INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the original text directly from the copy submitted. Thus, some dissertation copies are in typewriter face, while others may be from a computer printer.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyrighted material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is available as one exposure on a standard 35 mm slide or as a 17" × 23" black and white photographic print for an additional charge.

Photographs included in the original manuscript have been reproduced xerographically in this copy. 35 mm slides or 6" × 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
An attributional analysis of differences in rating type in a performance evaluation context: A use of verbal protocol analysis

Martin, Scott Lawrence, Ph.D.
The Ohio State University, 1987
PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark √.

1. Glossy photographs or pages ______
2. Colored illustrations, paper or print ______
3. Photographs with dark background ______
4. Illustrations are poor copy ______
5. Pages with black marks, not original copy √
6. Print shows through as there is text on both sides of page ______
7. Indistinct, broken or small print on several pages ______
8. Print exceeds margin requirements ______
9. Tightly bound copy with print lost in spine ______
10. Computer printout pages with indistinct print ______
11. Page(s) ______ lacking when material received, and not available from school or author.
12. Page(s) ______ seem to be missing in numbering only as text follows.
13. Two pages numbered ______. Text follows.
14. Curling and wrinkled pages ______
15. Dissertation contains pages with print at a slant, filmed as received ______
16. Other ____________________________

_______________________________

_______________________________

U.M.I
AN ATTRIBUTIONAL ANALYSIS OF DIFFERENCES IN RATING TYPE IN A PERFORMANCE EVALUATION CONTEXT: A USE OF VERBAL PROTOCOL ANALYSIS

DISSERTATION

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

By

Scott Lawrence Martin, B.A., M.A.

* * * * *

The Ohio State University

1987

Dissertation Committee:  Approved by
Richard J. Klimoski, Ph.D.
Robert S. Billings, Ph.D.
Philip J. Smith, Ph.D.
Robert J. Vance, Ph.D.

Richard Klimoski
Advisor
Department of Psychology
ACKNOWLEDGMENTS

I wish to offer my thanks to the entire dissertation committee for providing a stimulating yet consistently supportive environment throughout this project. I am indebted to my advisor, Rich Klimoski, for investing a great deal of time and effort in my academic and professional development. He is a truly divergent thinker whose insights were as profound as those generated by any other industrial psychologist I know.

I am also grateful to the remaining members of my dissertation committee. Bob Billings was consistently capable of recognizing the most significant issues of concern at every stage of this investigation. My thanks to Phil Smith, a cognitive engineer, whose clear thinking and encouragement were responsible for transforming a study that would have been limited to methodological contributions into an interesting piece of research. Bob Vance was especially helpful since he demonstrated not only the profound thinking revealed by his colleagues, but also applied the insights in providing solutions to problems that arose.

I was also fortunate to be assisted by several competent undergraduate research assistants: Lisa Cannon, Rick DeShon,
Randy Ferguson and Charles Richardson. They performed the painstaking tasks of transcription, coding verbal reports and data entry. I must also thank all the managers who took time out of busy schedules to serve as subjects in this study.

To my friends and colleagues at Ohio State, especially Larry and Tom, you have played a significant role in making my stay at Ohio State enjoyable.

My sincere love and thanks to my parents, Barbara and Don Martin, for they are ultimately more responsible for the completion of this project than anyone else. While their achievements in academic, professional, spiritual, social and family pursuits have been outstanding, the most significant criteria involve the extent to which one's positive qualities are instilled in others. And I have yet to meet anyone who could have done as fine a job in parenting as my parents have done.

To my brothers, Greg and Eric, thanks for your love and for being such enjoyable company.

To my dear friend, and loved one, Maureen—thank you for always understanding me. I owe you a great deal.
VITA

March 3, 1960 .................... Date of Birth,
Camden, NJ

1982 ........................ B.A., Psychology,
Stonehill College,
North Easton, MA

1984 ........................ M.A., Industrial Psychology,
University of New Haven,
West Haven, CT

FIELDS OF STUDY

Major Field: Industrial Psychology

Minor Field: Applied Statistics
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PAGE</strong></td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
</tr>
<tr>
<td>VITA</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
</tr>
<tr>
<td><strong>CHAPTER</strong></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
</tr>
<tr>
<td>II. LITERATURE REVIEW</td>
</tr>
<tr>
<td>Potential Rating Sources in Performance Appraisal</td>
</tr>
<tr>
<td>Cognitive Processing View of Performance Appraisal</td>
</tr>
<tr>
<td>Studying Cognitive Processes</td>
</tr>
<tr>
<td>Verbal Protocol Analysis</td>
</tr>
<tr>
<td>Concurrent Reports as Accurately Representing Cognitive Processes</td>
</tr>
<tr>
<td>Relevance of Verbal Reports</td>
</tr>
<tr>
<td>Potential Limitations of Verbal Reports</td>
</tr>
<tr>
<td>Recognition of Motivational Influences</td>
</tr>
<tr>
<td>Completeness of Verbal Reports</td>
</tr>
<tr>
<td>Effects of Verbalization on Task Performance</td>
</tr>
<tr>
<td>Procedures for Collecting Verbal Protocols</td>
</tr>
<tr>
<td>Analysis of Verbal Protocol Data</td>
</tr>
<tr>
<td>Verbal Protocol Analysis and the Performance Appraisal Models</td>
</tr>
</tbody>
</table>
VII. CONCLUSION . . . . . . . . . . . . . . . . . . . . . . . . . . . 204

APPENDICES

A. VERBAL PROTOCOL TRANSCRIPT FOR FIRST MANAGER . . . 205
   First Supervisor Evaluation . . . . . . . . . . . . . . . . 205
   Self-evaluation . . . . . . . . . . . . . . . . . . . . . . . . 207
   Second Supervisor Evaluation . . . . . . . . . . . . . . . 208
   Third Supervisor Evaluation . . . . . . . . . . . . . . . . 210

B. CODING CATEGORY USE REGARDLESS OF RATING TYPE . . 212

LIST OF REFERENCES . . . . . . . . . . . . . . . . . . . . . . . . . 213
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>49</td>
</tr>
<tr>
<td>2.</td>
<td>52</td>
</tr>
<tr>
<td>3.</td>
<td>53</td>
</tr>
<tr>
<td>4.</td>
<td>60</td>
</tr>
<tr>
<td>5.</td>
<td>75</td>
</tr>
<tr>
<td>6.</td>
<td>76</td>
</tr>
<tr>
<td>7.</td>
<td>161</td>
</tr>
<tr>
<td>8.</td>
<td>161</td>
</tr>
</tbody>
</table>

General Components of Performance Appraisal

Abstracted version of Ilgen & Feldman (1983)

