INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations 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 through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

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

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.
Kerkar, Shanta Purushottam

EFFECT OF INFORMATION DISPLAY FORMAT ON JUDGMENT

Rice University

Ph.D. 1984

University Microfilms International

300 N. Zeeb Road, Ann Arbor, MI 48106
RICE UNIVERSITY

EFFECT OF INFORMATION DISPLAY FORMAT ON JUDGMENT

by

SHANTA P. KERKAR

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

DOCTOR OF PHILOSOPHY

APPROVED, THESIS COMMITTEE:

William C. Howell
Professor of Psychology
Chairman

Michael J. Watkins
Professor of Psychology

David M. Lane
Associate Professor of Psychology

David W. Brady
Professor of Political Science

HOUSTON, TEXAS

APRIL 1984
EFFECT OF INFORMATION DISPLAY FORMAT ON JUDGMENT

SHANTA P. KERKAR

ABSTRACT

The multiple regression model has been applied to the study of judgment primarily through the paradigms known as policy capturing and multiple cue probability learning (MCPL). The former involves modeling the way the decision maker weights predictive information; the latter focuses on the process by which the decision maker acquires this strategy. In the first part of the paper, the logic underlying the two approaches is examined, and the research generated by each is discussed. Based on this comparative review, it is argued that neither approach has yet lived up to its potential because each has concentrated too narrowly on particular kinds of issues, at least some of which present serious logical difficulties. Policy capturing, for example, has sought to isolate judgment processes through multiple regression despite the fact that it is logically capable of capturing only predictions. For its part, MCPL research has fallen short by limiting itself to variables suggested by Brunswik's lens model. Better use could be made of the model if it were applied in a more functional manner—one in which it is used to index performance rather than to infer cognitive processes. The functional approach serves as a guiding philosophy for a series of experiments that are described in the second part of the paper. The primary question of interest in these experiments was how the format of displaying visual information affects subsequent judgments and decisions. Two types of displays,
numerical and graphical, were investigated using the policy capturing paradigm in the first three experiments and MCPL in the fourth one. The most consistent finding was that subjects' cue weighting differed reliably with type of format. Numerical policies tended to be less precise than graphical ones, but accuracy of predictions did not differ with display. The implications of these findings are discussed both from a practical and theoretical perspective.
ACKNOWLEDGMENTS

This research was supported by Engineering Psychology Programs, Office of Naval Research, Contract N00014-82-C-0001. I would like to thank Drs. William C. Howell, Michael J. Watkins, David M. Lane, and David W. Brady for their co-operation in serving as members of my committee. I am especially grateful to both Bill Howell and David Lane for contributing to my intellectual growth during my graduate years at Rice. I also appreciate my great fortune in having had the opportunity to work with Mike Watkins, a scientist of the highest caliber, to be judged only by his own standards. Finally, I recognize that this research would have not been completed without the loving "nonacademic" support provided by my family and friends.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>48</td>
</tr>
<tr>
<td>Method</td>
<td>49</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>53</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>65</td>
</tr>
<tr>
<td>Method</td>
<td>66</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>69</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>74</td>
</tr>
<tr>
<td>Method</td>
<td>76</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>78</td>
</tr>
<tr>
<td>Experiment 4</td>
<td>89</td>
</tr>
<tr>
<td>Method</td>
<td>92</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>97</td>
</tr>
<tr>
<td>General Discussion</td>
<td>122</td>
</tr>
<tr>
<td>Footnotes</td>
<td>129</td>
</tr>
<tr>
<td>Reference Notes</td>
<td>130</td>
</tr>
<tr>
<td>References</td>
<td>131</td>
</tr>
</tbody>
</table>
# LIST OF TABLES AND FIGURES

<table>
<thead>
<tr>
<th>Figure/Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>6</td>
</tr>
<tr>
<td>Table 1</td>
<td>8</td>
</tr>
<tr>
<td>Table 2</td>
<td>35</td>
</tr>
<tr>
<td>Figure 2</td>
<td>51</td>
</tr>
<tr>
<td>Table 3</td>
<td>54</td>
</tr>
<tr>
<td>Table 4</td>
<td>56</td>
</tr>
<tr>
<td>Table 5</td>
<td>64</td>
</tr>
<tr>
<td>Table 6</td>
<td>68</td>
</tr>
<tr>
<td>Table 7</td>
<td>70</td>
</tr>
<tr>
<td>Table 8</td>
<td>73</td>
</tr>
<tr>
<td>Table 9</td>
<td>80</td>
</tr>
<tr>
<td>Table 10</td>
<td>83</td>
</tr>
<tr>
<td>Table 11</td>
<td>95</td>
</tr>
<tr>
<td>Table 12</td>
<td>98</td>
</tr>
<tr>
<td>Table 13</td>
<td>100</td>
</tr>
<tr>
<td>Table 14</td>
<td>104</td>
</tr>
<tr>
<td>Table 15</td>
<td>105</td>
</tr>
<tr>
<td>Figure 3</td>
<td>107</td>
</tr>
<tr>
<td>Figure 4</td>
<td>108</td>
</tr>
<tr>
<td>Figure 5</td>
<td>110</td>
</tr>
<tr>
<td>Figure 6</td>
<td>112</td>
</tr>
<tr>
<td>Figure 7</td>
<td>114</td>
</tr>
<tr>
<td>Figure 8</td>
<td>115</td>
</tr>
<tr>
<td>Table 16</td>
<td>118</td>
</tr>
</tbody>
</table>
INTRODUCTION

Making a judgment or decision on the basis of multiple pieces of information is a fundamental cognitive activity. Given the technological sophistication of today's world, people are confronted with an ever increasing array of information that they must integrate into one judgment or choice. For example, a physician diagnoses a patient's illness using data such as X-ray reports, knowledge of the symptoms, a medical history, and a physical examination. Similarly, a college admissions committee selects applicants on the basis of various aptitude test scores, prior performance records, recommendations, and interviews. In all such cases the effectiveness of the judgment is dependent on how the decision maker interprets, integrates and differentially weights the intertwined bits of available information. Researchers have devoted considerable attention to the description of how people use the information at hand, as well as to the "normative" question of how they should handle it. Both emphases have relied heavily upon multiple regression techniques.

The first section of this paper evaluates critically the use of the multiple regression paradigm in judgment research and suggests ways in which its application to judgment and decision problems could be enhanced. The argument developed is that the multiple regression approach has not advanced our understanding of human judgment as much as was originally hoped because of an overcommitment to the search for illusive cognitive processes. An alternative functional approach is proposed in which task-judgment relations are explored within the
multiple regression paradigm. Subsequent sections describe a series of experiments investigating the effect of one particular task feature—the manner in which key information is displayed—on judgment using the advocated approach.

REGRESSION APPROACH IN JUDGMENT

The statistical model of regression has, of course, served psychology well for many years, notably as a tool for expressing the relationships between a set of predictors and some (usually behavioral) criterion. Its use in the context of human judgment research dates back at least 30 years, although the ideas developed initially by Brunswik and elaborated upon by Hammond and others (see, e.g., Hammond 1955) had a relatively limited impact upon the field until fairly recently (the last dozen or so years). The way in which the multiple regression model has typically been adapted to this sort of problem can best be illustrated by means of an example.

A human subject (the "judge" or decision maker) is required to evaluate each of a large set of multidimensional stimuli on some global scale (e.g., overall goodness, importance, worth). Usually each stimulus dimension ("cue") is quantified, thereby affording the judge a "decomposed" or profile definition of each stimulus. After a number of such judgments have been rendered, multiple regression analysis is used to unravel how information about the stimuli (viz., the presented cues) is used by the judge in producing an overall evaluation (Slovic & Lichtenstein, 1971). For instance, an admissions officer might be asked to predict success in graduate school for each of a set of
prospective students based on test scores and undergraduate grades. The extent to which his or her judgments are influenced by each of these cues is then estimated by regressing the predictions on the values of the cues, and observing the regression weight for each one. The more importance he or she tends to attach to a cue such as GRE test scores, the heavier should be the obtained weight for that cue.

This general paradigm has been adapted for use in judgment research in two complementary ways: the first is usually referred to as "policy capturing" and the second, as "multiple cue probability learning" (MCPL) (Slovic & Lichtenstein, 1971). In policy capturing, the focus is on the individual's method of combining and weighting information. As illustrated above, multiple regression is typically used to model or "capture" empirically the judgment "policy", although other mathematical equations have sometimes been used (Hoffman, 1960). The assumption is that the individual combines information in a fashion analogous to the mathematical model, and his or her weighting strategy is reflected in the specific parameter values obtained when his or her judgments are incorporated into the model. In the previous example, the admissions officer is assumed to combine predictive information in a linear fashion, and the regression weights are taken to indicate his or her subjective weighting of the various cues.

By contrast, MCPL focuses on the individual's acquisition of cue-criterion relationships that actually exist in the environment rather than on the idiosyncratic policy for weighting cues. Thus, while a policy equation is again derived from a series of judgments,
the emphasis is on how this policy and the judgments comprising it compare to those characterizing the environment. The MCPL approach, therefore, permits normative as well as descriptive use of multiple regression in analyzing human judgment. The logic on which MCPL is based was developed originally by Brunswik (1952, 1955) with the intent of analyzing behavior and judgment under uncertainty in ecologically representative (realistic) settings. Through his now-famous "lens model" (see below), he hoped to explain how the organism deals with uncertainty actually present in the environment (Slovic & Lichtenstein, 1971).

Returning to the graduate admissions example, the researcher using the MCPL paradigm not only describes how the admissions officer weights test scores and grades in judging the success of applicants (his or her policy); he determines the degree of correspondence between that policy and the empirical relationship of the cues to success criteria. This correspondence is taken as an index of how successfully the individual has learned the characteristics of the environment.

In most MCPL situations, the researcher does not rely on a subject's prior knowledge about environmental interrelations; typically, the subject makes a judgment about the criterion based on cues presented to him, following which he receives feedback about the actual criterion. The provision of explicit feedback enables the researcher to present cue-criterion values that conform to an underlying environmental model that includes specified cue-criterion functions and cue interrelations. Thus he can be sure that all
subjects are exposed to the same model of the environment. Also, since all parameters of the environmental model are known, it is easier to use the multiple regression model in analyzing the subject's performance.

**Lens Model.** The lens model was formulated by Brunswik (1952, 1955) in an attempt to separate the environmental and organismic components of the judgment situation and to specify their interrelationship. A diagram of the model is presented in Figure 1. The judge's task is to predict the criterion, Ye, on the basis of information cues $X_1, X_2, \ldots X_i$ about a stimulus. The left portion of the diagram represents the environment, and indicates the empirical relationship between each cue, $X_i$, and $Y_e$ (a function termed the "ecological validity" of $X_i$). The portion on the right side represents the subject whose judgment of the criterion is denoted by $Y_s$. The extent to which this judgment is influenced by each $X_i$ is also determined empirically, a function termed "cue utilization" of $X_i$.

Since any environment in which judgment is required would involve some ambiguity or uncertainty, we can assume that the cues do not predict the criterion perfectly. Hence the degree of environmental predictability can be expressed in terms of the following regression equation,

$$
\hat{Y}_e = \sum b_{ie} X_i
$$

Equation 1

The multiple correlation coefficient $R_e = r_{Y_e \hat{Y}_e}$ indicates the extent to which $Y_e$ can be predicted from a weighted combination of the cues.
Figure 1. Brunswik's lens model.
In an analogous manner, the subject's responses can be predicted by the following policy equation,

\[ \hat{Y}_s = \sum b_{is} X_i \] \hspace{1cm} \text{Equation 2}

The multiple correlation coefficient \( R_s = r_{Ys} \hat{Y}_s \) indicates the degree to which the subject's judgments can be captured by a linear combination of the cues. It might be noted parenthetically that the strategy of using multiple regression in the above equations minimizes the sum of squared deviations between \( Y_e \) and \( \hat{Y}_e \), and between \( Y_s \) and \( \hat{Y}_s \) respectively, and represents just one of the possible ways of determining environmental predictability and the judge's policy.

Within the lens model framework, the two principal measures used to express the quality of the judge's performance are known as the achievement index and the matching index. The former represents the correlation between the criterion and the subject's response or:

\[ r_a = r_{YeYs} \] \hspace{1cm} \text{Equation 3}

The matching index represents the correlation between the criterion values predicted from the environmental model and the subject's policy or:

\[ G = r_{YeYs} \] \hspace{1cm} \text{Equation 4}

A summary of these and several other common measures derived from the lens model is provided in Table 1.

RESEARCH ISSUES IN POLICY CAPTURING AND MCPL

Policy Capturing

Two focal issues in policy capturing research concern (1) use of the policy capturing methodology to reflect or describe the judgment
TABLE 1
Summary of correlations obtained from the lens model

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achievement ($Y_a$)</td>
<td>$r$ between criterion $Y_e$ and subject's responses $Y_s$</td>
</tr>
<tr>
<td>Matching coefficient ($G$)</td>
<td>$r$ between the best prediction of the subject's responses $\hat{Y}_s$ and best prediction of environment $Y_e$.</td>
</tr>
<tr>
<td>Ecological Validity</td>
<td>$r$ between criterion $Y_e$ and cue $X_i$.</td>
</tr>
<tr>
<td>Cue utilization</td>
<td>$r$ between cue $X_i$ and subject's responses $Y_s$.</td>
</tr>
<tr>
<td>Optimality coefficient</td>
<td>$r$ between subject's responses $Y_s$ and the optimal prediction of the environment $Y_e$.</td>
</tr>
<tr>
<td>Subject's multiple correlation ($R_s$)</td>
<td>$r$ between best prediction of subject's responses $\hat{Y}_s$ and actual responses $Y_s$.</td>
</tr>
<tr>
<td>Environmental multiple correlation ($R_e$)</td>
<td>$r$ between the best prediction of criterion $Y_e$ and the criterion $Y_e$.</td>
</tr>
</tbody>
</table>
"process", and (2) assessment of the accuracy with which the judge's policy can predict the criterion. Associated with the latter issue is the whole question of whether actual decision making (outcomes) can be improved by substituting the judge's policy for the judge himself. Needless to say, both these issues could be addressed in the same study (see, e.g., Dawes, 1971), but in a majority of the work to date, the focus has been on either the process or the outcome issue exclusively. The status of these two research domains will now be reviewed briefly.

Process Evaluation. Researchers who have sought to explicate the judgment process through use of multiple regression have implicitly assumed that judges integrate information in the manner depicted by the policy equation. That is, judges weight each cue differentially and combine this weighted product linearly in formulating a judgment. Moreover, the weights used by the judge remain constant across all stimulus profiles (Murphy, 1980).

Dawes (1971) reported a case study in which multiple regression was used to describe the decision process of the admissions committee in judging graduate school applicants. The admissions committee typically used four pieces of information—GRE scores, undergraduate GPA, quality of undergraduate institution, and letters of recommendation—to evaluate applicants on a 6-point scale that ranged from "reject now" to "offer fellowship". Dawes applied a multiple regression analysis to the rating data for a set of actual applicants and used the resulting regression weights in an effort to describe the
committee's rating process. Studies have used multiple regression similarly to describe the judgment process underlying a wide variety of tasks including evaluation of physical and mental pathology (Wiggins & Hoffman, 1968), personality characteristics (Knox & Hoffman, 1962), job performance (Zedeck & Kafry, 1977), teaching effectiveness (Marques, Lane, & Dorfman, 1979), stock or gamble attractiveness (Slovic, 1969; Slovic & Lichtenstein, 1968), and even quality of livestock (Phelps & Shanteau, 1978).

One problem with this logic is the general finding that individuals differ widely in their weighting strategies (Hoffman, 1960; Hoffman, Slovic & Rorer, 1968; Slovic & Lichtenstein, 1971). Thus if the equations are taken as literal representations of processes, one is left with the rather nonparsimonious conclusion that people approach judgments in highly idiosyncratic ways. For example, Hoffman et al. (1968) found considerable disagreement among nine radiologists regarding the signs indicative of malignant ulcers. The radiologists rated hypothetical ulcers, (described by the presence or absence of seven roentgenological cues), in terms of the likelihood of their being malignant. The median interjudge correlation among the overall ratings was only .38. There was some consensus on the diagnosticity of one cue, size: seven of the nine radiologists judged small ulcers as more likely to be malignant than large ones. However, in a follow-up study by Slovic, Rorer and Hoffman (1971), other experienced radiologists indicated that large ulcers were more likely to be malignant than small ones!
The tendency of multiple regression to highlight individual differences in judgment has led to the development of techniques for clustering judges in terms of the homogeneity of their regression equations (Adler & Kafry, 1980; Christal, 1968; Dudycha, 1970; Naylor & Wherry, 1965; Wherry & Naylor, 1966; Zedeck & Kafry, 1977). A variety of clustering techniques exist, but none has so far gained universal acceptance. All such techniques are based on the mistaken assumption that there are qualitatively different types of judges rather than a continuum along which judgment policies differ. But they have the practical virtue of providing a parsimonious means of describing similarities and differences between the policies of groups of judges (Adler & Kafry, 1980). They fail, however, to resolve the theoretical question of how adequate policy equations are for describing judgment processes.

Studies that rely exclusively on multiple regression to capture the rater's policy or to explore individual differences in weighting assume a priori that the judgment process is linear, a position that has not gone unchallenged (Anderson & Shanteau, 1977; Hoffman, 1968; Rapoport & Wallsten, 1972; Ynetma & Torgerson, 1961). Some investigators have attempted to evaluate the tenability of the linearity assumption empirically by introducing systematic nonlinearities into the judge's model, and then comparing the adequacy of various models (including the linear one) for describing his behavior (Einhorn, 1970; Slovic, 1969; Wright, 1979). To the extent that a nonlinear model provides a better "fit", the linearity
assumption is weakened; to the extent that any particular model proves superior to others, inferences may be drawn regarding the underlying judgment process. Before discussing this research further, however, it is necessary to consider what is meant by nonlinearity.

In general two types of nonlinearity may be distinguished (Slovic & Lichtenstein, 1971; Anderson, Deane, Hammond, McClelland, & Shanteau, 1981). One type occurs when a judge uses a cue in a curvilinear manner. Suppose, for example, that the judge is predicting test performance on the basis of anxiety level. It is likely that the judge might consider both extremely high and extremely low levels of anxiety more detrimental to performance than moderate anxiety, thereby implying a curvilinear relation between anxiety and performance. Such functions can be modeled by using polynomial regression values to develop the policy equation.

