sophisticated
techniques. It is reiterated that such a conclusion would be drawn more from opinion than fact; but if such a conclusion is valid, then parsimony would be considered as a positive characteristic of these four methods.

Based upon the findings of the present study and the author's past experience with these methods, it is suggested that the educational decision maker take a number of positive steps in structuring the decision making process:

1. Specify the criteria to be utilized in selecting decision alternatives.
2. Decide whether simple rankings or interval weights of these criteria are necessary for the decision.
3. Apply one of the suggested methods to these criteria.
4. On the basis of the weighted criteria, apply one or more of the suggested methods to the decision alternatives.
5. Select the alternative which optimizes the criteria.

Such a process will formalize the decision and, if used appropriately, should provide a much more satisfactory outcome. The methods under investigation in the present study are suitable for such a structured process. It should be cautioned, however, that the study does not address the question of the utility of these methods when compared to decisions which result from processes where the methods are not applied.

The present study has also considered the elapsed-time for method application as being an important variable and suggests that a few minutes difference in time may be critical. However, if decision-
making is to be effectively formalized, the decision maker must spend some time in the process. The choice is essentially his; but it is anticipated that "hindsight remorse" and default decisions can be reduced by the application of the suggested methods and decision process.

C. Recommendations

Two forms of recommendations are in order. First, recommendations regarding the application of the findings of the present study are called for. Secondly, recommendations for additional research along parallel lines of the present investigation are considered essential.

1. Application

Since the purpose of the present study is exploratory in nature, caution should be exercised in applying the findings. The study utilized a limited number of subjects in four decision settings. Based upon the present findings, it appears that an application of the five methods will produce essentially the same results. Thus, if one can assume equal reliabilities, three characteristics become critical in selecting a method for application. These characteristics are the time required for applying the method, the difficulty of the method and the perceived applicability of the method. The successive comparisons method may receive resistance from the potential users because of the increased amount of time and difficulty associated with its application. Perceived applicability could be described as a form of face validity. If a method is not considered to be appropriate, then its use will be rejected on those grounds until evidence can be provided to the contrary.
The five methods appear to be appropriate for weighting objectives, criteria and decision alternatives. However, the methods were independently applied by the individual subjects. Neither inter-rater reliability, nor consensus were considered to be conditions under study. The values assigned to each criterion were averaged across subjects. Therefore, no evidence has been provided regarding the use of the methods in group decision-making settings. Furthermore, the ranking method produces only ordinal information, while the balance of the methods produce more information than a simple ordering of the criteria. Thus, although ranking was considered to be one of the most applicable methods, the rating method may provide the user with more information for decision-making; this would depend upon the decision in question. If an ordering of alternatives is all that is necessary, then rating will provide more information than is required.

Interpretability may also be considered as an important characteristic. The ranking and rating methods provide directly interpretable results, while the paired comparisons and successive comparisons methods require an intermediate step for transformation of the data into an interpretable form. This step requires additional time and increases the probability for computational errors. The occurrence of such errors must be considered when determining the reliability of a method under field conditions.

Unequivocal advocacy of a method is not suggested by the findings of the present study. Future research is suggested in which further
comparisons of the methods are made under a variety of settings and conditions. It is hoped that definitive answers may be provided at that time.

2. Research

The results of the present study suggest a need for additional research within the educational discipline along at least three lines of inquiry. First, additional research regarding the question of method reliability should be continued. Particularly, a comparative study of method reliabilities designed to determine the superiority of a method or methods would be desirable. The need for a standard for comparison must be reconciled. Preferences lack a basis for determining the "true reliability" or generalizability of the outcome. Without such a standard, the validity of such methods can not be directly determined. In the present instance, an estimate of the "true preferences" of the judges was used as the standard for comparison. An application of these methods in a setting where a true weight of the criteria could be independently established would provide substantial additional evidence.

Another consideration along the lines of reliability is the relative consequence associated with making a decision error on the basis of the information provided by a weighting method. Since neither the individual inter-method reliabilities, nor the method convergent reliabilities were perfect, such errors would appear to exist. Is there a method which results in the smallest number of decision errors for a given number and kind of criteria, and a given form of decision?
Furthermore, do the errors which are incurred result from the legitimate use or misuse of the weighting methodologies? It is doubtful that a response to these questions can be provided without a substantial amount of additional research.

