Contrasting Models of Appraisal Judgments for Positive and Negative Purposes Using Policy Modeling

MATTHEW CHAMPAGNE AND MARY KAY STEVENSON

Purdue University

This article incorporates the areas of measurement and performance appraisal in an attempt to better represent the performance evaluation process. The technique called Policy Modeling was used to examine the effects of positive versus negative perceived purpose on the performance appraisals of work study students. Both group and individual analyses were performed to determine the subjective values of the attributes used to describe workers, the relative importance of each attribute in the evaluation and the interpretation of different rating scales. Most raters' appraisals were generally best represented by a configural averaging strategy, although the nature of the configural adjustment differed under conditions of reward and punishment. The appliability of policy modeling as a vehicle for performance appraisals are discussed. © 1994 Academic Press, Inc.

For over a half century, performance appraisal has been under the scrutiny of psychologists. Many studies have attempted to obtain an accurate method for measuring worker's job performance, but the solution has been elusive. Much recent research in this area had dealt with the modeling of performance evaluations as a process of the observation, encoding, storing, and retrieving of information (e.g. Bazerman, Beekun, & Schoorman, 1982; DeNisi, Cafferty, & Meglino, 1984; Feldman, 1981; Landy & Farr, 1980).

The models of DeNisi et al. (1984), Ilgen and Feldman (1983), and the suggestions of Williams, DeNisi, Blencoe, and Cafferty (1985) and Zedeck and Cascio (1982, 1984) all include the purpose of an evaluation as a component of the performance appraisal process. Landy and Farr (1981, 1983) reviewed studies in which there appeared to be an effect of purpose

This research was based on the master's thesis of the first author under the guidance of the second author. Portions of this article were previously presented at the annual meeting of the Midwestern Psychological Association, May, 1991, and at the conference of the Mathematical Psychology Association, August, 1992. Comments at both conferences were appreciated. Special thanks to Michael Birnbaum, David Schoorman, Alice Stuhlmacber, Howard Weiss, Derek Brown, and one anonymous reviewer for their helpful suggestions; to InJaE Myung for his skillful programming; and to Elise Amel for her timely assistance in the data analysis.

Reprint requests should be addressed to M. Champagne, Department of Philosophy, Psychology, and Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY 12180-3590.

0749-5978/94 $6.00
Copyright © 1994 by Academic Press, Inc.
All rights of reproduction in any form reserved.
on appraisal ratings, but concluded that further testing of this component was needed.

Since a variety of definitions for the term "purpose" exist, researchers have used different classification systems. Zedeck and Cascio (1982) attempted to determine if raters used different strategies of weighting and combining performance information across three administrative purposes: (1) the need for development (amount of training necessary), (2) recommendation for raise, and (3) the retention of an employee. Williams et al. (1985) used a different classification system for purpose comprised of three types of administrative decisions: promotion, merit raise and remedial training. In both the Zedeck and Cascio and the Williams et al. studies, differences were found, but purpose was not associated with different cognitive strategies as the authors had hypothesized. For example, in the latter study, subjects in the promotion and merit raise (positive) conditions did not differ significantly in the way they combined and weighted given information for an appraisal decision (i.e. their policy or strategy were not significantly different). But the subjects in these two conditions differed from those in the remedial training (negative) condition. Raters used similar policies for purposes that were given positive connotations by the experimenters and different policies for appraisals requiring negative consequences.

Zedeck and Cascio (1982) found evidence for the positive/negative dichotomy in administrative decisions. Raters appeared to use similar policies when outcomes were perceived to be positive (merit raises), but appeared to use different policies for "negative" conditions (additional training and possible dismissal). These results seem to suggest that raters do not cooperate by using different policies depending on the particular purpose chosen apriori by the experimenter. Instead, these results seem to support the more parsimonious hypothesis that raters use policies for performance appraisal that differ for positive and negative outcomes.

Little work of performance appraisals under "negative" purposes has been done. Past research has looked at raises, promotions and similar positive administrative decisions, but not on layoffs, terminations or demotions which also occur with some regularity in organizations.

Although evaluators vary their responses depending on the purpose of evaluation, the question of where purpose impacts on the judgment process remains. Several possibilities exist. Purpose could affect how raters view information about a worker, how they weight that information, or how they combine the information to form a judgment. Raters may use one strategy in their evaluation of workers for raises but make adjustments to that strategy if it is known that a cut in personnel is imminent. To tease apart where and under what conditions purpose affects a judgment, we must move beyond the present models of cognitive processes
used in performance appraisal research. We should follow the advice of Borman (1978) who suggested the development of new approaches for tapping human judgments about performance.

POLICY MODELING

One such method that is grounded in measurement theory is policy modeling (Stevenson, 1986; Stuhlmacher & Stevenson, 1993). Policy modeling is a method of examining how people interpret and combine information to make a judgment. It has the same objective as policy capturing (Hammond, 1965; Hobson & Gibson, 1983; Slovic & Lichtenstein, 1971) and Raiffa’s model (1982) which describe how individuals weight the importance of information used to make evaluations. However, policy modeling provides greater flexibility. This approach scales the levels of each characteristic used to describe a worker’s performance to reflect the views of the rater. For example, each rater may use their own interpretation about what “average” means to them. It is not necessarily the midpoint of a dimension. Policy modeling does not assume that raters add dimensions to arrive at their final rating, but can describe any strategy that a person may use to combine information (as long as the rater consistently uses the same strategy). Some common strategies include additive (e.g. multiple regression, expected utility, social judgment theory models), multiplicative (Anderson & Shanteau, 1970; Norman & Louviere, 1974; Norman & Singh, 1989; Shanteau, 1975), averaging (Birnbaum, 1974, 1976; Stuhlmacher & Stevenson, 1994), or configural (Birnbaum & Stegner, 1979, 1981; Birnbaum, Coffey, Mellers, & Weiss, 1992).

Like conjoint analysis, policy modeling assumes that the observed ratings are monotonically (not necessarily linearly) related to the performance attributes. The ordinal characteristics of the observed ratings are used to test the raters’ policy models. Through this flexibility, policy modeling can describe more precisely the value of the information to the rater, its importance in the final evaluation and the particular strategy that is used to combine this information, whether additive, averaging or configural.