A second type of nonlinearity occurs when cues are combined in a configurational manner. The notion of a configurual or patterned relation between predictors and criterion refers to "... the situation in which the indication of a given variable with respect to the criterion is not constant, but the weight, and possibly even the 'direction' (sign) of contribution of that variable, are functions of the values which the other predictor variables have taken on" (Meehl, 1954, p.132). Thus configurality implies that a judge's weighting of one item of information differs as a function of the nature of the other available information. In judging the threat posed by enemy forces, for example, a military commander might give substantial weight to a particular kind
of surveillance report only in the presence of certain political intelligence information. Note that a configural relation is formally different from a nonlinear one. To illustrate, \( Y = 2X_1 + 1.5X_2 + 3X_1^2 + 2X_2^3 \) represents a nonlinear relation but it is not configural because the dependency of \( Y \) on \( X_1 \) is invariant with respect to the values taken on by \( X_2 \) (or vice versa). On the other hand, a function such as \( Y = 2X_1 + 1.5X_2 + .5X_1X_2 \) is configural because the effect on \( Y \) of an increment in \( X_1 \) depends on the value of \( X_2 \). Both types of nonlinearity have been investigated.

In a study designed to compare the descriptive efficiency of three alternative models, Wiggins and Hoffman (1968) applied a linear, a quadratic, and a "sign" model to the diagnosis produced by 29 clinicians on the basis of MMPI (personality) profiles. The linear model simply used 11 MMPI scale scores (\( X_i \)) as predictors; the quadratic model added to these predictors their squared values (\( X_i^2 \)) and 55 cross-product terms (\( X_iX_j \)). The sign model was based on 70 diagnostic signs that were derived empirically from previous MMPI research and contained many nonlinear relations. The results, using a cross-validated multiple correlation to measure comparative goodness-of-fit, indicated that 13 subjects were best described by the sign model, 3 by the quadratic model, and 12 by the linear model. But, most importantly, there was very little gain in predictive superiority with either of the nonlinear models over the linear one.

In another context, however, Einhorn (1971) showed that subjects use a conjunctive strategy to express job preferences. That is, he
found that they value a job positively only if it meets a minimum
criterion on all relevant dimensions (e.g., pay, location); a linear
model, of course, would dictate compensating for low values on one
dimension with high values on others.

It should be obvious that a plethora of models can be used to
assess the judge's nonlinearity. To surmount the problem of selecting
and testing a variety of available models, some researchers have used
ANOVA designs to demonstrate configural effects (Hoffman et al., 1968;
Rorer, Hoffman, Dickman & Slovic, 1967; Slovic, 1969). Such studies
involve the construction of stimuli that are defined by factorial
combinations of cue levels. Thus the set of profiles viewed by the
judge include all cue combinations, and statistical independence of the
cue dimensions defining the "environment" is assured. If, then,
analysis of consequent judgments reveals significant interactions,
one can only infer that the judge is using a nonlinear policy; for it
is he, not the environment, that must be the source of the
nonlinearity. The results of this line of research have tended to show
that judges do, indeed, adopt nonlinear judgment strategies but the
increase in predictive power afforded by configural effects appears to
be small (Hoffman et al., 1968; Rorer et al., 1967; Slovic, 1969).

Briefly, then, several lines of research have used multiple
regression to describe the judgment process. That the process may, in
fact, be nonlinear has led to tests of configural and other nonlinear
models as alternatives to the linear model. While these studies have
demonstrated instances of nonlinear cue utilization among judges, the
amount of nonlinear variance accounted for by such models has been negligible. It does not appear, therefore, that the use of models other than multiple regression in policy capturing has resolved the question of how judgments are formed—the process issue.

Other approaches, notably those involving self-report or explicit expression of subjective weights have not fared much better. By and large, they have suggested that subjective weights do not agree well with the "policy captured" weights (e.g., Summers, Fletcher, & Taliaferro, 1970; Cook & Stewart, 1975). Consequently, one faces the dilemma of deciding which set of weights really describes the underlying judgment process. More often than not, researchers have used regression-based weights as an objective standard for evaluating subjective weights (Schmitt & Levine, 1977). Thus, elicitation of subjective weights has contributed little to settle the process issue in policy capturing.

Outcome Evaluation. It is not surprising that, lacking any empirical means to demonstrate that policy capturing truly describes the judgment process, some researchers have focused solely on outcomes—the success of the judge's model in predicting decision behavior (Einhorn, Kleinmuntz, & Kleinmuntz, 1979). Two related concerns have dominated this predictive emphasis, both centering around the application of the multiple regression approach. The first involves comparison of the model's proficiency with that of the judge himself (i.e., the "clinical vs. statistical" controversy); the second involves use of the model in decision aiding (i.e., "bootstrapping")
(Dawes & Corrigan, 1974). In both cases the research question is whether the judge's model or his actual judgments show greater efficiency in predicting the criterion. For example, suppose that success in graduate school is predicted from GRE and GPA scores. The subject's judgments are first modeled using multiple regression or some other mathematical equation. Then, for a new sample of cases, the correlation between the model's predictions and the criterion values (say faculty ratings of graduate success), are obtained. This cross-validated correlation from the model is compared to the correlation between the subject's judgments and the criterion values. The measure that correlates higher with the criterion obviously yields better predictions.

The clinical vs. statistical controversy raged during the 1950's and 1960's, and a voluminous literature comparing the validities of the judge and his model appeared (see Meehl, 1954, 1965; Goldberg, 1968a; Sawyer, 1966 for reviews). The findings from these studies have been clear-cut: The predictions of the model are at least as good as, and in most cases better than, the judge's predictions. This situation led Meehl (1965) to say "... a pretty strong case can be made for an overarching decision-policy to predict by actuarial methods" (p.32), and further, "... it would be difficult to mention any other domain of psychological controversy in which such uniformity of research outcome as this would be evident in the literature" (p.27).

The aiding application of statistical prediction involves actually replacing the decision maker with his model in order to pull up
("bootstrap") the validity coefficient. The principal reason for the superiority of this approach over unaided human judgment is the elimination of factors such as fatigue, boredom, and context effects which reduce the judge's reliability but do not affect the model. In effect, it allows the judge's basic strategy to be applied more consistently than he, himself, is able to do (Dawes, 1979; Goldberg, 1970).

It should be recognized that statistical and clinical prediction are not necessarily mutually exclusive approaches (Sawyer, 1966). There are two ways in which they may be integrated: clinical synthesis in which the decision maker is given the outcome of the statistical prediction and asked to improve upon it, and statistical synthesis, in which the decision maker's judgment is included as an additional predictor in the actuarial model. Neither type of synthesis, however, has proven very successful. For example, Goldberg (1968b) found that allowing judges to override an optimal solution in the analysis of MMPI profiles yielded poorer performance than that produced by the formula alone. Similarly, Einhorn (1972) had four medical experts rate nine characteristics of biopsies taken from a number of Hodgkin's patients as well as the overall severity of their disease. Then, after the patients died, he built and cross-validated two models for each doctor: one predicting longevity from the nine rated characteristics, and the other including overall rating as an additional predictor. For two doctors, the model with the overall rating resulted in better prediction; for the other two, the simpler model was better.
In light of findings that demonstrate the unquestionable predictive superiority of the judge's model, some researchers have taken the issue to a logical extreme. Questioning the utility of even the weights derived from human judgments (and hence the whole concept of "bootstrapping"), this school claims that any linear model—even a randomly weighted one—is as good as bootstrapped judgments (Dawes & Corrigan, 1974; Dawes, 1979). In support of this position, Dawes and Corrigan tested the efficacy of bootstrapping in five different contexts by comparing it with validities produced by actual judgments, a linear model with randomly chosen weights, a linear model with equal weights, and a linear model with optimal weights (i.e., regression model of actual cue-criterion relationships). Although bootstrapping proved superior to unaided predictions in all contexts, all the other linear models—including the randomly weighted one—surpassed or equalled its predictive accuracy.

It appears, then, that even the predictive application of multiple regression in policy capturing is of inconsistent practical value (Dawes, 1979; Einhorn & Hogarth, 1975; Nystedt & Magnusson, 1975; Von Winterfeldt & Edwards, 1973; Wainer, 1976). As Dawes and Corrigan put it, "the whole trick is to decide what variables to look at and then know how to add" (p.105).

Multiple Cue Probability Learning

In contrast to policy capturing, MCPL research has from the very beginning focused on processes—in particular, how subjects learn to use information in making predictions or judgments. As discussed
earlier, learning is indexed by the correspondence between multiple regression models of human response and the "true" nature of environmental relationships (e.g., the achievement and matching measures).

Recall that in a typical MCPL paradigm the subject is required to make quantitative judgments for a set of cues and is then given the criterion values as feedback. A common aim is to determine how the acquisition of relationships proceeds as a function of environmental variables such as number of cues, cue validities, cue intercorrelations, and cue-criterion functions forms. This paradigm has been applied to a host of judgment problems ranging from basic perception to complex social interactions (Brehmer, 1976; Brunswik, 1955; Hammond et al., 1980; Hammond, Rohrbaugh, Mumpower, & Adelman, 1977; Hammond, Stewart, Brehmer, & Steinman, 1975; Smedslund, 1955; Summers, 1962, 1969).

The most general conclusion that may be drawn from this research is that the learning of cue-criterion relationships is heavily dependent upon task features. People have little difficulty learning to use linear cues (Smedslund, 1955; Summers, 1962). For example, Summers (1962) used three cues—color, angle, and area of triangles—of different validities to predict the length of one of the sides of the triangle. All cues were linearly related to the criterion and there was no error of prediction. The results showed that subjects were responsive to the validities of the cues and learned to use them appropriately.
Studies have extended the above paradigm to include nonlinear cue-criterion relations (Brehmer, 1969; Hammond & Summers, 1965; Summers & Hammond, 1966). In a study by Hammond and Summers (1965) subjects predicted a hypothetical trait on the basis of two test scores, one of which was linearly related to it and another was nonlinearly related. Subjects generally showed a distinct tendency to depend only on the linear cue. The nonlinear cue was used, although less effectively, by subjects who were forewarned about the nonlinear relation. Summers and Hammond (1966) confirmed these findings and reached the conclusion that subjects can learn nonlinear relations very gradually, but only if they are aware of the existing nonlinearity.

It might be remarked that in most MCPL studies subjects' nonlinearity is measured with the C-coefficient (Hursch, Hammond & Hursch, 1964). C is the correlation between the residual which cannot be linearly predicted in the criterion and which cannot be linearly predicted in the judgment. If either of these residuals is random, C is zero (Slovic & Lichtenstein, 1971). Thus unlike policy capturing in which different nonlinear functions are often fitted to the subjects' responses, response nonlinearity is measured directly by using the C-coefficient.

Some studies have explored the effect of linearity in the environmental system ($R^2_e$) on learning. Very simply, $R^2_e$ represents the amount of variance in the criterion that can be predicted from a linear combination of the cues. It has been found that $R_s$ matches $R_e$ for intermediate values of $R_e$ (Brehmer & Lindberg, 1970).
Often, the environment is dynamic and thus subjects' adaptation to changes in environmental characteristics becomes a relevant question. Summers (1969) investigated subjects' sensitivity to changes in cue validities and function form relating cues to the criterion. Thus for a "cue shift" group the regression equation relating cues and criterion remained the same, but the specific relation between individual cues and the criterion was reversed over time (e.g., the equation $Y = aX_1 + bX_2$ was changed to $Y = aX_2 + bX_1$); for the "rule" shift group the nature of the functional relation between the cues and criterion was changed (e.g., $Y = aX_1 + bX_2$ was changed to $Y = cX_1 + dX_2$); the "complete shift" group had both cue and rule shifts, and finally the control group had no shifts. The control group performed the best and showed high achievement. The rule shift and complete shift groups, while not different from each other, performed the worst and the cue shift group was intermediate. It is apparent that it was easier for subjects to learn which cues to use (cue shift subjects) than to discover the functional rule relating cues to the criterion (rule shift subjects).

In tasks that have varied cue-intercorrelations, subjects have failed to take appropriate account of redundancies in correlated cues. Thus subjects' regression weights matched the cue-criterion correlations rather than cue-criterion regression weights (e.g., Naylor & Schenek, 1958).

Perhaps the most frequently cited task effect is that involving the way feedback is administered. In general, outcome feedback
(whereby subjects are informed of the correct criterion value after each judgment) produces relatively slow learning, whereas process feedback (whereby they are shown the corresponding cue utilization and ecological validity coefficients) is more effective (Todd & Hammond, 1965). In fact, just knowing the ecological validities—-in effect, the environmental rules—is highly beneficial (Slovic & Lichtenstein, 1971). Recently, also, Adelman (1982) has shown that task congruence can moderate the process-outcome discrepancy: for tasks in which the rules are fairly obvious, outcome feedback can be as useful as process feedback. The generalization that seems to be emerging, then, is that MCPL is heavily dependent upon the decision maker's ability to comprehend and apply the rules governing environmental uncertainty. This, in turn, rests heavily upon task features such as explicit instructions, task congruence, rule complexity, time pressure, and probably a host of others (Hammond, McClelland, & Mumpower, 1980; Goldsberry, 1983; Einhorn, 1980; Wickens & Scott, 1983).

Within the realm of interpersonal learning and conflict, the MCPL paradigm is applied in the following manner. In the initial "training" phase, pairs of subjects learn the judgment policies required by the experimenter. Each subject learns to use a different cue, perhaps in a different way (e.g., linear or nonlinear). In the second "conflict" phase, subjects are brought together to jointly predict the criterion. Often, a subject's individual judgments are obtained before the joint judgments are made. A lens model analysis of each subject's individual judgment and the pairs' joint judgment is used to study the mechanisms
whereby subjects learn from the task and from one another (Brehmer, 1976; Hammond et al., 1977; Slovic & Lichtenstein, 1971).

Brehmer (1976) summarized major results from studies on interpersonal conflict to conclude that although subjects can reduce systematic differences in their policies as they interact, they become increasingly inconsistent and thus overt discrepancies are not diminished at the end of the conflict phase. Moreover, subjects' agreement is dependent on characteristics of the task since they tend to adapt policies to the task structure: Tasks that are highly uncertain, include multiple cues or nonlinear cue-criterion relations lead to lower consistency and more conflict than do those that are highly predictable, include only one cue or have linear cue-criterion relations.

The results obtained from studies on interpersonal conflict are similar to those from learning studies. Overall environmental variables are important in determining how subjects learn to make judgments--individually and jointly-- based on multiple cues. However, researchers have typically limited themselves to the variables suggested by the lens model, a point that will be made more fully in the following sections.

CONCEPTUAL DIFFICULTIES

The principal conclusion to be drawn from the foregoing discussion is that use of multiple regression in MCPL research has demonstrated how environmental variables might affect different aspects of subjects' decision performance. Despite the usefulness of such findings,
however, there has been an overemphasis on the structural variables based on the lens model. On the other hand, use of multiple regression in policy capturing research has contributed more heavily to prediction than to delineation of underlying processes. Simple models have been developed that consistently outperform people, and "bootstrapping" has proven a useful decision aid, but there is still little understanding of how human "policy" is deficient or the circumstances under which it becomes most critically so.

This section discusses several conceptual problems that may account, in part at least, for the limitations so far encountered in multiple regression research. The line of argument to be developed is that: (1) "policy capturing" is not capable of truly capturing processes, only outcomes; (2) one need not capture processes to gain understanding; and (3) both "policy capturing" and MCPL would benefit from a more functional use of the multiple regression approach--one in which the model is used to index performance rather than to infer cognitive processes.

What is Captured in "Policy Capturing"?

Use of multiple regression models to "capture" the judgment policy has, as discussed in the last section, been directed toward two main objectives: describing the judgment process, and predicting decision outcomes. The latter objective is considerably easier to justify conceptually than the former, since the criterion for success is simply the magnitude of $R^2$. So long as the correspondence between the decision maker's judgments and his model's predictions is high, the
assumption that the model has "captured" something important about his behavior appears tenable. The better the prediction, the more confident one can be of the capture, and the more practical usefulness one can ascribe to the model. Indeed, one may even be able to verify the model's superiority over the unaided human if there is some means of describing the external environment. The only problem is that, empirically, the linear model is so robust. As Dawes (1975) has noted, what is captured may be more a reflection of the task than of the decision maker per se. In any case, so long as the goal is purely predicting outcomes, there exists a basis for evaluating models.

It is when the emphasis shifts from sheer prediction to describing underlying processes that the multiple regression approach encounters real difficulties. To satisfy this objective, one must first settle on a working definition of "process" (so that whatever stumbles into the net may be identified), and then have some basis for verifying the identification. As noted earlier, however, policy capturing research has made little headway on either score. Typically, the assumption has been that if the model predicts judgments well, it is a good representation of the cognitive process. That is, if the linear regression model accounts for a substantial amount of the variance in observed judgments, the process is assumed to be linear; if nonlinear terms add significantly to the explained variance, it is assumed to be nonlinear.

There are two problems with this approach. The first is logical: Accurate performance of the model is a necessary condition for
inferring process characteristics, but not a sufficient one (Anderson and Shanteau, 1977). Indeed, the literature examined in the last section clearly demonstrates that a linear model can do extremely well predicting results generated by a nonlinear process (i.e., the same robustness that makes a linear model so useful for predictive purposes makes it ineffective for descriptive purposes). There is simply no way to determine in a specific task whether the human judgment process is in fact linear, or whether it is nonlinear and the robustness of the model obscures that fact.

The same logical difficulty applies as well to the research involving empirical comparison of alternative models (e.g., Einhorn, 1971; Nystedt & Magnusson, 1975): Process differences might or might not be reflected in performance differences. However, this research highlights the second difficulty in the explained-variance rationale—a practical and empirical problem. If one hopes to pinpoint the judge's process by finding the model that produces the best empirical results, on what basis does one select the set of models for comparison? Testing all possibilities would be impractical, and sampling from among them incurs the risk of excluding the one "real" process. Results to date, as we have seen, have shown little difference in performance between the "best" and "worst" of models.

One possible way to circumvent the whole process-description problem is to adopt a rather liberal definition of the process concept. Some argue, for example, that a mathematical model is just a "paramorphic" representation of the judge's strategy—a simulation—not
the literal mental calculus (Hoffman, 1960; Dawes, 1971). In fact, Hoffman went so far as to define a cognitive process as "... a functional relationship which accounts for consistencies in response to divergent stimulus information patterns" (p.117). Viewed in this way, any model that is capable of predicting judgments from observed cues is "capturing a process". From there it is only a short logical step to the proposition that the best description of the process is the model that results in the best prediction, and we are back to the problems of the explained-variance criterion. Paramorphic definition, then, does little more than reduce the meaning of judgment process to that which is verified through prediction. What is captured is outcomes, not processes.

In view of the conceptual difficulties inherent in the use of multiple regression to capture judgment processes, therefore, it seems reasonable to conclude that this approach is all but futile. There may, however, be considerable practical value in using multiple regression to capture outcomes. For example, Murphy (1982) has demonstrated that regression models can show discriminant validity relative to unit-weighted models when predictor intercorrelations are low and judgments in various tasks are not correlated. To illustrate the discriminant validity approach, it is often the case that different academic specialties judge the same applicant for graduate school differently. The engineering department, for example, would certainly not give the same pattern of weights to verbal test scores, quantitative test scores, language course grades, math and science
course grades, and years spent abroad, as would the French department. Unit weighting, of course, would predict the same evaluation outcome for all departments. Suppose, however, that separate regression models are constructed, one for the French department and another for the engineering department. To the extent that the "French" model yields better predictions of the French department's subsequent judgments than does the "Engineering" model, and vice versa, discriminant validity would be high and evidence favoring the utility of a policy capturing approach would be strong.