Ilgen & Feldman (1983) model of performance appraisal

Framework of verbal protocol for the integration and judgment

Preliminary empirical framework for supervisor ratings

Preliminary empirical framework for self-ratings

Final empirical framework for supervisor ratings

Final empirical framework for self-ratings
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Preliminary verbal protocol coding scheme</td>
<td>91</td>
</tr>
<tr>
<td>2. Correlations between words per minute and experience</td>
<td>97</td>
</tr>
<tr>
<td>3. P-values for contrasts between self- and supervisor evaluations</td>
<td>102</td>
</tr>
<tr>
<td>4. Establishing the validity of verbal protocols</td>
<td>104</td>
</tr>
<tr>
<td>5. Reliability estimates for 10 managers</td>
<td>107</td>
</tr>
<tr>
<td>6. Reliability estimates for 21 categories</td>
<td>109</td>
</tr>
<tr>
<td>7. Complete ANOVA for composite categories for frequency data</td>
<td>114</td>
</tr>
<tr>
<td>8. Complete ANOVA for composite categories with proportion data</td>
<td>118</td>
</tr>
<tr>
<td>9. Restricted ANOVA for composite categories for frequency data</td>
<td>121</td>
</tr>
<tr>
<td>10. Restricted ANOVA for composite categories for proportion data</td>
<td>122</td>
</tr>
<tr>
<td>11. Complete ANOVA for composites with level for frequency data</td>
<td>126</td>
</tr>
<tr>
<td>12. Complete ANOVA for composites with job for frequency data</td>
<td>127</td>
</tr>
<tr>
<td>13. ANOVA and contrast for use of internal attributions</td>
<td>129</td>
</tr>
<tr>
<td>14. ANOVA and contrast for use of external attributions</td>
<td>134</td>
</tr>
<tr>
<td>15. ANOVA and contrast for the use of consistency/improvement</td>
<td>139</td>
</tr>
<tr>
<td></td>
<td>Topic</td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td>16.</td>
<td>ANOVA and contrast for use of consensus, standards composite</td>
</tr>
<tr>
<td>17.</td>
<td>Relationship between attributions and evaluations</td>
</tr>
<tr>
<td>18.</td>
<td>Relationship between attributional information and attributions</td>
</tr>
<tr>
<td>19.</td>
<td>Post-hoc chi-square for attributional information</td>
</tr>
<tr>
<td>20.</td>
<td>Post-hoc chi-square without attributional information</td>
</tr>
<tr>
<td>21.</td>
<td>Summary of results for primary hypotheses</td>
</tr>
<tr>
<td>22.</td>
<td>Correlations of positive with negative evaluations</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

Performance appraisal plays one of the most critical roles in an organization's human resource system (Latham & Wexley, 1981; Wexley & Klimoski, 1984). Formal performance evaluations affect such organizational decisions as promotions, demotions, dismissals, salary levels and training requirements. As a result of its importance to organizational functioning, considerable attention has been paid to the development of performance appraisal systems. Despite this effort, researchers and practitioners have been discouraged by the lack of progress in this area.

Traditionally, the immediate supervisor has been responsible for evaluating the employee by completing a performance appraisal form. In response to dissatisfaction due to the limitations of this approach, Barrett (1966) is often cited for having identified the employee him/herself and the employee's peers as alternative rating sources. A number of researchers have explored the nature of ratings derived from these alternative sources and have found that they don't necessarily agree with each other. Such results put the burden on the researcher to explain the discrepancies in ratings.
Unfortunately, attempting to account for differences in ratings by source has been problematic. One reason for this is that most of the research in this area has focused solely on the outcome of the rating process—the ratings themselves. This tendency has been a response to Thorndike's (1949) book on personnel selection. Thorndike emphasized various psychometric properties in order to properly evaluate a measurement instrument. His arguments resulted in an abundance of studies emphasizing psychometric characteristics.

As a result, many of the suggestions accounting for differences in rating sources have been based on somewhat indirect evidence. Researchers have proposed that differences in ratings can be accounted for by motivational as well as informational factors. While such suggestions are probably legitimate and valid, it appears that further investigation into this area is warranted.

A more recent trend in performance appraisal suggests the type of analysis that might be useful in exploring the differences in rating sources. This phase of research can be viewed as a response to disappointment with the continued emphasis on psychometric issues. Landy & Farr (1980) and Feldman (1981) argued persuasively for placing more emphasis on the cognitive processes of the rater as opposed to rating outcomes. This shift in focus resulted in a number of studies applying theory from social and cognitive psychology to performance appraisal.
In response to this shift in focus, this study placed an emphasis on the cognitive processes of the rater. However, this approach is currently being criticized (Banks & Murphy, 1985; Ilgen & Favero, 1985; Latham, 1986) for even further widening the gap between research and practice. Cognitive processing models (see Ilgen & Feldman, 1983; DeNisi, Cafferty, & Meglino, 1984) have been highly complex, hypothetical, and difficult to test and to use in practice. As a result, it has been argued that research in this area has not been very relevant to those in applied settings. It was, therefore, considered essential that this study be conducted in a field setting.

The primary purpose of this study was to examine differences in the processing of information in the context of self- and supervisor ratings. The study focused on the formation of attributions and the use of information which has been thought to result in attributions (consistency, consensus and distinctiveness). Thus, it was an initial attempt to conduct an attributional analysis by rating target in performance appraisal. The study also served to integrate some of the cognitive processing research with performance appraisal in an applied setting. Descriptive information was obtained from managers while in the process of completing an open-ended performance evaluation. The resulting information provided insight into the value of cognitive processing models for the practice of performance appraisal.
Another purpose of this study was to provide an initial assessment of the validity and utility of a technique yet to be applied to the domain of performance appraisal. This technique is referred to as "verbal protocol analysis" and has been used with success in the areas of decision making and problem solving. Since this study used verbal protocol analysis in a new domain where different factors may be operating, it was necessary to demonstrate that the method provides accurate information with respect to one's thoughts while evaluating performance.
CHAPTER II
LITERATURE REVIEW

POTENTIAL RATING SOURCES IN PERFORMANCE APPRAISAL

As discussed in the Introduction, the supervisor has been the traditional source of performance information. For example, Lacho, Stearns & Villere (1979) found that the supervisor was the only source of ratings in 78-93% of the appraisal systems surveyed. However, researchers have challenged this traditional approach in which the supervisor is the sole source of performance information for various reasons.

First, the manager may not always be in the best position to evaluate a subordinate. A lack of opportunity to observe the employee at work or a lack of familiarity with the job can result in inaccurate performance ratings. As researchers have pointed out (Landy & Guion, 1970; Landy & Farr, 1980), the relevancy of contact between rater and ratee is more important than the frequency of contact, particularly if regular contact is not directly related to the critical dimensions of performance.

Second, it may not be uncommon for managers to make administrative decisions first and then manipulate performance ratings
so they correspond with those decisions (see, for example, Teel, 1980; Warmke & Billings, 1979). In such cases, the appraisal is being used to justify previous administrative decisions.

Finally, a manager may be interested in different aspects of the employee's performance than other sources would be (Guion, 1965). A manager may be interested in technical knowledge and overlook other factors such as supervisory or interpersonal skills. As a result, the manager's perspective is more limited than it perhaps should be.

One obvious solution to the problems above is to find alternative appraisers who are able and willing to rate an employee's performance more accurately than the immediate supervisor. The search for alternative raters has resulted in a fair amount of research on self- and peer ratings.

There is evidence indicating that when groups of raters in different organizational positions relative to the employee rate performance the resulting correlations are fairly low (Baird, 1977; Holzbach, 1978; Klimoski & London, 1974; Landy & Farr, 1980; Lawler, 1967; Prien & Liske, 1962; Schneier & Beatty, 1978; Thornton, 1968). However, studies which have compared different rating sources have found few explanations that consistently account for these correlations. For instance, studies comparing supervisory and peer ratings have generally found that peers produce more lenient ratings than supervisors.
(Miner, 1968; Rothaus, Morton & Hanson, 1965; Schneier, 1977). However, studies of these two rater groups have shown that neither group consistently has better interrater reliability (Gordon & Medland, 1965; Klieger & Mosel, 1953).

In studies comparing supervisory and self-ratings, self-ratings are usually found to be more lenient (Kirchner, 1965; Parker, Taylor, Barrett & Martens, 1959), but this leniency effect has been shown to be slight (Kraiger, 1986) and, in some cases, nonexistent (Heneman, 1974). Thornton's (1968) review of the literature on self-appraisal concluded that self-ratings generally tend to show more leniency, less discriminant validity, less reliability, less agreement with other sources and less halo than ratings derived from other sources. Landy & Farr (1980), however, did not find such a definitive trend. Finally, in studies comparing supervisory ratings, peer ratings and self-ratings simultaneously, there is evidence that supervisory and peer ratings have more convergent and discriminant validity than self-ratings (Lawler, 1967).