However, the most useful information may be provided by research along other lines. For example, a second suggested focus of inquiry should be the comparison of the time required for application of the methods when the number and difficulty of the criteria are varied. Under the conditions of the present study, the paired comparisons and successive comparisons methods required slightly more time for application than the ranking and rating methods. Does this proportion of time change when the number and difficulty of the criteria are varied extensively? If so, then decisions regarding the applicability of the methods may also change. Thus, information is required for determining the applicability of the methods based upon the time required for their application, when the number and difficulty of the criteria differ from the present study. Following these lines, another question which may be asked is to what extent must the elapsed-times between methods differ in order to be of practical significance. How much time does a decision-maker allocate to a decision? If the time provided for decision-making is large, then differences which are measured in seconds could be considered inconsequential.

A final recommendation for additional research is the further testing of these methods under a variety of decision settings in education and the addition of other study variables, such as
interpretability, ordinal-interval decision needs, appropriateness of the methods for group decision-making and the error rates encountered by the application of the various methods. With the addition of such information, the utility of these methods can be determined and generalized across the settings of educational decision-making, and as a result, the decision-making processes in education can be improved.
APPENDIX A

THE STUDY INSTRUMENTS*

*Note: The ordering of the five method exercises within the instruments are random. The ordering of the exercises within the instruments contained in Appendix A do not necessarily represent those orderings used in the actual study instruments.
CRITERION WEIGHTING STUDY

Instructions

This study is designed to compare methods of weighting criteria for assessing the quality of evaluative information.

The state-of-the-art of educational evaluation has been a point of discussion by educators during recent years. A number of criteria have been proposed for assessing the quality of information generated by evaluation practices.*

Seven of these criteria have been listed below:

- **Internal Validity**: the information is a reliable and accurate representation of reality.
- **External Validity**: the information is generalizable to the population it purports to represent.
- **Objectivity**: the information is unbiased; replicable among independent observers.
- **Relevance**: the information is applicable to the decision to be served.
- **Credibility**: the source of information is believable.
- **Timeliness**: the information is provided at the time which it is needed for decision-making.
- **Efficiency**: the information is generated through a reasonable expenditure of time, money and resources.

Considering the characteristics which evaluative information should manifest, you are to assign relative weights to each of the seven criteria. Each of the following sections is designed to collect data for weighting these criteria by a different method. READ THE DIRECTIONS FOR EACH SECTION CAREFULLY. COMPLETE EACH SECTION BEFORE PROCEEDING TO THE NEXT SECTION. WORK QUICKLY, BUT AVOID MAKING MISTAKES. DO NOT REFER BACK TO PREVIOUS SECTIONS UPON THEIR COMPLETION.

* The criteria used in the present study are derived from a list developed by the Phi Delta Kappa National Study Committee on Evaluation.
RATING

Directions

The seven criteria are listed below. Rate these criteria by drawing a line from each criterion to a point on the adjacent scale which best represents the value of the criterion for assessing the quality of evaluative information. Criteria may be assigned similar values.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal Validity</td>
<td>10</td>
</tr>
<tr>
<td>External Validity</td>
<td>9</td>
</tr>
<tr>
<td>Objectivity</td>
<td>8</td>
</tr>
<tr>
<td>Relevance</td>
<td>7</td>
</tr>
<tr>
<td>Credibility</td>
<td>6</td>
</tr>
<tr>
<td>Timeliness</td>
<td>5</td>
</tr>
<tr>
<td>Efficiency</td>
<td>4</td>
</tr>
</tbody>
</table>

COMPLETE THIS SECTION BEFORE CONTINUING TO THE NEXT SECTION
SUCCESSIVE COMPARISONS

Directions

1. The seven criteria are listed below. Rank the relative value of the criteria in their order of importance by placing a number 7 in front of the criterion which you feel is most important for assessing the quality of evaluative information, 6 in front of the next most important criterion, etc. Continue until you have assigned the number 1 to the least important criterion.

___ Internal Validity
___ External Validity
___ Objectivity
___ Relevance
___ Credibility
___ Timeliness
___ Efficiency

2. The seven criteria have been randomly assigned to three groups which are listed below. One criterion has been randomly selected and assigned a numerical value of 1.00. This criterion, credibility, has been included in each of the three group listings.

Tentatively assign numerical values between 10.0 and 0.0 to each of the remaining two criteria in each of the groups. These numerical values should reflect initial estimates of their relative importance, where the larger the number the greater the relative value.