Policy modeling as described in this study is similar to the measurement framework of information integration theory (Anderson, 1982) in that parameters are estimated for the subjective values of the stimuli and the importance weights. However, with policy modeling, different assumptions are made about the response function and about the way to verify specific models. This approach is more similar to the modeling strategies and parameters introduced by Birnbaum and his colleagues (Birnbaum, 1974, 1976; Birnbaum & Stegner, 1979, 1981; Birnbaum et al., 1992), but differs in the assumptions made about the source of error and methods of estimating the response function.
An example of the policy modeling framework applied to performance appraisal is shown in Fig. 1. In this case, a worker is being evaluated for a possible reward based on three attributes of his or her work performance: level of efficiency ($\Phi_{\text{eff}}$), amount of responsibility ($\Phi_{\text{res}}$), and consistency of promptness ($\Phi_{\text{prm}}$). For example, $\Phi_{\text{eff}}$ might represent "this worker is extremely efficient" or $\Phi_{\text{prm}}$ might represent "this worker has never been late for work." Each rater has his or her own subjective value for a given level of performance. These subjective values are designated by $S_{\text{eff}}$, $S_{\text{res}}$, and $S_{\text{prm}}$.

The importance weights (represented by $W_{\text{eff}}$, $W_{\text{res}}$, and $W_{\text{prm}}$) indicate how much a particular attribute of performance influences the overall appraisal. Importance weights and subjective values provide conceptually different information about the rater's point of view (Stevenson, Bussemeier, & Naylor, 1990). The performance attributes are then combined to represent the rater's overall implicit appraisal of a worker (represented by $\Psi$). The tradeoff strategy used by the rater to combine the information might be represented by an additive, averaging, or some other combina-

---

**Fig. 1.** Policy modeling diagram for pay raise condition.
tion strategy. The response strategy represents the mapping of the estimated value of the worker onto the reward scale. Different purposes may result in different response strategies for the same sample of workers. These response functions must be identified in order to compare a rater's policy across rating scales. This identification will address the point made by Landy and Farr (1980) that little research to date has been done on the nature of the psychological scale of measurement implied by the given rating scale.

All of the parameters within the box in Fig. 1 represent cognitive processes that cannot be directly observed. The features outside the boxes represent observable attributes and responses. Our model does not investigate the process of searching for information in the work environment, but rather focuses on how that information is used to make important evaluative decisions.

ALTERNATIVE POLICY MODELS

Each of the following policy models represents a potential evaluation strategy. The additive model implies that the rater perceives the effect of each behavior as independent of the other behaviors. The additive model is expressed in Eq. (1).

\[ R = J [W_{\text{eff}}S_{\text{eff}} + W_{\text{res}}S_{\text{res}} + W_{\text{prm}}S_{\text{prm}}] \]  

This compensatory model states that the observed performance rating \( R \) is some monotonic function \( (J) \) of the sum of the attributes weighted by their perceived importance. The relative importance associated with efficiency, responsibility, and promptness of a worker are represented by \( W_{\text{eff}}, W_{\text{res}}, \) and \( W_{\text{prm}}, \) respectively, whereas \( S_{\text{eff}}, S_{\text{res}}, \) and \( S_{\text{prm}} \) represent the subjective values of these behavioral characteristics. When the response function is assumed to be monotonic, the additive strategies cannot be distinguished from multiplicative strategies because there is no difference in the rank order of their predictions unless negative or zero attribute values are evaluated (Stevenson et al., 1990). This model, with a linear response function, is typically assumed in policy capturing studies.

The relative weight averaging model is identical to the additive model when predicting the impact of all three attributes. However, if, as is often the case, one or more attributes are missing, the predictions of the averaging model are quite different than the additive model. According to the additive model, more information is always better as long as it is not negative. The impact of that information is the same whether the worker is viewed in a positive or less than positive way. However, the impact of new information on an averaging model depends on the current standing. Positive but mediocre ratings on a particular attribute will lower the over-
all appraisal if the current standing is high and may raise overall appraisal if the current standing is low. The relative weight averaging model is stated in Eq. (2):

\[ R = J \left[ \frac{W_o S_o + W_{eff} S_{eff} + W_{res} S_{res} + W_{prm} S_{prm}}{W_o + W_{eff} + W_{res} + W_{prm}} \right]. \] (2)

The initial subjective value is represented by \( S_o \), the initial weight is represented by \( W_o \), and the other subscripts are defined in Eq. (1). In the current application, a group of supervisors may be doing the yearly evaluation of a worker who has a high level of efficiency and is extremely responsible. While doing the evaluations, the supervisors may learn that this worker has been late for work once or twice. Although this is still a positive attendance record, the evaluation of this worker may be lowered since this information is less positive than their initial evaluation based solely on knowledge of the worker’s responsibility and efficiency. An additive model would predict that any positive information would increase the worker’s rating. The averaging model, however, would predict that if the supervisor has a high opinion of a worker on several attributes and the worker performs at an average level on another attribute, the supervisor’s evaluation of that worker would be lowered.

The advantages of testing between additive models vs. averaging models are twofold. First, the models make different predictions that have practical implications. For example, an applicant to graduate school who assumes that evaluations are determined by an additive model should submit more than the required number of letters of recommendation. If these extra letters will not be as positive as the best three, they might still improve the evaluation. On the other hand, the averaging model implies that the evaluation would be lowered as extra letters of moderate value are added. Second, the importance weight given to information can not be determined in an additive model (e.g. multiple regression) because the weights and the scale values of the information are confounded. The importance weights of a given attribute cannot be determined when using an observational or a factorial design (Schönemann, Cafferty, & Rotten, 1973; Stevenson et al., 1990). In order to estimate the importance of information in the final appraisal, information has to be omitted or manipulated in order to measure the effect of its omission on the ratings (Birnbaum, 1976; Birnbaum, Wong, & Wong, 1976).

A configural-averaging model (Birnbaum et al., 1992; Birnbaum & Stegner, 1979) is a more general averaging model. In this case, information is combined with an averaging process and parameters associated with the possible interaction of attributes are included. These parameters, called configural weights, may be positive or negative (or equal to zero
indicating no interaction). The interactions that are represented by the configural terms are shown in Fig. 2.