Even though these different sources may be equally valid (Dunnette & Borman, 1979; Landy & Farr, 1980; Vance, MacCallum, Coover & Hedge, 1987) it is useful to understand, for both theoretical and practical reasons, why rating groups tend to produce different ratings.

Various reasons have been proposed to account for the ratings generated by the different rating groups. First, there is
evidence which demonstrates that different raters attach importance to different performance dimensions or criteria (Miner, 1968; Schneier, 1977; Borman & Vallon, 1974; Landy, Farr, Saal & Freytag, 1976). Also, differences between rating sources have commonly been explained by differences in organizational level (Borman, 1974). Such differences between the rating types can be considered forms of "impression bias."

Second, different rater groups are motivated by different sources of bias. These differences can be attributed to "reporting bias." For instance, the supervisor may rate primarily because personnel decisions must be justified by performance ratings. A manager may rate harshly because of the need to feel more competent (Levinson, 1972). Peers may rate more leniently due to mutual respect and attraction. Employees themselves are motivated to get high ratings since evaluations will have personal consequences (Schneier, 1977).

Finally, rating sources may be influenced by their ability to rate performance accurately. Some rater groups may have more relevant and frequent opportunities to observe the employee's job performance. As mentioned, such observational opportunities are assumed to result in more accurate ratings.

The focus of this study is to provide further insight into the difference between self- and supervisor ratings. While there is no reason to doubt the previous suggestions with respect to rating group differences, there appears to be more
we can discover in this area. Some of the previous accounts for rating differences were somewhat indirect. That is, they were made independently of the data obtained during the rating process. This is because the data for many of these studies consisted of only the ratings themselves. There was no attempt to collect information on what prompted the ratings. As a result, attempts to explain differences in ratings were problematic. It is possible that by collecting these additional data we could discover differences in how information is processed in the context of self- and supervisory ratings. This study is designed to explore this possibility by recognizing a cognitive processing view of performance appraisal. The role cognitive processes play in performance appraisal is reviewed next.

COGNITIVE PROCESSING VIEW OF PERFORMANCE APPRAISAL

The most recent emphasis in the area of performance appraisal has been on the cognitive processing required of the rater. The rater is most often viewed as the immediate supervisor in this research so this presentation will take this perspective, but the general framework would presumably apply to other rating sources as well. The results of this study will be able to address this assumption. This cognitive perspective focuses on both the conscious and subconscious mental processes that affect a manager’s evaluation of a subordinate. These processes have been best viewed as taking place continuously over the
period of time in which a manager is responsible for the subordinate. Thus, as Ilgen & Feldman (1983) point out, performance appraisal should be viewed as a memory-based decision process rather than as a stimulus-based one. At the time the formal performance appraisal takes place, the appraiser must retrieve information from memory in order to arrive at an evaluation.

The entire sequence of cognitive processes that have been hypothesized to affect appraisals are complex and cyclical. Ilgen & Feldman (1983) and DeNisi et al. (1984) can be credited with developing models that outline the mental processes that can be expected to occur on the part of the appraiser. Although the models are elaborate, they essentially consist of three major components or sets of operations.

The first component involves the type of information acquired or attended to. The major theme regarding this operation is that potential evaluators are active and selective in attending to the performance of their subordinates. Before performance information about a ratee can be processed, it must be attended to and recognized as information. The manner in which a subordinate's behavior would be attended to is a function of the characteristics of the information seeker, the subordinate, and the organizational environment.

The second major cognitive operation required of appraisers involves the categorization and storage of information in memory. Evidence suggests that raters do not simply store raw
information, but rather store an interpretation of what is attended to. The interpreted representation is seen as being stored in a category with a similar prototype (or schema). This prototype or schema has additional implications for the appraisal process. A schema may serve to interpret ambiguous information. In addition, once a subordinate has been assigned to a schema, it may have implications for the information a supervisor recalls about a subordinate. The supervisor may only recall an overall representation of behavior in terms of the schema rather than the actual behavior.

The final component discussed in the cognitive processing model involves the retrieval and integration of information necessary to form a judgment for a formal appraisal. The issues involved here concern search procedures, weighting strategies, causal attributions, the use of comparative standards and explicit organizational demands. Causal attributions and attributional information, which occur during this third and final component of the cognitive processing model, will be discussed in further detail later as it is these processes which will be the focus of attention in this investigation.

One of the primary purposes of analyzing the cognitive processes underlying appraisals has been to improve the accuracy of performance evaluations. However, in order to achieve this objective, the researcher's goals are served in that it is essential that we provide a better understanding or theory of
the appraisal process. Specifically, the cognitive processing approach is designed to discover how assessments of performance are formed. A number of studies employing the cognitive processing approach have appeared in the literature (Banks, 1979; Cooper, 1981; Lord, 1985, 1986; Murphy, Balzer, Lockhart, & Eisenman, 1985; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982).

Unfortunately, few if any studies have attempted to examine cognitive processes in field settings. Thus, while the above cognitive processing studies may improve our understanding of decision making, they have done little to advance performance appraisal as it is practiced. Banks & Murphy (1985) assert that cognitive processing research will enhance the practice of performance appraisal only to the extent that it is studied in the field. Banks & Murphy claim that there are simply too many controls in the laboratory that would have a substantial effect on the cognitive processing of a rater. Furthermore, it can be argued that with limited knowledge in this area, we don't always know what we should be controlling for.

Ilgen & Favero (1985) provide a similar criticism of the recent "process" approach to performance appraisal. These authors also question the ability to generalize to applied settings from cognitive processing research conducted in the laboratory. Ilgen & Favero discuss three critical characteristics of performance appraisal that are not adequately represented in
the laboratory. First, researchers ignore the effect of past and future interaction between rater and ratee on appraisals. Second, laboratory settings rarely recognize the interdependence between the rater and ratee. Performance evaluations can have implications for the rater as well as the ratee. Finally, laboratory research must recognize that not only does a ratee's behavior have an effect upon ratings, but the outcomes of that behavior also have an effect upon a supervisor's evaluation.

It seems safe to conclude that Ilgen & Favero (1985) and Banks & Murphy (1985) agree that cognitive processing research as applied to performance appraisal would be best conducted in the field or in a laboratory in which boundary conditions have been significantly modified. The present investigation will follow this advice by attempting to better understand the cognitive processing that occurs for self- and supervisor ratings while conducting a performance appraisal in an applied setting.

STUDYING COGNITIVE PROCESSES

There are two main approaches that could be used to study the cognitive processes that lead up to a final appraisal. One approach is to conduct what Svenson (1979) calls a structural analysis of the final judgments or decisions. This approach, also referred to as policy capturing, typically focuses on the end result of a decision process and tries to relate the final decision to parameters characterizing the decision problem. A

Policy capturing, as applied to performance appraisal, typically involves providing raters with a large set of ratee performance profiles consisting of scores on major dimensions of performance. Raters are asked to review the profiles and assign each ratee an overall rating based on the information provided. Multiple regression analysis is often then used to determine the extent to which the overall ratings can be predicted by the information on the performance dimensions and to compute the relative importance of each dimension in determining overall ratings. The resulting regression equation provides the "captured policy" for each individual rater.

The general linear model has been effective in describing rater policies, as is demonstrated by consistently high multiple correlations (with R-square ranging from .40 to .90) in performance appraisal research using policy capturing (see Hobson & Gibson, 1983, for a review and precise estimates of the variance accounted for). The studies have been informative in indicating the performance dimensions that are more or less
important in determining overall ratings. However, there is agreement that the policy capturing approach does have limitations.

