A PSYCHOPHYSICAL STUDY OF PERFORMANCE RATINGS

by

MARIA DEL CARMEN MARTIN DOMINGO

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

MASTER OF ARTS

APPROVED, THESIS COMMITTEE:

Jay C. Thomas, Assistant Professor of Psychology
Chairman
William C. Howell, Professor of Psychology

David M. Lane, Assistant Professor of Psychology

HOUSTON, TEXAS
JUNE, 1982
ACKNOWLEDGMENTS

I would like to thank the members of my committee: Drs. Jay C. Thomas, William C. Howell, and David M. Lane for their helpful comments, suggestions, and significant contribution to my intellectual development during my graduate education at Rice University. Special thanks are due to my advisor, Jay C. Thomas, not only for his consistent and invaluable academic help, but also for his constant enthusiasm and moral support which allowed me to initiate and complete this thesis with total satisfaction and enjoyment.

My deepest appreciation is extended to Drs. Mike Watkins and Kevin Murphy. Their approach to scientific inquiry and their constant understanding and friendship provided me with the encouragement to overcome all the difficulties.

I would also like to thank the Banco de Bilbao and Fulbright Commission for enabling me to achieve my Master's Degree through their granting of my scholarship.

Finally, I want to dedicate this work to my family, especially to my parents and friends, here and in Spain. Without them, this thesis would never have been completed.
# 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>iii</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Method</td>
<td>20</td>
</tr>
<tr>
<td>Results</td>
<td>27</td>
</tr>
<tr>
<td>Discussion</td>
<td>54</td>
</tr>
<tr>
<td>References</td>
<td>63</td>
</tr>
<tr>
<td>Appendix 1</td>
<td>66</td>
</tr>
<tr>
<td>Appendix 2</td>
<td>67</td>
</tr>
<tr>
<td>Appendix 3</td>
<td>68</td>
</tr>
<tr>
<td>Appendix 4</td>
<td>69</td>
</tr>
<tr>
<td>Appendix 5</td>
<td>70</td>
</tr>
<tr>
<td>Appendix 6</td>
<td>72</td>
</tr>
<tr>
<td>Appendix 7</td>
<td>74</td>
</tr>
<tr>
<td>Appendix 8</td>
<td>76</td>
</tr>
<tr>
<td>Appendix 9</td>
<td>79</td>
</tr>
<tr>
<td>Appendix 10</td>
<td>80</td>
</tr>
<tr>
<td>Appendix 11</td>
<td>81</td>
</tr>
<tr>
<td>Appendix 12</td>
<td>82</td>
</tr>
</tbody>
</table>
INTRODUCTION

Improving the effectiveness of performance evaluation procedures has been an issue that has received constant attention by researchers due to the key role that the performance assessment plays in organizations. The use of ratings as primary criteria of performance evaluation and for many administrative decision making about employees has been generalized and, in fact, for many jobs, performance ratings are the only indicators of individual contributions to the organization. Unfortunately, the validity or accuracy of these ratings has been found to be questionable: raters may commit a variety of perceptual, judgmental, and/or response errors rendering inaccurate their evaluations of ratee's performance effectiveness. As a consequence, an enormous amount of research has been carried out in order to identify and to reduce the systematic and random distortions in the rating process.

Downey and Saal (1978) have reviewed the literature dealing either directly or indirectly with rating errors or other indices for rating quality. They have found that halo, leniency/severity and central tendency are the major criteria used for evaluating rating quality. Reliability has also been invoked as an index for rating quality; and other criteria for judging the adequacy of performance ratings have included the extent to which similarity error, contrast error, first impression error, and biases due to sex, race, and age are minimized. Downey and Saal (1978) have also found that while definitions of these types of errors were generally in agreement, only the methods for quantifying halo showed appreciable consistency. Their analysis revealed a lack
of agreement with regard to the operational definitions of central tendency, leniency and several other criteria for rating quality such as discrimination among raters and interrater reliability or agreement.

Most of the research of the past five decades has been focused on examining different strategies to reduce or eliminate these sorts of errors from performance ratings. One such strategy has been to provide rater training to reduce errors in performance evaluations. The analysis of the literature examining the effects of different training methods on reducing rating errors allows the conclusion that such raters training programs seem to reduce significantly halo effects, leniency, and contrast errors (Bernardin, 1978; Bernardin and Walter, 1977; Latham, Wexley, and Pursell, 1975; Ivancevich, 1979). However, it has also been shown that by reducing these sorts of errors, training does not necessarily improve rating accuracy (Borman, 1979). The assumption that the instrument used to elicit information has an effect on the accuracy of the ratings has led to the study of the potential effects of different ratings formats on the reduction of biases and errors by the raters. Comparison of the relative merits in reducing rating errors between the traditional graphic scales and the widely used Behavioral Observation Scales (BOS) has been debated by a number of authors. The review of the literature (Landy and Farr, 1980) suggests that there is no clear evidence that BOS are psychometrically superior (in terms of halo, leniency, and reliability) to the most traditional trait-oriented scales.

Although the research mentioned above has provided us with important findings in order to improve methods for reducing rating errors
and to improve rating instruments, the important question of whether or not ratings are an accurate measure of performance remains unanswered. As Borman (1977, 1978) and Gordon (1970) have pointed out, rating errors such as leniency and halo appear to be unrelated to the accuracy of performance evaluations and, consequently, the use of these errors as criteria for evaluating rating accuracy has to be questioned.

**Ratings Accuracy.**

If ratings quality is to be considered in terms of accuracy, two aspects need to be addressed: the content of the measurement, that is, what is to be assessed, and the definition and operationalization of accuracy. With regard to the content of the ratings, it would be necessary to differentiate whether personality traits, behaviors, outcomes or overall judgments are the categories upon which the measurement of accuracy is going to be based.

Research about the appropriateness of different performance evaluations contents is far from being conclusive. As Kane and Lawler (1979) have stressed, if personality traits are characteristics of the person rather than of the person's performance, regardless of the strength of their causal effect on the latter, then traits should not be taken as content in performance evaluation. However, if performance is shown to be a trait of the person, then the conclusion that traits would qualify as legitimate elements of performance evaluation would still be valid (Kavanagh, 1971).

Kane and Lawler (1979) have also pointed out that the choice between behaviors and outcomes as the content of performance assessment
should depend on the level of "equifinality" that characterizes a particular job. Equifinality is defined as "a characteristic of a job (or any other goal directed system) referring to the extent to which the job's objectives or desired outcomes can be reached by two or more equally acceptable paths" (p. 445). Therefore, if a particular job is not characterized by equifinality, performance evaluation should be based upon the extent to which the incumbents exhibit the appropriate behaviors. On the other hand, outcomes should be the content of performance evaluation in situations in which equifinality prevails in a job, since by definition no one set of behaviors is inherently preferable to the behaviors comprising the other equally acceptable routes.

However, Murphy, Martin, and Garcia (1982) have shown that performance ratings based on behavioral observation scales are highly correlated with trait-oriented ratings in situations involving recall. This finding suggests that when performance ratings are based upon long term, as is the case in most of the performance evaluation situations, behavioral ratings represent inferences about the frequencies of different behavior rather than observations of behavior; therefore, trait judgments become crucial cues in those inferences. Consequently, the controversy of whether or not behaviors are more appropriate measures of performance than personality traits or outcomes remains far from being solved.

The definition and operationalization of ratings accuracy has been mainly addressed within the research on person perception; therefore the study of accuracy has been mainly based on the judgments of personality traits rather than on behaviors. Cronbach (1955) has provided one of the main contributions to the study of judgmental accu-
racy. He called attention to a number of methodological problems associated with the measurement of perceptual accuracy and showed that the overall index of judgmental accuracy was made up of several different accuracy components which he labeled Elevation, Differential Elevation, Stereotype Accuracy, and Differential Accuracy. Cronbach (1955) suggested that these components should be analyzed separately due to the fact that they might yield uncorrelated accuracy scores.

After Cronbach's paper, attention has been paid to specify the type of accuracy which is more relevant for addressing a particular research question. Given that the Differential Accuracy (DA) component reflects the extent to which a rater is able to predict difference among others on each trait or item considered separately, DA has been suggested to be the most appropriate component to assess the accuracy of performance judgments (Sechrest and Jackson, 1961; Borman, 1977, 1979).

As Scheneider, Hastorf, and Ellsworth (1979) have pointed out, an accuracy measure is based on differences between predicted and actual scale positions. Therefore, in order to evaluate the accuracy of performance ratings, it is necessary to obtain a "true" score measure of performance. This "true score has been generally developed by expert raters selected to evaluate the effectiveness of different levels of performance displayed on videotapes (Borman, 1977; Murphy et al. 1982). If interjudge agreement is high, and if expert ratings agree with the intended levels of performance, mean expert ratings are taken as "true" score indicators of performance.

So far, the research has focused on the accuracy of the final
judgments. However, more recently, Feldman (1981), Landy and Farr (1980), and Decotiis (1977) have suggested that performance evaluation should be considered as a specific case of more general cognitive processes, and they have emphasized the need for a better understanding of the cognitive processes involved in a rating task. The attention to and recognition to relevant information, the organization and storage of this information, and the recall and integration of information into some sort of final judgments are cognitive operations that the raters have to perform, and that need to be investigated.

Given that the raters are asked to base their judgments on the observation of performance, it seems more appropriate to examine rating accuracy in terms of how accurate raters are in the perception, recognition and recall of specific behavioral events. Within this framework, attention is focused not on the accuracy of the evaluation of performance but on the accuracy in observation, recognition and recall of performance and its relationship with the evaluation of that performance by the raters.