A positive configural weight describes a converging interaction. In Fig. 2A, the difference between a low efficiency worker and a high efficiency worker is greater when the worker has a low level of responsibility than if the worker has a high level of responsibility. This converging pattern is predicted with a positive value for the configural weight for efficiency and responsibility \( (CW_{ee}) \) in the model.

A negative configural weight describes a diverging interaction. In Fig. 2B, the difference between a low promptness worker and a high promptness worker is greater at high levels of efficiency than at low levels. This diverging pattern is predicted with a negative value for the configural weight for efficiency and promptness \( (CW_{ep}) \). Figure 2C illustrates no interaction between the levels of performance and responsibility. The configural weight for responsibility and promptness \( (CW_{rp}) \) in this case is zero. Parallel patterns of the attributes is described with the relative weight averaging model introduced earlier.

There have been different ways of defining configurality in judgment models. Birnbaum and Stegner (1979, 1981) defined the configural terms as an adjustment based on the similarity of the scale values of the attributes in their range model:

\[
R = J \left[ \frac{W_0 S_o + W_{eff} S_{eff} + W_{res} S_{res} + W_{prm} S_{prm}}{W_o + W_{eff} + W_{res} + W_{prm}} \right] + w_{c} |S_{\text{max}} - S_{\text{min}}|.
\]  

(3)

The average of the attributes is adjusted according to the discrepancy in the information denoted by \( W_c |S_{\text{max}} - S_{\text{min}}| \). The largest and smallest scale values are represented by \( S_{\text{max}} \) and \( S_{\text{min}} \), respectively, and \( W_c \) is the configural weight parameter. When three attributes are presented, the configural weight is applied to the maximum difference in scale values regardless of what attributes are present. In a three variable design, Birnbaum uses one configural weight for all the attribute combinations. The averaging model, therefore, is simply a special case of the range model where \( W_c = 0 \).

In the present study, we defined configurability based on the range model but applied the configural weights to the interaction of each pair of attributes. The equation for two attribute judgments is exactly the same as the range model. When three attributes were described, we assumed that three potential adjustments were possible. In this model, we assume that if the values of the attributes are interpreted relative to each other, the same process would apply when two or three attributes are given. Equation (4) describes this configural averaging model:
Fig. 2. Examples of possible configural adjustments. (A) converging, (B) diverging, (C) no interaction.

\[
R = \frac{W_o S_o + W_{eff} S_{eff} + W_{res} S_{res} + W_{prm} S_{prm}}{W_o + W_{eff} + W_{res} + W_{prm}}
\]

\[
\begin{align*}
&= CW_{\text{eff}} | S_{\text{eff}} - S_{\text{res}} | \\
&+ CW_{\text{eff}} | S_{\text{eff}} - S_{\text{prm}} | \\
&+ CW_{\text{prm}} | S_{\text{res}} - S_{\text{prm}} |
\end{align*}
\]  \tag{4}

The weights and subjective values are defined in the same way as the averaging model. \(CW_{\text{eff}}\) represents the configural weight associated with the interaction between levels of efficiency and levels of responsibility, \(CW_{\text{eff}}\) represents the configural weight associated with the interaction between levels of efficiency and promptness and \(CW_{\text{prm}}\) is the configural weight associated with responsibility and promptness. If the estimated configural weight parameters are zero, the model is identical to the relative weight averaging model in Eq. (2).

Evidence of interactions that can be represented by configural weighting has been obtained in a variety of contexts, including judgments by expert judges (Ogilvie & Schmitt, 1979; Wiggins & Hoffman, 1968), tenured faculty judging non-tenured faculty (Brannick & Brannick, 1989), and in situations where the cost of an incorrect decision for graduating MBA or industrial engineering students would be high (Einhorn, 1971).

Two general hypotheses about differences between positive and negative purpose of appraisal are advanced in this article. The first deals with the effect of purpose on the tradeoff strategies and the parameters of the models. The second deals with the type of rating scale used.

Past research has discussed the importance of the purpose of the appraisal and has shown differences in the way raters evaluate workers when this variable is manipulated. We propose one classification based on the purpose of administrative decisions. That is, given the same information about the same workers, raters will use different policies depending
on whether the purpose of the appraisal is perceived to lead to a positive or negative outcome for the worker. Although these policies may involve different interpretations of the information (Fig. 1: $S_{\text{eff}}$, $S_{\text{res}}$, $S_{\text{prm}}$) or different weighting strategies across performance attributes (Fig. 1: $W_{\text{eff}}$, $W_{\text{res}}$, $W_{\text{prm}}$), the purpose is expected to have its greatest impact on the tradeoff strategies between performance attributes. More specifically, when appraising workers for a positive purpose, raters’ judgments are expected to be adequately described by a relative weight averaging model. Budget cuts forcing a layoff or the reduction of a worker’s pay or some other negative purpose should be aversive to most raters and they might be hesitant to reduce pay or terminate a worker. Raters may be more unwilling to punish unless their are multiple indicators of poor performance. A configural strategy using interaction terms to account for this adjustment may be a more accurate model of appraisals for negative consequences.

Consider a simple two factor study of performance appraisal. If the experimenter finds a diverging interaction in the ratings, this pattern may be represented either by a negative configural weight component or a simple averaging model and a log response function. The configural weights may tradeoff with the form of the response function in accounting for the pattern of data. We believe that the rating scales used for the positive and negative purposes may be used differently. For example, when evaluating workers’ performance, subjects may award raises evenly along the rating scale (represented by a linear function) but may be reluctant to impose the strongest negative measures unless the worker’s performance is unusually poor (represented by a negatively accelerated function). We believe that it will be more difficult for raters to punish workers than to reward them, and that both the information and the response scales will be used differently so that purpose will affect the configural weights as well as the response function. Both models stated in Eqs. (3) and (4) will be tested.

**METHOD**

**Stimuli**

A set of hypothetical workers were generated from a combination of three levels of each of three characteristic behavior dimensions: efficiency, promptness, and responsibility. The levels of the dimensions were as follows:

**Efficiency.** (1) Usually needs additional time to finish work, (2) likely to finish most of the work on time, (3) always finishes work on time.