The primary limitation concerns the ability of regression models to improve our understanding of the cognitive processes underlying decision making. Svenson (1979) notes that decision making cannot be understood simply by studying final decisions. Hoffman (1960) originally suggested that regression analysis could be used as a representation of human judgment. However, as Dawes & Corrigan (1974) note, Hoffman did not imply that the actual processes of individuals involves weighting various variables and adding them up. Thus, regression analysis was not meant to capture the processes by which decisions are made. Pitz & Sachs (1984) state that a simple test of a linear model is rarely sufficient when one is interested in the processes being used. Hoffman (1960) stressed that the presence of intercorrelated pieces of information about some decision object makes it impossible to know how accurately the regression model captures the underlying cognitive processes. Anderson (1969) suggests that high correlations can occur despite an "incorrect model" since two or more decision models may be algebraically equivalent while representing radically different underlying decision processes. In general, there is agreement that policy capturing does not precisely describe the manner in which raters process or combine performance information in making overall ratings.
Another related limitation of policy capturing approaches is reflected in Naylor & Carroll's (1969) suggestion that nonlinear strategies are available to decision makers and that they will use them when the situation is appropriate. While nonlinear or configural processes could be considered, attempting to evaluate all possible combinations could easily become overwhelming, if not impossible, in many situations. Dawes & Corrigan (1974) claim that linear regression has provided a good fit in various decision making contexts because researchers have selected conditions which would insure a good fit by the linear model.

An additional limitation to the use of policy capturing procedures is that they operate under the assumption of complete information search (Svenson, 1979). Raters are provided with all of the information that is to be used in arriving at a decision (Billings & Marcus, 1983; Einhorn & Hogarth, 1981). Thus, policy capturing does not provide the rater with the freedom of accessing information independently. As Billings & Scherer (in press) clearly illustrate, the type of response required of the decision maker will have significant implications for the amount of information actually used. Their study indicates that requiring subjects to make judgments as opposed to choice decisions results in more information search. That is, all relevant information is not used under all conditions. The use of policy capturing to study decision making ignores
the amount of information searched and the sequence of the search that precedes the decision (Pitz & Sachs, 1984).

In response to the limitations of the policy capturing approach, Payne (1976) suggested the need for an alternative approach to provide more insights into decision making strategies. He adopted process tracing procedures as have been used in the study of human problem solving (Newell & Simon, 1972) to further explore human decision behavior. Unlike policy capturing which focuses on the outcome of the decision process, process tracing approaches observe the predecisional behavior in analyzing a situation by tracing the steps leading up to a decision (Payne, Braunstein, & Carroll, 1978). Process tracing techniques are designed to provide detailed data on the ongoing psychological processes that lead up to a particular choice or judgment. Process tracing directly assesses what information was accessed to form a judgment and the order in which the information was accessed. Such information would be an appropriate response to Wexley & Klimoski’s (1984) call for mapping the strategies of real-time cognitive processing in performance appraisal.

This study will attempt to document the cognitive processing of appraisers using verbal protocol analysis. Verbal protocol analysis is one type within a set of process tracing techniques. It provides process information by allowing subjects to report continuously on the intermediate inputs and outputs
of their cognitive processing. This technique has yet to be applied to performance appraisal processes, but would appear to be quite useful. A discussion of verbal protocol analysis is presented in the next section.

As mentioned, there are different techniques which can be categorized as process tracing methods. Aside from verbal protocol analysis, the recording of eye movements and the use of information display boards have been two of the more popular methods of tracing psychological processes. Schechtman & Ford (1987) have used the information display board in a performance appraisal context to examine the influence of appraisal purpose and prior knowledge on information acquisition. Banks (1979, 1982a, 1982b) has developed another approach for measuring processes and has applied it to performance appraisal simulations in the laboratory. In these studies, subjects watch the target behavior and operate a button-pressing apparatus and explain what prompted any button-pressing behavior. Cronshaw & Lord (1987) have also successfully used a button-pressing methodology to study ongoing cognitive responses to leadership behavior.

Each of the above techniques has distinct advantages and could be applied to the study of performance appraisal. However, verbal protocol analysis was viewed as superior for the study of performance appraisal processes in this investigation. In order to record eye fixations, button-presses, or information selected from a display board, all relevant information
must be in the physical presence of the subject at the time the decision is being made. This, of course, would be impossible in a field setting. Even if it were possible, it would threaten the external validity of the study since as Ilgen & Feldman (1983) point out, performance appraisal is best viewed as a memory-based task. However, this is not to imply that verbal reports are not without their limitations. Self-reports are a reactive form of data collection and might be expected to have an impact on the responses being obtained (Folger & Belew, 1985). A complete discussion of verbal protocol analysis is presented next.

VERBAL PROTOCOL ANALYSIS

Verbal protocol analysis has become increasingly popular in the study of decision making since the seminal work of Newell, Shaw, & Simon (1958) and Newell & Simon (1972). Verbal protocol analysis involves asking subjects to "think aloud" while they are in the process of solving a problem or making a decision. The outcome of the verbal behavior is then treated as a record of the sequence of thoughts involved in the decision as it occurred. Verbal protocols have been used in a variety of problem solving or decision processes and include applications to making parole decisions (Carroll & Payne, 1977), consumer choices (Bettman, 1974), the selection of living quarters (Payne, 1976), investment trust decisions (Clarkson, 1963),
decisions in aircraft accident investigations (Braunstein & Coleman, 1967), and simulations of complex managerial decision making (Schweiger, Anderson, & Locke, 1985). While the collection of verbal reports is conceptually straightforward, it is an extremely difficult procedure to conduct in practice.

Before discussing verbal protocol analysis in detail, it is important to recognize the distinction between verbal protocols and what was called introspection in psychology's early years. The method of collecting data through verbal reports was common for early experimental psychologists (Svenson, 1979). Wundt and Titchener employed introspection as the primary method of investigating psychological processes. Introspection was virtually abandoned as the behaviorists criticized the method for its lack of objectivity (Marx & Hillix, 1963). It seems rather ironic, but Payne et al. (1978) indicate that Watson was able to accept verbal reports (behavior) as data when they were verifiable and repeatable.

It does, however, appear that the apparent similarity between verbal protocols and introspection discouraged modern psychologists from using verbal protocol analysis (Hayes, 1968). Thus, it is important to clarify the distinctions between verbal protocols and introspection (see Payne et al., 1978). First, in protocol analysis subjects are naive about the processes under investigation. Introspection, however, uses subjects that are extremely knowledgeable about the theo-
ries and hypotheses under investigation. In addition, those providing verbal protocol data are instructed to report data which they are currently attending to. This is not the case with introspection.

Another characteristic of verbal protocol analysis that sets it apart from introspection is that the verbal reports occur simultaneously as the subject performs on the task. It is common in psychology to collect verbal reports after the performance of interest has already been completed. This distinction is critical and will be discussed later in the context of how verbal protocol data should be collected. The next section will review research on the use of verbal reports to accurately represent cognitive processes.

CONCURRENT REPORTS AS ACCURATELY REPRESENTING COGNITIVE PROCESSES

The theory or model upon which the technique of verbal protocol is based has significant implications for the types of information (inputs and outputs to cognitive processes) that can be accurately reported and for the collection of verbal protocol data. The theory (see Ericsson & Simon, 1980; Ericsson & Simon, 1984) underlying verbal protocol analysis proposes that information recently acquired by the central processor is kept in short-term memory and is directly accessible for verbal reporting. In contrast, information from long-term memory must
be located, retrieved, and transferred to short-term memory before it can be reported. As a result, information recalled from long-term memory is not a direct reflection of one's thought processes. Thus, verbal protocol analysis is designed to provide a running account of the contents of short-term memory. This implies that verbal reports must be collected from subjects concurrently as they are performing the task of interest.