Can cue importance be measured?

When multiple regression is used to monitor changes in judgment policy, particularly relative to a task-defined optimum, the MCPL paradigm replaces that of policy capturing. It will be recalled that many of the current research issues in MCPL center around the efficacy of various ways of providing a decision maker with information for improving his performance. How, for example, should he be informed that he is not weighting predictive cues properly, or that his strategy for combining them is non-optimal? And what are the critical task features?

The principal conceptual difficulty in the MCPL applications lies in measuring the weights or importance that a decision maker attaches to cues. It is in these terms, after all, that the decision maker's policy is described and evaluated (i.e., against "optimal" weights). Unfortunately, however, multiple regression does not provide a single
estimate of cue importance, and the several alternative indexes may or
may not agree depending upon the underlying task structure.

The three common indexes of cue importance include (1) the simple
correlation between a cue and criterion, (2) the semipartial
correlation (correlation obtained when all other cues are partialled
out of each cue), and (3) the standardized regression coefficient
associated with a cue. If cues are orthogonal, all the above measures
are equivalent. However, when cues are intercorrelated, the different
measures of importance may be totally contradictory, a point that led
Darlington (1968) to conclude "... the notion of 'independent
contribution to variance' has no meaning when predictor variables are
intercorrelated" (p. 169). In real world decision contexts, of course,
cues are generally correlated. For example, a military attack is
likely to be preceded by troop build-ups, increased density of
communication, hostile political statements, and other related events.
If one wishes to preserve a degree of realism in a MCPL task, such
complexities must be preserved or built in; otherwise, the task will
appear contrived and illogical. The result, of course, is that the
researcher is often forced to choose between unrealistic tasks for
which cue importance can be calculated unambiguously, and realistic
tasks for which it cannot.

Recently, an alternative to the conventional indexes has been
proposed that overcomes some of these measurement problems (Lane,
Murphy, and Marques, 1982). It is the raw score regression weight, a
measure that represents the expected change in judgment per unit of cue
change and remains invariant over all but the most extreme levels of correlation among cues. While a definite improvement over the others, however, even this index has its limitations. To be used legitimately, three conditions must be satisfied: (1) all variables that correlate with the criterion are included in the regression equation (or are uncorrelated with the variables in the equation), (2) the criterion has no effect on the predictors, and (3) the relation between criterion and predictors is additive and linear (or the appropriate nonlinear terms are included in the equation). Most of these conditions can be met in laboratory applications of MCPL; to what extent the same can be said for measuring the decision maker's cue weighting policy in natural settings remains to be determined.

TOWARD A FUNCTIONAL APPROACH

The conceptual problems just described have limited the usefulness of the multiple regression approach in decision research. It is the present contention, however, that--serious though they may be--the difficulties are not insurmountable; that, for the most part, they stem from overconcentration on certain issues and neglect of others; and that, to realize the full potential of multiple regression paradigms, researchers must approach them from a new perspective.

The difficulties inherent in capturing processes with multiple regression are serious only if one assumes, as most have done, that true understanding proceeds in no other way. Preoccupation with inaccessible processes may actually hinder understanding by diverting attention away from important systematic relationships that are present
in the data. Watkins (1981) has made this argument rather convincingly with regard to an analogous difficulty in the field of human memory. Hypothesized "encoding", "storage", and "retrieval" processes have proven as difficult to verify and isolate as the operations involved in judgment. Watkins suggests a more functional approach—one that relies heavily on operational definitions and on the search for systematic relationships between task characteristics and behavior (i.e., judgment, choice, or whatever aspect of performance is of interest).

In the domain of judgment research, few people have even attempted to define "process" operationally. Hoffman (1960), as we saw, operationalized it in terms of mathematical models that best predict judgment. But testing models advances our understanding of judgment little more than does the assumption that the multiple regression model actually represents the judgment process. Watkins' (1981) argument for a functional approach would seem as well taken in the context of human judgment as in that of memory.

The present argument, then, is that our practical understanding of judgment would be furthered by a redirection of effort away from process description and toward discovery of task-judgment relationships. On the independent variable side, this requires some taxonomizing or structuring of task characteristics, and several such efforts have already appeared (e.g., Hammond, 1980). On the dependent variable side, it calls for measures indicating the internal consistency and external validity of the decision maker's judgments.
As we saw earlier, both the MCPL and policy capturing paradigms offer suitable indexes for internal consistency. Only the former, however, typically includes any reference to an external criterion. From a practical standpoint it is obviously useful to know not only what weighting strategy a decision maker uses, but also how accurate it is. The hypothetical stimuli constructed for laboratory research usually have no external basis for establishing the "correctness" of a judgment. Real world situations may provide actual outcomes (hence a means of verifying the accuracy of judgments), but these are generally confounded by information not available to the decision maker. Thus he may succeed despite using a "poor" strategy, or fail despite using a "good" one (Einhorn, 1980; Lichtenstein, Fischhoff, & Phillips, 1977).

The multiple regression approach does afford ways of indexing the quality of judgment policies—even though it may not yield a true representation of the mental processes involved. If, for example, it can be established that the environment is characterized by nonlinear cue-criterion relationships, the adequacy of the decision maker's policy can be assessed by testing the appropriate interaction terms. Similarly, if the "true" ecological importance of particular environmental cues can be established, comparison of these with the decision maker's raw-score regression weights will indicate the accuracy of his performance.

Thus what is emphasized within the functional approach proposed here is neither the mere capturing of outcomes for practical applications in decision aiding nor the construction of process models
that reflect vague underlying processes. Instead, indexes based on the
multiple regression model are regarded as measures of consistent
behavioral tendencies so that progress can be made toward understanding
as well as improving human judgments. Task characteristics such as
those identified by Hammond are manipulated systematically with an eye
toward concomitant changes in raw-score regression weights, policy
linearity, "achievement," "control," or the various other measures
afforded by the multiple regression model. That, then, is the essence
of the functional approach which is being advocated here. It might be
pointed out that some research consistent with this philosophy has
appeared (see Slovic & Lichtenstein, 1971 for a review); still, the
bulk of the work has been concentrated on either process or outcome
issues as discussed earlier.

Research on MCPL has provided some insight into the way people
learn to integrate multiple sources of information into judgments. In
general, however, it has been limited to tasks and issues involving
correlations—the various relationships depicted in the lens model
(see Figure 1). Attention has invariably centered around questions
such as how closely the subject's perception of cue-criterion
correlations agrees with the ones constructed by the experimenter, how
this correspondence changes over time, and how different feedback modes
affect the acquisition function. Very little attention is usually
accorded the nature of the cues themselves or the context in which they
are observed. An alternative, which is being advocated here, is to
regard the lens model as a useful source of behavior indices, and to
explore how these indices change as a function of systematic task manipulations (i.e., the functional approach). Potential candidates for such functional analysis include variables suggested by recent taxonomic efforts and by research domains that are not traditionally included within the MCPL framework.

One methodological issue deserves comment. In general, investigators have been concerned solely with describing subjects' performance in relative terms. Thus, in policy capturing or MCPL, $R^2$'s or linear consistency of policies is reported (e.g., Knox & Hoffman, 1962) without any reference to its component measures—$SS_y$, $SS_\hat{y}$, or $SS_e$. A definition of these terms is provided in Table 2. As will be shown later in the paper, the magnitude of $R^2$'s varies directly with obtained differences in $SS_\hat{y}$ and $SS_e$. Thus, the same level of $R^2$'s could be obtained from different combinations of $SS_\hat{y}$ and $SS_e$, obviously implying different types of judgment performance. For example, the magnitude of $SS_e$ reflects the degree of precision in judgments (viz., smaller $SS_e$ implies greater precision) and $SS_\hat{y}$ reflects the degree of linear cue usage. Similarly, measures of accuracy such as achievement, matching coefficient, optimality coefficient are correlational. Although correlational data yield useful knowledge regarding the relative accuracy of subjects' predictions of the criterion, there are many real world situations where absolute accuracy of predictions is crucial. For example, a doctor prescribing certain dosages of medicine to his patients has to strive for absolute accuracy; after all, large amount of a dosage, although proportional to the "true"
### TABLE 2

**Definition of components of R^2s**

<table>
<thead>
<tr>
<th>SSŷ</th>
<th>sum of squared deviations of subject's predicted responses (Ŷs) from their mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE</td>
<td>sum of squared deviations of the residual in subject's responses that could not be predicted from a weighted combination of the cues (Ys - Ŷs) from their mean.</td>
</tr>
<tr>
<td>SSy</td>
<td>sum of squared deviations of subject's responses (Ys) from their mean or (SSŷ + SSE)</td>
</tr>
</tbody>
</table>
requirement (and thus accurate relatively) might prove lethal! The present contention is that absolute measures of performance are equally important and should be included in conjunction with relative measures in an analysis of performance.

**IMPORTANCE OF TASK CHARACTERISTICS IN JUDGMENT**

The review of MCPL studies illustrates that the task structure significantly influences how successfully the individual will learn to make judgments. However, these results cannot be readily generalized to the policy capturing paradigm. MCPL provides insight into what decision makers can do under a specific task structure. On the other hand, since policy capturing shifts its emphasis to what the decision maker, in fact, does, the conclusions from MCPL studies provide an upper bound on performance, but cannot be applied directly to actual decision behavior. Moreover, researchers using the policy capturing paradigm have concentrated largely on capturing processes or outcomes and there has been little systematic work relating task characteristics to judgments in that paradigm.

What appears to be necessary then is to identify and classify relevant task components that affect judgment—the taxonomic approach. Recently, Hammond has proposed a Cognitive Continuum theory (1980) that provides such a taxonomic framework.

**A Framework for Evaluating Task Influences**

It might be mentioned at the outset that Hammond's theory is useful for testing the influence of task characteristics on decision behavior, but it concentrates on vague cognitive processes to explain
task-judgment relations, a position that is not entirely consistent with the functional view propogated here. It should be noted that any references to processes are made to present Hammond's arguments with completeness and are not intended to contradict the functional approach.

The basic premise of the theory is that there are various modes of cognitive functioning that can be ordered on a continuum that ranges from "intuitive" to "analytic" processing. The notion of a continuum replaces the conventional analysis-intuition dichotomy, in which the two are seen as functioning as rival or competing forms of cognition. In keeping with the traditional view, Hammond recognizes that the distinction between analysis and intuition is clear-cut. Indeed, analysis signifies a "step-by-step, conscious, logically defensible process of problem solving" (p.8), whereas intuition signifies the opposite: it is a process that "somehow permits the achievement of an answer, solution, or idea without the use of a conscious, logically defensible, step-by-step process" (p.8). Hammond departs from tradition, however, by suggesting that the two modes can be brought to bear in varying degrees on particular judgment tasks. In fact, most cognitive functioning is not solely analytical or intuitive; rather it represents quasi-rational functioning—the mixture of analysis and intuition—that varies systematically over the continuum. Such quasi-rational functioning is alternatively referred to as "common sense" or "bounded rationality".
The mode of cognitive functioning that is invoked, according to Hammond, is determined by three broad categories of task features (complexity of the task, ambiguity of the task, and form of task presentation). Thus, for example, an unfamiliar task in which the subject is given minimal feedback about the task outcome (viz., an ambiguous task) is more likely to induce intuition whereas its antithesis (a familiar one that provides feedback) is more likely to induce analytic processing. The other specific task characteristics similarly pull cognition toward one pole or the other. To complete the picture, these hypothesized modes of cognitive functioning are seen to have direct behavioral consequences. Intuitive processing results in judgments that have low predictability over time, that cannot be easily verbalized, and have brief response times. In contrast, analytic processing (which, by definition, is based on an explicit rule) results in judgments that show a high degree of predictability and retraceability and that require more response time.

Drawing on the continuum theory logic, it is possible to make predictions about the kinds of performance that might be expected under particular sets of task conditions. For example, a complex task containing a multiple set of cues that are highly correlated should induce intuitive processing and produce judgment characterized by low predictability, lack of retraceability and brief response times.

While it must be recognized that continuum theory is cast in a very general form and that its key "analysis" and "intuition" concepts all but defy operational definition, it does offer a good point of
departure for studying the effect of task characteristics on judgment and decision behavior. Among other things, it provides a logic for deciding which task-judgment relations are most worthy of exploration. One task variable suggested by this framework which has considerable practical relevance is the manner in which information is displayed to the decision maker. Since empirical evidence relating display format to judgment is relatively sparse (see the following section), a series of studies was undertaken to address this issue. The four resulting experiments constitute the empirical contribution of this paper.

**Display Format in the Performance of Visual Tasks**

Processing coded information is a very common-place activity—so much so, in fact, that people rarely give it any thought. Reading maps, music, road signs, daily market reports, even the spoken or written word seems so natural that it is easy to forget that somewhere, somehow decisions were made to represent particular concepts with particular symbols. However, the rapid evolution of the computer has alerted everyone to the importance of how information is represented (the awkwardness of early computer languages, for example, alienated an entire generation of users). It is therefore of practical as well as theoretical value to understand which display characteristics affect the speed and manner in which the displayed information is processed. Engineering psychologists have directed a considerable research effort toward this end, especially in the context of human/machine interface design.
The effect of display format has generally been studied using one of the following five types of tasks: (1) visual search, (2) question answering, (3) reading, (4) subjective ratings, and (5) decision making/problem solving (Tullis, 1982).

**Visual Search.** Visual search tasks require subjects to indicate the presence or absence of a target in a display. The display might involve individual numbers, letters, symbols, or more complex arrays of words or alphanumeric codes. In all such cases, the subject might be asked to detect a target from a predefined stimulus set, or to locate the position in which it appears on the display. Reaction time is usually the measure of primary interest, although accuracy is also typically recorded. Williams (1965), for example, used a scanning task that required subjects to pick pairs of nonidentical numbers from an array of pairs of numbers. Scanning was considerably faster for a horizontal than for a vertical arrangement (cf., Coffey, 1961). In another study, Klemmer and Frick (1953) had subjects reproduce the location of dots that were presented on the inside of either a blank square or a square that was divided into a matrix by grid lines. They found no effect of inclusion or omission of grid lines.

**Question Answering.** The question-answering technique entails asking subjects a variety of questions about the displayed information. The questions range in complexity from simple ones that can be answered by retrieving one item from the display (e.g., "Does the target appear in location X", in essence a visual search task) to complex ones that involve integration of several displayed items (e.g.,
counting the number of targets present in the display, verifying similarities or differences, etc.).

In an early series of studies by Hitt, Schutz, Christner, Ray, & Coffey (1961), alphanumerical, trend, and cartographic displays were investigated to develop general design criteria for intelligence display formats. The primary finding from these studies was that the type of task to be performed determines the efficacy of a particular type of format. For example, numeral coding was superior to color coding only for identification of symbols but not for counting or verification tasks (Hitt et al., 1961, Experiment IV). In another investigation of display format, Grace (1966) measured the effectiveness of three printout formats--"Verbal" (information presented in words), "Data Block" (information presented in tabular form), and "Eidoform" (information presented in maplike form). Subjects answered questions on the content of the display with greater accuracy when information was presented in the Verbal or Data Block formats than the Eidoform format. Wright (1968) varied the form of currency conversion tables and concluded that a fully redundant table that required no additional arithmetic operations was used most efficiently. In another study Tullis (1981) compared four formats (viz., narrative, structured, black and white graphics, and color graphics) and found that accuracy of responding did not significantly vary with format but that response times were shorter for the two graphic formats than for the narrative or structured formats.
**Reading.** Reading speed and comprehension are the two measures used to evaluate how display format affects reading. The variables that are typically investigated and found to affect reading include spacing of lines, character density, and typography or letter type (e.g., Kolers, Duchnicky, & Ferguson, 1981; Poulton & Brown, 1968; Tinker, 1955; Tinker & Patterson, 1928).

**Subjective Ratings.** In addition to measuring speed and/or accuracy of subjects' responses to different types of display, another more direct technique is to measure subjects' qualitative ratings of the display. Thus subjects are asked to rate each display on overall quality and other dimensions (e.g., Christie, 1981; Grace, 1966), rank order their preferences (Vitz, 1966), or rate the similarity of pairs of display (Siegel & Fischl, 1971).

**Decision Making/Problem Solving.** Another technique used to evaluate display format effects is to study subjects' problem solving or decision making performance as a function of format manipulations (e.g., Bettman & Kakkar, 1977; Ward & Jenkins, 1965, Wright, 1977). Giancchinelli and Lantz (1978) found that a "filtered" display that eliminated irrelevant stimulus information produced greater accuracy in classification of stimuli than a "nonfiltered" display. The fact that increased data density in a display could be detrimental to performance was exemplified in another study of problem solving by Baker and Goldstein (1966). In their study a "sequential" display that presented only the response alternatives permissible at a time resulted in shorter times to learn the solution than a "batch" display that
presented all permissible and nonpermissible response alternatives. Instead of using the traditional measures such as speed or accuracy in evaluating display effectiveness, Silver, Jones, and Landis (1966) used a complex measure of "decision quality" and demonstrated significant effects of structurally different visual displays. Goldsmith and Schvaneveldt (1982) argued that in tasks such as MCPL that require subjects to integrate multiple sources of information, a display that involves dimensions perceived as an integrated whole (integral dimensions) would facilitate learning relative to a display that involves dimensions perceived individually (separable dimensions). They compared two graphical displays—an integral display (e.g., one in which cue values were represented by width and height of rectangles) and a separable display (e.g., one in which cue values were represented as bar graphs) and demonstrated that superior performance resulted from the former display. Wickens and Scott (1983) recently compared verbal and spatial-graphical display formats in a tactical decision making task. Their task essentially required subjects to integrate a number of information sources that varied in diagnosticity and reliability and then make likelihood judgments concerning the tactical maneuvers in effect. The spatial display mode increased accuracy of decisions relative to the verbal mode. Rather than displaying the information to be judged in different forms, Hammond (1971) used a verbal or a graphical display to provide subjects feedback about cue weights and cue-criterion functions in a MCPL task. The graphical feedback mode produced more efficient learning than the verbal one.
In short, there are many specific findings involving display format effects but only a few broad generalizations. One that seems to be emerging is that increased information density produces systematic decrements in performance (e.g., see Tullis, 1982 for a review). Other generalizations pertain to theoretical explanations introduced to account for observed format differences. One good illustration is Wickens' principle of stimulus-central processing-response (S-C-R) compatibility. On the basis of this principle Wickens and Scott (1983) argued that their tactical decision task (which demanded spatial/analog processes in working memory) was better served by a spatial display than a verbal one because the former—being more "compatible" with the required cognitive process—placed less demand on the human's processing capacity.