**Responsibility.** (1) Needs close supervision, has to be repeatedly told to do things, (2) needs hourly supervision, can follow instructions, (3) needs very little supervision, works independently.
Promptness. (1) Does not come into work when scheduled on occasion, (2) sometimes is 15 minutes late for work, (3) never late for work.

Design

Subjects were presented with 63 hypothetical student worker profiles consisting of 27 \((3 \times 3 \times 3)\) combinations of efficiency \(\times\) responsibility \(\times\) promptness; 27 stimuli involving all two-way combinations of efficiency and responsibility, efficiency and promptness, and responsibility and promptness; and 9 stimuli with each level of each dimension presented individually. This design was chosen to provide data to estimate both the subjective values and the importance weight parameters (Birnbaum, 1976, 1980; Norman, 1976). The three attribute, two attribute, and single attribute profiles were mixed and presented in a different pseudo-random sequence to each subject.

Instructions

Subjects were told that they would be rating the performance of students in the work study program. The work study program had been chosen because of its familiarity and relevancy to the subjects. Subjects were told that they were giving the end-of-semester evaluations in order to determine the starting pay and hours levels for each student for the following semester. They were asked to use four alternative worker evaluation forms. These forms consisted of two evaluations for positive purposes (pay raise, increase hours) and two evaluations for negative purposes (pay cut, decrease hours).

For the pay raise condition, subjects rated the workers on a 1–9 scale where a "1" rating meant that this student should receive "no raise" and a "9" rating meant this student should receive a "40 cent per hour raise." Intermediate ratings were marked on the scale in increments of five cents per hour raise. In the increase hours condition they used a 1–9 scale where a "1" rating meant this worker should not be given additional hours", a "9" rating indicated that this student should be allowed to work "8 additional hours"; intermediate ratings were in increments of 1 h. Subjects were reminded that student workers want to work as many hours as their schedule permits which results in a larger paycheck.

For the pay cut condition, subjects used a 1–9 scale where a "1" rating meant a "40 cent per hour cut" in pay and a "9" rating meant "no cut" in pay. Intermediate ratings were marked in the scale in increments of five cents per hour cut in pay. In the decrease hours condition, subjects used a 1–9 scale, in which "1" meant this worker should have "8 hours cut" from their schedule (leaving the student with no hours, that is, fired), a "9" indicated that "no hours" should be taken away from this student
(leaving the student with 8 h/week). Intermediate ratings were in increments of a 1-h cut per week in time scheduled to work.

Procedure

Each subject completed two sessions, administered 1 week apart. Half of the subjects were randomly chosen to complete the two positive tasks first; the other half completed the two negative tasks first. Subjects were also counterbalanced in assignment to order of task within purpose. All subjects rated the same 63 profiles, randomly presented, in each of the four (two positive and two negative) tasks.

Each subject was individually instructed before each task. Subjects were told to base their judgment on the information presented on each trial. For example, profiles containing only efficiency and responsibility information could be readily explained as a situation in which workers kept track of their own hours so that promptness was not recorded or unknown. For the negative tasks, subjects were reminded that funds for the work study program are limited and there are many students desiring to be in the program. By decreasing workers’ pay or hours, they are freeing up resources for others to get an opportunity to work.

Subjects were given eight practice trials before each condition to check their understanding of the instructions. Each profile (paired with the appropriate rating scale) was presented randomly, one at a time on a computer screen. Subjects responded by typing their choice, 1 through 9, followed by the backslash key. At this point, subjects had the opportunity to change their response. Only after typing the backslash key twice on the same response would their evaluation be recorded. At the conclusion of the second session, subjects filled out a questionnaire and were debriefed in full.

Subjects

Thirty-nine students (21 men and 18 women) in introductory psychology courses at a large midwestern university volunteered as subjects in partial fulfillment of their class requirement. Four students were dropped from analyses: two did not follow instructions, ignored much of the information and gave nearly every worker the same rating and two others took significantly less time than the other subjects and gave obviously random responses in at least one condition. All remaining subjects were familiar with the work study program and nearly one-third were presently employed or had been employed at one time by the program.

RESULTS

In order to determine whether an additive or averaging process would be more accurate in representing the data patterns for individual subjects,
a diagnostic set of points was plotted for each subject for each combination of attributes in each condition. An example of this diagnostic pattern for a representative subject is shown in Fig. 3. The solid lines connect points that represent evaluations based on the attributes of responsibility and promptness. The dashed line represents the subject’s evaluation when responsibility was the only information that was available. If these lines were parallel, an additive model could account for these responses. If the dashed line crosses the solid lines it indicates that when subjects evaluated a worker that had a high level of responsibility and a moderate (but still positive) level of promptness, the evaluation was lower than if responsibility was the only information available. For the additive model, information about promptness could raise or lower the evaluation but the effect would be constant across the levels of responsibility. Each subject was classified using this method and the patterns indicated that an adding model was not appropriate. And averaging model may be used to better represent their judgments.

The next step was to determine which model would best represent these judgments in each condition. In order to test the models, the mean response was computed across subjects and within conditions. The most accurate model was then fit to each individual subject’s data in order to estimate the parameters for each subject. These parameters were then analyzed statistically to test the various hypotheses about the impact of purpose on the various components of the judgment process.

First, the configural model of Birnbaum and Stegner using one configural weight and a linear response function (Eq. (3)) was fit to the mean

![Figure 3](image_url)  
**Fig. 3.** Diagnostic fit of representative subject. Solid lines represent evaluations based on responsibility and promptness attributes. Dashed line represents evaluation based on responsibility alone.
evaluations for each condition. Although a large percentage of variance was accounted for by the model for the conditions of positive pay (93.49%), positive hours (93.86%), negative pay (98.75%) and negative hours (98.52%), there were systematic deviation in the fits. The ratings when all three attributes were present are shown in Fig. 4 for both positive and negative pay conditions.¹ The lines represent the predictions of the model and the 27 points represent the observed responses for all three-way combinations of attributes. The model is unable to account for the bilinear fan pattern apparent in the data. This pattern is not present in the negative condition resulting in a better fit. Sum of squared residuals for the range model is 648 for the positive condition and 143 for the negative.