As a method of collecting data on psychological processes, verbal protocol analysis has received considerable support (Ericsson & Simon, 1980; Payne et al., 1978; Ericsson & Simon, 1984; Schweiger, 1983). Its primary feature is its ability to provide a dense number of data points within a given period of time. This renders it particularly useful for assessing process theories.

However, while protocol analysis has been proposed as a useful technique, a number of issues must be addressed concerning how accurately verbal reports reflect the actual information used in cognitive processing. The issues fall under two major headings and concern the relevance of verbal reports to cognitive processes and the completeness of the verbal reports. The issue of relevance will be discussed in the remainder of this section and the completeness of verbal reports will be presented in the following section.
Relevance of Verbal Reports

The relevance of cognitive information to verbal reports refers to whether the verbalizations are closely related to one's underlying thoughts. This is in essence a question of validity. The criticism often aimed at verbal reports is that they are or can be epiphenomenal. That is, that they are generated independently of the cognitive processes that produce performance. The model or theory upon which protocol analysis is based would indicate that verbalizations reported immediately from short-term memory are not epiphenomenal, but are relevant to the cognitive processes underlying performance.

Ericsson & Simon (1984) consider three criteria for determining whether verbalizations are pertinent to the cognitive processes required for performance. First, the verbalizations should be relevant to the given task. Newell & Simon (1972) indicate that for many kinds of tasks (performance appraisal would appear to be included), an a priori task analysis would indicate what types of information are relevant to task performance. This criterion would not be useful for tasks such as day-dreaming or free association.

The second criterion claims that verbalizations, to be considered pertinent, should be logically consistent with the verbalizations that immediately precede and follow them. If a sequence of verbalizations are not related to each other and consistent, then it is possible that they could have been pro-
duced by random processes. This would, of course, argue against the pertinence of verbal reports.

The final criterion involves long-term memory and argues that a subset of the information attended to (and therefore held in short-term memory) during task performance will be remembered. When subjects are thinking aloud, much of this information will be remembered and available for subsequent retrieval.

Those supporting verbal protocols argue that if verbalizations meet the three criteria above, they could only be produced by cognitive processes similar to those used for performing the task, and it would not be parsimonious to assume two different cognitive processing mechanisms when one would suffice. An even stronger argument against epiphenomenality would hold that it would be highly implausible for a subject to solve a problem (e.g. a performance appraisal) as well while thinking aloud about a different problem (e.g. a training course on performance appraisal) as while thinking aloud on the same problem.

A review of research demonstrating the relevance of verbal reports will now be presented. The type of task will determine what criteria will be used to demonstrate the relevance of verbal reports. For instance, if task information is perceptually available to the subject, establishing the correspondence between verbal reports and the stimuli attended to is a rela-
tively easier undertaking. While the three criteria discussed above could be used, a more objective approach is available. Verbal reports can be compared to the information attended to based on recording eye fixations. A number of researchers (Winikoff, 1967; Deffner, 1983; Geiselman and Belleza, 1977; Russo and Rosen, 1975; cited in Ericsson & Simon, 1984) have used this approach. The results provide strong evidence for the correspondence between the information that is physically attended to and the information that is reported verbally. This provides direct support for the relevance of verbal reports.

Evidence supporting verbal protocols has also been found for tasks that require the generation and retrieval of information. In these tasks, stimuli are not perceptually available so the criteria of relevance and consistency must be relied on. Even tasks within this category vary in the extent to which information must be recalled from memory.

In anagram type tasks, memory retrieval is essential in locating the desired word. Sargent (1940) and Mayzner, Tresselt, & Helbock (1964) provide clear evidence that word hypotheses and cues verbalized during anagram tasks are relevant to stimuli presented and the goal of the task.

Another type of task that would require less retrieval of information from memory is mathematical problem solving. A great deal of research has indicated that subjects' verbal
reports, while solving such problems, correspond closely to the sequence of steps required to solve the problem (Sokolov, 1972; Krutetskii, 1976; Newell & Simon, 1972; Larkin, McDermott, Simon, & Simon, 1980; Paige & Simon, 1966).

The final type of task to be discussed involves concept attainment. Such problems do not require retrieval of information from long-term memory. All of the relevant information is presented to the subject and he/she is to determine some underlying concept. A number of researchers have found substantial evidence indicating the pertinence of verbal reports to information presented to subjects and to their behavior during the concept attainment task (Coltheart, 1971; Frankel, Levine, & Karpf, 1970; Karpf, 1973; Karpf & Levine, 1971; Schwartz, 1966; Cahill & Hovland, 1960; Bourne, Goldstein, & Link, 1964).

Ericsson & Simon (1984) discuss studies which rely on the long-term memory criterion to demonstrate the pertinence of verbal reports. These studies involve the use of mnemonic codings in an attempt to memorize a list of items. If the same mnemonic that is reported verbally at the time of presentation is also reported at the time of recall, it is unlikely that the verbalizations are epiphenomenal. Ericsson & Simon provide evidence of substantial agreement among successive reports of mnemonic encodings and are able to demonstrate that the mnemonic encoding is in fact recalled rather than generated by the actual item to be recalled.
Ericsson & Simon (1984) discuss findings in three other research areas that indicate the pertinence of verbal reports to ongoing psychological processes. First, subjects' reports about the methods they use in maze learning tasks have been shown to be highly correlated with performance and types of errors on those tasks. Second, there is research that indicates that response latencies expected in various tasks correspond with verbal reports. For example, more cognitive information would be reported on during tasks with longer expected response latencies. Third, computer models have been developed to capture the sequence of required steps to properly complete a variety of different types of tasks. The sequence of steps for a given task is developed through a careful task analysis. The task analysis is conducted independently of individuals' protocols. However, comparisons between the computer models and verbal protocols have been impressive. These analyses provide substantial evidence that verbalized protocols correspond closely with the information that has to be considered for solving problems.

Thus, the research presented throughout the above section clearly indicates that individuals are able to produce verbal reports that are pertinent to ongoing cognitive processes. This does not, however, imply that there are no limits on what can be reported with accuracy. The remainder of this section and the following section will be used to delineate these conditions.
Potential Limitations of Verbal Reports

The first issue to be discussed in this section has clear implications for the accuracy or faith a researcher may have in interpreting verbal reports. Simon (1979) indicates that in order to interpret verbal protocols it must be assumed that subject's internal encoding of information allows it to be expressed easily in words. Thus, it would be ideal if all of the stimuli of interest were verbal in nature. This is clearly not the case when evaluating performance. While some of the ratee's behavior may be verbal, much of it is not. However, it certainly doesn't appear that verbal descriptions of nonverbal work behavior would be so difficult as to restrict the use of protocol analysis. It is not uncommon for employees and supervisors to verbally describe the work behavior of others.

However, to produce a verbalization from a thought, lexical items must be selected for each relation and entity. In addition, a syntactical form or order of verbalization must be selected. In most cases, the selection of lexical items should not be problematic (Ericsson & Simon, 1984). Most items have a unique, readily accessible name. However, where an object lacks a readily accessible name, additional cognitive processing would be required. This additional cognitive processing might interfere with attempts to report on other items of information being held temporarily in short-term memory. This would create a problem in terms of the completeness of verbal
reports. In cases where an appropriate label cannot be generated, the subject may settle on using a poor substitute or task irrelevant characteristics. In such cases, the relevance of verbal reports to ongoing cognitive processes is clearly threatened.