Despite these generalizations, most evidence to date suggests that while many alternative ways of displaying information to the decision maker can have an important bearing on performance, the precise nature of the effect depends on the type of task performed.

**Display format effects in policy capturing.** A common feature of the studies reviewed above is the specification of external criteria to evaluate performance differences as a function of format manipulations. Very simply, formats that result in quicker responses, fewer errors, or are preferred by its users are deemed more efficient relative to the comparison formats. However, a variety of situations exist in the real world where an external performance criterion is not available or is infeasible. This is particularly true in the case of judgment or
decision tasks. In such cases what seems important is to monitor how format might affect internal consistency or other aspects of the decision maker's judgments. As pointed out in an earlier section, the policy capturing paradigm is a useful approach to the study of decision making behavior in such situations. But there is little empirical evidence regarding the effects of display format within the policy capturing paradigm.

An early study by Knox and Hoffman (1962) examined the effect of profile format on subjects' judgments of a person's intelligence and sociability. The cue values were displayed graphically as T-scores (with a mean of 50 and standard deviation of 10) or as percentile scores. It was found that subjects responded "... not only to the underlying meaning of the scores, but to the position of the points on the profile in some absolute sense" (p.19). Thus the percentile scores that had more extreme values produced more variable judgments than T-scores that tended to appear "squeezed in". Percentile scores resulted in more reliable judgments and higher values of Rs. The regression weights, however, did not differ between formats. As the authors pointed out, the subjects were not statistically sophisticated and thus their greater familiarity with the percentile scores may have contributed to a more precise handling of those data than T-scores.

Anderson (1977) reported another study that compared judgment policies of teacher quality using verbal and numerical profiles. The linear consistency of verbal profiles was significantly less than that
of numerical profiles, but the pattern of weights for the two cue formats was similar.

One methodological point related to both of these studies concerns the use of standardized regression weights to compare cue weighting under different formats. By definition, standardized weights measure the amount of change in the criterion (Y) in standard deviation (SD) units as a function of one SD unit change in the predictor (Xi). Although independent of scale, measuring the change in the criterion in terms of SD units (as standardized weights do) may not reflect actual changes in cue weighting in some situations. Suppose that the cue weighting as measured by raw score regression weight (bi) increases as a function of some experimental manipulation. Such an increase results in a direct increase in the SD of the criterion (SDy). Thus when standardized regression weight (Bi) is used, the increase in the raw score weight is offset by an increase in SDy and the change in cue weighting may not be apparent, as is indicated by Equation 5.

\[ Bi = bi \left( \frac{SDx}{SDy} \right) \]  \hspace{1cm} \text{Equation 5}

As Lane et al. (1982) have argued, the raw score regression weight adjusted by the SD of the cues may be a more appropriate index when cue weight is to be measured independent of the scale of measurement of cues.

Although Anderson (1977) did not report the SD of judgments under the two formats, Knox and Hoffman (1962) found judgments to be more variable in response to percentile scores than T-scores. It is therefore not immediately obvious whether the failure to find
differences in standardized weights in these studies indeed implies that there were no differences in cue weighting or whether it resulted from the insensitivity of the measure used to detect these differences.

Studies of policy capturing, then, have offered inconclusive evidence on whether display format affects subjects' cue weighting. Also the comparison formats used were not readily interpretable (e.g., Knox & Hoffman, 1962) or involved verbal and numerical displays (e.g., Anderson, 1977). The purpose of the present studies was to compare the effect of two relatively common formats, numerical and graphical representation of cues, on subjects' judgments and decisions. These formats represent two broad classes of structured displays; moreover, subjects' familiarity with them makes them easily understood. Of course, as the earlier review illustrated, similar formats have been shown to affect decision accuracy in other contexts but their effect has not been explored in the multiple regression paradigm. The primary goal of the present research, then, was to determine the effect of display format on cue weighting and other aspects of subjects' policy when no external optimization criteria were available. The first three experiments addressed the issue directly using the policy capturing paradigm. Since format produced a reliable change in subjects' judgments, the generality of the effect was explored further using the MCPL paradigm in Experiment 4. It might be noted that additional task variables were manipulated in conjunction with format in each specific experiment, and that the task scenario was varied between experiments. Scenarios that had "face validity" and that possessed features relevant
to the requirements of the experimental paradigm under consideration were chosen. Further discussion of these methodological features is reserved for the detailed account of each experiment.

EXPERIMENT 1

Given that information regarding display format effects in a policy capturing paradigm is sparse, this experiment was simply an attempt to determine whether a gross format difference (numerical vs. graphical display) would affect either judgments or choices based upon identical input data. Subjects were required to process multidimensional stimuli that were displayed numerically and graphically and their subsequent responses under both formats were compared.

Subjects performed two types of decision tasks—judgment and choice. A number of investigators have suggested that the type of response required—judgment and choice— influences how people process information (e.g., Einhorn & Hogarth, 1981; Hammond et al., 1980; Payne, 1982), and thereby produces substantially different kinds of decision behavior. In a judgment task the subject is typically required to assign values to individual alternatives to reflect their psychological worth (in terms of ratings on a scale or in terms of the amount of money a subject would pay for an alternative) whereas in a choice task he has to select one or more alternative/s that is/are preferred from a set of alternatives. For example, evaluating the quality of cars available on the market on the basis of information such as size, m.p.g., cost, etc. constitutes a judgment whereas
selecting a car that is most preferred from a set of cars constitutes a choice. Both judgment and choice are clearly interdependent in that choosing from a set of alternatives may well entail judging them with respect to several dimensions; nevertheless, making a choice involves explicit consideration of utilities (Edwards & Tversky, 1967), a dimension that is not necessarily involved in judgment. In view of these considerations, both types of response were examined for possible display effects.

METHOD

Task. The basic task required subjects to select or rate suitability of applicants for the job of a secretary (a task that most subjects find both meaningful and realistic). More specifically, subjects were presented with profiles of information about hypothetical applicants which were comprised of four dimensions: intelligence, motivation, skill, and experience. Each profile was represented in one of two ways: as a set of numerical scores (numerical format) or as a set of bar graphs (graphical format).

All subjects performed a rating task and a choice task under the two display formats. The rating task simply required them to review one applicant profile at a time and rate it on a suitability scale that ranged from 1 (extremely low) to 10 (extremely high). Each subject rated 100 profiles under each format condition. In the choice task, two applicant profiles were presented together and subjects had to indicate which one of the two applicants they considered more suitable.
for the job. Fifty applicant profile-pairs were presented for choice under each format condition.

Stimuli. Four sets of 100 applicant profiles (designated as p, q, r, and s) were produced by a multivariate normal generator such that values on the four cues were not intercorrelated. A multivariate array of deviates in the range of 0 to 1 was produced. The deviates were further transformed such that the actual values that defined the four cues (intelligence, motivation, skill, and experience) were sampled from populations with means (and SDs) of 25 (15), 5 (2), 10 (3), and 3 (.5) subject to the constraint that the cue values ranged between 1-50, 1-10, 1-25, and 1-5 respectively.

Sets p and q were used in the rating task and sets r and s in the choice task. The profiles in each set were printed on unlined, continuous paper; only one profile with an applicant number appeared on each page in the rating task (sets p and q), but two profiles designated as applicant A and applicant B were presented on the same page in the choice task. Each set of profiles was represented numerically and graphically in separate booklets.

An illustration of the two formats for the rating task is shown in Figure 2. It might be noted that the four cue values were represented as raw scores in the numerical display condition. Under the graphical format the length of the bar for each cue indicated its value and a single asterisk alongside the bar represented the highest possible value on that particular dimension. Note also that the four cues were represented on identical physical scales. Now, since the range of
Numerical Display

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intelligence</td>
<td>35</td>
</tr>
<tr>
<td>Motivation</td>
<td>6</td>
</tr>
<tr>
<td>Skill</td>
<td>16</td>
</tr>
<tr>
<td>Experience</td>
<td>3</td>
</tr>
</tbody>
</table>

Graphical Display

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ</td>
<td>*******************</td>
<td>*</td>
</tr>
<tr>
<td>MOTIV</td>
<td>****************</td>
<td>*</td>
</tr>
<tr>
<td>SKILL</td>
<td>*******************</td>
<td>*</td>
</tr>
<tr>
<td>EXP</td>
<td>*******************</td>
<td>*</td>
</tr>
</tbody>
</table>

Figure 2. Illustration of the numerical and graphical formats.
values for the cues was varied, the maximum value for each cue was
different and the raw scores had to be transformed in order for them to
be represented on equivalent scales. Each of the scales was
constructed such that one unit on it corresponded to one unit on the
cue intelligence. Consequently, the remaining three cues
(motivation, skill, and experience) were scaled upwards by factors of
5, 2, and 10. To illustrate how the scores were transformed, suppose
that the numerical scores on intelligence, motivation, skill, and
experience were 35, 6, 16, and 3 respectively. For the graphical
representation these scores would be represented as 35, 30, 32, and 30
units on a common scale (see Figure 2).

Design. A simple 2 x 2 design was used with display format and
type of decision task as the two within-subjects variables. Thus
display format, which consisted of numerical or graphical
representation of information, was combined factorially with type of
decisions required of subjects, viz., a rating or a choice task.

Both rating and choice tasks under a particular format were
performed in a block such that half of the subjects were presented with
a numerical block of profiles followed by graphical ones and the
opposite was true for the remaining subjects. The order in which the
rating and choice tasks were performed within each of these blocks was
counterbalanced. Half of the subjects were presented with two
particular sets of profiles in numerical form (sets p and r) and the
other two sets in graphical form (sets q and s) whereas the reverse
assignment was used for the remaining subjects. A total of eight conditions was thus generated (see Table 3).

**Subjects.** Forty-eight subjects were recruited from undergraduate psychology courses at Rice University. They either received $4.00 or course credit in exchange for their participation. An equal number of subjects was assigned randomly to one of the eight conditions.

**Procedure.** Subjects participated in the experimental sessions individually or in groups of 3-6 for about one hour. They were given detailed procedural instructions, with special attention paid to the characteristics of the cues and the way they were represented under the two display formats. Subjects then performed the rating and choice task under each format in a sequence determined by the condition to which they were assigned.

Subjects rated each profile paced by a beeper tone that sounded at 8-second intervals; they were allowed 16 seconds to choose from each pair of profiles. Both the rating and choice responses were written on separate response sheets.

**RESULTS AND DISCUSSION**

The effects of format were evaluated separately for the two response modes and therefore will be described separately in the following sections.

**Rating Task**

Of the 100 profiles rated under each format, 10 at either end were used as buffer profiles, and responses to these were not analyzed. The
### Table 3

Eight conditions used for counterbalancing various factors (Experiment 1).

<table>
<thead>
<tr>
<th>Successive Block</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) $N_{R_P}$ $N_{C_r}$ $G_{R_q}$ $G_{C_s}$</td>
</tr>
<tr>
<td>(2) $N_{C_r}$ $N_{R_P}$ $G_{C_s}$ $G_{R_q}$</td>
</tr>
<tr>
<td>(3) $G_{R_q}$ $G_{C_s}$ $N_{R_P}$ $N_{C_r}$</td>
</tr>
<tr>
<td>(4) $G_{C_s}$ $G_{R_q}$ $N_{C_r}$ $N_{R_P}$</td>
</tr>
<tr>
<td>(5) $N_{R_q}$ $N_{C_s}$ $G_{R_P}$ $G_{C_r}$</td>
</tr>
<tr>
<td>(6) $N_{C_s}$ $N_{R_q}$ $G_{C_r}$ $G_{R_P}$</td>
</tr>
<tr>
<td>(7) $G_{R_P}$ $G_{C_r}$ $N_{R_q}$ $N_{C_s}$</td>
</tr>
<tr>
<td>(8) $G_{C_r}$ $G_{R_P}$ $N_{C_s}$ $N_{R_q}$</td>
</tr>
</tbody>
</table>

Note: The single letters denote type of format (Numerical/Graphical) and the subscripts denote type of decision task (Rating/Choice) and set of information ($p$, $q$, $r$, $s$).
buffer profiles at the beginning were included to familiarize subjects with the task and to allow them to develop a consistent rating strategy; those at the end were included to reduce any effects of inattentiveness that might occur toward the end of a session (Lane et al., 1982). For every subject, a separate policy equation was obtained for the numerical and graphical displays by regressing each type of judgment on the four cues.

The raw score regression weights obtained for the four cues under each condition are an index of the subjects' weighting of those cues under that particular display condition. Thus, one way of determining the effect of format on decision strategies is simply to compare the regression weights using a repeated measures ANOVA. However, since the cues were presented on different scales, it was considered essential to make the raw score weights comparable by multiplying them by the SD of each cue. These adjusted weights represent the magnitude of change in the criterion produced by a change of one SD unit in the predictor (cue) (Lane et al., 1982). The mean adjusted weights are shown in Table 4.

Both the effect of cue and the cue x format interaction were significant, $F (3, 141) = 90.84$ and $8.10$ respectively, $p$ in each case $< .0001$. The main effect of format, however, was not significant, $F (1, 47) < 1$. Clearly subjects weighted the four cues differently—intelligence received the highest weight followed by motivation, skill, and experience. Although the average weight for the cues did not differ between format, the magnitude of weights attached
Table 4

Mean raw score regression weights (adjusted) for the four cues under the numerical and graphical formats (Experiment 1).

<table>
<thead>
<tr>
<th>Cues</th>
<th>Intelligence</th>
<th>Motivation</th>
<th>Skill</th>
<th>Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical Format</td>
<td>1.12</td>
<td>.57</td>
<td>.37</td>
<td>.03</td>
</tr>
<tr>
<td>Graphical Format</td>
<td>.97</td>
<td>.70</td>
<td>.33</td>
<td>.15</td>
</tr>
</tbody>
</table>

Note: Adjusted regression weights were obtained by multiplying the raw score weights and standard deviations of cue values to equate scale differences among the four cues.
to each cue changed reliably with format. Individual t-tests were conducted to test for differences in the weights for each of the four cues. The weight for intelligence was reliably smaller for the graphical than for the numerical display, \( t(47) = 2.29, p < .05 \); those for motivation and experience were reliably larger, \( t(47) = 2.60, p < .02 \) and \( 4.90, p < .01 \) respectively. The slight increase in the weight for skill was not significant, \( t(47) < 1 \). What is particularly noteworthy about the observed changes is that the fourth cue, experience, which had virtually no impact on judgment under the numerical format did receive some weight when displayed graphically. Coupled with this increase, the decrease in the highest weight (for intelligence) produced a more even distribution of graphical cue weights than did the numerical display.

Interestingly, the finding that the graphical display resulted in a tendency to weight cues evenly is consistent with Hammond's Cognitive Continuum theory. Hammond has suggested that a pictorial or graphical display is generally intuition-inducing and should therefore produce more equal cue weighting than the analysis-inducing numerical display. While the cue weights obtained in this study were far from equal (as is evidenced from a significant effect of cues), the trend was in the right direction and, of course, the theory does not require that any graphical display induce purely intuitive processing. It will be recalled that most task features are seen to induce quasi-rational functioning that differs in the level of intuition and analysis brought to bear on a judgment. Presumably the particular graphical format used
here may have been more intuition- than analysis-inducing relative to the numerical format, thereby providing greater, yet far from complete, equalization of cue weights. While, as noted earlier, Hammond's theory has deficiencies as an explanatory device, the present data are consistent with it.

The suggestion that a graphical format leads subjects to consider and weight all cues should be tempered by the observation that differences in the scales used to represent cues was confounded with format in the present study. To obtain maximum sensitivity in detecting format differences, the numerical cues were presented on different scales, each with its own mean and SD whereas the graphical cues were presented on identical scales and thus were transformed for presentation. Such rescaling of cues directly affected their variability. Consequently cues with lower variability (viz., motivation, skill, and experience) may have appeared to be more "scattered" in the graphical format thereby inflating their cue weights relative to the numerical format. This possibility, of course, represents an alternative explanation of the results—particularly with respect to the finding that experience, which had the lowest variance, was weighted substantially more heavily under the graphical than the numerical display. This confound was addressed directly in Experiment 2.

Any difference between the displays would likely be obscured in a comparison of mean weights if some subjects weighted the cues in one direction under one display and in the reverse direction under the
other one, while the remaining subjects did just the opposite. Therefore, the effect of display on the regression weights was tested individually for all subjects. This was accomplished by coding format as an additional predictor (see Cohen & Cohen, 1975, for a discussion of contrast coding) and then regressing judgments on the following 9 predictors: the four cues, coded vector representing format and the four cross-products of the cue values and coded vector. Out of 48 subjects, 38 showed a significant effect of format, $p < .01$. An overall probability value was obtained by combining individual probability values (Winer, 1971, p. 49) and the combined test was significant, $\chi^2(96) = 606.83$, $p < .001$. A similar test was conducted to evaluate the pooled effect of the interaction between format and each of the four cues, and once again, the combined test for cue x format interaction was significant, $\chi^2(96) = 778.66$, $p < .001$. At the individual level, there was a significant interaction for 24 subjects at $p < .01$ and for 9 subjects at $p < .05$. These findings merely confirm the earlier conclusion that format significantly affects subjects' cue utilization.

Besides regression weights (which index subjects' cue utilization), another useful descriptive measure is the linear consistency of individual policies viz., the squared multiple correlation or $R^2$'s obtained from regressing judgments on the four cues. The difference between $R^2$'s obtained from the numerical and graphical policies (0.64 vs. 0.69) was not significant, $F(1, 47) = 2.69$, $p > .10$. However, numerical judgments tended to be more variable than
graphical judgments. The mean SDs for numerical and graphical judgments were 1.74 and 1.58 respectively, $t(47) = 3.63$, $p < .001$. Note that $R^2$s = $SS\hat{y} / (SSy)$ or $(SS\hat{y} + SSE)$ and this proportion did not differ significantly between formats. On the other hand, the greater variability in numerical judgments implies a larger SSy for those judgments compared to graphical ones. A comparison of the two components of $R^2$s—SS$\hat{y}$ and SSE—for the two formats thus seemed necessary. However, since the distribution of SS tends to be skewed, the square roots of SS$\hat{y}$ and SSE were compared. Although $\sqrt{SSE}$ was greater for numerical than for graphical policies (8.90 vs. 7.51), $t(47) = 3.59$, $p < .001$, the difference in $\sqrt{SSy}$ (12.19 vs. 11.61) failed to reach statistical significance, $t(47) = 1.07$, $p > .10$. It is therefore apparent that the variability in numerical judgment occurred primarily through a larger amount of error associated with this format than with the graphical format. That is, numerical judgments showed less precision than graphical ones.