(a) (b) (c)
___ Efficiency ___ Relevance ___ Internal Validity
___ External Validity ___ Objectivity ___ Timeliness
1.00 Credibility 1.00 Credibility 1.00 Credibility

CONTINUE TO NEXT PAGE
For each group, apply the following procedures:

3. Based upon the numerical values which you have assigned to each criteria, let O₁ represent the most valued criterion, O₂ the next most valued criterion and O₃ the least valued criterion in a group. Compare the most valued criterion (O₁) to the combination of the two remaining criteria (O₂ and O₃) and ask the following question:

Which has more value, the one most important criterion or the sum of the two less important criteria?

- If you feel that the one criterion is of more value, then the numerical value of this criterion should be greater than the sum of the numerical values of the remaining two.

- If you feel that the one criterion is of equal value to the remaining two, then the numerical value of this criterion should be equal to the sum of the numerical values of the remaining two.

- If you feel that the one criterion is of less value than the combination of the remaining two, then the numerical value of this criterion should be less than the sum of the numerical values of the remaining two.

Adjust the relative values of the criteria, except credibility which must retain a value of 1.00, in accordance with your decision above. Record your adjustments below. Repeat this procedure for each group.

(a) Efficiency (b) Relevance (c) Internal Validity
(b) External Validity (c) Objectivity (c) Timeliness
1.00 Credibility 1.00 Credibility 1.00 Credibility

CONTINUE TO NEXT PAGE
4. List below the ranks obtained for each criterion in Step 1 and the computed values derived in Step 3. Rank the computed values by their absolute magnitude, where 7 represents the greatest numerical value and 1 the smallest numerical value.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Rank</th>
<th>Computed Value</th>
<th>Rank of Computed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal Validity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>External Validity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objectivity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Credibility</td>
<td></td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Timeliness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficiency</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If the above rank orders differ, reconsider the ranking computed in Step 1. If these original ranks are considered to be correct, then proceed again through Steps 2 and 3 of the procedure until consistent results are obtained. If the computed rankings are considered to be correct or the rank orders do not differ, then proceed to Step 5.

CONTINUE TO NEXT PAGE
5. Normalize the computed values by applying the following method:

Derive the sum of the seven computed values. Divide each of the computed values by this total and list these values under the "normalized value" column below.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Computed Value</th>
<th>Normalized Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal Validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>External Validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objectivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Credibility</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Timeliness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficiency</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>1.0</strong></td>
</tr>
</tbody>
</table>

COMPLETE THIS SECTION BEFORE CONTINUING TO THE NEXT SECTION
COMPLETE-PAIRED COMPARISONS

Directions

Each criterion has been matched with each of the other six criterion. For each pair, circle the criterion which is of more importance in assessing the quality of evaluative information. Complete each pair.

- Internal Validity - Relevance
- Internal Validity - Efficiency
- Relevance - Credibility
- Timeliness - Internal Validity
- Efficiency - Timeliness
- External Validity - Objectivity
- Objectivity - External Validity
- Relevance - Timeliness
- Efficiency - Internal Validity
- Internal Validity - Objectivity
- External Validity - Credibility
- Timeliness - External Validity
- Relevance - Objectivity
- Efficiency - Objectivity
- Objectivity - Internal Validity
- Relevance - Efficiency
- Credibility - Timeliness
- Efficiency - Relevance
- Credibility - Objectivity
- Internal Validity - External Validity
- Efficiency - Objectivity
- Objectivity - Internal Validity
- Relevance - Objectivity
- Relevance - Efficiency
- Efficiency - Objectivity
- Internal Validity - External Validity
- External Validity - Timeliness
- External Validity - Efficiency
- Timeliness - Credibility
- Credibility - Relevance
- Internal Validity - Timeliness
- Credibility - External Validity
- Timeliness - Efficiency
- Relevance - Internal Validity
- Efficiency - Credibility
- Objectivity - Timeliness
- Efficiency - External Validity
- Objectivity - Credibility
- Internal Validity - Credibility

COMPLETE THIS SECTION BEFORE CONTINUING TO THE NEXT SECTION
PARTIAL-PAIRED COMPARISONS

Directions

Each criterion has been matched with each of the other six criteria. For each box, place the number of the criterion in the pair which is of more importance for assessing the quality of evaluative information.