Literature on person perception and impression formation has addressed the study of the effects of inference making on recall of behaviors (Carlston, 1980; Hastie, 1980). Nevertheless, there is a systematic lack of research on the performance evaluation literature addressing the analyses of the processes by which raters recognize and recall specific behaviors on which they base their judgments and how and to which extent this information determine the judgments to be made.
Only recently, Murphy, Garcia, Kerkar, Martin, and Balzer (1981) have studied the relationship between observational accuracy and accuracy in performance evaluation. Using procedures similar to Borman's (1977), they obtained measures of the four accuracy components described by Cronbach (1955) and correlated accuracy scores indicating accuracy in observing behaviors with accuracy in evaluating performance. Their findings suggest that accuracy in observation is related in different ways to accuracy in performance evaluation. However, one shortcoming of this study was the fact that the number of behaviors displayed in each videotape was not controlled, and when the raters were asked to report actual counts of each of the behaviors, they showed extremely low levels of interrater agreement. Therefore, the final scale used to rate the occurrence of behaviors ranged from Never (0) to once or twice (1) to throughout the entire lecture (5). Thus, the expert ratings level of agreement, rather than the true number of behaviors, was the criteria against which raters accuracy of behavioral observation was compared.

In an attempt to give more light on this issue, the present study addressed the analysis of the accuracy in performance observation and recognition and its relationship with performance evaluation. The approach taken followed the psychophysical methods. This approach is based on the fact that in a performance evaluation situation, a rater can be considered as a measuring instrument who is asked to judge stimuli (the ratee's performance), and this point of view is parallel to the classic psychophysics.

In a broad sense, psychophysics can be defined as the quantitative
branch of the study of perception, examining the correspondence between the magnitude of stimulus properties as assessed by the instruments of physics and as assessed by the perceptual systems of people (Baird and Noma, 1978). Thus, while psychophysics is concerned with relationships between physical and psychological magnitudes, the aim of this study was concerned with relationship between performance observation and the evaluation of that performance.

In psychophysics human beings are considered as measuring instruments interpreting physical stimuli impinging upon sensory systems from the surrounding environment. It is assumed that the responses or judgments made by the human perceptual system may be systematically analyzed. Therefore, the psychophysical theory provides us with a set of statements that describe how an organism processes stimulus information. Moreover, the psychophysical models allow us to know "how" a person behaves as a measuring instrument and to predict "what" a person will actually do under carefully specified conditions.

Among the different ways to study raters as measuring instruments, two psychophysical models were chosen based on their appropriateness for the purpose of the present study: the Signal-Detection Theory as applied to recognition memory, and the psychophysical function as applied to the relationship between observation and performance evaluation.

The Signal-Detection Theory

The Signal-Detection model for recognition memory, widely used in the learning and memory literature, seems to be an appropriate approach to the study of rating accuracy of observable behaviors. The model enables us to derive an estimate of the amount and kind of information
in memory on which subjects base recognition judgments.

The Signal-Detection Theory (TSD), which first appeared in 1954 (Peterson, Birdsall, and Fox, 1954) was originated within both statistical decision theory and electrical engineering. However, when considered from a psychological perspective, the theory can be viewed as a natural consequence of classical psychophysics and Thurstonian scaling. Thurstone's contributions to scaling and psychophysics extended Fechner's classical approach (Fechner, 1966) by devising methods to find psychophysical scales value for complex stimuli even if corresponding physical measurements were unknown. As Baird and Noma (1978) have pointed out, the similarity between classical and Thurstonian approaches lies in the assumptions made by both models: (1) that a single stimulus evokes a normal distribution of internal events, and (2) that stimuli give rise to overlapping internal distributions. The innovative aspect of TSD was to provide a different approach to the analyses of judgments by emphasizing cognitive factors that act independently of the subject's sensory ability to distinguish stimuli. The most important aspect of the theory was the distinction and separation of a judgment into two components, one sensory and one cognitive. Thus, an observer is considered (1) as a sensor, i.e. his/her sensitivity to discriminate among stimuli, and (2) as a decision maker, i.e. the effects of his/her values and expectations on his/her judgments.

The initial development of the TSD within the psychological experimentation was due to Tanner and Swets (1954) who applied the theory to the analysis of human sensory and perceptual processes (Green and Swets, 1966; Swets, 1973). Egan (1958) applied TSD to measure verbal
retention and, since then, the techniques of TSD have been used in a va-
riety of analyses of human memory.

The application of TSD to the analysis of behavioral observation
accuracy depends on conceiving of behaviors as signals the subject must
detect in a recognition task. In a basic performance recognition situa-
tion, the subject is given an observational interval during which he/she
observes another person's behavior. After the interval, the subject is
presented with a list of behaviors which he/she has to recognize as
having occurred during the observational period. The decision problem in
recognition is analogous to a psychophysical detection situation in
which the subjects make a binary yes (the behavior occurred) or no
(the behavior did not occurred) response based on the strength of a con-
tinuous input.

Given the subject's "yes" or "no" responses to occurrence of be-
haviors, a decision matrix like that shown in Fig. 1 can be constructed.
Such a matrix is generally the basic datum for any experimental
condition considered by TSD. A Hit (H) occurs when the subject correct-
ly identifies that a behavior occurred, and a False Alarm (FA) occurs
when the subject incorrectly identifies that a behavior occurred when
actually it did not.

Fig. 2 illustrates the probability density distributions showing
how likely an observation is to have resulted from presentation of a
listed behavior that did not occur in the observational interval (left
distribution), or from a listed behavior that was presente in the obser-
vational interval (right distribution).
Fig. 1 The four possible outcomes of a rater's decision and their conditional probabilities.

Fig. 2 Concepts in the Signal-Detection theory as applied to behavioral recognition memory.
Fig. 2 shows the two values that represent the sensory (d') and the decision (β) aspects. The difference between the means of the two distributions is a measure of their distance on the familiarity or strength continuum. In the TSD, the distance between the means of the two distributions in standard-deviation units becomes a measure known as d-prime (d'), which is essentially a measure of how sensitive the subjects are to the difference between behaviors that have or have not occurred. Thus, d' is considered a measure of "true" sensitivity and it represents the information that is in memory (Klataky, 1980). That is, the separation of the two distributions is an indication of the level of sensory discrimination possible, and the term "sensitivity" refers as to how well two stimuli can be discriminated. Therefore, the higher the value of d', the easier will be for the subject to discriminate between stimuli of the two distributions.

The second theoretical value is called beta (β), which is the criterion upon which the subject bases the decision. The value of the response criterion beta will be determined by the rater's expectancies concerning the behavior being rated and the payoff matrix, which takes into account the values of Hits and Correct Rejections, and the costs of Misses and False Alarms.

The use of TSD as a model for the analyses of perceptual accuracy as applied to rating tasks in the industrial field has been overlooked by the researchers in the performance evaluation area. However, TSD offers a highly potential value to study certain problems of performance evaluation in industry. More specifically, Baker and Schuck
(1975), applying TSD to Gordon's data (Gordon, 1970), have shown the appropriateness of the model for the study of behavioral recognition accuracy.

Gordon (1970) has studied the effects of the correctness of the behaviors rated on the accuracy of the raters. Previous work in the area of perceptual accuracy (Harris, 1968; Jacobs and Vandeventer, 1968) has shown that accuracy appeared to be related to the correctness of the input. However, limitations of these studies due to small sample size and differences in the number of correct and incorrect samples viewed do not allow to conclude that identification of correct stimuli is easier than the identification of incorrect stimuli. In order to specifically test this hypothesis, Gordon (1970) carried out a more systematic and controlled study. In this study the subjects were managers of life insurance agencies and were given scripts containing lines the "agent" (actors playing the role and recorded on video tape) should say in order for his sales approach to be considered correct. The manager's task was to view the simulated sales situation and to judge, by means of a rating checklist, whether the "agent's" behavior was correct or incorrect. It was found that 88.0% of the agent's correct behaviors were identified, whereas only 73.8% of his incorrect behaviors were judged to be incorrect. This difference was statistically significant and was called the Differential Accuracy Phenomenon (DAP), which can be defined as a False Alarm rate greater than the rate of Misses. That is to say, raters are more accurate in observing correct behaviors than incorrect behaviors.
The implications of this finding are important; consequently, the question of whether or not DAP is a generalizable phenomenon or whether it is peculiar to the rating situation employed in Gordon's study, needs to be investigated. Based on the analysis of the perceptual errors underlying the DAP, Gordon (1972) suggested that this phenomenon might be an indication of the favorability with which the raters view the ratee. In order to rule out this alternative explanation he carried out a second experiment in which he specifically tested the hypothesis that the DAP is due to a predisposition on the part of the rater to make favorable judgments of the ratee. It was found that the ability to rate accurately was not impaired by the tendency to view the ratee in a generally favorable way that allows us to draw the conclusion that DAP is unrelated to the favorability of the rater's general impression of the ratee.

In order to investigate the factors underlying the DAP, Baker and Schuck (1975), applying TSD reanalyzed Gordon's (1970) data. The results of the analysis suggested that differing rater expectancies or payoff matrices resulted in the DAP effect for some dimensions and the converse for others. Moreover, no relation was found between DAP and the difference in the difficulty of the perceptual task involved in recognizing correct and incorrect behaviors.