The analyses discussed so far assume that subjects' policies could be described solely in terms of a linear model. In view of previously cited evidence of occasional nonlinearity (e.g., Einhorn, 1971; Wiggins & Hoffman, 1968), a quadratic and a configurational model were also applied to subjects' judgments. In addition to the four cues (Xi) and coded format vector, the quadratic model included as predictors the four squared values of cues ($X_i^2$) and their interaction with the coded vector; the configurational model included 11 cross-products of cues ($XiXj$) and their interactions with the coded vector. Ten subjects showed a
significant quadratic and configural effect of cues, four subjects showed a significant quadratic effect, and 14 showed a significant configural effect, all at $p < .05$. However, the amount of increase in these subjects' $R^2$s achieved by using a nonlinear policy was rather small—the linear model alone accounted for 91.90% and 90.80% of the total variance accounted for by the quadratic and configural models respectively. Thus it appears that these subjects could be described largely in terms of a linear model even though they showed some nonlinearity.

For present purposes the nonlinearity of subjects' policies is not as important as the extent to which it varies as a function of display format. For each of the four cues, therefore, subjects who showed a significant interaction of format with the quadratic effect of that cue at $p < .05$ were isolated for further analyses. Note that these data are included merely for descriptive purposes since they pertain to a relatively small number of subjects and represent idiosyncratic rules for those subjects. Despite the individual nature of the data, there were some consistent trends in the quadratic usage of the cue perceived as most important, i.e., intelligence. Of the 10 subjects who showed a significant interaction of intelligence with format, nine showed a more pronounced quadratic form on graphical judgments than numerical ones. Thus the quadratic function describing their graphical judgments was more positively accelerated than the one describing their numerical judgments. In other words, higher values of intelligence produced more change in subjects' judgments than did lower ones under the graphical
than the numerical display. There were no interpretable trends
describing the quadratic use of the remaining cues as a function of
format. It suffices to say that four subjects showed a significant
interaction of motivation with format, four more of skill with
format, and two of experience with format, all at $p < .05$.

Since nonlinearity was also assessed using a configural model, one
can examine whether configural usage of cues varied systematically
with display format. However, testing each of the 11 configural cue x
format interactions for significance would result in an inflated Type I
error rate. Therefore subjects who showed an overall configural cue
x format interaction at $p < .05$ were selected and their individual
data were then examined further for any specific changes in configural
cue use with format. Only eight subjects showed an overall significant
interaction: Of the 11 configural cue x format interactions examined
for each of these subjects, only two interactions were significant for
one subject at $p < .05$. Consequently these data did not appear to
justify interpretation.

**Choice Task**

The primary question of interest here was whether choice
performance differed significantly with format. Since there was no
external criterion available to score choice accuracy, performance
could not be evaluated directly in terms of percentage of correct
choices. What was done instead was to use subjects' numerical and
graphical rating policies to predict choices. This was done by
applying "policy captured" weights to stimulus values of each pair of
choice profiles to determine the profile that would be chosen on the basis of the individual policies. These predicted choices were in turn compared to actual choices under the two formats to obtain "accuracy" measures under a consistent condition (e.g., "numerical" policy based choices compared to actual "numerical" choices) or an inconsistent one (e.g., "numerical" policy based choices compared to actual "graphical" choices). These accuracy scores were analyzed in a 2 x 2 ANOVA design with format and consistency of rating policy as the two within-subjects variables. The mean accuracy scores are reported in the first two rows of Table 5.

The effect of format and the interaction between format x consistency of policy were not significant, $F(1, 47) = .58$ and 1.60, $p = .45$ and .21 respectively. This suggests that despite the differences in subjects' rating policies under the two formats, they predicted choices with similar levels of accuracy. There was, however, a significant effect of consistency, $F(1, 47) = 9.43$, $p < .01$.

Although the differences were extremely small, a consistent policy predicted slightly better than an inconsistent one. This implies that subjects' rating and choice behavior were more similar when information was displayed in identical than in different formats. Thus while numerical and graphical cues were processed differently, the same display mode induced similar kinds of processing for both rating and choice tasks.

In order to establish the efficacy of regression-based judgment models in predicting choices, it was necessary to compare the accuracy
Table 5

Choice Accuracy based on a comparison of subjects' actual choices under the two formats and choices predicted from their regression models and a unit-weighted model.

<table>
<thead>
<tr>
<th></th>
<th>Graphical Format</th>
<th>Numerical Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistent rating policy</td>
<td>83.96</td>
<td>83.54</td>
</tr>
<tr>
<td>Inconsistent rating policy</td>
<td>80.63</td>
<td>82.86</td>
</tr>
<tr>
<td>Unit-weighted model</td>
<td>77.55</td>
<td>80.42</td>
</tr>
</tbody>
</table>
of such predictions to those of another baseline model. A simple unit-weighted model was used for this purpose. Predictions from a unit-weighted model were obtained separately for the numerical and graphical profiles and compared to subjects' actual choices under those formats. The resulting accuracy scores are reported in the third row of Table 5. For both graphical and numerical formats, the consistent rating policy predicted better than a unit-weighted model, $t (47) = 5.06$ and $2.66$, both $p < .05$. However, the inconsistent policy predicted better than the unit-weighted model only for the graphical format, $t (47) = 2.35$, $p < .05$; for the numerical format, the difference was not reliable, $t (47) = 1.95$, $p > .05$. The finding that the consistent policy for both formats fared reliably better than the inconsistent one corroborates the conclusion that subjects processed stimulus profiles similarly (regardless of task) under a particular display. It also implies that the regression policy did capture something important about the subject's behavior under one particular display format.

In summary, the major conclusion to be drawn from this study is that display format does induce differences in the way people handle predictive data.

**EXPERIMENT 2**

The findings of Experiment 1 suggested a difference in pattern of cue weighting for numerical and graphical formats. More specifically, there appeared to be a tendency for the graphical format to produce a more even weighting of cues than the numerical format. Whether this
pattern of cue weighting resulted from inherent differences in the manner in which cues were processed or from the scale differences underlying the two formats was unclear. Of course, such features are themselves an aspect of display formatting, although it was not the aspect to which Experiment 1 was primarily addressed.

Therefore, Experiment 2 sought to remove the confounding of scale with display format effects. The design was similar to that of Experiment 1 except that subjects performed only the rating task under the numerical and graphical formats. Moreover, both numerical and graphical information were presented using comparable scale units. The main purpose of this experiment, then, was to evaluate the effect of format on the judgment of otherwise strictly equivalent information.

METHOD

Materials and Design. Subjects performed a rating task similar to the one described in Experiment 1. They reviewed profiles of hypothetical applicants for the job of secretary and rated them on a suitability scale that ranged from 1 (extremely low) to 10 (extremely high). Each profile contained information about the applicant’s intelligence, motivation, social skill, and typing ability.

Two sets of 100 profiles (p, q) were generated in a manner identical to that in Experiment 1 except that values on all 4 dimensions—intelligence, motivation, skill, and typing ability—were sampled from populations with a mean of 25 and SD of 15. The scores were generated randomly subject to the constraint that they ranged
between 1-50. Both profile sets were represented numerically and graphically.

The format in which the information was displayed was varied within subjects: Under the numerical format the cue values were presented as numerical scores, whereas under the graphical format they were presented as horizontal bar graphs (refer to Figure 2). The presentation of the cues under the numerical and graphical cues was the same as Experiment 1 except that units on the various graphical scales were of identical size. The sequence in which the numerical or graphical information was displayed was counterbalanced such that half of the subjects rated numerical followed by graphical profiles and the other half rated them in the reverse order. The specific sets of profiles (p and q) used in the numerical and graphical conditions were rotated so that they were represented in both formats. Such counterbalancing resulted in four conditions that are described in Table 6.

**Subjects.** Twenty subjects from Rice University served in the experiment for course credit toward undergraduate psychology courses or for pay. They participated in the experimental sessions individually or in groups of 3-6. The assignment of subjects to the conditions described in Table 6 was randomized with the restriction that an equal number of subjects appear in each condition.

**Procedure.** After initial instructions regarding the task, subjects received individual booklets in which the profiles to be rated were printed. The sequence in which the numerical and graphical
Table 6

Four conditions used to counterbalance various factors (Experiment 2).

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>(N_p)</td>
</tr>
<tr>
<td>(2)</td>
<td>(G_q)</td>
</tr>
<tr>
<td>(3)</td>
<td>(N_q)</td>
</tr>
<tr>
<td>(4)</td>
<td>(G_p)</td>
</tr>
</tbody>
</table>

Note: The single letters denote type of format (Numerical/Graphical) and subscripts denote set of information \((p, q)\).
profiles were presented and also the specific set of profiles reviewed was determined by the condition to which the subject was assigned. Subjects reviewed and rated each of the 100 profiles in the numerical and graphical format at the rate of eight seconds per profile. A 5-minute rest interval was interposed between the rating of the two sets.

RESULTS AND DISCUSSION

Again as in Experiment 1, subjects rated 100 profiles under each format and judgments to 10 buffer profiles at either end were not analyzed. Thus every subject's policy equation for the two formats was based on judgments to the remaining 80 profiles. The mean raw score regression weights obtained through policy capturing are shown in Table 7. An ANOVA applied to these weights showed a marginal effect of format, $F (1, 19) = 3.68, p = .007$; a significant effect of cue, $F (3, 57) = 4.33, p = .008$; and a significant cue x format interaction, $F (3, 57) = 2.88, p = .04$.

The weights attached to the four cues were quite different, but the average weight across cues differed only marginally between formats. Despite the lack of a main effect of format, there was a reliable difference in the way subjects weighted the individual cues under the two formats. Thus, format again produced a differential weighting of cues. In order to describe this interaction precisely, individual t-values were obtained on the difference in the mean weights under the two formats for each cue. Only one of the comparisons was significant: the weight for motivation displayed graphically was
Table 7

Mean raw score weights for the four cues under the numerical and graphical formats (Experiment 2).

<table>
<thead>
<tr>
<th>Cues</th>
<th>Intelligence</th>
<th>Motivation</th>
<th>Social Skill</th>
<th>Typing Ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical Format</td>
<td>.82</td>
<td>.73</td>
<td>.46</td>
<td>.78</td>
</tr>
<tr>
<td>Graphical Format</td>
<td>.75</td>
<td>.94</td>
<td>.57</td>
<td>.93</td>
</tr>
</tbody>
</table>

Note: Since the scales had equal SDs, no adjustment was necessary for the analyses. The means in the Table are, however, multiplied by the SDs to make them comparable to those of Experiment 1.
reliably larger than that displayed numerically, $t(19) = 2.53$, $p < .05$. However, the format differences for intelligence, skill, and typing ability failed to achieve significance, $t(19) = 1.05$, $p > .10$, $1.87$, $p < .10$ and $1.46$, $p > .10$ respectively. Thus the cue x format interaction does not seem to reflect a consistent tendency across subjects to weight particular cues more heavily under one format than under the other.

As Table 7 shows, both the magnitude of the weights and their order of importance differed between formats. Moreover, the pattern of weights obtained from Experiment 1 was not replicated in these data. More specifically, the numerical format did not result in extreme differences in cue weighting and the graphical one in a more even weighting of cues. Rather, the weighting of individual cues differed with format, but in no simply described pattern.

The effect of display was analyzed individually for each subject by including as additional predictors in his policy equation a coded format vector and its interaction with the four cues. The tests of the effect of format and cue x format interaction based on aggregated individual probability values were both significant, $\chi^2(40) = 129.09$ and $137.56$, $p < .001$ respectively. At the individual level, nine of the 20 subjects showed a significant effect of format and nine showed a significant cue x format interaction, all at $p < .05$. These analyses serve to corroborate the conclusion that subjects weighted cues differently for the two displays.
Turning to the consistency of judgments ($R^2$s), numerical policies were less consistent than graphical ones (0.54 vs. 0.67); this difference was reliable, $F(1, 19) = 8.13, p < .01$. The SDs of the judgments, however, did not differ (1.90 vs. 1.89), $t(19) < 1$. $R^2$s was broken down into its two components, $SS^y$ and SSE, and a separate comparison of $\sqrt{SS^y}$ and $\sqrt{SSE}$ was made for the two formats. The mean $\sqrt{SS^y}$ for the numerical and graphical judgments were 12.12 and 13.72, with a t-test on this difference showing $t(19) = 1.86, p < .10$; $\sqrt{SSE}$ for numerical judgments was larger than that for graphical (11.21 vs. 9.47), $t(19) = 2.57, p < .02$. It thus appears that the lower consistency of numerical judgments resulted largely from greater error in those judgments than in the graphical ones. The finding that the numerical format produced lesser precision than the graphical one parallels that of Experiment 1. A summary of $R^2$s and its component measures for the two Experiments is provided in Table 8.

As in Experiment 1, a quadratic and a configural model were applied to subjects' judgments to test for nonlinearity of policies. Out of 20 subjects, seven showed a significant quadratic and configural effect of cues and four showed a significant configural effect, all at $p < .05$. Although these subjects appeared to use the cues nonlinearly, the linear model alone accounted for a substantial proportion of the variance accounted for by the quadratic and configural models (87.98% and 85.27% respectively).

The data were examined further to determine whether there were any systematic relationships between the nonlinear--quadratic or
<table>
<thead>
<tr>
<th></th>
<th>Experiment 1</th>
<th></th>
<th></th>
<th>Experiment 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Numerical</td>
<td>Graphical</td>
<td>Numerical</td>
<td>Graphical</td>
<td>Numerical</td>
<td>Graphical</td>
</tr>
<tr>
<td>$R^2_s$</td>
<td>.64</td>
<td>.69</td>
<td>.54</td>
<td>.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>1.74</td>
<td>1.58</td>
<td>1.90</td>
<td>1.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sqrt{SSy}$</td>
<td>12.19</td>
<td>11.61</td>
<td>12.12</td>
<td>13.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sqrt{SSe}$</td>
<td>8.90</td>
<td>7.61</td>
<td>11.21</td>
<td>9.47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
configural—use of cues for the two formats. However no consistent trends emerged, largely due to the small number of interactions observed between format and quadratic/configural values of cues. Parenthetically it might be noted that only three subjects showed an interaction at $p < .05$.

Summarizing the first two experiments, it appears that format does influence the manner in which people weight cues, but the nature of this influence is not simply described. It may, in fact, be quite idiosyncratic. Nonetheless, one generalization does emerge: judgment is less consistent under the numerical format, and this is attributable chiefly to the lower precision of numerical judgments relative to graphical ones.

EXPERIMENT 3

The previous experiments suggest that the format in which information is displayed influences subjects' judgments. There is some evidence (see below) suggesting that a graphical form of presentation encourages a more holistic type of processing than does a numerical display. Since holistic processing would require that all cues are presented simultaneously, this experiment explored whether cue weighting differences under the two formats would persist when cues are presented sequentially rather than concurrently.

There is a large body of literature on the effect of structural properties of stimuli on perceptual tasks (Garner, 1974). One of the consistent findings from these studies is that some stimulus dimensions are perceived holistically (integral dimensions), while others are
perceived individually (separable dimensions). For example, the height and width of a rectangle are combined holistically to produce perception of rectangular area (Felfoldy, 1974; Garner & Felfoldy, 1971; Lockhead, 1979). Later investigations have generalized this result to include decision tasks, more particularly in the MCPL paradigm (Goldsmith & Schvaneveldt, 1982; Wickens & Scott, 1983).

It was postulated that the graphical format in the present experiments might produce a tendency toward a holistic perception of cues, whereas the numerical format might produce a more serial form of processing. If, then, cues were presented sequentially rather than simultaneously, the holistic perception of graphical cues would be largely eliminated and, as a consequence, so would the difference between the graphical and numerical formats.

The present experiment, therefore, involved a sequential presentation of cues under numerical and graphical formats. One additional manipulation involved the number of cues presented to subjects. Previous investigators (e.g., Einhorn, 1971) have found that increasing the "information load" increases the difficulty of integrating cues and thus is detrimental to performance. In a sequential presentation of cues, subjects are obliged to rely heavily on memory in making their judgments or choices, thus exacerbating the difficulty. By varying the number of cues, therefore, the interest was to provide an adequate range of task difficulty for the appearance of any potential format effects.
METHOD

Materials and Design. As in Experiments 1 and 2, subjects rated multidimensional stimuli on a global dimension. However, the present task involved teaching effectiveness judgments instead of personnel selection/rating tasks used previously. The primary reason for this change was to explore the generality of display effects in another realistic judgment context, while preserving the formal properties of the task. The stimuli consisted of profiles of hypothetical college instructors whose performance was described with respect to either four or six cues and values of the cues were displayed either numerically or graphically. The design, then, involved two variables—number of cues (four or six cues) that was manipulated between-subjects and combined factorially with display format (numerical vs. graphical) that was manipulated within-subjects.

A set of 200 profiles was generated in a manner identical to the one described in Experiment 1 except that (1) six cue values were generated per profile, and (2) all cues were sampled from populations with a mean of 30, a SD of 10, and a range of 1-60. The six cues describing the profiles were designated as information imparted in course, arousal of interest, presentation style, knowledge of the field, rapport with students, and clarity of course requirements. Subjects in the four-cue condition were presented with only a subset of the six cues described earlier; moreover, the exact subset of cues sampled differed among subjects. The order in which information under the two formats was displayed was counterbalanced in both the four- and six-cue conditions. Thus half of the subjects rated numerical profiles
followed by graphical profiles and the reverse was true for the remaining subjects.

The numerical and graphical presentation of cues was similar to Experiments 1 and 2. As in Experiment 2, all cues had identical scale units so that no transformation of the cues was necessary for graphical presentation. For every subject 100 profiles were chosen randomly for graphical presentation and the remaining 100 profiles were presented numerically. The order of presentation of the cues (for the four- and six-cue conditions) and the selection of four of the six cues (for the four-cue condition) was similarly randomized individually. Any given subject, however, reviewed the same cues in a specific order under both types of display format.

Subjects. The experiment was conducted in individual sessions that lasted for about an hour. Twenty subjects, enrolled in undergraduate psychology courses at Rice University, participated in the experiment in exchange for course credit or pay.

Procedure. The subject was seated in a cubicle before the screen of a TRS-80 (Model II) microcomputer. After initial instructions regarding the task and procedure, the profiles of instructors were displayed on the screen of the computer, one profile at a time. Each profile was presented in the following manner. First the words "Instructor #" appeared on the screen along with the number of the profile being rated. This message served primarily as a preparatory signal for subjects to attend to the incoming information and also to distinguish one profile from another. Then the cues were
presented successively in an order that was determined randomly for every subject: Each cue was presented with its label and value for 2 seconds and the subject saw either four or six cues depending on group assignment. After all cues were presented, the instructions "Rate the instructor on a scale of 1 to 10" were displayed on an otherwise blank screen. The subject then proceeded to write down his/her rating on a separate response sheet. After the response had been recorded, the experimenter depressed a programmed key on the computer keyboard to present the next profile. Thus, although the time of presentation of cues was controlled, subjects' responses were essentially self-paced. Subjects rated numerical profiles followed by graphical profiles or vice versa depending on the condition to which they were assigned. A brief rest period intervened between the ratings of two sets of profiles.