<table>
<thead>
<tr>
<th>INTERNAL VALIDITY 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXTERNAL VALIDITY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OBJECTIVITY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RELEVANCE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CREDIBILITY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIMELINESS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFFICIENCY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

COMPLETE THIS SECTION BEFORE CONTINUING TO NEXT SECTION
RANKING

Directions

The seven criteria are listed below. Rank the relative value of these criteria by placing a number 1 in front of the criterion which you feel is most important for assessing the quality of evaluative information, 2 in front of the next most important criterion, etc. Continue until you have assigned the number 7 to the least important criterion.

___ Internal Validity
___ External Validity
___ Objectivity
___ Relevance
___ Credibility
___ Timeliness
___ Efficiency

COMPLETE THIS SECTION BEFORE CONTINUING TO THE NEXT SECTION
CRITERION WEIGHTING STUDY

Method Evaluation

You have just participated in a study which utilized five methods for weighting criteria. Please respond to the following questions regarding these methods.

1. Which method was easiest to use? (Circle one)
   a. ranking
   b. rating
   c. partial-paired comparisons
   d. complete-paired comparisons
   e. successive comparisons

2. Which method was most difficult to use? (Circle one)
   a. ranking
   b. rating
   c. partial-paired comparisons
   d. complete-paired comparisons
   e. successive comparisons

3. Which method do you feel is most applicable to your decision-making needs? (Circle one)
   a. ranking
   b. rating
   c. partial-paired comparisons
   d. complete-paired comparisons
   e. successive comparisons

THANK YOU FOR YOUR TIME AND COOPERATION
CRITERION WEIGHTING STUDY

Instructions

This study is designed to compare methods of weighting criteria for refunding the continuation of ESEA Title III exemplary projects in Oregon. Listed below are ten criteria which have been proposed as being relevant to such a refunding decision:

**TRANSPORTABILITY:** The extent to which the project's products and/or processes can be physically transported to adopter sites.

**INNOVATIVENESS:** The degree to which the project's processes resolve a problem through a unique approach and/or improve an existing practice within the target area.

**GENERALIZABILITY:** The magnitude of an existing or predicted target population for which the project has been designed.

**COST:** The amount of resources required for developing, installing and maintaining the project.

**SIGNIFICANCE:** The degree to which the project focuses upon an important problem.

**EVIDENCE:** The quality of the information used to document the outcomes of the project.

**CLARITY:** The extent to which the project processes have been explicitly defined.

**MAINTENANCE:** The probability that the project processes can be continued through local funding sources.

**EFFECTIVENESS:** The extent to which the project satisfactorily impacts upon the defined problem area.

**COMPATIBILITY:** The extent to which the project's goals, processes and outcomes are congruent with the existing values and attitudes of the potential adopter population.

Considering the characteristics which exemplary projects should manifest, you are to assign relative weights to each of the ten criteria. Each of the following sections is designed to collect data for weighting these criteria by a different method. READ THE DIRECTIONS FOR EACH SECTION CAREFULLY. COMPLETE EACH SECTION BEFORE PROCEEDING TO THE NEXT SECTION. WORK QUICKLY, BUT AVOID MAKING MISTAKES. DO NOT REFER BACK TO PREVIOUS SECTIONS UPON THEIR COMPLETION.
RATING

Directions

The ten criteria are listed below. Rate these criteria by drawing a line from each criterion to a point on the adjacent scale which best represents the importance of the criterion for determining the refunding of exemplary projects. Criteria may be assigned similar values.

TRANSPORTABILITY • -10 (Maximum Importance)
INNOVATIVENESS • -9
GENERALIZABILITY • -8
COST • -7
SIGNIFICANCE • -6
EVIDENCE • -5 (Moderate Importance)
CLARITY • -4
MAINTENANCE • -3
EFFECTIVENESS • -2
COMPATIBILITY • -1
0 (No Importance)

COMPLETE THIS SECTION BEFORE CONTINUING TO THE NEXT SECTION
RANKING

Directions

The ten criteria are listed below. Rank the relative importance of each of these criteria by placing a number 1 in front of the criterion which you feel should be the most important consideration for project refunding, a 2 in front of the next most important criterion, etc. Continue until you have assigned the number 10 to the least important criterion.