As it has been pointed out earlier, the innovative aspect of the TSD was the separation of a judgment into a sensitive (d') and a decision (\(\beta\)) aspects. Therefore, as Baker and Schuck (1975) have stressed, according to the TSD there are three factors that might
determine whether or not DAP will exist: (1) the expectancies of the rater about the behaviors being rated, (2) the values and costs in the rater's payoff matrix, and (3) the level of difficulty of the rating task.

According to the above mentioned findings, one possible explanation for the DAP to occur might be that the subjects in the Gordon's study were provided, prior to the observational task, with a set of scripts which contained the lines the "agent" ought to say exactly as if he made his approach correctly, but the incorrect behaviors were not provided. Therefore, it might be possible that the script with correct behaviors served to subjects as cues producing a change in their expectancies and, consequently, a selective attention effect towards the correct behaviors. The present study was addressed to rule out this alternative explanation.

The Psychophysical Function.

As it has been suggested earlier, the psychophysical function provides us with an appropriate approach to the study of the relationship between observation and evaluation of performance. Stevens (1972) investigating the relationship between stimuli and responses found that, for many sensory phenomena, the psychological impression of the physical intensity follows a simple mathematical function: the power function described as

\[ J = k I^p \]

where \( J \) is the Judgment of psychological magnitude, \( I \) is the physical Intensity, \( p \) is the power, the size of the exponent that governs the
relationship between physical and psychological magnitude, which determines the curvature of the power function, and $k$ is an arbitrary constant, simply serving to get the psychological magnitude into the actual numbers used by the subject. When plotted in log-log coordinates (logarithmic scales on both axes), the graph of a power function becomes a straight line where $p$ is the slope. In logarithm, the power law equation becomes:

$$\log J = p \log I + \log K$$

Stevens (1972) has summed up the relation between stimulus and sensory response in the following statement: "Equal stimulus ratios produce equal subjective ratios". Essentially, the power law, known as the Steven's law, says that whenever an observer matches any continuum to any other continuum, he/she will produce a new power function = a straight line in log-log coordinates. Moreover, it predicts that the new matching function will have a slope (exponent) equal to the ratio between the two original slopes.

The applicability of the Steven's law as a magnitude estimation procedure in experiments involving human judgments has received considerable support (Stevens, 1972; Selling and Wolfgang, 1964). However, the psychophysical law has received little attention in the area of performance evaluation. Only Whitlock (1963) has addressed the analysis of the relationship between observation and evaluation of performance following the psychophysical methods. In accordance with Steven's law, Whitlock hypothesized the existence of a lawful relation between the number of performance specimens observed for an individual during an observational period and the overall evaluation of that per-
formance. The findings of his several studies allowed him to conclude that the judged quality of performance grows as a function of the ratio of the number of effective to ineffective performance specimens observed. Consequently, he concluded that the proper expression for the relation between performance observation and evaluation can be described by the Stevens's law in the following terms:

\[ R = \lambda S^n \]

where "R" is the evaluation of performance, "S" is the number of specimens observed and "\( \lambda \)" and "n" are constants.

The Whitlock's findings have been surprisingly overlooked by the researchers on performance evaluation, but its implications for the study of judgment processes are of great importance, considering the similarity of this approach to the process model of performance rating proposed by Landy and Farr (1980). As Whitlock has stressed "anything less than a perfect correlation between appropriate performance measures and measures of evaluation means that some attribute of the observer or the performance observed or their interaction is significantly influencing the transition from observation to evaluation of performance" (p.22).

One shortcoming of Whitlock's studies is the fact that subjects were asked to check the performance specimens that they had observed on the part of the ratees. However, the "true" performance specimens of the ratees were not controlled, and the accuracy of the raters observations was not examined. It has to be pointed out that, according to Whitlock's findings, the ratio of correct to incorrect behaviors will determine the slope of the psychophysical function. Therefore, if
Raters tend to commit more False Alarms than Misses errors for correct behaviors as Gordon (1970, 1972) has found, then the ratio of correct to incorrect behaviors would be higher and, the evaluation ratings would also be higher. This would imply that the level of accuracy would affect the evaluation ratings, and this phenomenon would be reflected in the size of the psychophysical function.

Another aspect not considered in Whitlock's studies is the possibility that the relation between performance observation and evaluation of that performance might be described in terms of another psychophysical function, as the logarithmic function, known as the Fechner's law. In contrast to the Stevens's law, the Fechner's law predicts that a "geometric increase in the stimulus is accompanied by an arithmetic increase in the sensation" (Baird and Noma, 1978, p.87). According to this formulation, if the evaluation of performance is plotted against the logarithm of the ratio of the number of behavioral specimens observed, the result would be a straight line. Therefore, in order to determine which function better describes the relation between performance observation and evaluation, the results obtained with the Stevens's law should be compared with the results that might be obtained with the Fechner's law.

The present study was designed to overcome the above mentioned limitations and was specifically addressed:

1. To study the effects of the instructions on the accuracy of recognition for correct and incorrect behaviors.
2. To examine whether the judged level of performance grows as a power function of the "true" ratio of the number of specimens of effective
to ineffective performance.

3. To explore the relationship between observational accuracy and performance evaluation.

The following hypotheses were tested:

1. DAP is not a phenomenon to be generalized to all kinds of situations, but only when raters are provided with a list of correct behaviors prior to the observation of performance.

   1a. The converse of DAP would be found when raters are provided with a list of incorrect behaviors prior to the observation of performance.

2. The level of accuracy in recognizing behaviors is higher when subjects are provided with a list of behaviors to be identified than when no cue at all is given prior to the observation of performance.

3. The proper expression for the relation between "true" level of performance and evaluation of that performance follows the Stevens's law which states that "Equal stimulus ratios produce equal subjective ratios".

   3a. Differences in observational accuracy would yield to differences in the slopes of the psychophysical function.

Hypotheses 1 and 2 dealt directly with recognition accuracy in performance observation, and hypotheses 3 addressed the relationship between observation and evaluation of performance.

Although the terms of "appropriate" and "inappropriate" behaviors appear to be more adequate to describe a performance situation, in order to replicate the situation specified by Gordon (1970, 1972), the terms "correct" and "incorrect" behaviors were used throughout the study.
**METHOD**

Development of Behavioral Incidents and Videotapes. The performance of a Computer Programming Consultant was chosen for the rating task. Dimensions or aspects representative of a Computer Programming Consultant's performance were developed by experts. The final dimensions were: (1) Reception to the student, (2) Attention to the problem, (3) Empathy, (4) Use of the Manual, (5) Explanation of how to solve the problem, (6) Confirmation that the student has understood, (7) Answering to questions, and (8) Further help. For each dimension two examples of behaviors were developed: one correct and one incorrect. A pilot study was carried out in order to assure that the behavioral incidents were clear examples of correct and incorrect behaviors. Appendix 3 shows the final list of correct and incorrect behaviors for each dimension.

Scripts for ten videotaped situations were prepared. The ten films represented situations in which a student came in to the Consultant's office presenting a problem with the computer. Real students played the roles of the "Student". The role of the Consultant was always played by the same person who was an actual Computer Programming Consultant in the Social Science Computing Laboratory (SSCL). The setting for each of the ten situations was the SSCL. Final versions of each of the ten films ran approximately three minutes.

The dialogue for each film was written in such a way that different proportions of what the Consultant did was correct and incorrect in terms of the task standards. Table 1 contains a summary of the structure of the ten films in terms of correct and incorrect behaviors,
<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Films</th>
<th>Practice</th>
<th>Practice</th>
<th>Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Reception to the student</td>
<td>I</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>2. Attention to the problem</td>
<td>I</td>
<td>C</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>3. Empathy</td>
<td>C</td>
<td>I</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>4. Use of the Manual</td>
<td>I</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>5. Explanation</td>
<td>I</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>6. Assurance of understanding</td>
<td>I</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>7. Questions</td>
<td>I</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>8. Further Help</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of behaviors without practice films</th>
<th>1/7</th>
<th>5/3</th>
<th>4/4</th>
<th>7/1</th>
<th>6/2</th>
<th>2/2</th>
<th>2/6</th>
<th>3/5</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct behaviors</td>
<td>I</td>
<td>I</td>
<td>C</td>
<td>I</td>
<td>C</td>
<td>I</td>
<td>C</td>
<td>I</td>
<td>4</td>
</tr>
<tr>
<td>Incorrect behaviors</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1: Correct and Incorrect behaviors presented on each film.
the order they were presented to the subjects.

The first two films were practice films. These were included to enable the subjects to get acquainted with the requirements of the rating task and to become familiar with the rating material. Only the last eight films were taken into account for the data analysis.

The total number of behaviors of interest in each film was eight, except in one film, in which the total number was four. The ratios of correct to incorrect behaviors were either 7/1, 6/2, 5/3, 4/4, 2/2, 3/5, 2/6, or 1/7.

As it can be seen from Table 1, the total number of correct and incorrect behaviors over the eight films were equal. There were thirty correct behaviors and thirty incorrect behaviors.

Once the ten situations were filmed, a pilot study was carried out in order to assure that the behaviors were clearly perceived.

Rating Scales. Two types of rating instruments were developed: a rating scale to measure the performance of the Consultant, and four behavioral recognition checklists. The first rating scale was a Likert-type scale which asked raters to evaluate the Consultant's overall performance for each film, on a scale ranging from Very Ineffective (1), to Very Effective (5) (Appendix 9). All the checklists randomly contained the same 29 items. There were 16 items corresponding to the correct and incorrect behaviors that the Consultant should or should not perform in order to consider his approach to be done perfectly. Moreover, 13 of the items were neutral behaviors included in the checklists as distractors. Four different
checklists were developed in order to control the effect of the order of presentation of the behaviors. These checklists asked the subjects to answer with "Y" (yes) or "N" (no) whether they had observed those behaviors in the Consultant. Appendix 8 shows the checklist for the films presented in order two, six, and ten.