RESULTS AND DISCUSSION

The principal issue addressed in this experiment was the effect of sequential presentation of cues on subjects' judgments under numerical and graphical display formats. The expectation was that the sequential procedure would eliminate the display effect.

Of the 100 profiles rated under each format, ratings of 10 buffer profiles at either end were not analyzed. Both numerical and graphical policies for each subject were then obtained by regressing the 80 judgments on four or six cues depending on the number of cues the subject was presented with. The raw score regression weights from the
numerical and graphical policies were then analyzed in two ways: for order effects and for specific cue effects.

The first type of analysis pertained to the weights attached to the sequential position of successively presented cues. Note that order of presentation of specific cues (e.g., information imparted in course or presentation style) was randomized individually so that the effect of order was independent of cue over subjects. Therefore, the format and order effects were tested together in separate ANOVAs applied to the four- and six-cue data. The mean weights for the cues are presented in Table 9.

Looking first at the four-cue ANOVA, the means for the numerical and graphical formats were .64 and .52 respectively, a difference that was significant, $F(1, 9) = 11.01$, $p < .01$. This implies that sequential presentation of the four cues did not eliminate entirely the effect of format. However, the cue x format interaction was not significant, $F(3, 27) < 1$. Given that this interaction defined format difference under a simultaneous presentation of cues (Experiments 1 and 2), the failure to find it in the present data deserves consideration. It suggests that holistic processing of graphical cues was eliminated largely under a sequential presentation of cues. But, obviously, such a conclusion can only be tentative due to the inherent danger in accepting the null hypothesis. One secondary result concerns temporal difference in processing of cues. Turning to the four-cue condition in Table 9, cues in positions 3 and 4 tend to be weighted more heavily than those in positions 1 and 2 and thus provide
Table 9

Mean raw score regression weights for the successively presented cues.

<table>
<thead>
<tr>
<th>Number of Cues</th>
<th>Format</th>
<th>Order of Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Four</td>
<td>Numerical</td>
<td>.55</td>
</tr>
<tr>
<td></td>
<td>Graphical</td>
<td>.38</td>
</tr>
<tr>
<td></td>
<td>$\bar{X}$</td>
<td>.46</td>
</tr>
<tr>
<td>Six</td>
<td>Numerical</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>Graphical</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>$\bar{X}$</td>
<td>.32</td>
</tr>
</tbody>
</table>

Note: The means in the Table are adjusted by the SDs of the cues to make the data consistent with those from Experiments 1 and 2.
a hint of a recency effect—a tendency for recent information to be weighted more heavily than earlier information. This observation was only partially supported statistically and the effect of order was marginally significant, $F(3, 27) = 2.55, p = .08$. Of course, it might be recognized that the variance of the weights is likely to be large in such an analysis. Note that the weighting of cues would differ due to their inherent importance. For example, knowledge of the field might be weighted more highly in judging teaching effectiveness than clarity of course requirements. Now since order of presentation of specific cues was individually randomized, the mean for each cue-position would represent an aggregate of cues with different importance thereby increasing the variance (which constitutes the error variance). The failure to find a reliable effect of order, therefore, does not preclude completely the idea that some order effects were present. Perhaps, including cues that were perceived as being equally important might have accentuated these effects. There was virtually no evidence that order effects differed with respect to format: the format x order interaction was nonsignificant, $F(3, 27) < 1$.

The six-cue ANOVA also failed to reveal a significant cue x format interaction, $F(5, 45) = 1.12, p = .37$, thereby supporting the claim that a sequential presentation eliminated the holistic processing of graphical cues. However, the main effect of format found in the four-cue condition was absent here, $F(1, 9) < 1$. Exactly why this
should occur is not clear. There was also no suggestion of the presence of order effects, $F(5, 45) = 1.23$, $p = .31$.

The analyses discussed so far determined whether the processing of cues was affected by the format in which they were presented and their temporal ordering. The second analytic approach was based on cues irrespective of order, the purpose being to establish whether the particular cues were weighted differently under the two formats. This analysis was possible only under the six-cue condition since the four-cue condition did not provide all subjects with the same subsets of cues. Since, the effect of format or the cue x format interaction were not significant, $F(1, 9) < 1$ and $F(5, 45) = 1.07$, $p = .39$, the regression weights were collapsed across format and these means are presented in Table 10. As is apparent from Table 10, there was clearly an effect of cue, $F(5, 45) = 3.77$, $p = .006$. As one would expect, some cues were weighted more heavily than others.

In sum, it appears that a sequential presentation of cues did not affect the differential weighting of cues under the two formats—the cue x format interaction present consistently in both the previous experiments was eliminated in this one. However, the numerical format still differed in overall cue weighting from the graphical format in the four-cue condition. Thus there is evidence that some processing differences as a function of format persisted even under a sequential presentation of cues, a point made even more apparent in the individual analyses to which we now turn.
Table 10

Mean raw score regression weights (collapsed across format) for the specific cues in the six-cue condition.

<table>
<thead>
<tr>
<th>Information imparted in course</th>
<th>Arousal of interest</th>
<th>Presentation style</th>
<th>Knowledge of field</th>
<th>Rapport with students</th>
<th>Clarity of requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.47</td>
<td>.56</td>
<td>.35</td>
<td>.32</td>
<td>.25</td>
</tr>
</tbody>
</table>
The effect of display was tested individually for every subject. It will be recalled from Experiments 1 and 2 that this analysis is done by regressing each subject's judgments on the following predictors: the individual cues, coded format vector, and interaction of each cue with the format vector. Thus subjects' judgments in the four-cue condition were regressed on 9 predictors (4 cues, coded format vector, and the interaction of the 4 cues with the format vector) and those in the six-cue condition were regressed on 11 predictors (6 cues, coded format vector, and interaction of the 6 cues with format) to obtain individual policy descriptions.

Of the 10 subjects in the four-cue case, eight showed a significant effect of format, \( p < .01 \); in the six-cue case, five showed a significant effect at \( p < .01 \) and 2 at \( p < .05 \). A combined test based on the individual probabilities was significant for both four- and six-cue conditions, \( \chi^2(20) = 137.70 \) and 109.94 respectively, both \( p < .001 \). Similar tests were conducted to test whether the overall interaction of format with cues was significant; the combined tests again reached significance; \( \chi^2(20) = 88.94, p < .01 \) for the four-cue and \( \chi^2(20) = 42.80, p < .05 \) for the six-cue case. At the individual level, five of 10 subjects showed a significant interaction \( (p < .05) \) under the four-cue condition; by contrast, only one of 10 did so under the six-cue condition.

The individual analyses of the data suggest that format produced a difference in overall cue weighting for more subjects than would be expected by chance in the case of both four and six cues. However,
the results of ANOVA showed an effect of format only for the four-cue condition. Thus one can conclude that the highly idiosyncratic nature of the judgment policies obscured the effect of format in the six-cue condition. Turning to the cue x format interaction—present consistently in Experiments 1 and 2—there is a slight indication that it was present, but it appeared to be limited only to a few subjects.

The consistency or $R^2$'s obtained from subjects' policies was compared in a 2 x 2 ANOVA design, with number of cues (four vs. six cues) as a between-subjects variable and format (numerical vs. graphical format) as a within-subjects variable. The mean $R^2$'s for the four- and six-cue conditions was .67 and .57 respectively and the decline in consistency as the number of cues increased from four to six was significant, $F(1, 18) = 8.74$, $p = .008$. However, neither the effect of format nor the number of cues x format interaction was significant, $F(1, 18) = 1.90$ and $1.62$, $p = .19$ and .22. The finding that an increase in information load affects $R^2$'s is supported by previous studies (e.g., Anderson, 1977; Billings & Marcus, 1983; Einhorn, 1971). However, lowered consistency could result either from a decrease in cue usage due to the greater amount of processing load imposed by additional cues (measured by $SS^\hat{Y}$) or an increase in random error (measured by $SSE$). Looking at these components, only $\sqrt{SS^\hat{Y}}$ differed for the four- and six-cue conditions (11.12 vs. 8.92), $F(1, 18) = 4.33$, $p = .05$; the difference between $\sqrt{SSE}$ was not reliable (7.74 vs. 7.58), $F(1, 18) < 1$. Format did not affect $\sqrt{SS^\hat{Y}}$ or $\sqrt{SSE}$ significantly and the number of cues x format interaction also failed
to approach significance for both measures. These findings suggest that subjects who had a larger set of cues to process (six-cue condition) tended to use the information less completely than those who had a smaller set (four-cue condition), consequently lowering the linear consistency of their policies. Whether the sequential presentation of cues imposed an additional memory load and caused a greater decrement in consistency for the former condition relative to a simultaneous presentation is not possible to determine from these data.

A quadratic and a configural model were used to describe nonlinearity of subjects' judgments. The quadratic model included either four or six cues (Xi), coded format vector, squared values of the cues (Xi²), and the interaction of format and Xi². Thus the quadratic model in the four-cue condition was based on 13 predictors and that in the six-cue condition on 19 predictors. The configural model was based on the cues (Xi), the coded format vector, all two-way interactions of cues (XiXj), and the interaction of format and (XiXj) resulting in 17 and 37 predictors for the four- and six-cue conditions respectively. It might be noted that for the four-cue condition a complete model using all higher-order interactions (in addition to two-way interactions) did not result in a significant increase in R² compared to a model that was limited to two-way interactions: A test of usefulness of the complete model (Darlington, 1968) did not approach significance (p < .05) for any of the subjects, with an aggregate of individual probabilities showing χ²(20) = 27.29, p < .10. A complete model based on all possible interactions
for the six-cue condition involves an extremely large number of predictors and affects the experimentwise Type I error rate (Cohen & Cohen, 1975). Thus a model comprising all two-way interactions was considered sufficient to test for configurality.

Considering first the quadratic model, five subjects from the four-cue condition and six from the six-cue condition showed an overall quadratic effect of cues at \( p < .05 \). However, the linear model alone accounted for a large percentage of the variance accounted for by the quadratic models in the two conditions (95.85% and 89.26% respectively). The same held true for the configural model. The number of subjects who showed a significant configural use of cues was five for the four-cue and two for the six-cue conditions. Again, the linear model accounted for 94.41% and 73.86% of the variance accounted for by the configural model. It is evident from these findings that both the quadratic and configural models did not add substantially to the explained variance and the linear model was adequate in describing subjects' policies.

To evaluate whether nonlinear use of cues changed systematically under the two formats, pooled tests for overall quadratic cue x format interaction and configural cue x format interaction were conducted. Five of the 10 subjects in the four-cue condition and one in the six-cue condition showed a significant quadratic cue x format interaction at \( p < .05 \). A significant configural cue x format interaction was observed for four subjects in the four-cue and three in the six-cue conditions.
The data from these subjects were examined further to determine whether there was any commonality in the way format interacted with quadratic or configural use of cues. For the most part the observed interactions were idiosyncratic and no consistent trends emerged. However, there was one exception. All subjects with a reliable overall quadratic cue x format interaction showed an interaction of the fourth cue and format at $p < .05$. None of the other individual cues interacted similarly with format. This finding implies that quadratic usage of the most recent information (viz., cue 4) varied with format, at least for those subjects who showed an overall reliable interaction. Unfortunately the exact nature of the effect was not apparent from the data.

To summarize, a sequential presentation of cues eliminated the format x cue interaction evident in Experiments 1 and 2. While this finding obviously is consistent with the idea that a graphical display induces a more holistic processing approach, its conclusiveness is diminished by the fact that format still tended to produce some processing differences, especially under the lighter processing load (four cues). A number of different kinds of analyses failed to produce any coherent explanation of the main effect of display, although all indications are that it is highly idiosyncratic across subjects. It is thus clear that a sequential presentation of cues did not eliminate format differences completely, although it may have destroyed one consistent aspect of processing in the graphical case—"holistic" processing of cues.
EXPERIMENT 4

The primary conclusion that can be drawn from the findings of the first three experiments is that numerical and graphical presentation of cues produced differences in cue weighting. More specifically, it appeared that a graphical format encouraged holistic processing under a simultaneous presentation of cues. In contrast, under a sequential presentation the weighting of cues differed on average under both formats. In addition, the numerical format produced less precision in judgments than the graphical format (Experiments 1 and 2); the other measures of performance (e.g., nonlinearity of policy) failed to show any consistent differences with format. Consequently, these experiments provide a useful demonstration of display effects in a policy capturing paradigm. However there remain questions regarding the practical implications of the nature of the induced difference.

The present study extended the results of the previous experiments to a MCPL paradigm. It will be recalled from the earlier discussion in the Introduction that besides affording measures of different aspects of subjects' policies obtained from the policy capturing studies, the MCPL paradigm provides information about other important aspects of performance. In particular, it allows an evaluation of the extent to which policy-based measures relate to external "optimal" criteria through measures like achievement or the matching coefficient. Since there was no information regarding how performance accuracy is affected by variations in format in the earlier experiments, it was not possible to assess the relative merits of the formats under
consideration. By using the MCPL paradigm such information could be obtained, a finding that would have considerable practical importance.

The basic variable manipulated in this experiment, then, was display format. Thus subjects were again presented with numerical and graphical cues as in the previous experiments and they made evaluative judgments based on the available information. They were subsequently given feedback about the actual criterion value that they were judging. Since the criterion values were based on an explicit model, the lens model measures could be obtained to index subjects' learning of the model under the two formats.

Another variable manipulated in the experiment was the type of weighting rule that related cues to the criterion in the optimal environmental model. It has been found that the weighting rule influences significantly the ease with which subjects can learn the optimal model. In general, linear combination rules or unit weighting rules produce quicker learning than nonlinear or differential weighting rules (e.g., Einhorn & Hogarth, 1975; Goldsberry, 1983; Summers & Hammond, 1966). Given that there was a consistent difference in cue weighting as a function of display in the previous experiments, it was felt that manipulating the nature of the weighting rule might interact with display and produce different patterns of learning.

Two other issues, although of secondary concern, warrant comment. First, most researchers using the MCPL paradigm have focused solely on correlations to measure accuracy of performance (e.g., achievement, matching coefficient). As emphasized earlier in the paper, there are
situations in which absolute accuracy of predictions is crucial. In the present study, then, deviation measures analogous to the conventional correlational measures were also used to measure absolute accuracy of performance.

Second, subjects' cue weighting was evaluated in the context of the phenomenon of conservatism. Conservatism simply implies that subjects do not change their judgments sufficiently to match the optimal values when they encounter new information. Thus, even if they change in the optimal direction, they act as if the data are less diagnostic than they really are. Although conservatism has been well documented in the normative study of decision making (Edwards, 1968; Rapoport & Wallsten, 1972), it has not been an explicit concern in MCPL research. The data obtained from the MCPL paradigm do permit an assessment of the generality of the phenomenon and the conditions under which it may/may not occur. The present study, therefore, explored how conservatism would be affected by a combination of display format and ease of weighting rule.

Turning to the methodological issue of how conservatism can be measured in correlational tasks, Brehmer and Lindberg (1970) have proposed that it can be measured by the relation between be and bs (i.e., the slopes of the regression lines relating the criterion values and judgments to the cues). Subjects tend to be conservative to the extent that they do not change their judgments in response to changes in cue values as much as is warranted optimally (viz., bs is smaller than be). Peterson, Hammond, and Summers (1965) found that subjects
were conservative in weighting the most valid of three cues, but tended to overweight the cue with lowest validity.

**METHOD**

**Task.** As in most MCPL studies, subjects were required to predict criterion values from separate information sources (cues) on several trials and following each prediction they received feedback about the actual criterion. Their task was to learn the cue-criterion relations such that their predictions matched the actual criterion values as the trials progressed. The actual scenario used here was designed on the basis of the following defining features: (1) unfamiliarity—subjects would not have preconceived notions regarding cue weighting that might influence their learning of the designated environmental model in unpredictable ways, (2) face validity—subjects would accept the task as realistic, and (3) ease of quantification—the to-be-presented cues and criterion could be easily quantified. The requirement that the task be unfamiliar and novel to subjects automatically ruled out the task scenarios used in the previous experiments. What was used instead was a military threat diagnosis task.3

Each subject assumed the role of a naval task force commander. He or she was given a map that illustrated a hypothetical base from which forces might be deployed and the location in that vicinity where the enemy forces prevailed. Their task was to monitor the following four pieces of information on several trials: movement information (information pertaining to ship density and traffic of ships around
enemy territory), EMT reconnaissance (electromagnetic reconnaissance or information obtained from electronic devices such as radar), threat propaganda (reports of threat summarized from local media--radio and press--and local intelligence sources), and instability report (report of instability in that area communicated directly from Headquarters).

Subjects studied information from these four sources on each trial and made a judgment concerning the strength of force they wished to deploy in view of potential threat of enemy attack. They were then shown the optimal strength of force they should have deployed (viz., criterion values) had they used the information (cues) correctly. In essence, subjects had to determine the relation between the four cues and the criterion values to match their own response to the optimal one.

**Stimuli.** Two sets of 210 stimuli were generated using different optimal weighting rules--an unequal or equal weighting of the cues--in determining the criterion values. Each set had four predictors (cues) that were sampled from populations with means of 25 and SDs of 3.5 and, further, were uncorrelated with each other. All cues ranged between 0-50. The target $R^2$ or the extent to which the criterion could be predicted from a weighted combination of cues was .80 and values of the criterion had a mean of 100 and SD of 35 in both sets (the values ranged from 0-200). The unequal weights were chosen such that their magnitude decreased ordinarily from the first cue to the fourth: the target weights were 7, 5, 2.5, and 1 respectively in the unequal condition and 4.5 in the equal condition. The unequal and equal
weights were displayed numerically or graphically depending on the group to which the subject had been assigned.

**Design.** As in the previous experiments, the primary variable of interest was display format. Display format, specifically numerical (N) or graphical (G) format, was crossed with type of optimal weighting rule, an unequal (U) or equal (E) weighting of the cues, in a 2 x 2 factorial design. Each of the resulting combinations was the basis for an experimental group. Thus subjects saw a numerical or graphical display of cues weighted unequally (NU and GU groups) or equally (NE and GE groups) in determining the criterion (see Table 11 for the complete design).

Each group responded to 210 numerical or graphical stimuli that had an unequal or equal weighting of four cues. Although the cue values in a stimulus set were displayed in the same order for all subjects, the names assigned to them (viz., movement info., CIMT reconnaissance, threat propaganda, and instability report) were randomized individually for every subject.

The first 10 stimuli were used as a practice set and judgments to these were not analyzed. The remaining 200 stimuli were split into four blocks of 50 stimuli each of which served as a level of an additional within-subjects variable (learning) for all analyses.

**Subjects.** Thirty-two subjects were assigned randomly to one of four experimental groups under the constraint that the groups be of equal size (n=8). Subjects participated in individual experimental
Table 11

Description of the four experimental groups that resulted from a manipulation of display format and type of optimal weighting rule.