The configural averaging model that describes the interactions of each pair of attributes (Eq. (4)) was fit to the mean responses using a linear response function. The percentage of variance accounted for remained high for the positive pay (96.77%), positive hours (96.53%), negative pay (98.95%), and negative hours (99.00%) conditions, but systematic deviations were still evident for the positive outcome evaluations. The three attribute graphic fits for this model are shown in Fig. 5. The greatest improvement can be seen in the positive pay condition where the sum of squared residuals has been reduced to 321 (SSR in negative pay condition was 120). But there is still substantial systematic overprediction of some combinations of attributes and underprediction of others using this model.

In Fig. 6, the relationship of the observed values and the linear response function is shown for all four conditions. There are systematic deviations evident in the ratings that were done to reward the student workers but these deviations were not present for the ratings that were focused on reducing the pay or hours of the student workers. It is apparent that the subjects were not using positive response scales in the same way as the negative response scales.

An averaging model with no configural weights and a nonlinear response function was also fit to the mean responses. The nonlinear function was fit using basis splines (de Boor, 1978; Stevenson, 1986), a technique introduced by Winsberg and Ramsay (1979) to keep the function monotonic by using the integral of the spline function. Five parameters were used to estimate the form of the response function.² The use of this model did not improve the fit in the negative conditions (no decrease in

¹ Although Figs. 4, 5, and 7 illustrate only the two pay conditions, results were similar for both hours conditions. All graphs may be obtained from the first author.

² This technique differs from MONANOVA since an exact number of parameters are estimated to describe the shape of the function. Note that if the function is linear, all of the parameters will have a constant value.
Fig. 4. Mean responses across subjects for all 27 combinations of the three attribute responses are represented by circles. Lines represent the predicted fit of the range model (Eq. (3)). Numbers represent the levels of each attribute. (A) Positive pay condition; (B) negative pay condition.
FIG. 5. Mean responses across subjects for all 27 combinations of the three attribute responses are represented by circles. Lines represent the predicted fit of the configural averaging model with a linear response function. Numbers represent the levels of each attribute. (A) Positive pay condition. (B) Negative pay condition.
sum of squared residuals), but there was much improvement in the positive conditions (see Table 1 for listing of indices of fit). It appears that the incorporation of non-linearity into the rating scale was necessary for positive purposes but not for negative purposes.

Finally, the configural averaging model based on the two way interaction terms (Eq. 4) was fit to the means for each condition using the spline parameters that allow the response function to be monotonic but non-linear. Although the global fit statistic did not change very much for any of the conditions, positive pay (98.98%), positive hours (98.72%), negative pay (99.00%), and negative hours (99.06), the residual patterns improved. The three way interaction fits are shown in Fig. 7. Note that the predicted patterns most accurately reflect the observed response patterns. The form
of the response functions for each condition are shown in Fig. 8. Although the ratings for negative purpose, regardless of the actual evaluation scale are nearly linear, the response function for the positive purpose, regardless of the actual response scale are positively accelerated. Raters were not willing to give rewards until the workers reached a particular level of performance based on their own policy value. In the negative condition ratings were clustered closer to the less negative end of the scale and the decrease hours and cut in pay scales were linearly related to the policy values. These results confirm one aspect of the appraisal process that is affected by purpose, the use of the response scale differs across purpose.

**Individual Analyses**

In order to test the hypotheses concerning the other components of the measurement model, the configural averaging model represented in Eq. (4), was fit to each subject for all four conditions using the nonlinear response parameters. Due to the anticipated individual differences among raters, it was more efficient to use the more general model with the spline parameters. If the configural weight parameters were not necessary, they would be estimated to be zero. If the response function is linear then the estimated spline parameters would be approximately equal. Evidence from the group analyses supported the expectation that more subjects evaluating workers for the negative outcomes would be represented by a configural averaging model, but subjects evaluating workers for a positive outcome would be sufficiently described without the configural weight adjustments (a relative weight averaging model). In order to assess individual differences across purposes and/or the response scales, the parameters from the individual analyses were compared across conditions.
Fig. 7. Mean responses across subjects for all 27 combinations of the three attribute responses are represented by circles. Lines represent the predicted fit of the configural averaging model with splines used to allow non-linearity in the response function. Numbers represent the levels of each attribute. (A) Positive pay condition, (B) negative pay condition.
The fit of the general configural model (Eq. (4)) varied across subjects but was accurate for the majority of the subjects. Table 2 shows the fit distributions for each condition. Individual fits for this model ranged from $R^2 = .640$ to $.978$ for the positive conditions and $.675$ to $.996$ for the negative conditions.

A $2$ (purpose) $\times$ $2$ (rating scale) $\times$ $3$ (configural weight interaction) Hotelling’s $T$ statistic was computed on the configural weight parameters estimated for the individual subjects to determine which, if any, configural weights differed significantly from zero in each condition. The results are shown in Table 3. The $T^2$ statistic was significant ($p < .05$) for all four conditions indicating at least one configural weight was not zero. A post-hoc analysis of the 95% confidence intervals surrounding the means of the
TABLE 2
FREQUENCY DISTRIBUTIONS OF $R^2$% FOR Raters IN EACH CONDITION USING THE
CONFIGURAL AVERAGING MODEL

<table>
<thead>
<tr>
<th>$R^2$ interval</th>
<th>Pos hours</th>
<th>Pos pay</th>
<th>Neg hours</th>
<th>Neg pay</th>
</tr>
</thead>
<tbody>
<tr>
<td>.60 to .69</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>.70 to .79</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>.80 to .89</td>
<td>13</td>
<td>13</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>.90 to .89</td>
<td>19</td>
<td>18</td>
<td>20</td>
<td>13</td>
</tr>
</tbody>
</table>

configural weights showed that only the configural weight for the efficiency by promptness interaction ($CW_{e/p}$) in both positive hours and pay tasks was significantly different from zero. But in the two negative tasks, both $CW_{e/p}$ and the configural weight for the responsibility by promptness interaction ($CW_{r/p}$) had 95% confidence intervals below zero. Mean values of the configural weight for the efficiency by responsibility interaction ($CW_{e/r}$) were not significantly different from zero in any condition and were not needed for the policy models.