Sample and Procedures. 68 graduate and undergraduate students participated in the experiment. The participation was voluntary, or in change for course credit, or money. Those subjects who participated for money were paid $4.00. 52 subjects were students attending Rice University, 11 were enrolled in different departments in the Medical Center, four were enrolled at Houston Community College, and one subject was attending University of Houston.

All subjects were asked to view, in the same order, the set of ten films presenting situations of the Computer Programming Consultant's performance. The subjects were randomly assigned to one of four groups, depending on whether or not a list of behaviors was given prior to the observational task: (1) No list - subjects were not provided with any list of behaviors, (2) List of correct behaviors - subjects were provided with a list of correct behaviors that the Consultant should perform, (3) List of incorrect behaviors - subjects were provided with a list of incorrect behaviors that the Consultant should not perform, or (4) List of correct and incorrect behaviors - subjects were provided with a list of correct and incorrect behaviors that the Consultant should and should not perform respectively. Appendices 4, 5, 6, and 7 show the instructions given to the subjects
in groups 1, 2, 3, and 4 respectively. Appendices 1, 2, and 3 contain the list of behaviors given to groups 2, 3, and 4 respectively. In those groups in which subjects were provided with a list of behaviors, special care was taken that the subjects read carefully and got familiar with the list of behaviors. After the subjects had seen each film, they were required to complete the checklist and to evaluate the performance of the Consultant on that particular film.

**Scoring procedure.** In order to determine the accuracy of each subject's observations, the ratings were scored by comparing them with the key used to develop the actual scripts for each of the films. This key contained the actual behaviors displayed by the Consultant on each of the eight dimensions. Two scoring procedures were used to determine the accuracy of all the ratings for each subject: the first procedure used was to score the two kinds of errors: Misses and False Alarms. These errors were scored for both correct and incorrect behaviors. The second procedure used was to score the Hits for correct and incorrect behaviors also. These procedures allowed computation of two accuracy scores: one score for the accuracy of the ratings of behaviors which, according to the key, were actually performed correctly, and one score for the accuracy of ratings of behaviors which were actually performed incorrectly. These accuracy scores were averaged across the eight films.

**Data Analysis.** As it was mentioned earlier, two procedures were followed to compute accuracy scores. One procedure was based upon the errors (False Alarms and Misses) for correct and incorrect behaviors. There-
fore, in order to test the hypotheses dealing with recognition accuracy (hypotheses 1, 1a, and 2), two different repeated measures ANOVAs of behaviors (2), by groups (4) were conducted, one ANOVA with accuracy scores based on errors, and another ANOVA with accuracy scores based on hits.

In order to examine the relationship between observational accuracy and performance evaluation, the mean evaluation for each film averaged across all subjects and across the four groups were calculated. Moreover, the mean of the perceived number of correct and incorrect behaviors for each film averaged across all subjects and across the four groups were also computed, and with this data the mean ratios of correct to incorrect behaviors were obtained for each film. The "true" and perceived ratios of correct to incorrect behaviors were plotted on log-log paper against the mean evaluation for each film. Similarly, in order to compare the Stevens's law with the Fechtner's law, the mean evaluation for each film was plotted against the logarithm of the ratios of correct to incorrect behaviors.

To analyze the difference in the slopes of the psychophysical function, regression equations were computed for all subjects, in which the dependent variable was log (Evaluation) as a function of log (ratio of correct to incorrect behaviors). Taken the beta weights obtained in this analysis as dependent variables, two ANOVAs of groups (4) were conducted: one ANOVA to examine the difference in the slopes among the four groups, and another ANOVA to examine the differences in the intercepts among the four groups also.
An ANOVA of groups (4) by films (8) was conducted in order to examine the differences in evaluation among the four groups.
Analyses of Variance for Accuracy. The results of the ANOVA based on errors showed the following pattern. The main effect of groups was significant, $F, (3,64)=3.63, p < .01$. Table 2 shows the mean percentage of errors for the four groups. The highest percentage of total errors (24.13) corresponds to group 3, which was provided with a list of incorrect behaviors. The lowest percentage of total errors (17.25) corresponds to group 2, which was provided with a list of correct behaviors. The mean percentage of total errors for group 1, which was not provided with any list of behaviors, and group 4, which was provided with a list of correct and incorrect behaviors, was the same (23.13). As shown in Fig. 3, taking into account errors, group 2 was the most accurate, group 3 was the least accurate and groups 1 and 4 showed the same level of accuracy. A Newman-Keuls test (Winer, 1971) was carried out in order to make a post-hoc comparison among group means. The results were not able to show which means differed from each other.

The main effect of behaviors was also significant, $F(1,64)=23.47, p < .01$. Table 3 shows the mean percentage of errors for correct and incorrect behaviors. The mean percentage of total errors for correct behaviors was 25.29, while the same statistic based on incorrect behaviors was 18.62. Fig. 4 shows that, contrary to DAP, accuracy for correct behaviors is lower than accuracy for incorrect behaviors.

The main effect of errors was not significant, $F(1,64)=.11, p > .05$. Therefore, the mean percentage of False Alarms (21.76), and the mean
### Table 2
Mean percentage of errors - Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>FAC</th>
<th>MISC</th>
<th>FAI</th>
<th>MISI</th>
<th>Total errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - No list</td>
<td>20.39</td>
<td>28.23</td>
<td>23.53</td>
<td>20.39</td>
<td>23.13</td>
</tr>
<tr>
<td>2 - List correct</td>
<td>20.39</td>
<td>21.96</td>
<td>17.25</td>
<td>9.41</td>
<td>17.25</td>
</tr>
<tr>
<td>3 - List incorrect</td>
<td>25.09</td>
<td>32.94</td>
<td>20.39</td>
<td>18.82</td>
<td>24.31</td>
</tr>
<tr>
<td>4 - List correct-incorrect</td>
<td>23.53</td>
<td>29.80</td>
<td>23.53</td>
<td>15.68</td>
<td>23.13</td>
</tr>
</tbody>
</table>

FAC - False Alarms for correct behaviors  
MISC - Misses for correct behaviors  
FAI - False Alarms for incorrect behaviors  
MISI - Misses for Incorrect behaviors

### Table 3
Mean percentage of errors - Behaviors

<table>
<thead>
<tr>
<th>Behaviors</th>
<th>Errors</th>
<th>Total errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>FA: 22.35, MISS: 28.23</td>
<td>25.29</td>
</tr>
<tr>
<td>Incorrect</td>
<td>FA: 21.17, MISS: 16.07</td>
<td>18.62</td>
</tr>
</tbody>
</table>
Figure 3. Mean Errors–Groups
Figure 4. Mean Errors-Behaviors
percentage of Misses (22.15) were not significantly different.

The interaction between behaviors and errors was significant, \( F(1,64)=9.96, p < .01 \) (Fig. 5). Table 3 shows that the mean percentage of Misses for correct behaviors is higher than the mean percentage of False Alarms (22.35) for correct behaviors. However, the mean percentage of Misses (16.07) for incorrect behaviors is lower than the mean percentage of False Alarms (21.27) for incorrect behaviors.

The interaction between groups and behaviors was not significant, \( F(3,64)=1.22, p > .05 \). The interaction between groups and type of errors was not significant, \( F(3,64)=1.45, p > .05 \). The interaction between groups, type of errors, and behaviors was not significant, \( F(3,64)=.10, p > .05 \).

The results of the ANOVA based upon hits showed the following pattern. The main effect of groups was significant, \( F(3,64)=3.95, p < .01 \). Table 4 shows the mean percentage of hits for the four groups. The highest percentage of total hits (88.62) corresponds to group 2, which was provided with a list of correct behaviors. The percentage of hits for group 1, which was not provided with any list of behaviors, was 81.56; the percentage of hits for group 4, which was provided with a list of correct and incorrect behaviors was 80.00. The lowest percentage of hits corresponds to group 3 (78.43) which was provided with a list of incorrect behaviors.

As shown in Fig. 6, taking into account hits, group 2 was also the most accurate, and group 3 was the least accurate. Similar to the results obtained with errors, the Newman-Keuls test did not show which group means differed from each other.
Figure 5. Mean Errors vs. Behaviors
Figure 6. Mean Hits–Groups
Table 4
Mean percentage of Hits - Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Behaviors</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Incorrect</td>
<td>Total</td>
</tr>
<tr>
<td>1 - No list</td>
<td>76.86</td>
<td>86.27</td>
<td>81.56</td>
</tr>
<tr>
<td>2 - List correct</td>
<td>81.56</td>
<td>95.68</td>
<td>88.62</td>
</tr>
<tr>
<td>3 - List incorrect</td>
<td>73.72</td>
<td>83.13</td>
<td>78.43</td>
</tr>
<tr>
<td>4 - List correct- incorrect</td>
<td>73.72</td>
<td>86.27</td>
<td>80.00</td>
</tr>
</tbody>
</table>

Table 5
Mean percentage of Hits - Behaviors

<table>
<thead>
<tr>
<th>Behaviors</th>
<th>Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>76.47</td>
</tr>
<tr>
<td>Incorrect</td>
<td>87.84</td>
</tr>
</tbody>
</table>
The main effect of behaviors was also significant, $F(1,64)=35.60, p<.01$. Table 5 shows that the mean percentage of hits for incorrect behaviors (87.84) is higher than the same statistic based on correct behaviors (76.47), Fig. 7.