____ TRANSPORTABILITY
____ INNOVATIVENESS
____ GENERALIZABILITY
____ COST
____ SIGNIFICANCE
____ EVIDENCE
____ CLARITY
____ MAINTENANCE
____ EFFECTIVENESS
____ COMPATIBILITY

COMPLETE THIS SECTION BEFORE CONTINUING TO THE NEXT SECTION
SUCCESSIVE COMPARISONS

Directions

1. The ten criteria are listed below. Rank the relative value of the criteria in their order of importance by placing a number \( 10 \) in front of the criterion which you feel is most important for determining the refunding of exemplary projects, \( 9 \) in front of the next most important criterion, etc. Continue until you have assigned the number \( 1 \) to the least important criterion.

- Transportability
- Innovativeness
- Generalizability
- Cost
- Significance
- Evidence
- Clarity
- Maintenance
- Effectiveness
- Compatibility

2. The ten criteria have been randomly assigned to three groups which are listed below. One criterion has been randomly selected and assigned a numerical value of \( 1.00 \). This criterion, evidence, has been included in each of the three group listings.

Tentatively assign numerical values between \( 10.0 \) and \( 0.0 \) to each of the remaining three criteria in each of the three groups. These numerical values should reflect initial estimates of their relative importance, where the larger the number the greater the relative importance.

<table>
<thead>
<tr>
<th></th>
<th>(a) Generalizability</th>
<th>(b) Transportability</th>
<th>(c) Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>Clarity</td>
<td>Cost</td>
<td>Compatibility</td>
</tr>
<tr>
<td>(b)</td>
<td>Effectiveness</td>
<td>Innovativeness</td>
<td>Maintenance</td>
</tr>
<tr>
<td>(c)</td>
<td>Evidence</td>
<td>Evidence</td>
<td>Evidence</td>
</tr>
</tbody>
</table>

CONTINUE TO NEXT PAGE
For each of the three groups (a, b and c), apply the following procedures:

3. Based upon the numerical values which you have assigned to each criterion, let $O_1$ represent the most important criterion, $O_2$ the next most important criterion, $O_3$ the next, and $O_4$ the least important criterion in the group. Compare the most valued criterion ($O_1$) to the combination of the three remaining criteria ($O_2$ and $O_3$ and $O_4$) and ask the following question:

- Which is of more importance, the one most important criterion or the sum of the three less important criteria?

  - If you feel that the one criterion is of more importance, then the numerical value of this criterion should be greater than the sum of the numerical values of the remaining three.
  
  - If you feel that the one criterion is of equal value to the remaining three, then the numerical value of this criterion should be equal to the sum of the numerical values of the remaining three.

  - If you feel that the one criterion is of less importance than the combination of the remaining three, then the numerical value of this criterion should be less than the sum of the numerical values of the remaining three.

Adjust the value of $O_1$ (your most important criterion in the group), in accordance with your decision above. Should $O_1$ represent "evidence", adjust the values of the balance of the criteria. Repeat this procedure for each group and record your adjustments below.

(a) Generalizability
(b) Transportability
(c) Significance

Clarity
Cost
Compatibility

Effectiveness
Innovativeness
Maintenance

1.0 Evidence
1.0 Evidence
1.0 Evidence

CONTINUE TO NEXT PAGE
4. Now, compare $O_2$ (your second most important criterion in each group) with the combined importance of the two remaining criteria, $O_3$ and $O_4$.

Again, ask the following question:

Which is of more importance, the one more important criterion or the sum of the two less important criteria?

- If you feel that the one criterion is of more importance, then the numerical value of this criterion should be greater than the sum of the numerical values of the remaining two.

- If you feel that the one criterion is of an importance equal to the remaining two, then the numerical value of this criterion should be equal to the sum of the numerical values of the remaining two.

- If you feel that the one criterion is of less importance than the combination of the remaining two, then the numerical value of this criterion should be less than the sum of the numerical values of the remaining two.

Adjust the relative values of the criteria, except "evidence" which must retain a value of 1.0, in accordance with your decision above. Repeat this procedure for each group and record your adjustments below.