The selection of syntactical form can be characterized in a similar manner. Thoughts can be expressed syntactically in many ways. Where there is no accepted form or pattern, a great deal of variability might be expected across individuals. However, in many situations sentence development will follow easily and directly from one's thoughts. A model of verbalizing in the context of performance appraisal will be presented later and it is important to recognize that lexical and syntactical encoding are necessary for such verbalizations.

RECOGNITION OF MOTIVATIONAL INFLUENCES

There is one final issue to be discussed that has direct implications for the relevance or validity of verbal reports. This issue involves motivational factors to deliberately distort verbal reports of one's cognitive processes. Motivational factors have been ignored in research and reviews on protocol analysis presumably because such factors are largely irrelevant to problem solving tasks conducted in the laboratory. There would be little, if any, justification for why subjects would or would not prefer to report certain information in such set-
tings. Thus, cognitive factors serve as a sufficient explanation for verbal reports in this context.

However, the use of protocol analysis in the context of performance appraisal in an applied setting clearly requires the recognition of motivational influences. Managers will recognize that they are being asked to provide self-relevant information that will have implications for how they are perceived by others. In this context, managers could be expected to influence, consciously or unconsciously, how they are viewed by other people. In addition to being concerned with the experimenter's immediate impressions, the manner in which a supervisor completes a performance appraisal is of real concern to governmental agencies, the organization in which he/she functions, his/her immediate superiors, as well as his/her subordinates. The ramifications of performing inappropriately from the perspective of one or more of these parties can not be taken lightly. It is possible, then, that verbal reports would be generated in a manner consistent with the desires of interested parties at the expense of reporting on what one is actually attending to. Thus, it seems clear that motivational influences must be recognized in the context of this study.

In response to this recognition of motivational influences, this study must build in features to minimize these biasing factors. Suggestions for minimizing self-enhancement tendencies can be found in literature from industrial, social, and
Within industrial psychology, research in this area has focused on self-assessments of ability and performance and has placed an emphasis on subject's ability to provide accurate self-evaluations. Social psychological research has emphasized impression management processes in the context of social interactions. Finally, research on self-reporting in the area of counseling psychology can be found in literature on self-disclosure (Jourard, 1964). Self-disclosure refers to verbally communicating information about oneself to another. Although the information can be either positive or negative, most theoretical and research interest has focused on information that might be perceived as negative.

Many of the suggestions used to encourage honesty in self-reporting can be found in more than one of the research areas. For instance, one popular approach for inducing individuals to report honestly is to insure anonymity, or at least confidentiality, for the respondents. This suggestion has been supported in industrial (Mabe & West, 1982) and social (Baumeister, 1982; Schlenker, 1980) psychological research. This feature must be present in this study.

Another obvious approach for reducing self-enhancement is to use instructions that explicitly request that subjects report honestly on their cognitive processes. This rather straightforward approach has been successful in research in social psychology (Gergen, 1965; Healy, 1971; Schneider & Eustis, 1972;
Schneider, Eustis, Manzolati, Miller, & Gordon, 1971). Thus, such instructions would appear to be useful and will be included in this research.

Research from industrial psychology suggests that subjects' experience in providing self-reports will enhance the accuracy of the reports (Mabe & West, 1982). This provides support for the need to allow subjects to self-report on a series of practice exercises. Instructions to subjects during these practice exercises can also incorporate specific comments that might further serve to enhance the accuracy of the verbal reports. Research in industrial and social psychology (Mabe & West, 1982; Schlenker, 1975; Schlenker, 1980; Wortman, Costanzo, & Witt, 1973) indicates that self-enhancement will be reduced to the extent that self-reports can be verified. While one's cognitive processes during performance appraisal cannot be directly verified, indirect methods of determining whether information is being reported from short-term memory are available. These methods include evaluating the number of words reported per minute, the types of verbs used, and the content of the protocol (these issues will be discussed later in further detail). The instructions to the subjects will indicate that the experimenter can determine whether information is being reported appropriately. Furthermore, this point will be made apparent throughout the practice exercises as the experimenter will most likely have to intervene and modify the manner in which the subject is reporting his/her thoughts.
Two other techniques presented in the self-disclosure literature will be used by the experimenter to elicit honest verbal reports. Social approval has been found to enhance self-disclosure (Taylor, Altman, & Sorrentino, 1969). Thus, during the practice exercises, comments expressing the experimenter's approval will follow information that has been properly reported from short-term memory. A second technique which is characteristic of verbal protocol analysis and counseling psychology is probing (Vondracek, 1969). When there is a prolonged pause during a subject's verbal report or when a subject is reporting inappropriately, the experimenter will probe the subject to facilitate proper reporting.

Research on the concept of impression management has found a relationship between individuals' self-esteem and their need to present themselves in a favorable manner. In particular, people with low self-esteem have been found to be more interested in providing favorable self-presentations (Jones, 1973; Mettee & Aronson, 1974). With regard to this research, this suggests attempting to enhance the manager's self-esteem particularly with respect to his/her performance appraisal practices. A set of statements by the experimenter suggesting that "when it comes to evaluating your own performance or the performance of your subordinates you are the expert" would presumably be useful in supporting the manager's self-esteem.
One final suggestion found in the self-disclosure literature may be useful for reducing self-enhancement biases. A number of researchers (see Cozby, 1973, for a review) have indicated that disclosure by the interviewer is apt to elicit greater disclosure by the subject. In addition to inducing this reciprocal self-disclosure, Cozby notes that the interviewer who discloses is rated as more trustworthy and more positively in general than the researcher who does not disclose information.

While techniques identified to facilitate self-disclosure and accurate self-presentation in social contexts may be employed in this study, a more influential factor appears to be present in organizational settings than is the case for the typical counseling or social interaction scenario. Counseling involves assisting individuals overcome personal inhibitions in the face of potentially severe interpersonal anxiety. Impression management involves people’s desires to project an image which is consistent with the identity they wish to maintain. While interpersonal anxiety and impression management are present in industrial settings, the respondent in this study would clearly have additional concerns. Inappropriate behavior in the work place often has implications for one’s financial status and job security. Thus, the rater might perceive serious external consequences for describing his/her performance appraisal procedure as one that could be viewed as inappropriate. Whether such consequences would be real or only imagined...
in this context would not be relevant; the perceived consequences and their impact upon the verbal report would be identical.

Thus, the organization might be expected to have a more significant influence on one's honesty in providing verbal protocols. As Schlenker (1980) points out, in order to induce subjects to report honestly, they should not view any personal gain in self-enhancement. Therefore, managers should not feel they will be punished for reporting honestly and should believe that honest self-reports will be of some benefit to them. It seems clear that the sponsoring organization must be sincere (and must convey its sincerity) in desiring candid verbal reports only for the benefit of further developing its performance appraisal system.

While the above factors will be useful in gathering accurate verbal reports, there are certainly patterns of behavior in the context of performance appraisal that individuals would not attempt to distort for motivational reasons. Prescriptions do not exist for every component in the process of conducting performance appraisals. There are many permutations of gathering and evaluating information that would be viewed as acceptable. Motivational factors should be minimal for these aspects of the process. Also, while many managers receive training on performance appraisal procedures, it is unlikely that an entire group of managers would recall and have a similar working
knowledge of how performance appraisal ought to be conducted. Particularly since prescriptions on how one ought to complete performance evaluations may differ depending upon whose perspective the rater relies upon for the prescriptions. In sum, while motivational factors will be present, it does not appear that such influences would preclude making significant discoveries in the area of performance appraisal.

**Completeness of Verbal Reports**

This section is related closely to the previous sections on the accuracy of verbal reports. In those sections, an attempt was made to demonstrate that under conditions where verbal reports are claimed to be accurate, they do in fact correspond to the underlying cognitive processes. This section is designed to discuss the limits on the cognitive information that can be reported with accuracy. That is, there is certain information that individuals can not accurately represent through verbal reports. In addition, under conditions where individuals can provide verbal reports, they are not capable of reporting everything that is attended to.