The interaction between groups and behaviors was not significant, $F(3,64)=0.38, p>.05$.

As it can be observed, in terms of accuracy the results of the ANOVA based upon errors were similar to the ANOVA based upon hits. Results from both analyses show that group 2 is the most accurate and that group 3 is the most inaccurate. However, while no differences in accuracy were found between groups 1 and 4 in terms of errors, group 1 shows to be more accurate than group 4 when hits are taken into account. Moreover, both ANOVAS showed that the level of accuracy for incorrect behaviors is higher than for correct behaviors.

**Beta values.** Following the TSD, the decision criterion ($\beta$) provides a different approach to the analysis of the differences in accuracy between correct and incorrect behaviors. Hits and False Alarms rates were used in order to calculate beta values for correct and incorrect behaviors, according to Hochhaus (1972). As shown in Table 6, the beta values for correct behaviors are higher than for incorrect behaviors. These results seem to suggest that subjects used a stricter criterion for correct behaviors than for incorrect behaviors. As Baker and Shuck (1975) have pointed out, DAP is the consequence of a lax criterion, while the converse of DAP is the consequence of a strict criterion. Consequently, the difference in beta values for
Figure 7. Mean Hits-Behaviors
correct and incorrect behaviors agrees with the fact that, in this study, subjects were shown to be more accurate with incorrect behaviors than with correct behaviors.

**True and perceived ratios vs. Evaluation.** Table 7 shows the mean evaluation for each film averaged across all subjects. Fig. 8 shows the plot on log-log paper of the "true" and perceived ratios of correct to incorrect behaviors against the mean evaluation for each film averaged across all subjects.

Tables 8 and 9 shows the mean evaluation for each film averaged across the four groups. Figs. 9, 10, 11, and 12 show the plot on log-log paper of the "true" and perceived ratios of correct to incorrect behaviors against the mean evaluation for each film averaged across the four groups. As it can be seen in all the cases, a good
### Table 7
Ratios of Behaviors and Evaluation of Films
All subjects

<table>
<thead>
<tr>
<th>Films - Order of presentation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>True correct behaviors (C) / True incorrect behaviors (I)</td>
<td>7/1</td>
<td>4/4</td>
<td>5/3</td>
<td>1/7</td>
<td>2/2</td>
<td>2/6</td>
<td>6/2</td>
<td>3/5</td>
</tr>
<tr>
<td>True ratio C/I</td>
<td>7.000</td>
<td>1.000</td>
<td>1.667</td>
<td>0.143</td>
<td>1.000</td>
<td>0.333</td>
<td>3.000</td>
<td>0.600</td>
</tr>
<tr>
<td>Perceived correct behaviors (C')</td>
<td>6.044</td>
<td>3.324</td>
<td>4.971</td>
<td>0.794</td>
<td>4.118</td>
<td>1.279</td>
<td>5.809</td>
<td>2.074</td>
</tr>
<tr>
<td>Perceived incorrect behaviors (I')</td>
<td>1.485</td>
<td>4.603</td>
<td>2.691</td>
<td>6.956</td>
<td>1.603</td>
<td>5.721</td>
<td>2.191</td>
<td>6.235</td>
</tr>
<tr>
<td>Perceived ratio C'/I'</td>
<td>4.744</td>
<td>0.982</td>
<td>2.015</td>
<td>0.124</td>
<td>2.887</td>
<td>0.228</td>
<td>2.687</td>
<td>0.372</td>
</tr>
</tbody>
</table>
### Table 8

**Ratios of Behaviors and Evaluation of Films**

**Group 1 and Group 2**

<table>
<thead>
<tr>
<th>Films - Order of presentation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived correct behaviors (C)</td>
<td>6.294</td>
<td>2.941</td>
<td>5.059</td>
<td>0.824</td>
<td>4.000</td>
<td>1.235</td>
<td>6.000</td>
<td>2.118</td>
</tr>
<tr>
<td>Perceived incorrect behaviors (I)</td>
<td>1.529</td>
<td>4.882</td>
<td>2.176</td>
<td>6.941</td>
<td>1.647</td>
<td>5.765</td>
<td>1.882</td>
<td>6.235</td>
</tr>
<tr>
<td>Ratio C/I</td>
<td>4.990</td>
<td>0.672</td>
<td>2.729</td>
<td>0.131</td>
<td>2.926</td>
<td>0.213</td>
<td>3.533</td>
<td>0.392</td>
</tr>
<tr>
<td><strong>Group 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived correct behaviors (C)</td>
<td>6.176</td>
<td>3.529</td>
<td>5.471</td>
<td>0.824</td>
<td>4.118</td>
<td>1.294</td>
<td>5.588</td>
<td>2.235</td>
</tr>
<tr>
<td>Perceived incorrect behaviors (I)</td>
<td>1.765</td>
<td>4.647</td>
<td>2.588</td>
<td>7.000</td>
<td>1.647</td>
<td>5.824</td>
<td>2.647</td>
<td>6.412</td>
</tr>
<tr>
<td>Ratio C/I</td>
<td>4.534</td>
<td>1.100</td>
<td>2.276</td>
<td>0.123</td>
<td>2.922</td>
<td>0.222</td>
<td>2.498</td>
<td>0.373</td>
</tr>
</tbody>
</table>
Table 9  
Ratios of Behaviors and Evaluation of Films  
Group 3 and Group 4

<table>
<thead>
<tr>
<th>Films - Order of presentation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived correct behaviors (C)</td>
<td>6.176</td>
<td>3.235</td>
<td>4.706</td>
<td>0.765</td>
<td>4.176</td>
<td>1.353</td>
<td>5.824</td>
<td>2.000</td>
</tr>
<tr>
<td>Perceived incorrect behaviors (I)</td>
<td>1.059</td>
<td>4.588</td>
<td>2.882</td>
<td>6.765</td>
<td>1.588</td>
<td>5.412</td>
<td>2.059</td>
<td>6.176</td>
</tr>
<tr>
<td>Ratio C/I</td>
<td>5.875</td>
<td>1.048</td>
<td>1.773</td>
<td>0.128</td>
<td>2.756</td>
<td>0.261</td>
<td>3.006</td>
<td>0.365</td>
</tr>
<tr>
<td>Evaluation</td>
<td>4.000</td>
<td>2.647</td>
<td>3.294</td>
<td>1.382</td>
<td>3.382</td>
<td>1.676</td>
<td>3.559</td>
<td>1.912</td>
</tr>
<tr>
<td>Group 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived correct behaviors (C)</td>
<td>5.529</td>
<td>3.588</td>
<td>4.647</td>
<td>0.765</td>
<td>4.176</td>
<td>1.235</td>
<td>5.824</td>
<td>1.941</td>
</tr>
<tr>
<td>Perceived incorrect behaviors (I)</td>
<td>1.588</td>
<td>4.294</td>
<td>3.118</td>
<td>7.118</td>
<td>1.529</td>
<td>5.882</td>
<td>2.176</td>
<td>6.118</td>
</tr>
<tr>
<td>Ratio C/I</td>
<td>4.612</td>
<td>1.107</td>
<td>1.578</td>
<td>0.113</td>
<td>3.255</td>
<td>0.217</td>
<td>2.811</td>
<td>0.357</td>
</tr>
</tbody>
</table>
Figure 8. Mean evaluation as a function of the ratios of correct to incorrect behaviors. Log-log coordinates. (Stevens's scale) All subjects.
Figure 9. Mean evaluation as a function of the ratios of correct to incorrect behaviors. Log-log coordinates (Stevens's scale) Group 1.
Figure 10. Mean evaluation as a function of the ratios of correct to incorrect behaviors. Log-log coordinates (Stevens's scale).
Figure 11. Mean evaluation as a function of the ratios of correct to incorrect behaviors. Log-log coordinates (Stevens's scale).

Group 3.

Ratios of Correct to Incorrect Behavior (C/I)
Figure 12. Mean evaluation as a function of the ratios of correct to incorrect behaviors. Log-log coordinates (Stevens's scale) Group 4.
fit to a straight line is perceived, as predicted by the Stevens's law. However, as shown in Figs. 13, 14, 15, 16, and 17, when the mean evaluation is plotted against the logarithm of the ratios of correct to incorrect behaviors, the function shows also a good fit to a straight line, in agreement with the prediction based on Fechner's law.

**Test of slopes and intercepts in the psychophysical function.** Regression equations were computed for all subjects. The dependent variable was log(Evaluation) as a function of log(Ratio of correct to incorrect behaviors). The \( t \) test for the individual regression equations showed to be significantly different from zero in 56 cases. The median and the interquartile range of the correlation between log(Evaluation) and log (Ratio of correct to incorrect behaviors) for the four groups are shown in Table 10.

The results of the ANOVA for the slopes showed a significant main effect of groups, \( F(3,64)=4.18, p<.01 \). Table 11 shows that the highest mean of slopes corresponded to group 2, and the lowest mean corresponded to group 1. A Tukey test (Winer, 1971) was carried out in order to make a post-hoc comparison among group means. The results showed a significant difference between group 2 (list of correct behaviors) and the other three groups.

The results of the ANOVA for the intercepts showed that the intercepts of the psychophysical function were not significantly different among the four groups, \( F(3,64)=0.86, p>.05 \).