(a)  (b)  (c)

| Generalizability | Transportability | Significance |
| Clarity         | Cost             | Compatibility |
| Effectiveness   | Innovativeness   | Maintenance   |
| 1.0 Evidence    | 1.0 Evidence     | 1.0 Evidence  |

CONTINUE TO NEXT PAGE
5. List below the ranks obtained for each criterion in Step 1 and the computed values derived in Step 4. Rank the computed values by their absolute magnitude, where 10 represents the greatest numerical value and 1 the smallest numerical value.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Initial Rank</th>
<th>Computed Value</th>
<th>Rank of Computed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transportability</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
<tr>
<td>Innovativeness</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
<tr>
<td>Generalizability</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
<tr>
<td>Cost</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
<tr>
<td>Significance</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
<tr>
<td>Evidence</td>
<td>______ 1.0</td>
<td>______</td>
<td>______</td>
</tr>
<tr>
<td>Clarity</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
<tr>
<td>Maintenance</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
<tr>
<td>Effectiveness</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
<tr>
<td>Compatibility</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
</tbody>
</table>

If the above rank orders differ, reconsider the ranking initially computed in Step 1. If these initial ranks are considered to be correct, then proceed again through Steps 2 through 4 of the procedure until consistent results are obtained. If the computed rankings are considered to be correct or the rank orders do not differ, then proceed to Step 6.
6. Normalize the computed values by applying the following method:

Derive the sum of the ten computed values. Divide each of the computed values by this total and list these values under the "normalized value" column below.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Computed Value</th>
<th>Normalized Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transportability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Innovativeness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generalizability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Significance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evidence</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Clarity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintenance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effectiveness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compatibility</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td>1.0</td>
</tr>
</tbody>
</table>

COMPLETE THIS SECTION BEFORE CONTINUING TO NEXT SECTION
**Directions**

Each criterion has been matched with each of the other nine criterion. For each pair, circle the criterion which is of more importance in determining project refunding. Complete each pair.

<table>
<thead>
<tr>
<th>Significance - Effectiveness</th>
<th>Compatibility - Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalizability - Cost</td>
<td>Generalizability - Clarity</td>
</tr>
<tr>
<td>Effectiveness - Cost</td>
<td>Compatibility - Generalizability</td>
</tr>
<tr>
<td>Transportability - Innovativeness</td>
<td>Generalizability - Transportability</td>
</tr>
<tr>
<td>Effectiveness - Maintenance</td>
<td>Significance - Compatibility</td>
</tr>
<tr>
<td>Cost - Significance</td>
<td>Transportability - Effectiveness</td>
</tr>
<tr>
<td>Maintenance - Effectiveness</td>
<td>Innovativeness - Significance</td>
</tr>
<tr>
<td>Maintenance - Cost</td>
<td>Transportability - Compatibility</td>
</tr>
<tr>
<td>Compatibility - Evidence</td>
<td>Generalizability - Innovativeness</td>
</tr>
<tr>
<td>Compatibility - Significance</td>
<td>Clarity - Generalizability</td>
</tr>
<tr>
<td>Transportability - Evidence</td>
<td>Maintenance - Significance</td>
</tr>
<tr>
<td>Clarity - Cost</td>
<td>Evidence - Innovativeness</td>
</tr>
<tr>
<td>Transportability - Clarity</td>
<td>Effectiveness - Compatibility</td>
</tr>
<tr>
<td>Innovativeness - Compatibility</td>
<td>Generalizability - Evidence</td>
</tr>
<tr>
<td>Innovativeness - Clarity</td>
<td>Transportability - Cost</td>
</tr>
<tr>
<td>Innovativeness - Maintenance</td>
<td>Transportability - Maintenance</td>
</tr>
<tr>
<td>Clarity - Significance</td>
<td>Cost - Compatibility</td>
</tr>
<tr>
<td>Compatibility - Maintenance</td>
<td>Maintenance - Generalizability</td>
</tr>
<tr>
<td>Evidence - Significance</td>
<td>Effectiveness - Significance</td>
</tr>
<tr>
<td>Evidence - Maintenance</td>
<td>Maintenance - Innovativeness</td>
</tr>
<tr>
<td>Compatibility — Innovativeness</td>
<td>Clarity — Effectiveness</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Maintenance — Clarity</td>
<td>Evidence — Generalizability</td>
</tr>
<tr>
<td>Innovativeness — Evidence</td>
<td>Compatibility — Transportability</td>
</tr>
<tr>
<td>Compatibility — Clarity</td>
<td>Significance — Innovativeness</td>
</tr>
<tr>
<td>Significance — Transportability</td>
<td>Clarity — Evidence</td>
</tr>
<tr>
<td>Effectiveness — Evidence</td>
<td>Maintenance — Evidence</td>
</tr>
<tr>
<td>Evidence — Clarity</td>
<td>Cost — Transportability</td>
</tr>
<tr>
<td>Innovativeness — Cost</td>
<td>Clarity — Innovativeness</td>
</tr>
<tr>
<td>Significance — Generalizability</td>
<td>Evidence — Compatibility</td>
</tr>
<tr>
<td>Cost — Clarity</td>
<td>Effectiveness — Transportability</td>
</tr>
<tr>
<td>Generalizability — Significance</td>
<td>Significance — Clarity</td>
</tr>
<tr>
<td>Compatibility — Cost</td>
<td>Effectiveness — Innovativeness</td>
</tr>
<tr>
<td>Cost — Maintenance</td>
<td>Clarity — Maintenance</td>
</tr>
<tr>
<td>Innovativeness — Transportability</td>
<td>Maintenance — Transportability</td>
</tr>
<tr>
<td>Evidence — Transportability</td>
<td>Cost — Innovativeness</td>
</tr>
<tr>
<td>Transportability — Generalizability</td>
<td>Cost — Evidence</td>
</tr>
<tr>
<td>Evidence — Effectiveness</td>
<td>Effectiveness — Generalizability</td>
</tr>
<tr>
<td>Generalizability — Compatibility</td>
<td>Generalizability — Effectiveness</td>
</tr>
<tr>
<td>Clarity — Transportability</td>
<td>Maintenance — Compatibility</td>
</tr>
<tr>
<td>Effectiveness — Clarity</td>
<td>Cost — Effectiveness</td>
</tr>
<tr>
<td>Significance — Evidence</td>
<td>Significance — Maintenance</td>
</tr>
<tr>
<td>Innovativeness — Effectiveness</td>
<td>Significance — Cost</td>
</tr>
<tr>
<td>Innovativeness — Generalizability</td>
<td>Compatibility — Generalizability</td>
</tr>
<tr>
<td>Transportability — Significance</td>
<td>Evidence — Cost</td>
</tr>
<tr>
<td>Clarity — Compatibility</td>
<td>Generalizability — Maintenance</td>
</tr>
</tbody>
</table>
PARTIAL-PAIRED COMPARISONS