This difference in the contributions of the configural weights between positive and negative conditions shows some support for the hypothesis that the purpose of the appraisal would influence the way subjects used the information. It appears that when subjects appraised workers to reduce hours or pay, promptness had more impact on the ratings when the worker was described as highly efficient rather than inefficient. Information about promptness also had more impact on subjects' ratings if the worker was described as being of high responsibility rather than irresponsible. When the workers were appraised for rewards, the subjects used the promptness information differently depending on the level of the worker's efficiency. The configural averaging model with the efficiency by promptness configural weight which best describes both positive pay and hours conditions will be referred to as the one configural weight model (CW1 model). The configural averaging model which includes both a configural weight adjustment for the efficiency by promptness relationship and the responsibility by promptness relationship that is most representative of both the negative hours and pay purposes will be referred to as the CW2 model.

**Individual Differences**

Although the CW1 model was found to best represent raters who were rewarding workers and the CW2 model best represented rates who punished workers, there were individual differences. The distributions of the configural weights for the responsibility by promptness ($CW_{r/p}$) interaction for each appraisal purpose are shown in Figure 9. The distributions of the configural weights for the negative purposes show more variance and more negative configural weight values than the distributions of configu-
TABLE 3
Hotelling's T Statistic

<table>
<thead>
<tr>
<th>Condition</th>
<th>T-stat</th>
<th>CW e/r</th>
<th>CW e/p</th>
<th>CW r/p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive hours</td>
<td>12.07*</td>
<td>.025</td>
<td>-.058*</td>
<td>-.025</td>
</tr>
<tr>
<td>Positive pay</td>
<td>21.92**</td>
<td>.035</td>
<td>-.052*</td>
<td>-.038</td>
</tr>
<tr>
<td>Negative hours</td>
<td>52.77**</td>
<td>-.025</td>
<td>-.080*</td>
<td>-.177*</td>
</tr>
<tr>
<td>Negative pay</td>
<td>46.90**</td>
<td>-.018</td>
<td>-.121*</td>
<td>-.076*</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01.

dural weights for the positive purpose conditions. The distributions for the positive conditions show more configural weights equal to zero. The negative pay condition appears to have a bimodal distribution of configural weights. Subjects were more likely to interpret promptness as a function of responsibility when decreasing hours or pay than when increasing hours or pay. The same pattern was found in the distributions of configural weights for the efficiency by promptness (CW_{e,p}) interactions. The

![Graph](image_url)

**Fig. 9.** Distribution of responsibility by promptness configural weights for all subjects. (A) Hours tasks and (B) pay tasks.
strategy differences related to the purpose of the appraisals are evident at the individual level.

A Kolmogorov-Smirnov test\(^3\) was computed to find if the distributions of configural weights were equivalent between purposes and between rating scales. The results showed that the configural weight distributions for responsibility by promptness (as shown in Fig. 9) differed for reward and punishment purposes \((D_{m,n} = .30, p < .01)\) regardless of the response scale. No differences were found between rating scales within either purpose with this test.

It was concluded that subjects evaluated the workers differently under negative and positive purposes. Under punishment conditions, the configural weight parameter \(CW_{r/p}\) was negative for the majority of the subjects. This reflects the subjects' views that the impact of promptness is greater when the worker is also high in responsibility than when the worker is low in responsibility when cutting hours or pay for most subjects. Under reward conditions, the configural weight parameter \(CW_{r/p}\) was approximately zero for a majority of the subjects. This indicates that when rewarding workers, the impact of promptness did not depend on the workers' level of responsibility.

**Subjective Values**

A 2 (purpose) \(\times\) 2 (rating scale) repeated measures ANOVA was computed on the subjective values for the medium level of promptness (the other levels were fixed) and a 2 (purpose) \(\times\) 2 (rating scale) \(\times\) 3 (level of attribute) repeated measures ANOVA was computed on the subjective value parameters for efficiency and responsibility. When subjects were evaluating work performance to reward workers their interpretation of the performance characteristics was more positive than when they were evaluating work performance to punish workers (efficiency: \(F(1,34) = 23.19, p < .001\); responsibility \((1,34) = 10.02, p < .001\); promptness: \(F(1,34) = 16.70, p < .001\)). In general, the main effects were significant for levels of efficiency: \(F(2,68) = 160.97, p < .001\), and responsibility: \(F(1,68) = 149.81, p < .001\). There were no significant differences in the rating scales or interactions for efficiency or promptness, but responsibility appeared to be interpreted differently as a function of the rating scale: \(F(1,34) = 7.46, p < .01\). In general, these results indicate that the raters' subjective values of the attributes differed depending on the purpose of the appraisal, but did not differ for the most part for the rating scales.

Policy modeling can also be used to test the assumption that the attribute manipulations are subjectively distinct. Table 4 lists the number of

---

\(^3\) This test compares the cumulative distributions of the samples and is sensitive to differences in central tendency, dispersion, or skewness (Siegel & Castellan, 1988).
TABLE 4
TALLY OF NUMBER OF SUBJECTS PERCEIVING THE LEVELS OF PERFORMANCE ON EACH DIMENSION TO BE EQUALLY SPACED

<table>
<thead>
<tr>
<th>Policy capturing assumption:</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) - (b) = 0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of subjects in each condition

<table>
<thead>
<tr>
<th></th>
<th>Negative</th>
<th></th>
<th>Positive</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hours</td>
<td>Pay</td>
<td>Hours</td>
<td>Pay</td>
</tr>
<tr>
<td>Efficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a = b</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>a &gt; b</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>a &lt; b</td>
<td>24</td>
<td>22</td>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>Responsibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a = b</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>a &gt; b</td>
<td>17</td>
<td>14</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>a &lt; b</td>
<td>15</td>
<td>16</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Promptness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a = b</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>a &gt; b</td>
<td>17</td>
<td>15</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>a &lt; b</td>
<td>13</td>
<td>17</td>
<td>26</td>
<td>23</td>
</tr>
</tbody>
</table>

subjects: (1) whose subjective values were equally spaced between levels for each attribute, (2) who felt the greatest subjective difference was between the first and second levels of each attribute, and (3) who felt the greatest subjective difference was between the second and third levels of each attribute.