**Analysis of Variance for Evaluation.** In order to analyze whether dif-
Figure 13. Mean evaluation as a function of the logarithm of the ratios of correct to incorrect behaviors. (Fechner's scale) All subjects.
Figure 14. Mean evaluation as a function of the logarithm of the ratios of correct to incorrect behaviors. (Fechner's scale) Group 1.
Figure 15. Mean evaluation as a function of the logarithm of the ratios of correct to incorrect behaviors. (Fechner's scale)
Figure 16. Mean evaluation as a function of the logarithm of the ratios of correct to incorrect behaviors (Fechner's scale) Group 3.
Figure 17. Mean evaluation as a function of the logarithm of the ratios of correct to incorrect behaviors (Fechner’s scale)
Group 4.
Table 10
Correlations between log(Evaluation) and log (ratio of correct to incorrect behaviors)

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>Q1</th>
<th>Q3</th>
<th>Q range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - No list</td>
<td>.87</td>
<td>.77</td>
<td>.93</td>
<td>.16</td>
</tr>
<tr>
<td>2 - List correct</td>
<td>.89</td>
<td>.86</td>
<td>.95</td>
<td>.09</td>
</tr>
<tr>
<td>3 - List incorrect</td>
<td>.87</td>
<td>.77</td>
<td>.91</td>
<td>.14</td>
</tr>
<tr>
<td>4 - List correct-incorrect</td>
<td>.88</td>
<td>.80</td>
<td>.93</td>
<td>.13</td>
</tr>
</tbody>
</table>

Table 11
Means of Slopes

<table>
<thead>
<tr>
<th>Group</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - No list</td>
<td>0.2674</td>
</tr>
<tr>
<td>2 - List correct</td>
<td>0.3951</td>
</tr>
<tr>
<td>3 - List incorrect</td>
<td>0.2773</td>
</tr>
<tr>
<td>4 - List correct-incorrect</td>
<td>0.2705</td>
</tr>
</tbody>
</table>
ferences in accuracy would yield to differences in performance evaluation, an ANOVA of groups (4) by films (8) was conducted. No significant differences in evaluation for each film were found among the four groups.
DISCUSSION

The results of the present study clearly suggest that DAP is not a phenomenon to be generalized to all kinds of situations. Correct and incorrect behaviors were not rated with the same degree of accuracy. A differential accuracy phenomenon was found, but in a contrary direction to Gordon's findings. In Gordon's studies (1970, 1972) it was found that raters were more accurate with correct than with incorrect behaviors. In the present study, the difference in accuracy between the observations of correct and incorrect behaviors was statistically significant in terms of both hits and errors. However, the direction of this difference indicated that raters were more accurate with incorrect behaviors than with correct behaviors. The magnitude of this effect was 11.37% for hits and 6.67% for errors.

It was hypothesized that DAP would occur only when raters are provided with a list of correct behaviors prior to the observation of performance. The results of both ANOVAS, for hits and errors, did not support this hypothesis. The interaction between groups and behaviors was not significant, suggesting that there is no tendency for raters who were provided with a list of correct behaviors to be more accurate with correct behaviors than with incorrect behaviors.

As stated before, in this study it was found that raters are more accurate with incorrect behaviors than with correct behaviors; therefore, our results appeared consistent, at first glance, with the hypothesis that the converse of DAP would be found when raters are provided with a list of incorrect behaviors prior to the observation.
of performance.

However, the hypotheses about recognition accuracy were based upon the assumption that the differences in accuracy for correct and/or incorrect behaviors were due to the fact that the presentation of a list of behaviors prior to the observational task served the raters as cues, producing a selective attention effect towards the behaviors (correct and/or incorrect) provided in the list. According with this hypothesis, no DAP or converse of DAP would have been found for group 1, which was not provided with any list of behaviors, nor for group 4 which was provided with a list of correct and incorrect behaviors. Given that the converse effect of DAP was found for all groups, that is, raters in all groups were more accurate with incorrect than with correct behaviors, the hypothesis that the cause of DAP, or the converse of DAP, is the list of behaviors provided to the raters before the observational task, received no support in this study.

The hypothesis that the level of accuracy in recognizing behaviors is higher when subjects are provided with a list of behaviors to be identified than when no cue at all is given prior to the observation of performance was not confirmed. According to this hypothesis, it would have been expected that group 1, which was not provided with any list of behaviors before the observational task, showed the lowest level of accuracy. The results of both ANOVAS, for hits and errors, suggested that raters who were provided with a list of correct behaviors were the most accurate in recognizing behaviors, while raters who were provided with a list of incorrect behaviors were the most inaccurate. These findings suggest that the level of accuracy does not
depend on the amount, but rather, on the kind of information provided to the subjects prior to the observational task.

With regard to the relationship between level of performance and evaluation of that performance, the plot on log-log paper of the perceived ratios of correct to incorrect behaviors against the evaluation of that performance showed a good fit to a straight line. It has to be pointed out that the fit to a straight line was true for all cases except for the film which had a "true" ratio of two correct to two incorrect behaviors. This result might be explained by the fact that the perceived ratio for this film was 4.118 correct to 1.603 incorrect behaviors. Moreover, the level of accuracy for this film was clearly lower than for the rest of the films. These results suggest the possibility that the intended "true" ratio of correct to incorrect behaviors was not actually "true". However, it might also be possible that in a performance situation in which few behaviors are displayed and the correct and incorrect behaviors are equal in number, raters tend to perceive more correct behaviors and, therefore, give higher evaluations.

The plot on log-log paper of the "true" ratios of correct to incorrect behaviors for each film against the evaluation of those films was consistent with the hypothesis that the proper expression for the relation between "true" level of performance and evaluation of that performance follows the Stevens's law which states that "Equal stimulus ratios produce equal subjective ratios". Although not perfect, similarly to the plot of the perceived ratios, a good fit to a straight line was obtained in all cases. However, these findings do not allow one to conclude that the Stevens's law describes better than the Fechner's
law the relation between performance observation and evaluation, given that similar results were obtained when the mean evaluation was plotted against the logarithm of the ratio of correct to incorrect behaviors.

The hypothesis that differences in observational accuracy would yield to differences in the slopes of the psychophysical function was partially confirmed. It was hypothesized that group 2, which was provided with a list of correct behaviors, would show a DAP effect and, therefore, the slope of the psychophysical function would be higher for this group than for the rest. The results did not support the hypothesis that group 2 was more accurate for correct than for incorrect behaviors. However, it was found that this group, which had the highest level of accuracy, showed the highest slope of the psychophysical function. This finding seems to suggest that the higher the level of accuracy, the higher the evaluation.

Group 3, which was provided with a list of incorrect behaviors, appeared to be the least accurate. However, the lowest slope of the psychophysical function corresponded to group 1, which was not provided with any list of behaviors. Consequently, these findings seem to suggest that the kind of information (correct and/or incorrect behaviors) provided to the raters prior to the recognition task, has an effect on the level of accuracy in recognizing behaviors. However, both the amount and kind of information provided to the raters prior to the recognition task affects the evaluation ratings. Raters showed a tendency to rate higher in evaluation when they knew the correct behaviors that the ratees should perform than when raters did not have a standard against which to compare the performance of the ratee.
Implications and Suggestions for further research

Previous studies (Gordon 1970, 1972) have found that raters are more accurate with correct than with incorrect behaviors. Our findings suggest that, at least, in some situations, raters are more accurate with incorrect than with correct behaviors. Therefore, these results seem to indicate that generalizations about the effect of the correctness of behaviors on rating accuracy are far from being conclusive.

Our results showed evidence to support the contention that differences in accuracy for correct and incorrect behaviors are unrelated to the information provided to the raters prior to the observational task. The implications of these findings are important given the extended use of ratings as primary criteria in many areas of psychological research and, more specifically, in performance evaluations. Behavioral Observation Scales (BOS) have been recommended (Latham and Wexley, 1977), and their use has been widely generalized because of their emphasis on performance-related behaviors rather than on inferences about ratees. The fact that raters seem to be more accurate with correct behaviors in some situations, and more accurate with incorrect behaviors in others, will have direct effect on the accuracy of the ratings based on BOS. Hence, additional research would be needed in order to determine the factors that affect the accuracy for the correctness of the behaviors.

As Gordon (1970) has suggested, DAP or the converse of DAP might be determined by the rating situation, the type of rater or the behavior rated. There is some evidence suggesting that one of
those factors might be the distinctiveness or meaningfulness of the behaviors to be rated. As it was mentioned earlier, Baker and Schuck (1975) reanalyzing Gordon's data, did not find a consistent DAP effect. They found a DAP effect for some dimensions but the converse of DAP for others. This finding point out to the possibility that for some dimensions the correct behaviors were more distinctive or meaningful while for other dimensions the incorrect behaviors showed these characteristics.

Research on memory and learning provide us with evidence stressing the effects of the kind of stimulus material on recognition memory. Murdock (1974) has reported an unpublished experiment conducted at Toronto by Elizabeth Wells in which the stimulus material was either CVC (consonant-vowel-consonant) or CCC (consonant-consonant-consonant) trigrams. She found that $d'$, that is, the measurement of retention, was higher for CVC than for CCC trigrams. Moreover, Martin and Melton (1970) compared the retention of low-, medium- and high-meaningful CVC and CCC trigrams. Similarly to Elizabeth Wells' findings (Murdock 1974, p.54), they found that the Hits for CVC trigrams were higher than for CCC trigrams and that this difference decreased with meaningfulness.