The types of information that individuals are unable to represent by means of verbal report will be discussed first. This discussion follows directly from the theory underlying verbal protocol analysis. As stated, the theory holds that only information that is attended to and held in short-term memory
can be accurately reported. Thus, information that has not been attended to cannot be reported verbally.

The first limit on reportable information involves recognition processes. Ericsson & Simon (1984) argue that recognition processes occur while leaving very little information in short-term memory. Generally, individuals tend to recognize familiar faces and objects directly without storing in short-term memory the characteristics extracted from the stimuli and used in recognizing them. Such processes can occur with more complex patterns and relations also. Generally, the extent to which automatic, recognition processes will invalidate verbal reports depends upon the demands of the particular task (Kellogg, 1982). For tasks, or portions of tasks, that rely on recognition processes, there will be no information available for the subject to report. There is no reason to believe that such recognition processes would not be relevant to certain portions of performance appraisal. Thus, it would be inappropriate to ask for verbal reports on aspects of performance appraisal that rely on recognition processes since only the final product would be left as a trace in short-term memory.

To further explore the implications of recognition processes for performance appraisal, we can examine how such recognition processes might develop. There appears to be a relationship between experience with a task and an increase in automatic, recognition processes (Dean & Martin, 1966; Ryan, 1970; Shiff-
rin & Schneider, 1977). In the context of performance appraisal, a rater's experience in conducting performance evaluations overall and with the individual who is the ratee at the time of the verbal report might have implications for the verbal protocols collected.

Another limitation on what individuals are capable of reporting involves general underlying theories or rules that they may have used during task performance. This information is simply not held in short-term memory. When asked to report on such general theories of performance, even though subjects may be unaware of any underlying theories they will often generate a rational response that is unrelated to the rules that actually guided their behavior (Ericsson & Simon, 1980; Ericsson & Simon, 1984; Philip Smith, personal communication). Thus, subjects should not be asked about the general procedures they use when performing a task. This inferential procedure should be left to the researcher based on data reported continuously and directly from short-term memory.

One final type of information that individuals cannot report accurately involves retrospective reports of earlier cognitive processes. This assertion follows directly from our knowledge of the capacity of short-term memory. Information remains in short-term memory only momentarily until it is replaced by subsequent information or decays due to the passing of time.
The general theory or model of cognitive processing assumes that a subset of the information attended to during task performance is left in long-term memory. If subjects are asked to provide a retrospective report on an activity just completed, it appears that much of the information held in long-term memory can be retrieved based on remaining cues in short-term memory (Ericsson & Simon, 1984). However, after a brief period of time or following other intervening activities, one's ability to recall information attended to while previously performing a task becomes highly suspect (Smirnov, 1973; Smedslund, 1969). All available data on subjects' memory of their cognitive processes during previous activities shows their recall to be poor (Ericsson & Simon, 1984). As a result, while retrospective reports may be useful in providing supplemental information, it is clear that concurrent verbal reports are essential to accurately map ongoing cognitive processes.

The other major issue to discuss with respect to the completeness of protocols is subjects' inability to report everything that they attend to or that passes through short-term memory. The main problem is that thought proceeds much faster than speech. The time required for attending to information during task performance may be only a fraction of the time required for verbalization. Newell & Simon (1972) and Payne et al. (1978) have recorded verbal protocols which contained approximately two words per second. While the number of words
generated will vary depending on the task, two words per second is fairly low when one considers the types of thought processes that can occur in a matter of seconds.

Ericsson & Simon (1984) also indicate that some individuals are better able to vocalize their thoughts than others. This lends additional support for the need to provide a brief training session for subjects in order to minimize such individual differences. One's level of experience with the given task would also influence the completeness of the verbal reports. It is expected that more complete reports would be provided by those less familiar with the task because information will come into the subject's attention at a slower rate (Ericsson & Simon, 1984). Thus, another reason for documenting the subject's experience with the ratee and with conducting performance evaluations.

In sum, although verbal protocols allow for collecting a high temporal density of cognitive processing data, it is clear that verbal reports do not provide a complete record of a subject's state of knowledge at every moment (Newell & Simon, 1972). Protocols, like other methods used in the behavioral sciences, will only provide partial information about the cognitive processes of an individual.
EFFECTS OF VERBALIZATION ON TASK PERFORMANCE

One final issue related to the validity of verbal protocols is whether or not the reporting of verbal protocols interferes with problem solving or decision making on the task being performed. This issue is a critical one. If the verbal protocol affects task performance, then it will lead to problems of interpretation since the effects of the verbal report will be confounded with the effects of the cognitive processes as measured by the protocol.

Several researchers (Reitman, 1970; Russo, 1978; Svenson, 1979) have suggested that collecting verbal reports has the potential for being obtrusive to the process of performing the task. The argument is that the additional verbal responses required by the subjects demands mental effort which detracts from the effort that could be devoted to performing the task of interest. A criticism of a different type has been made by Lord (1986) and suggests that asking subjects to think aloud may cause them to adopt more "rational" ways of doing tasks than they would do otherwise.

In response to such claims, Ericsson & Simon (1984) indicate that the effect of verbalization on task performance depends upon the level or type of verbalization required. Verbal reports can require subjects to describe, explicate, or label information that they are presently attending to. Such verbalizations do not bring new information into the subjects atten-
tion. On the other hand, verbalization can be used to allow subjects to explain their thought processes. This involves more than a recoding of information and requires subjects to associate this information with earlier thoughts and information.

The descriptive type of verbalization is what is recommended in conducting verbal protocol analysis. There is a great deal of research indicating that providing such descriptive verbal reports does not have an adverse effect on task performance. Roth (1965) and Karpf (1973) examined the effect of verbalizing on problem solving tasks and found verbalizations had no effect on time to solution or on the number of problems solved correctly. Carroll & Payne (1977) assessed the effect of verbal reports on the process of making parole decisions. The results indicate that there were no differences between protocol and nonprotocol groups for the information requested, speed of decision or final judgment. In fact, some research has indicated that descriptive verbal reports may improve performance on some tasks (Dansereau & Gregg, 1966; see Ericsson & Simon, 1984, for a review). In a study more relevant to this research, Schweiger (1983) examined the effect of verbal reporting on problem solving and decision making performance for a computer-simulated executive business game. The results led the researchers to conclude that it was improbable that the collection of simultaneous verbal protocols had any effect on performance.
Thus, it appears that under certain conditions verbal protocols can be collected without having an adverse impact on performance. In the form of guidance, Simon (1979) notes two conditions which should occur to insure that verbalizations do not affect task performance. First, the instructions given to the subject should clearly indicate that the subject is to assign first priority to the task. In the context of this study, subjects should focus on conducting a performance evaluation and not on the verbal protocol. While this may not be extremely difficult, it would require careful instructions and a brief training session. Second, the "thinking aloud" instructions must be "bland" in that they don't request specific kinds of information. If the subject is asked to be selective, he/she will be forced to direct attention to this screening process. This is another characteristic that distinguishes introspection from protocol analysis. Protocol analysis requests that a subject report all thoughts that occur to him/her rather than only those thoughts that are perceived as relevant to a particular theory or hypothesis. This would also require a brief period of training in order to provide verbal reports in this manner.