Only a small percentage of the subjects perceived the levels of any attribute to have equally distinctive levels (the efficiency/positive pay cell was the exception). There was strong agreement among subjects for the efficiency attribute. Most of the subjects (regardless of the purpose of the appraisal) perceived the low and medium level of efficiency more similar to each other than the high performance level. The perception of responsibility and promptness depended on the purpose to the appraisal. When judging workers for rewards, two-thirds of the subjects perceived workers who needed close supervision and hourly supervision as more similar than workers needing very little supervision. When judging performance to cut pay or hours, subjects were less consistent. Half of the subjects felt workers who needed close supervision were more similar to workers who needed hourly supervision. The other half perceived workers who needed hourly supervision to be more similar to workers who needed little supervision.

The same pattern was obtained for the promptness attribute. When evaluating the workers for a reward, the majority of the subjects per-
ceived the best performance ("never late") as more distinctive than the less favorable levels of performance. When judging the workers for a negative purpose, the subjects were less consistent.

Importance Weights

Whereas subjective values indicate how levels of the same attribute compare, importance weights indicate how much a particular attribute affects ratings relative to the other information. The importance of an attribute can be inferred from the change in slope of the observed mean ratings of workers described by different combinations of attributes. This is illustrated for each condition in Fig. 10. The dashed line represents the mean ratings for workers described by the lowest level of responsibility in combination with efficiency and by the lowest level of promptness in combination with efficiency. The bottom line represents a combination of efficiency with lowest levels of responsibility and promptness. The pat-

Fig. 10. Mean responses across subjects for efficiency alone (dashed line) and in combination with lowest levels of other attributes. All four conditions.
terns of ratings appear similar across positive tasks and across negative tasks, but differ between purposes. Efficiency was more important when cutting pay or hours than when increasing pay or hours. The variance in the mean ratings for this attribute alone is quite pronounced. Adding negative information produced a bigger change in ratings to cut pay or hours than to reduce the reward level.

The relative change in slopes indicate how important information is to the subjects. Regardless of the purpose, subjects lowered their ratings when they were given more unfavorable information about the worker. This indicates that the subjects did not assume poor performance when the information was not available. For both purposes, promptness was more important (i.e. produced a greater change in the mean rating) to the raters than responsibility. The slope for efficiency and low promptness is less than the slope for efficiency and low responsibility. Therefore, the importance weights for promptness are relatively higher than the importance weight for responsibility. This pattern is similar in the positive conditions, but less pronounced.

Since these weights are unique up to a multiplicative constant they were normalized so that comparisons could be made across conditions. Normalized importance weights for the three attributes were computed for each subject and condition by dividing each weight parameter by the sum of weights. The mean and standard deviation of the normalized weights for each attribute is shown by condition in Table 5. A tally of each individual’s highest valued attribute is also shown in Table 5. The frequency distribution of the most important attribute is quite consistent across purpose and rating scales. Raters found promptness to be the most important attribute (especially in the negative conditions), efficiency was next most important and responsibility was least important in all conditions.

It was also of interest whether the distributions of the normalized importance weights for each attribute differed between purpose and task. For the promptness attribute, the relative weights estimated from both negative tasks show more variance and greater extreme values than the relative weights derived from the positive tasks. A Kolmogorov-Smirnov test was used to identify differences in purpose and task for relative importance weights of all three dimensions. The distribution of normalized weights for promptness differed significantly between positive and negative conditions (largest cumulative difference, $D_{m,n} = .357$, $p < .001$). That is, promptness was more important to more raters when they had to punish workers than when rewarding workers. No differences in the distributions for efficiency and responsibility between purposes were found. Cumulative distributions did not differ between pay and hours tasks for any of the relative importance weights indicating that type of response scale did not affect the rater’s judgments as to the importance of each attribute.


<table>
<thead>
<tr>
<th></th>
<th>Negative</th>
<th></th>
<th>Positive</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hours</td>
<td>Pay</td>
<td>Hours</td>
<td>Pay</td>
</tr>
<tr>
<td>Efficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.26</td>
<td>0.30</td>
<td>0.27</td>
<td>0.30</td>
</tr>
<tr>
<td>SD</td>
<td>(0.13)</td>
<td>(0.13)</td>
<td>(0.14)</td>
<td>(0.14)</td>
</tr>
<tr>
<td>Responsibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.30</td>
<td>0.24</td>
<td>0.23</td>
<td>0.26</td>
</tr>
<tr>
<td>SD</td>
<td>(0.15)</td>
<td>(0.12)</td>
<td>(0.10)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>Promptness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.47</td>
<td>0.50</td>
<td>0.36</td>
<td>0.34</td>
</tr>
<tr>
<td>SD</td>
<td>(0.17)</td>
<td>(0.18)</td>
<td>(0.12)</td>
<td>(0.13)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Individual (Number of subjects placing greatest importance on)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficiency</td>
<td>10 8 12 12</td>
</tr>
<tr>
<td>Responsibility</td>
<td>8 3 5 6</td>
</tr>
<tr>
<td>Promptness</td>
<td>17 24 18 17</td>
</tr>
</tbody>
</table>

Response Functions

Earlier it was reported that subjects did not use the 9-point Likert rating scale, whether for hours or pay, as being a linear scale in the positive conditions. The group response function was positively accelerated when rewarding workers and approximately linear when punishing workers. In order to determine if each subject’s response function matched the group response function, each individual’s response function was assessed. Results showed that 32 of the 35 subjects’ responding strategies were accurately represented by a positively accelerated function in both positive pay and hours tasks. In the negative tasks, 25 subjects in decrease hours and 28 in cut pay conditions were best represented by a linear function. The majority of the individual response strategies corresponded to the response function described for the fit of the group means.

DISCUSSION

The manipulation of the purpose of the appraisal may be similar to manipulating point of view. Birnbaum and Stegner (1979) and Birnbaum et al. (1992) compared the judgments of buyers and sellers and found that the utility of the gambles varied as a function of the point of view. In particular, both studies concluded that the point of view could be represented by different configural weights. In the current study, the purpose of the appraisal determined the magnitude and distribution of configural weights for individual subjects. Subjects who were making the appraisal
in order to decrease pay or hours were more likely to interpret one characteristic of the worker's behavior differently depending on another piece of information. Subjects who were evaluating the workers in order to increase pay or hours were more likely to average the information without any configural adjustment.