The above findings suggest that the level of accuracy in behavioral recognition might not depend on the correctness of the behaviors to be rated, but on the meaningfulness or salience of those behaviors. This hypothesis would explain the apparently contradictory results obtained in the present study and in Gordon's studies (1970, 1972). It might be possible that in Gordon's studies the cor-
rect behaviors were more salient than the incorrect behaviors, while in our study the level of distinctiveness was higher for incorrect than for correct behaviors. Further research would be needed in order to rule out this alternative explanation.

Another practical implication of our findings point to the content of rater training programs. So far, workshops of this sort in industrial settings have been addressed to train raters to reduce or avoid rating errors (i.e. contrast, halo, and leniency). As Fay and Latham (1982) have pointed out, rater training programs, rather than concentrate on avoiding rating errors, should emphasize the importance of increasing raters skills in observing and recording behavior and the development of specific examples of effective and ineffective behaviors.

As it was mentioned earlier, little is known about the relationship between ratings accuracy and performance evaluation. The results of this study support previous findings (Murphy et al, 1981) with regard to the existence of a relationship between behavioral accuracy and performance evaluation. In the present study, no significant differences among groups was found with regard to the overall evaluation. However, the fact that the differences in the slopes of the psychophysical function were significant seems to point to the existence of a relationship between accuracy and overall evaluation. At the same time, this finding suggests that the relationship is influenced by the information the raters have about the standards against which the ratees performance should be compared. Raters who
only know the standards of effective performance are more accurate in observing behaviors and also tend to give higher evaluations. However, raters who only know the standards of ineffective performance are the least accurate and they tend to give lower evaluations, but not as low as raters who do not have any standard against which to compare the raters' performance.

One shortcoming of the present study stems from the fact that the number of behaviors to be observed were very few and the ratios obtained were not compared with equal ratios obtained from a larger number of correct and incorrect behaviors. According to Whitlock (1963), a ratio of 5, obtained from 50 correct to 10 incorrect behaviors, would lead to the same evaluation as a ratio of 5, obtained from 500 correct to 100 incorrect behaviors. It would imply that the amount of behaviors displayed by a person does not affect the rater's evaluation of that person. The implications of this assumption are important and, therefore, further research would be needed in order to investigate whether equal ratios obtained from different numbers of correct and incorrect behaviors lead to similar performance evaluations.

The weight ascribed to different performance incidents is another variable that might affect the relationship between behavioral observation and evaluation of that performance. Given two sets of observations in which the number of correct behaviors are equal, the set containing behaviors which are weighted higher would result in higher evaluation. Therefore, it appears of great importance to determine
to which extent the weight ascribed to correct and incorrect behaviors affect the level of evaluation.

Finally, this study was focused on recognition accuracy and its relationship to performance evaluation. In real settings, performance evaluations take place after a long period of time, during which the recall of the observed behaviors plays a crucial role. Consequently, further research should be addressed to investigate the effect of time on the accuracy with which raters recall behaviors and on the evaluation of those behaviors.

Taking into account the limitations of a study of this sort, it is considered that our results provide a further step in the understanding of some of the human processes underlying performance ratings.
REFERENCES


Theory to clarify problems of evaluating performance in
industry. Organizational Behavior and Human Performance, 1975,

Bernardin, H. Effects of rater training on leniency and halo errors
in student ratings of instructors. Journal of Applied
Psychology, 1978, 63, 301-308.

Bernardin, H. & Walter, C. Effects of rater training and diary keeping
on psychometric errors in ratings. Journal of Applied Psy-
chology, 1977, 62, 64-69.

Borman, W. Consistency of rating accuracy and rating errors in the
judgment of human performance. Organizational Behavior and

Borman, W. Exploring the upper limits of reliability and validity in
63, 135-144.

Borman, W. Format and training effects on rating accuracy and rater

Carlston, D.E. Events, inferences, and impressions. In R. Hastie,
T.M. Ostrom, E.B. Ebbesen, R.S. Wyer, D.L. Hamilton, &
D.E. Carlston (Eds.), Person Memory: The cognitive basis
of social perception. Hillsdale, N.J.: Lawrence Erlbaum,
1980.

Cronbach, L. Processes affecting scores on "understanding of others"
and "assumed similarity". Psychological Bulletin, 1955,
52, 177-193.

DeCotiis, T.A. An analysis of the external validity and applied rele-
vance of three rating formats. Organizational Behavior and
Human Performance, 1977, 19, 247-266.

Downey, R.G. & Saal, F.E. Evaluating human judgment techniques.
Paper prepared for American Psychological Association Meeting,

Egan, J.P. Recognition memory and the operating characteristic. (Techn.
Laboratory, Indiana University, 1958.


APPENDIX 1

LIST OF CORRECT BEHAVIORS THAT THE CONSULTANT SHOULD PERFORM

RECEPTION OF THE STUDENT:
. When the Student comes in, the Consultant stops what he is doing immediately.

ATTENTION TO THE PROBLEM:
. When the Student presents the problem with the printout, the Consultant attends to it immediately.

EMPATHY:
. The Consultant shows empathy towards the Student, making statements like: "Don't worry, it is common to have these problems, 'It is a difficult language', etc.

USE OF THE MANUAL:
. The Consultant brings the Manual to the Student, and explains how to use it and where the answer to the problem is.

EXPLANATIONS:
. The Consultant explains how to solve the problem with the program.

ASSURANCE OF UNDERSTANDING:
. The Consultant assures himself that the Student has understood by asking him/her if he/she has understood.

ADDITIONAL QUESTIONS:
. When the Student asks additional questions, he answers the questions very clearly.

FURTHER HELP:
. The Consultant offers further help to the Student by saying to come back and giving him/her his office telephone number just in case he/she needs it.
APPENDIX 2

LIST OF INCORRECT BEHAVIORS THAT THE CONSULTANT SHOULD NOT PERFORM

RECEPTION OF THE STUDENT:
. When the Student comes in, the Consultant makes the Student wait.

ATTENTION TO THE PROBLEM:
. When the Student presents the problem with the printout, the Consultant asks the Student to come later.

EMPATHY:
. The Consultant shows lack of empathy towards the Student, making statements like: "You don't know anything," "You don't take notes," "you sleep in class," etc.

USE OF THE MANUAL:
. The Consultant points out the Manual without explaining how to use it and where the answer to the problem is.

EXPLANATIONS:
. The Consultant writes down the solution without explaining how to solve the problem with the program.

ASSURANCE OF UNDERSTANDING:
. The Consultant doesn't assure himself that the Student has understood by asking him/her if he/she has understood.

ADDITIONAL QUESTIONS:
. When the Student asks additional questions, the Consultant answers the questions very vaguely.

FURTHER HELP:
. The Consultant refuses further help to the Student, telling him/her not to come back or sending him/her to the Professor.
APPENDIX 3

LIST OF CORRECT AND INCORRECT BEHAVIORS THAT
THE CONSULTANT SHOULD AND SHOULD NOT PERFORM

RECEPTION OF THE STUDENT:
When the Student comes in, the Consultant stops what he is doing immediately.

ATTENTION TO THE PROBLEM:
When the Student presents the problem, the Consultant attends to it immediately.

EMPATHY:
The Consultant shows empathy towards the Student making statements like: "Don’t worry, it is common to have these problems," "It is a difficult language," etc. The Consultant shows lack of empathy towards the Student making statements like: "You don’t know anything," "You don’t take notes," "You sleep in class," etc.

USE OF THE MANUAL:
The Consultant brings the Manual and explains how to use it and where the answer to the problem is.

EXPLANATIONS:
The Consultant explains how to solve the problem with the program.

ASSURANCE OF UNDERSTANDING:
The Consultant assures himself that the Student has understood by asking him/her if he/she has understood.

ADDITIONAL QUESTIONS:
The Consultant answers additional questions very clearly.

FURTHER HELP:
The Consultant offers further help to the Student by saying to come back and giving him/her his office telephone number just in case he/she needs it.

When the Student comes in, the Consultant makes the Student wait.

When the Student presents the problem, the Consultant asks the Student to come later.

The Consultant points out to the Manual without explaining how to use it and where the answer to the problem is.

The Consultant writes down the solution without explaining how to solve the problem with the program.

The Consultant doesn’t assure himself that the Student has understood by asking him/her if he/she has understood.

The Consultant answers additional questions very vaguely.

The Consultant refuses further help to the Student by telling him/her not to come back or sending him/her to the Professor.
APPENDIX 4

INSTRUCTIONS GIVEN TO SUBJECTS IN GROUP 1

COMPUTER PROGRAMMING CONSULTANT EVALUATION

As part of their training most of the Graduate and Undergraduate students are expected to be familiar with Computer Programming and, therefore, the role of the Computer Programming Consultant is a key one in helping the students to solve computer problems.

In this study we are interested in looking at the factors that determine the effectiveness of the Computer Programming Consultant's performance. We have identified a number of behaviors that the Consultants typically may or may not engage in. We are interested in the extent to which a given behavior is associated with good or poor performance.

You will be asked to view a number of short videotaped situations and then to indicate which of the Consultant behaviors you observed and to evaluate the Consultant's performance. Each of the tapes you will see represents a situation in which a Student presents some problem to the Computer Programming Consultant of the Social Science Computer Lab (SSCL).