<table>
<thead>
<tr>
<th>Display format</th>
<th>Type of weighting rule</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Numerical</td>
<td>Unequal</td>
<td>NU</td>
</tr>
<tr>
<td>2. Graphical</td>
<td>Unequal</td>
<td>GU</td>
</tr>
<tr>
<td>3. Numerical</td>
<td>Equal</td>
<td>NE</td>
</tr>
<tr>
<td>4. Graphical</td>
<td>Equal</td>
<td>GE</td>
</tr>
</tbody>
</table>
sessions that lasted for an hour. They received course credit or pay in exchange for their participation.

Procedure. The subject was seated in a cubicle before a TRS-30 (Model II) microcomputer. The initial instructions described the task, explained the nature of the cues and emphasized the importance of the subject's learning the cue-criterion relation in order to minimize the discrepancy between his own and the optimal response.

Each trial began with a brief presentation of the trial number on the screen. The four cues along with their numerical and graphical values were then presented. The display format of cues and the specific set of stimuli (cues weighted unequally vs. equally) used were determined by the experimental group to which the subject was assigned. In addition to the cues, the instructions "Your Response?" were printed on the screen of the computer requiring the subject to study the cues and respond. The subject's response indicated the number of units he/she wished to deploy given the information before him/her, and it could range anywhere from 0 to 200. The subject was told to write down the response on a separate sheet and also call it out aloud so that the experimenter could display it on the computer screen via the keyboard. Although instructed to respond quickly, subjects were allowed to answer at their own pace. After the subject's response appeared on the screen the optimal response was displayed. Irrespective of the display format, both the subject's and the optimal response were presented numerically. Once the optimal response appeared, subjects had six seconds to study the information on the
screen—it included the cues, the subject’s response and the optimal response. Subjects performed 210 such trials.

RESULTS AND DISCUSSION

Unlike the previous experiments, the present one was based on an explicit environmental model, and it was of interest to compare various aspects of subjects’ performance to external "accuracy" criteria. A number of correlational (see Table 1 for a summary) and other indices were computed which will now be discussed.

Regression Weights (accuracy). One direct way to measure the extent to which subjects learned the optimal cue-criterion relations is to compare the regression weights obtained from the subject’s policy to the optimal weights. The greater the discrepancy between the subject’s and optimal weights, the lower is the subject’s accuracy in weighting the cues. For each of the four experimental groups, the total signed and unsigned deviations of the raw score weights from the optimal weights was computed across the four cues and these deviations are shown in Table 12. The signed deviations provide information about any consistent tendencies toward underestimation (negative sign) or overestimation (positive sign) of optimal weights; the unsigned deviations indicate the absolute magnitude of error in the subject’s weights compared to the optimum.

Considering first the signed deviations, an ANOVA revealed significant effects for both display and weighting rule, \( F (1, 28) = 7.41 \) and \( 14.38, \) \( p < .01 \) and \(.001 \) respectively. As is apparent from Table 12, the signed error was lower for the numerical relative to the
Table 12

Total deviations of raw score regression weights (summed across all four cues) derived from subjects' policies and the optimal policies for the four experimental groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>Signed Deviation</th>
<th>Unsigned Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NU</td>
<td>-4.48</td>
<td>7.29</td>
</tr>
<tr>
<td>GU</td>
<td>-7.69</td>
<td>9.39</td>
</tr>
<tr>
<td>NE</td>
<td>-8.82</td>
<td>9.14</td>
</tr>
<tr>
<td>GE</td>
<td>-11.36</td>
<td>11.49</td>
</tr>
</tbody>
</table>
graphical format and for the unequal relative to the equal weighting rule. A systematic decrease in error occurred across learning blocks, $F (3, 84) = 11.24$, $p = 0.00$. None of the interactions was significant except for a rule x block interaction, $F (3, 84) = 4.68$, $p = .005$. For the equal weighting rule the amount of error continued to decrease after the first two blocks ($-11.26$, $-11.75$, $-9.73$, $-7.60$), whereas the unequal weighting group—although superior overall—showed a decline in error from the first block to the second, but not much improvement thereafter ($-8.10$, $-5.45$, $-5.40$, and $-5.39$).

It appears, then, that subjects showed a consistent tendency to underestimate the optimal cue weights and thus were overall quite conservative. The findings that subjects performed better under the numerical format than the graphical format and under the unequal than the equal rule are amplified when the amount of error is considered separately for the four cues. Since the labels attached to the cues were varied randomly, the individual cues are referred to as cue 1, cue 2, etc. The numbers pertain to the position of the cue on the screen; thus cue 1 was topmost on the screen, was followed by cue 2 and so on. It might be noted that in the case of the unequal weighting rule, the numbers correspond to the order of importance of cues (i.e., cue 1 had the heaviest optimal weight, cue 2 the second heaviest and so on).

Both the signed and unsigned deviations for each cue are presented in Table 13. In the case of signed deviations, there was a significant effect of cue, $F (3, 84) = 23.37$, $p = .000$. Both display and weighting rule interacted significantly with cue, $F (3, 84) = 4.23$
Table 13

Mean deviations of subjects' regression weights from optimal weights broken down by cue.

<table>
<thead>
<tr>
<th>Group</th>
<th>Signed Deviations</th>
<th></th>
<th></th>
<th></th>
<th>Unsigned Deviations</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cue 1</td>
<td>Cue 2</td>
<td>Cue 3</td>
<td>Cue 4</td>
<td>Cue 1</td>
<td>Cue 2</td>
<td>Cue 3</td>
<td>Cue 4</td>
</tr>
<tr>
<td>NU</td>
<td>-2.14</td>
<td>-2.01</td>
<td>-1.27</td>
<td>.94</td>
<td>2.26</td>
<td>2.18</td>
<td>1.58</td>
<td>1.27</td>
</tr>
<tr>
<td>GU</td>
<td>-4.64</td>
<td>-1.97</td>
<td>-1.34</td>
<td>.27</td>
<td>4.64</td>
<td>1.97</td>
<td>1.53</td>
<td>1.24</td>
</tr>
<tr>
<td>NE</td>
<td>-1.82</td>
<td>-1.72</td>
<td>-3.32</td>
<td>-1.95</td>
<td>1.96</td>
<td>1.90</td>
<td>3.32</td>
<td>1.96</td>
</tr>
<tr>
<td>GE</td>
<td>-2.86</td>
<td>-2.35</td>
<td>-3.52</td>
<td>-2.62</td>
<td>2.86</td>
<td>2.46</td>
<td>3.55</td>
<td>2.62</td>
</tr>
</tbody>
</table>
and 25.90, \( p = .008 \) and \( .000 \); and so did learning block, \( F (9, 84) = 2.05, p = .04 \). None of the three-way or four-way interactions approached significance. The error scores for the four cues under the numerical display were \(-1.98, -1.87, -2.29, \) and \(-.51 \); the graphical display showed quite a different pattern, the error scores being \(-3.75, -2.16, -2.43, \) and \(-1.17 \). While a systematic tendency to underestimate the optimal weight for each cue prevailed, what is most noteworthy is the lowest error for the fourth cue under both formats. Given the long standing reading habits of the subjects (i.e., reading from top to bottom), the minimum error on the fourth cue could be construed as indirect evidence for some variation of serial processing in which recency advantage predominates. However, the idea that a recency effect was observed can only be tentative since the cues were presented concurrently. After all, the same effect could be interpreted as a primacy effect under the assumption that information was processed from bottom to top. Nevertheless there was a clear difference in the way the fourth cue was processed compared to the others.

Turning to the significant weighting rule x cue interaction, an interesting pattern of results emerged. The unequal rule produced error scores of \(-3.39, -1.99, -1.31, \) and \(.60 \) for the four cues and the equal rule produced error scores of \(-2.34, -2.04, -3.42, \) and \(-2.28 \). It is evident that subjects' conservatism in cue weighting is directly related to the cue's validity. Thus for the unequal rule groups (for whom cue validities were varied), the cue with the highest validity was most underweighted and the degree of underestimation declined with the
cues' validities. In fact it decreased so much that the cue with the lowest validity was slightly overweighted. These findings are consistent with those of Peterson, Hammond, and Summers (1965). In contrast, cues with equal validities produced similar magnitude of error.

An ANOVA on the total unsigned deviations almost exactly replicated the pattern of results obtained for total signed deviations (see Table 12). Although the effect of display was significant, $F(1, 28) = 4.76$, $p = .04$, the effect of weighting rule was only marginally so, $F(1, 28) = 3.73$, $p = .06$. The numerical format produced less error than the graphical. The amount of error declined across learning blocks, $F(3, 84) = 12.02$, $p = 0.00$, and a significant rule x block interaction occurred, $F(3, 84) = 4.41$, $p = .006$. Groups with equal weighting rule showed a steady decline in error after the first two blocks (11.38, 12.03, 9.82, 8.03); however, for those with the unequal rule most learning occurred within the first two blocks and performance on the third and fourth blocks was similar to that on the second (10.28, 7.94, 7.51, 7.64).

When the unsigned deviations were broken down further by cue (see Table 13), there was a significant effect of cue, $F(3, 84) = 11.09$, $p = .000$. Moreover, there were significant interactions of display x cue, $F(3, 84) = 6.02$, $p = .001$, weighting rule x cue, $F(3, 84) = 17.67$, $p = .000$, and display x weighting rule x cue, $F(3, 84) = 3.09$, $p = .032$. The unsigned errors for the four numerical cues were 2.11, 2.04, 2.45, and 1.62 whereas the errors for the graphical cues
were 3.75, 2.22, 2.54 and 1.93. Again there appears to be a tendency for the fourth cue to produce the least magnitude of error for both displays. Such an observation merely confirms the earlier conclusion that there were significant differences in processing the cues.

The cue weighting accuracy was inversely related to cue validity. Thus when the weighting rule was unequal and the cues decreased in validity from the first cue to the fourth, the amount of error also decreased—the mean error for the four cues was 3.45, 2.08, 1.56, and 1.26. The equal weighting rule did not produce such a systematic decline and the error scores for all four cues were similar—the means were 2.41, 2.18, 3.44, and 2.29. These findings largely parallel those obtained with the signed deviation measure.

As mentioned earlier, in addition to correlational measures it is useful to index absolute accuracy of performance using deviation measures. A summary of the correlational and deviation measures is provided in Tables 14 and 15. The data are collapsed across display format because the effect of that variable or its interaction with other variables was not significant for any of these measures.

**Achievement (r_{ve,ve}).** The effect of display was not significant, $F(1, 28) = 1.99, p = .17$. However, there was a significant effect of weighting rule and learning block, $F(1, 28) = 5.05, p = .03$ and $F(3, 84) = 9.93, p = 0.00$ in each case. Overall, the unequal rule groups performed better than the equal rule groups (0.61 vs. 0.49). Although learning occurred across blocks (see Figure 3), there was a rule x block interaction, $F(3, 84) = 4.42, p = .006$. As is
Table 14

Summary of Correlational measures used to index subjects' performance.

<table>
<thead>
<tr>
<th>Performance Index</th>
<th>Optimal weighting rule</th>
<th>Learning Block</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achievement</td>
<td>Unequal</td>
<td></td>
<td>.47</td>
<td>.61</td>
<td>.68</td>
<td>.69</td>
</tr>
<tr>
<td></td>
<td>Equal</td>
<td></td>
<td>.47</td>
<td>.45</td>
<td>.55</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td>( \bar{x} )</td>
<td></td>
<td>.47</td>
<td>.53</td>
<td>.61</td>
<td>.59</td>
</tr>
<tr>
<td>Matching Coefficient</td>
<td>Unequal</td>
<td></td>
<td>.80</td>
<td>.87</td>
<td>.89</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td>Equal</td>
<td></td>
<td>.79</td>
<td>.71</td>
<td>.79</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td>( \bar{x} )</td>
<td></td>
<td>.79</td>
<td>.79</td>
<td>.84</td>
<td>.89</td>
</tr>
<tr>
<td>Optimality</td>
<td>Unequal</td>
<td></td>
<td>.51</td>
<td>.68</td>
<td>.73</td>
<td>.73</td>
</tr>
<tr>
<td>Coefficient</td>
<td>Equal</td>
<td></td>
<td>.51</td>
<td>.46</td>
<td>.57</td>
<td>.61</td>
</tr>
<tr>
<td></td>
<td>( \bar{x} )</td>
<td></td>
<td>.51</td>
<td>.57</td>
<td>.65</td>
<td>.67</td>
</tr>
</tbody>
</table>

Note: The group means are collapsed across the numerical and graphical display formats.
Table 15

Mean deviations used to index the absolute accuracy of subjects' performance.

<table>
<thead>
<tr>
<th>Performance Index</th>
<th>Optimal weighting rule</th>
<th>Learning Block</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Learning Block</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Ye - Ys</td>
<td>Unequal</td>
<td>25.60</td>
<td>21.41</td>
<td>23.01</td>
<td>22.74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Equal</td>
<td>29.01</td>
<td>27.53</td>
<td>25.82</td>
<td>25.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\bar{X}$</td>
<td>27.31</td>
<td>24.47</td>
<td>24.42</td>
<td>23.87</td>
<td></td>
</tr>
<tr>
<td>$\bar{Y}$ - $\bar{Y} s$</td>
<td>Unequal</td>
<td>16.24</td>
<td>11.86</td>
<td>14.66</td>
<td>15.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Equal</td>
<td>18.73</td>
<td>20.28</td>
<td>15.50</td>
<td>12.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\bar{X}$</td>
<td>17.48</td>
<td>16.07</td>
<td>15.08</td>
<td>13.88</td>
<td></td>
</tr>
<tr>
<td>$\bar{Y}$ - $\bar{Y} s$</td>
<td>Unequal</td>
<td>22.41</td>
<td>17.72</td>
<td>19.81</td>
<td>20.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Equal</td>
<td>25.69</td>
<td>25.69</td>
<td>23.11</td>
<td>20.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\bar{X}$</td>
<td>24.05</td>
<td>21.71</td>
<td>21.46</td>
<td>20.02</td>
<td></td>
</tr>
</tbody>
</table>

Note: The group means are collapsed across the numerical and graphical display formats.
apparent from Figure 3, the unequal rule groups showed a steady improvement in performance across blocks, especially from the first to the second block (.47 to .61). In contrast, the equal rule groups showed uneven performance across blocks. None of the other interactions approached significance.

Mean deviation of Ye and Ys (Ye - Ys). Although the effect of display was not significant, F(1, 28) < 1, there was again a significant effect of weighting rule, F(1, 28) = 6.13, p = .02, and of learning block, F(3, 84) = 8.36, p = .00. The unequal rule groups showed lower deviations than the equal rule groups (23.19 vs. 26.84). The mean error for both groups continued to decline across blocks (see Figure 4). None of the interactions was significant; only the rule x block interaction approached significance, F(3, 84) = 2.54, p = .06.

To the extent that the achievement measure or in absolute terms, Ye - Ys, provide information about the correspondence between subjects' predictions and the criterion values, they indicate the subjects' accuracy or "hit" rate in determining environmental outcomes. However, actual environmental outcomes are often determined by extraneous factors other than the information cues on which the judgments are based. Given, then, that there is unpredictable error present in the environmental system, achievement or Ye - Ys provide imperfect measures of judgmental accuracy. After all, perfect achievement or zero Ye - Ys is nearly impossible when there is any degree of error present. Since there was some error involved in the environmental model used in the
Figure 3. Mean achievement of the weighting groups as a function of learning block.
Figure 4. Mean deviation of Ye − Ys of the weighting groups as a function of learning block.
present experiment, the limitations of achievement and deviation indices should be borne in mind when interpreting the data.

The higher achievement or lower error on $\bar{Y}_e - \bar{Y}_s$ produced by the unequal weighting rule indicates that the unequal rule was easier than the equal rule. Although some learning occurred for both groups across blocks, the unequal rule groups learned faster and better than the equal groups.

Matching Coefficient ($G = r_{\bar{Y}_e, \bar{Y}_s}$). The matching coefficient ($G$) has often been referred to as the "knowledge" measure—it indicates the correspondence between subjects' policies and the optimal policy. The assumption is that if a person has mastered cue-criterion relations in the environment perfectly, this knowledge would be reflected in the match between his/her policy and the environmental model depicting these relations.

The only significant effect for this measure was that of learning block, $F(3, 84) = 2.65, p = .05$. Thus there was some evidence of an increase in the matching coefficient across blocks (see Figure 5).

Mean deviation of $\bar{Y}_e - \bar{Y}_s$ ($\bar{Y}_e - \bar{Y}_s$). There was a slight tendency for the deviations to be lower for the numerical relative to the graphical format (14.24 vs. 17.01) and the effect of display on the deviation of $Y_e$ and $Y_s$, was marginal, $F(1, 28) = 3.37, p = .08$.

The effect of weighting rule was not significant, $F(1, 28) = 2.15, p = .15$. The deviations showed a systematic decrease across blocks and this effect was significant, $F(3, 84) = 6.43, p = 0.00$. Of all the interactions, only the rule x block interaction was
Figure 5. Mean matching coefficient of the weighting groups as a function of learning block.
significant, $F(3, 84) = 15.27, p = 0.00$. Thus, the unequal rule
groups showed a slight decline in error from the first block to the
second, but the error increased thereafter. The exact opposite was
true for the equal rule groups: Although the error increased slightly
on the second block, it showed a steady decline on the remaining blocks
(see Figure 6).

Except for a significant effect of learning block, indicating that
subjects were indeed learning the cue-criterion relations, none of the
the other effects was significant. Thus neither display nor weighting
rule affected directly subjects' acquisition of cue-criterion
relations. This conclusion should, however, be tempered by the fact
that the absolute index tells a slightly different story. The marginal
effect favoring numerical display suggests that absolute predictions of
the criterion values (based on optimal policy) were better achieved
with numerical than graphical policies. Also, equal rule groups showed
greater absolute acquisition of predictions of the criterion across
blocks than did the unequal groups.

**Optimality Coefficient ($r_{e-Ys}$).** Display did not significantly
affect the optimality coefficient, $F(1, 28) = 1.05, p = .31$. The
unequal rule groups fared better than the equal rule groups; the group
means were .66 and .54 and this difference was significant, $F(1, 28)
= 4.23, p = .05$. The increase in the optimality coefficient across
blocks was significant, $F(3, 84) = 11.82, p = 0.00$. The rule x
block interaction was also significant, $F(3, 84) = 4.94, p = .003$.
The unequal rule groups showed a rapid improvement from the first
Figure 6. Mean deviation of $\hat{\mathbf{e}} - \hat{\mathbf{s}}$ of the weighting groups as a function of learning block.
block to the second and a little thereafter, the equal rule groups actually declined in performance on the second block, but continued to show an improvement after that (see Figure 7).