The purpose of the appraisal also had an effect on the response function. A positively accelerated function mapped the evaluations onto the two positive scales, but there was a linear relationship between the evaluations and the two negative scales. Figure 6 shows the fit of the configural averaging model assuming a linear response scale. The mean responses show deviations from a straight line fit in both positive conditions. Figure 8 shows the same model with the same data, but includes b-spline parameters to adjust the prediction to account for the subjects' perceptions of the response scale. This flexibility improved the fit for positive purposes. The response function mapping the predicted performance appraisals onto the observed rating scale was positively accelerated for both positive response scales and approximately linear for both negative responses scales. The majority of the subjects' individual response functions matched that representing the group.

Atkin and Conlin (1978) have suggested that the rating scale used should be studied in conjunction with how it may affect a rater's strategy. This study found that there was little difference in raters' strategies whether a rating scale consisting of hours or pay was used, but there were differences when the rating scales represented different purposes.

Differences in relative weights were found depending on purpose of appraisal. The promptness of a worker was the most important attribute for the raters when assigning rewards, but efficiency and responsibility received moderately strong weights. When raters had to punish workers, promptness received relatively more weight than efficiency and responsibility and significantly more weight than when rewarding workers. These characteristics were consistent across both hours and pay rating scales.

Individual analyses indicated that the raters' subjective values of the levels of the attributes differed between conditions where they were to reward the workers and conditions where they were to punish them, but their subjective values were approximately the same for both hours and pay rating scales within each condition. The mean subjective values for each attribute were lower under punishment conditions than under reward conditions.

Although the configural averaging model was successful in the current study, other alternatives exist. Norman and his colleagues have had success in applying the multiplicative model to situations involving managerial decision making (Norman & Singh, 1989) and judging probabilities
(Norman & Louviere, 1974). However, the multiplicative model did not
fit our data as well as the configural averaging model.

Policy capturing, a technique that employs the multiple regression
model, has often been used in performance appraisal research. In the
present study, policy capturing (represented by Eq. (1) using a linear
response function) was compared to the averaging model for both group
and individual analyses. The mean fit statistic for both positive tasks \( R^2 \)
approximately .35 were typical of those found in the literature (Slovic &
Lichtenstein, 1983) and the mean fit for both negative tasks was lower \( R^2 \)
approximately .17. Individual fits ranged from \( R^2 = .028 \) to .65. When
additional parameters representing quadratic and cubic terms were added
to the model to equal the number of parameters in the configural averag-
ing model, the fit was an improvement over Eq. (1), but the resulting beta
weights were either not significant or uninterpretable. The simple main
effects for each attribute were large and negative. It appears that the
restrictive assumptions of multiple regression did not allow the model to
"capture" the subjects' underlying cognitive processes (Norman & Lou-
viere, 1974; Champagne & Stevenson, 1992).

According to the responses to post-study questions, the task was be-
lievable and interesting to the subjects. More often than not, the next
experimental session was delayed while subjects gave unsolicited opin-
ions about the behaviors of the "workers" and related stories of their own
experiences in the work study program. Most subjects were surprised at
the absence of a uniform appraisal system within the university and many
gave their own thoughts on what was and what was not a fair reward
system—even after being debriefed about the nature of the study.

A lab setting was necessary for this study to allow for control over
manipulations and the stimuli to be presented. Policy modeling has never
been applied to a performance appraisal prior to this study. Its success in
real organizations to identify the strategies of real supervisors will depend
both on its success in the laboratory and on a successful method of finding
the most relevant behavior characteristics of workers. Although some
might criticize the "paper people" nature of the study (Gorman, Glover
& Doherty, 1978), we believe the relevance and believability of this study
to the subjects, and their high involvement in the task provides a reason-
able context to study the judgment process. Levin, Louviere, and Schep-
anski (1983) provided a number of examples of lab studies similar to the
present study demonstrating that responses to judgments in the labora-
tory are meaningfully related to factors external to the task and are pre-
dictive of subsequent non-laboratory decisions. Field studies will be ne-
cessary to adapt the methods tested in the laboratory to the organization.

In this study, policy modeling has addressed some of the underlying
cognitive processes suggested by researchers or incorporated in recent
cognitive models such as how objective information is perceived by a rater (subjective values of attributes) and how information is integrated (by use of combination strategies). DeNisi et al. (1984) have suggested that a major issue of performance appraisal studies concerns the relative weights assigned to different items of information which form the evaluation. Policy modeling allows the assessment of individual differences in regards to the importance of a given attribute. The last step of the DeNisi et al. (1984) model is the conversion of a rater's decision onto some response scale. Policy modeling allows any monotonic conversion that a rater may use, although a majority of the subjects in the present study had similar responding strategies within each purpose.

In order to solve the performance appraisal problems in the field, the focus of the system must shift from reliance on the perceptual accuracy and memory of human observers to more objective measurement techniques. We believe that policies can only be established when the biasing factors are eliminated from the information available, which necessitates real policies to be established with a fictitious worker sample. When the key attribute are defined, more systematic, objective measurement techniques will be needed to gather unbiased information about the workers. The policy model can then be used to combine the information according to the objectives of the organization. If humans are asked to do this task, irrelevant information is very likely to bias the ratings. The policies can be clearly articulated to the workers. Therefore, if the purpose of the appraisal is to match work performance with outcomes or feedback we believe that Policy Modeling can be a useful tool to guide the process and eliminate the contamination of the evaluation with irrelevant information.

Researchers have spent a great deal of time examining the relationship between rater characteristics and criteria of rating effectiveness. As discussed earlier, few solid findings have been found. Instead of using the difference between subjects and "expert" ratings as a dependent measure of accuracy, a more useful approach might be a detailed description of the combination strategies and parameters representing the patterns of ratings. Subjects are not required to evaluate their strategies, but only to provide a carefully defined sample of ratings. With meaningful parameters, comparisons can be made among personnel, across purposes and for different positions. We hope that the utility of this approach has been established in this study.

REFERENCES


**RECEIVED:** March 5, 1992