We ask you to be as careful as possible in observing the Consultant's performance. Your ratings will provide us important feedback in order to improve the effectiveness of the Consultant in our Social Science Computer Lab. If you have any questions, do not hesitate to ask.
APPENDIX 5

INSTRUCTIONS GIVEN TO SUBJECTS IN GROUP 2

COMPUTER PROGRAMMING CONSULTANT EVALUATION

As part of their training most of the Graduate and Undergraduate students are expected to be familiar with Computer Programming and, therefore, the role of the Computer Programming Consultant is a key one in helping the students to solve computer problems.

In this study we are interested in looking at the factors that determine the effectiveness of the Computer Programming Consultant's performance. We have identified a number of behaviors that Consultants typically may or may not engage in. We are interested in the extent to which a given behavior is associated with good or poor performance.

You will be asked to view a number of short videotaped situations and then to indicate which of the Consultant's behaviors you observed and to evaluate the Consultant's performance. Each of the tapes you will see represents a situation in which a Student presents some problem to the Computer Programming Consultant of the Social Science Computer Lab (SSCL).

Before you view the videotapes, you will be given a list of the correct behaviors that the Consultant should perform in order to consider his approach to be made perfectly. These behaviors have been determined to be correct in that they help lead to a successful consultation and a positive learning experience by the student.

We ask you to be as careful as possible in observing the Consultant's performance. Your ratings will provide us important feedback in order
to improve the effectiveness of the Consultant in our Social Science Computer Lab. If you have any questions, do not hesitate to ask.
APPENDIX 6

INSTRUCTIONS GIVEN TO SUBJECTS IN GROUP 3

COMPUTER PROGRAMMING CONSULTANT EVALUATION

As part of their training most of the Graduate and Undergraduate students are expected to be familiar with Computer Programming and, therefore, the role of the Computer Programming Consultant is a key one in helping the students to solve computer problems.

In this study we are interested in looking at the factors that determine the effectiveness of the Computer Programming Consultant's performance. We have identified a number of behaviors that Consultants typically may or may not engage in. We are interested in the extent to which a given behavior is associated with good or poor performance.

You will be asked to view a number of short videotaped situations and then to indicate which of the Consultant's behaviors you observed and to evaluate the Consultant's performance. Each of the tapes you will see represents a situation in which a Student presents some problem to the Computer Programming Consultant of the Social Science Computer Lab (SSCL).

Before you view the videotapes, you will be given a list of the incorrect behaviors that the Consultant should not perform in order to consider his approach to be made perfectly. These behaviors have been determined to be incorrect in that they impair a successful consultations and a positive learning experience by the student.

We ask you to be as careful as possible in observing the Consultant's performance. Your ratings will provide us important feed-
back in order to improve the effectiveness of the Consultant in our Social Science Computer Lab. If you have any questions, do not hesitate to ask.
APPENDIX 7

INSTRUCTIONS GIVEN TO SUBJECTS IN GROUP 4

COMPUTER PROGRAMMING CONSULTANT EVALUATION

As part of their training most of the Graduate and Undergraduate students are expected to be familiar with Computer Programming and, therefore, the role of the Computer Programming Consultant is a key one in helping the students to solve computer problems.

In this study we are interested in looking at the factors that determine the effectiveness of the Computer Programming Consultant's performance. We have identified a number of behaviors that Consultants typically may or may not engage in. We are interested in the extent to which a given behavior is associated with good or poor performance.

You will be asked to view a number of short videotaped situations and then to indicate which of the Consultant's behaviors you observed and to evaluate the Consultant's performance. Each of the tapes you will see represents a situation in which a Student presents some problem to the Computer Programming Consultant of the Social Science Computer Lab (SSCL).

Before you view the videotapes, you will be given a list of the correct and incorrect behaviors that the Consultant should and should not perform in order to consider his approach to be made perfectly. These behaviors have been determined to be correct or incorrect in that they help lead to or impair a successful consultations and a positive learning experience by the student.
We ask you to be as careful as possible in observing the Consultant's performance. Your ratings will provide us important feedback in order to improve the effectiveness of the Consultant in our Social Science Computer Lab. If you have any questions, do not hesitate to ask.
APPENDIX 8

CHECKLIST

IDENTIFICATION: _______________________________ Tape # ______

Please mark with "Y" (yes) or "N" (not), if you observed in the Consultant the following behaviors:

. The Consultant looked tired. ________

. The Consultant asked the Student to which Department he belonged to. ________

. The Consultant brought the Manual to the Student and told him/her how to use it and where the answer to the problem was. ________

. The Consultant offered further help to the Student, by saying to come back and giving him/her his office telephone number just in case he/she needed it. ________

. The Consultant pointed out to the Manual without explaining how to use it and where the answer to the problem was. ________

. The Consultant placed elbow on desk. ________

. The Consultant smiled. ________

. The Consultant assured himself that the Student had understood by asking him/her if he/she had understood ________
The Consultant talked about personal things.

When the Student came in, the Consultant made the Student wait.

The Consultant walked around.

The Consultant answered additional questions asked by the Student very clearly.

The Consultant frowned frequently.

The Consultant showed lack of empathy towards the Student making statements like: "You don't know anything," "You don't take notes," "You sleep in class," etc.

The Consultant didn't assure himself that the Student had understood by asking him/her if he/she had understood.

The Consultant answered additional questions asked by the Student very vaguely.

The Consultant sat down all the time.

The Consultant explained to the Student how to solve the problem with the program.

The Consultant used his hands while talking.

When the Student presented the problem with the program the Consultant looked at it immediately.
The Consultant showed empathy towards the Student making statements like: "Don't worry, it is very common to have problems like these," "It's a difficult language," etc.

When the Student presented the problem with the program, the Consultant asked him/her to come later.

The Consultant spoke with a steady pace.

The Consultant refused further help to the Student by telling him/her not to come back or sending him/her to the Professor.

The Consultant didn't ask the Student who was his/her Professor.

When the Student came in, the Consultant stopped what he was doing immediately.

The Consultant cleared his throat several times.

When the Student came in, the Consultant was reading a magazine.

The Consultant wrote down the solution to the problem without explaining it.
CONSULTANT EVALUATION FORM

IDENTIFICATION: ________________________________ Tape # _____

Rate the tape you have just seen using the following scale:

1 2 3 4 5
Very Ineffective Average Effective Very
Ineffective Effective

79
### Table 12

Anova results - Errors

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>2.91</td>
<td>3</td>
<td>0.97</td>
<td>3.63</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Error</td>
<td>17.12</td>
<td>64</td>
<td>0.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behaviors</td>
<td>4.25</td>
<td>1</td>
<td>4.25</td>
<td>23.47</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>G x B</td>
<td>0.67</td>
<td>3</td>
<td>0.22</td>
<td>1.22</td>
<td>n.s.</td>
</tr>
<tr>
<td>Error</td>
<td>11.59</td>
<td>64</td>
<td>0.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Errors</td>
<td>0.01</td>
<td>1</td>
<td>0.01</td>
<td>0.11</td>
<td>n.s.</td>
</tr>
<tr>
<td>G x E</td>
<td>0.60</td>
<td>3</td>
<td>0.20</td>
<td>1.45</td>
<td>n.s.</td>
</tr>
<tr>
<td>Error</td>
<td>8.88</td>
<td>64</td>
<td>0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B x E</td>
<td>2.88</td>
<td>1</td>
<td>2.88</td>
<td>9.96</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>G x B x E</td>
<td>0.09</td>
<td>3</td>
<td>0.03</td>
<td>0.10</td>
<td>n.s.</td>
</tr>
<tr>
<td>Error</td>
<td>18.53</td>
<td>64</td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 13
Anova results - Hits

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>2.90</td>
<td>3</td>
<td>0.97</td>
<td>3.95</td>
<td>&lt; .01</td>
</tr>
<tr>
<td>Error</td>
<td>15.71</td>
<td>64</td>
<td>0.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behaviors</td>
<td>6.18</td>
<td>1</td>
<td>6.18</td>
<td>35.60</td>
<td>&lt; .01</td>
</tr>
<tr>
<td>G x B</td>
<td>0.20</td>
<td>3</td>
<td>0.07</td>
<td>0.38</td>
<td>n.s.</td>
</tr>
<tr>
<td>Error</td>
<td>11.12</td>
<td>64</td>
<td>0.17</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX 12

Table 14
Anova results - Evaluation

<table>
<thead>
<tr>
<th>Film</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Groups</td>
<td>86.76</td>
<td>3</td>
<td>28.92</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>3635.29</td>
<td>64</td>
<td>56.80</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Groups</td>
<td>64.71</td>
<td>3</td>
<td>21.57</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>2835.29</td>
<td>64</td>
<td>44.30</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Groups</td>
<td>63.24</td>
<td>3</td>
<td>21.08</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>3105.88</td>
<td>64</td>
<td>48.53</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Groups</td>
<td>89.34</td>
<td>3</td>
<td>29.78</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>1367.65</td>
<td>64</td>
<td>21.37</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Groups</td>
<td>86.40</td>
<td>3</td>
<td>28.80</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>2335.29</td>
<td>64</td>
<td>36.49</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Groups</td>
<td>172.06</td>
<td>3</td>
<td>57.35</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>3058.82</td>
<td>64</td>
<td>47.79</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Groups</td>
<td>30.51</td>
<td>3</td>
<td>10.17</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>3076.47</td>
<td>64</td>
<td>48.07</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Groups</td>
<td>198.16</td>
<td>3</td>
<td>66.05</td>
<td>1.41</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>2994.12</td>
<td>64</td>
<td>46.78</td>
<td></td>
</tr>
</tbody>
</table>