Directions

Each criterion has been matched with each of the other nine criteria. For each box, place the number of the criterion in the pair which is of more importance in determining the refunding of exemplary projects.

<table>
<thead>
<tr>
<th>TRANSPORTABILITY (1)</th>
<th>INNOVATIVENESS (2)</th>
<th>GENERALIZABILITY (3)</th>
<th>COST (4)</th>
<th>SIGNIFICANCE (5)</th>
<th>EVIDENCE (6)</th>
<th>CLARITY (7)</th>
<th>MAINTENANCE (8)</th>
<th>EFFECTIVENESS (9)</th>
<th>COMPATIBILITY (10)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
<td>(8)</td>
<td>(9)</td>
<td>(10)</td>
</tr>
</tbody>
</table>

COMPLETE THIS SECTION BEFORE CONTINUING TO THE NEXT SECTION
CRITERION WEIGHTING STUDY

Method Evaluation

You have just participated in a study which utilized five methods for weighting criteria. Please respond to the following questions regarding these methods.

1. Which method was easiest to use? (Circle one)
   a. ranking
   b. rating
   c. partial-paired comparisons
   d. complete-paired comparisons
   e. successive comparisons

2. Which method was most difficult to use? (Circle one)
   a. ranking
   b. rating
   c. partial-paired comparisons
   d. complete-paired comparisons
   e. successive comparisons

3. Which method do you feel is most applicable to your decision-making needs? (Circle one)
   a. ranking
   b. rating
   c. partial-paired comparisons
   d. complete-paired comparisons
   e. successive comparisons

THANK YOU FOR YOUR TIME AND COOPERATION


Brisken, Lawrence E. "A Method of Unifying Multiple Objective Functions," Management Science, XII no. 10 (June 1966), B406-416.


Macguire, Thomas O. "Value Components of Teachers' Judgements of Educational Objectives," \textit{A-V Communication Review}, I (1968), 63-86.


Miller, George A. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capability for Processing Information," \textit{The Psychological Review}, LXIII no. 2 (March 1956), 81-97.


---


---


---