Mean deviation of \( \hat{Ye} \) and \( Ys \). (\( \hat{Ye} - \bar{Ys} \)) An ANOVA of the deviations of \( \hat{Ye} \) and \( Ys \) paralleled the results based on the optimality coefficient. The effect of display was nonsignificant, \( F(1, 28) < 1 \). Again there was a significant effect of weighting rule, \( F(1, 28) = 4.58, p = .04 \) and the unequal rule group showed less error than the equal rule groups (19.99 vs. 23.63). A decrease in error was observed across learning blocks and this difference was significant, \( F(3, 84) = 9.38, p = 0.00 \). A significant interaction of rule x block occurred, \( F(3, 84) = 9.03, p = 0.00 \). The equal rule groups showed a decrease in error across all blocks. In comparison, the amount of error for the unequal rule groups dropped from the first block to the second and then continued to rise again (see Figure 8).

It was pointed out earlier that achievement and \( \hat{Ye} - \bar{Ys} \) are imperfect measures of accuracy because they are affected by the magnitude of unpredictable error in the environmental system. Since the subject is really incapable of predicting the error, his actual response accuracy should be compared to the predictions of the environmental model—as measured by the optimality coefficient or deviation of \( \hat{Ye} - \bar{Ys} \). If the subject has indeed acquired the cue-criterion relations present in the environment and applies this knowledge reliably, then the optimality coefficient would be high (or the deviation \( \hat{Ye} - \bar{Ys} \) low). Also, if \( G \) is high, implying that the
Figure 7. Mean optimality coefficient of the weighting groups as a function of learning block.
Figure 8. Mean deviation of $\hat{Y}_e - Y_s$ for the weighting groups as a function of learning block.
subjects' acquisition of cue-criterion relations (knowledge) is high, the discrepancy of the optimality coefficient from G indicates the inconsistency or unreliability of subjects' application of their knowledge. Indeed as is apparent from Table 14, the optimality coefficient was always lower than G indicating that there was some unreliability in subjects' responses, a point elaborated on later in the section.

In general it was evident that the unequal rule groups performed better with respect to both the optimality coefficient and the deviation of \( \hat{\alpha} \) - \( \hat{\beta} \) in comparison with the equal rule groups. However, the former group showed a learning effect only on the relative measure (viz., the optimality coefficient); in contrast, the latter group showed learning across blocks on both the relative and absolute measures.

To summarize the findings based on the correlational measures and the corresponding deviations, three major observations seem justified. First, all measures reflected significant learning across blocks. Thus the task was obviously learnable and performance on it showed significant improvement over time. Secondly, format did not affect subjects' efficiency in predicting the criterion. Only the absolute knowledge measure (\( \hat{\alpha} \) - \( \hat{\beta} \)) showed a marginal effect in favor of the numerical format, indicating that for the most part both formats produced equivalent learning of the optimal model. Finally, the unequal weighting rule proved to be easier to learn than the equal weighting rule. Learning of the equal weighting rule progressed
unevenly and on average was poorer (resulted in lower achievement, optimality coefficient and greater error on \( \bar{Y}_e - \bar{Y}_s \) or \( \bar{Y}_e - \bar{Y}_s \)) than the equal weighting rule.

The analyses considered thus far compared various aspects of subjects' policies to some optimum. However, there is another measure that indexes subjects' performance without explicit reference to any criterion—it is the consistency or \( R^2 \)'s obtained from subjects' policies. It will be recognized, of course, that \( R^2_e \) in the environmental model places a ceiling on the values that \( R^2 \)'s can take although \( R^2 \)'s can still be evaluated independently of \( R^2_e \). It will also be recalled that \( R^2 \)'s in combination with \( C \) influences measures such as achievement or the optimality coefficient. In fact within the MCPL paradigm, \( R^2 \)'s is known as the "control" measure—it denotes the extent to which subjects can apply consistently their acquired knowledge of the environment.

**Consistency or \( R^2 \)'s.** Since the effect of display format on consistency failed to reach significance, \( F(1, 28) = 1.27, p = .27 \), the mean consistency scores were collapsed across format (see Table 16). Note that the maximum value that \( R^2 \)'s can take is determined by \( R^2_e \), which in this case was .80. Subjects in the unequal rule groups showed greater consistency than those in the equal rule groups (.57 vs. .44). The difference in their means was significant, \( F(1, 28) = 7.23, p = .01 \). The consistency in policies of all subjects tended to increase across learning blocks, \( F(3, 84) = 19.28, p = 0.00 \). Despite this overall trend, there was a significant
Table 16

Consistency or $R^2$s obtained from subjects' policies

<table>
<thead>
<tr>
<th>Optimal weighting rule</th>
<th>Learning Block</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Unequal</td>
<td>.40</td>
<td>.58</td>
<td>.66</td>
<td>.64</td>
</tr>
<tr>
<td>Equal</td>
<td>.40</td>
<td>.39</td>
<td>.47</td>
<td>.49</td>
</tr>
<tr>
<td>$\bar{X}$</td>
<td>.40</td>
<td>.49</td>
<td>.56</td>
<td>.57</td>
</tr>
</tbody>
</table>

Note: The group means are collapsed across the numerical and graphical display formats.
rule x block interaction, $F(3, 84) = 6.33, p = .001$. The equal
rule groups showed an increase in consistency only after the second
block. However, the unequal rule group continued to rise in
consistency from the first block to the third and then showed a slight
drop on the fourth block.

Given that consistency differed between the unequal- and equal
rule subjects and the knowledge (G) measure did not, the difference in
achievement or optimality coefficients observed for those groups was
primarily a result of difference in consistency. Thus the type of
weighting rule affected how consistently subjects applied it rather
than how easily they acquired it.

Some subsidiary analyses concern comparisons of the components of
$R^2$s--variance of judgments, variance of predicted judgments, and
variance of error. However, instead of comparing variances, standard
deviations that tend to have less skewed distributions were compared.
A comparison of SD$\hat{y}$ showed greater variability in numerical judgments
relative to graphical (31.24 vs. 24.44), $F(1, 28) = 20.98, p = .000$. The effect of weighting rule was not significant, $F(1, 28) < 1$. The mean SD$\hat{y}$ for the numerical and graphical format was 22.28 and
16.70, showing a reliable effect of format, $F(1, 28) = 11.60, p = .002$. The unequal weighting rule groups showed a slightly larger SD$\hat{y}$
than the equal one (21.07 vs. 17.91) and this difference was marginally
significant, $F(1, 28) = 3.70, p = .064$. Turning to SDe, the means
for numerical and graphical formats were 20.43 and 17.44 and this
difference approached significance, $F(1, 28) = 3.80, p = .061$. 
The unequal weighting rule produced reliably lower SDe than the equal rule (17.15 vs. 20.71), $F(1, 28) = 5.37, p = .03$.

It thus appears that the larger SDy for numerical judgments compared to graphical ones was offset by a large SDy resulting in similar levels of consistency for the two formats. The lower consistency for the equal weighting rule occurred primarily due to a larger error in those judgments relative to the unequal weighting rule.

The results of this experiment substantiate the findings from the previous three experiments that format affects subjects' weighting of cues. Moreover, the present data permitted an analysis of the accuracy of weighting and suggested further that the numerical format produces more optimal weighting of cues than the graphical format. However, format proved to be relatively ineffective in influencing the other measures of learning. Subjects' accuracy in predicting the criterion as measured by achievement or the optimality coefficient (or the corresponding deviation measures) and also the knowledge and control measures did not differ with respect to format. There was a marginal effect of format on the $\bar{Y}_e - \bar{Y}_s$ deviation indicating that numerical policies might predict values of the environmental model better than graphical policies. Given the robustness of differences in cue weighting of numerical and graphical cues, the failure to find format effects in efficiency of learning or accuracy of predictive judgments is puzzling. Perhaps, providing numerical criterion feedback served to equate format differences by prompting subjects to "translate" graphical cues into numbers. It might thus be useful to test this
possibility in a replication of this study using identical display formats for both cue presentation and provision of feedback.

The unquestionable superiority of performance produced by the unequal weighting rule indicates that the equal weighting rule does not necessarily imply a cognitively simple model even though it could be mathematically simple. To extend the argument that Einhorn (1971) has made in the context of mathematical vs. cognitive simplicity of linear/nonlinear models, an unequal weighting rule might be cognitively simplified by concentrating fully on the most important cues and relatively neglecting the others. However, the equal weighting rule has to be compensatory and such a strategy cannot be employed without a considerable loss in accuracy of prediction. This argument is best exemplified when cues that are negatively correlated are to be integrated. Suppose that two cues, one with a high scale value and another with a low value had to be integrated using an unequal and an equal weighting rule. In the former condition, subjects can focus on the most important cue and achieve some degree of accuracy because the contribution of the neglected cue to prediction is not that high; however, the equal weighting rule would have to compensate for high values on one cue for low values on the other and thus ignoring any one cue would result in a substantial loss in predictive accuracy. It might be the case then that when subjects have to simplify cognitive load (due to time pressure or increased information), the unequal weighting rule can be used more efficiently than the equal one.
In summary, format differences in this experiment stemmed largely from cue weighting and were not apparent in any other learning measures. Subjects with the unequal weighting rule performed better than those with the equal weighting rule, but type of weighting rule did not interact significantly with format.

GENERAL DISCUSSION

A recurrent theme in the literature on judgment and decision making has been that task characteristics modulate human judgment and decision behavior (Einhorn & Hogarth, 1981; Slovic, Fischhoff, & Lichtenstein, 1977; Howell & Burnett, 1975). Hammond (1980) has recently proposed a Cognitive Continuum theory in which he outlines key structural features of the task that have an important bearing on judgment and decision performance. Display format is one variable suggested by the theory and in view of its relevance for the design of decision systems, it was chosen as the focal point of the present research.

There is a considerable amount of empirical evidence relating display format of visually presented information to human performance. One conclusion that has emerged from these studies has been that efficacy of a particular format over comparison formats is determined substantially by the type of task performed. Thus, for example, identification of symbols has been shown to benefit from numerical coding relative to color coding, but counting and comparison tasks do not appear to show such an enhancement (Hitt et al., 1961). The type of tasks typically investigated in studies of format have included
perceptual and decision tasks involving an explicit performance criterion (e.g., speed, accuracy). Thus formats have been evaluated primarily with respect to their effect on speed or accuracy of responding. However, there are a lot of judgment/decision situations in which a clear-cut performance criterion does not exist. There is a dearth of research on display effects in such situations. Some measure of progress was achieved toward this end in the first three experiments.

Two types of displays—numerical and graphical—were investigated. Both represent common classes of structured displays, are readily interpreted by untrained subjects, and are likely to be encountered in real world decision contexts, especially given the growing trend toward an extensive use of computers. The most pervasive finding with regard to the chosen displays was the substantial difference in the weighting of cues produced by the numerical relative to the graphical format. Although all experiments essentially required subjects to process multidimensional stimuli and make global evaluative judgments, the specific task scenario was varied across experiments. Thus subjects judged applicant suitability (Experiments 1 and 2), teaching effectiveness (Experiment 3), and the strength of force required to combat enemy attack (Experiment 4). Despite this variation, the consistent differences in cue weighting observed for the two formats underscore the generality of the effect.

Previous investigators using the multiple regression approach have failed to report or find any evidence for format-related differences in
the weighting of cues (e.g., Anderson, 1977; Goldsmith & Schvaneveldt, 1982; Knox & Hoffman, 1962; Wickens & Scott, 1983). Most of these studies have, however, used standardized weights that may be insensitive in detecting actual changes in cue weighting. Consequently, a more powerful alternative measure—the raw score regression weight—was used in the present analyses. The significance of the obtained results lends credence to the argument made by Lane et al. (1982) that the raw score weight is more appropriate than standardized weights for measuring cue importance.

As noted at the outset, integrating multiple cues to predict a criterion state is a common decision activity. Generally, in such situations what is of immediate practical consequence is the decision maker's actual predictions than a description of his policy (e.g., G, cue weighting). Thus from a purely practical standpoint, it is relevant to consider how the manipulated task characteristic might affect actual decision accuracy. As was apparent from the achievement index and the optimality coefficient (or $\bar{Y}_e - \bar{Y}_s$ and $\overline{Y_e - \bar{Y}_s}$), subjects learned to predict the criterion with similar levels of accuracy under both displays (Experiment 4). However, since the graphical format appeared to induce holistic processing of cues or at least a different form of processing compared to the numerical format, accuracy could conceivably vary with format under the right conditions (e.g., stress due to time pressure or information load). Thus additional considerations such as speed of making judgments, subjective ratings of preference, or ease of implementation might
perhaps serve to discriminate between the relative merits of the formats. The present data, then, do not speak to the issue of how a practical choice between the formats might be made.

Of course, it will be recognized that although the decision maker's actual response may be a more immediate concern than policy-based measures, the latter might be of considerable relevance in some display design contexts. Consider in particular displays designed for training decision makers on some variation of the MCPL task. Thus novice decision makers might be trained to use the policies of "experts" in that field (e.g., Slovic, 1969) or decision makers might have to learn cue-criterion relations in the environment (Hammond, 1971). In any event, displays in such training contexts are designed to give people feedback about their own policies in conjunction with the optimal policies so as to produce maximum learning. Recall that Experiment 4 provided only outcome feedback (viz., criterion values) and did not provide explicit process feedback (viz., cue weighting information). Previous studies have showed that learning based on process feedback is quicker and more optimal than that based on only outcome feedback (Adelman, 1982; Hammond, 1971); moreover there is another body of literature suggesting that positive feedback results in accelerated learning relative to negative or no feedback (Ammons, 1956; Annet, 1969). Taken together, these findings raise the interesting possibility that a process feedback manipulation might interact with format: Thus the numerical display would result in a more positive process feedback (due to the clear-cut superiority of numerical cue
weighting) and enhance learning at a greater pace than the graphical format. Although speculative, this suggestion merely serves to underscore the fact that the observed cue weighting differences under the two formats could have considerable practical implications.

An important methodological point that has been virtually ignored to date and that was addressed in the present research needs to be emphasized. It concerns $R^2$s or consistency of subjects' policies. $R^2$s refers to the proportion of variance in actual judgments that is accounted for by the variance of predicted judgments (that are based on a weighted combination of cues). Typically, studies have reported either $R^2$s alone (e.g., Einhorn, 1971; Knox & Hoffman, 1962) or $R^2$s along with variance of judgments without explicating their relationship (e.g., Anderson, 1977). As was illustrated by the present research, the obtained differences in $R^2$s as a function of the experimental manipulation can be explained better by also testing for differences between its components--variance of actual judgments (or, alternatively, SSy), variance of predicted judgments (or SSŷ), and variance of error (SSe). Thus, for example, in Experiment 4 although the levels of consistency were similar for the numerical and graphical formats, the former resulted in greater cue usage (SSŷ) as well as greater error (SSe) than the latter. Similarly in Experiment 2, it was evident that the lower consistency of numerical policies stemmed largely from the higher magnitude of error in those judgments compared to graphical ones. It should be thus obvious that the observed level of consistency can be explained more clearly by examining the
combination of $SS_g^2$ and $SSE$ on which it is based. In Einhorn's (1971) study, he raised the question of whether the low level of linear consistency in his high information-load condition resulted from nonlinearity or presence of error in judgments, but he did not explicitly test his hypotheses. It is likely that low consistency could have occurred due to a relatively large $SS_g^2$ and large $SSE$ and not necessarily due to nonlinearity of judgments. Indeed, a separate comparison of linear cue usage ($SS_g^2$) and error ($SSE$) would have provided insight into exactly how subjects processed additional information.

In sum, the purpose of the present research was two-fold: It explored the effect of one basic task variable--format of displaying information--along with other variables (viz., response mode, number of cues, type of weighting rule) on judgment and decisions. The obtained findings are significant both with respect to their practical implications for display design and also theoretically with respect to how people process information in making judgments or decisions. In addition to its empirical contribution, the research served to illustrate the functional approach advocated here. Based on the argument that it is all but futile to capture "processes" or "outcomes" using multiple regression, the model was regarded primarily as a useful source of behavioral indices and systematic task manipulations were introduced to reveal its effect on important aspects of decision behavior. Clearly, there is a long way to go. Several other variables need to be investigated from the functional perspective to achieve
better understanding of the factors that influence judgment and
decision making. Nevertheless, the present research represents an
effort in that direction.
FOOTNOTES

1. Technically, Norman Anderson's work on Information Integration Theory is subsumed under the regression approach because of its use of ANOVA which is a special case of multiple regression (Cohen, 1968). However, that research will be excluded from the present discussion. Information Integration theory seeks to discover "cognitive algebra", i.e., general quantitative laws that govern cognitive functioning and it involves specialized methodology that is not directly relevant to the theme of this paper. The interested reader may see Anderson (1974, 1978) for reviews.

2. Since one of the two subjects' linear policies accounted for only 64% of the configural variance, the magnitude of the mean (based on n = 2) in the six-cue condition is rather small.

3. I am grateful to John Cason for his help in devising the task scenario.
REFERENCE NOTES


REFERENCES


Goldberg, L. R. Seer over sign: The first "good" example? Journal of Experimental Research in Personality, 1968, 3, 168-171. (a)

Goldberg, L. R. Simple models or simple processes? Some research on clinical judgments. American Psychologist, 1968, 23, 483-496. (b)


Summers, D. A., & Hammond, K. R. Inference behavior in multiple cue
tasks involving both linear and nonlinear relations. *Journal of

Summers, S. A. The learning of responses to multiple weighted cues.
*Journal of Experimental Psychology, 1962, 64, 29-34.*

Tinker, M. A. Prolonged reading tasks in visual research. *Journal of
Applied Psychology, 1955, 39, 444-446.*

Tinker, M. A., & Patterson, D. E. Influence of type form on speed of

Todd, F. J., & Hammond, K. R. Differential feedback in two multiple
cue probability learning tasks. *Behavioral Science, 1965, 10,
429-435.*

Tullis, T. S. An evaluation of alphanumerical, graphic, and color

Vitz, P. C. Preference for different amounts of visual complexity.
*Behavioral Science, 1966, 11, 105-114.*

Wainer, H. Estimating coefficients in linear models: It don't make no

Ward, W. C., & Jenkins, H. M. The display of information and the
judgment of contingency. *Canadian Journal of Psychology, 1965,
19, 231-241.*

Watkins, M. J. Human memory and the information processing metaphor.
*Cognition, 1981, 10, 331-336.*

Wherry, R. J., Sr., & Naylor, J. C. Comparison of two approaches - JAN
and PROF - for capturing rater strategies. *Educational and

*Journal of Abnormal Psychology, 1968, 73, 70-77.*

Williams, C. M. Horizontal versus vertical display of numbers. *Human
Factors, 1966, 8, 237-238.*

Winer, B. J. *Statistical Principles in Experimental Design.* New

Wright, P. Using tabulated information. *Ergonomics, 1968, 11,
331-343.*