PROCEDURES FOR COLLECTING VERBAL PROTOCOLS

With regard to collecting verbal protocol data, it is apparent that extreme care must be taken. As indicated in previous sections, a major issue with respect to verbal data is whether
they should be collected during the actual performance of the task or afterward. As discussed, it is clear that protocols should be collected during performance of the task. Typically, the subjects' verbalizations are tape-recorded as they perform the task. Descriptions provided after performance on a task are problematic as they confound recall processes, current knowledge, and past knowledge which makes reliable inferences from the protocol difficult (Payne et al., 1978).

Nisbett & Wilson (1977) provide an impressive review arguing against subject's ability to verbally report upon higher mental processes. However, one particular problem with the relevance of this assertion to protocol analysis is that much of the evidence is based on verbal reports collected after the task of interest was already performed (Ericsson & Simon, 1980). Thus, many of the verbal reports used in the Nisbett & Wilson review are confounded in a way that would not be the case for verbal protocol analysis.

The model upon which verbal protocol analysis is based and much of the previous discussion has additional implications for the collection of verbal report data. First, there will be a number of mental processes which subjects will be unable to report because they never exist in short-term memory. Individuals are not aware of all of their mental processes; certain operations occur below the level of consciousness. Thus, there should be no attempt to collect such data. The Nisbett & Wil-
son (1977) review discussed above based some of their conclusions on subject's inability to report on subconscious processes. Verbal protocol analysis is designed only to collect data on processes which the subject is able to attend to.

Second, the theory of protocol analysis has direct implications for the aspects of the performance appraisal process which can be assessed using this technique. Only information which is directly attended to and therefore stored in short-term memory during the performance appraisal process can be collected using verbal protocol analysis. The specific components of the cognitive processing model which can be assessed through the use of protocol analysis will be discussed in the following section.

Third, and finally, much of the previous discussion indicates that subjects' previous experience with the task may have implications for what is verbally reported. That is, as an individual gains experience with a task, he/she may respond to certain stimuli automatically or without awareness. If this occurs, these automatic processes will not be stored in short-term memory and will not be reported in the way prescribed by verbal protocol analysis. Thus, raters experience in conducting performance appraisals must be considered in the context of this study.
ANALYSIS OF VERBAL PROTOCOL DATA

With regard to analysis of protocols, the first step is to make a complete transcript of the verbal reports given by each subject. The second step typically involves breaking the protocols up into short phrases (Newell & Simon, 1972). Since verbal protocols are qualitative in nature, an external coding scheme is often necessary in order to analyze the data (Schweiger et al., 1985). The coding scheme must be relevant to the purpose of the study. Svenson (1979) asserts that a coding scheme cannot be properly developed without a model or theory of the decision processes of interest. Fortunately, as was discussed earlier, there exist models of what processes are expected to occur when evaluating performance. These models will serve as the framework for developing the coding scheme. Once the verbal protocols have been satisfactorily coded, they become useful in a number of ways.

Payne et al. (1978) argue that verbal protocol data are extremely useful in exploratory research. The data are viewed as valuable during early phases of investigating decision processes. Thus, verbal reports would appear to be quite useful in the context of performance appraisal since our knowledge of cognitive processing in this area could be described as fairly primitive.

Verbal protocol data will also provide useful information relative to a model or theory (Simon, 1979). Descriptive sta-
tistics based on the coded protocols will provide information relevant to the theory or model of interest. While such descriptive information could not be used for rigorous model testing, Simon (1979) claims that such data are useful for the testing of theories. His assertion is that theory testing is often indirect in nature. A theoretical model is postulated, observable consequences are deduced from the model, and these consequences are compared with empirical data. To the extent that our model accounts for the observed data, it receives additional support. What makes Simon’s assertion particularly relevant to protocol analysis is his claim that our confidence in a theoretical explanation will be greater as the number of points of contact between the theory and empirical observations increases. Verbal protocol analysis is particularly useful in this manner as it provides the researcher with a temporally dense set of data points to evaluate the theory of interest.

VERBAL PROTOCOL ANALYSIS AND THE PERFORMANCE APPRAISAL MODELS

As was previously discussed, protocol analysis is designed only to collect information which a decision-maker attends to and which, therefore, is stored in short-term memory. Thus, it is essential to focus on aspects of the performance appraisal process which the appraiser is expected to have awareness of. As Nisbett & Wilson (1977) point out, if subjects are asked to report on information they do not have awareness of they will
respond based on a priori, causal theories of what they think should be occurring. But, as was mentioned, raters will not be asked to provide specific information so this problem should be minimized.

The Ilgen & Feldman (1983), DeNisi et al. (1984), and Larson (1984) models of performance appraisal provide theoretical support for this investigation into the cognitive processing of the two rating sources (self and supervisor) in the context of performance appraisal. With respect to identifying precisely which aspects of the cognitive processing models occur within the subject's awareness, the Ilgen & Feldman model is the most useful model since it does specify what information the rater would be aware of. The DeNisi et al. and Larson models, however, do not provide such guidance. It is, therefore, difficult to determine what information in these two models is attended to by the rater. However, the processes addressed in this study are sufficiently robust as to not be affected by this lack of consistency.

In terms of the three general components of the performance appraisal models (see Figure 1), this investigation will focus on the integration and judgment phase of the appraisal process. This component of the performance evaluation process has been selected for several reasons.

First, since this research involves an initial investigation into cognitive processing in an applied setting, it seemed
somewhat unrealistic to expect a single study to provide significant insight into the entire sequence of appraisal processes (e.g. attention, recognition, categorization, retrieval, etc.).

Second, it appeared that an investigation of the integration and judgment phase would provide the most insight into the cognitive processing of the two rating types in performance appraisal. This conclusion was reached because relatively more aspects of this phase are hypothesized to occur within the subject's awareness than is the case for the other components of the model.

Finally, the time during which integration and judgment are likely to take place renders the process more readily accessible. That is, the processes that occur during the attention and categorization phases are likely to be more informal, sporadic and continuous over time than is the case for integration
and judgment. The integration and judgment processes are more easily accessed since they are apt to occur during a more restricted and identifiable period of time.

In terms of the specific information that the subject would be aware of within the integration and judgment phase, the Ilgen & Feldman (1983) model is the most complete and will be presented to provide an overall framework for the processes hypothesized to occur during performance appraisal. However, this investigation will focus on the formation of causal attributions and on the use of attributional information in the context of the broader appraisal process. The DeNisi et al. (1984) and Larson (1984) models will only be mentioned to the extent that they discuss attributional information relevant to this investigation.


The Ilgen & Feldman model will now be discussed. A simplified version of the integration and judgment phase of this model is represented in Figure 2. These authors provide excellent guidance in identifying what information the manager will be aware of as he/she is forming a judgment. The integration and judgment phase is expected to begin when there is a demand from the environment for an evaluation. The manager's awareness of this demand will trigger the following processes (only those aspects that the manager would have awareness of will be identified). This demand for an appraisal will lead to an aware-
ness of the person to be appraised as well as the perceived situation. "Active search routines" will also be enacted which involve three methods of locating additional information about the person and the situation. The "unitizer" directs attention to smaller units of the appraisee's behavior. The "canvasser" provides information regarding the situation, setting, and base rate. Finally, the "conceptualizer" is used to search one's memory for less salient, but relevant category prototypes or causal schemata. It is important to point out that while all of the above types of information will be stored in short-term memory, most of it will be in the form in which it has been recalled or retrieved from memory. Thus, the initial objective information has been distorted as a result of perceptual, categorizational, and retrieval processes.

In addition to the above information, the appraiser should also be aware of causal attributions, affective responses, and predictions of behavior. The rater is also hypothesized to be aware of comparative standards used as well as organizational factors that would affect ratings. Finally, the appraiser is aware of attempts to verify or support his/her appraisal decision. Figure 3 provides the complete model as presented by Ilgen & Feldman (1983).

With respect to the DeNisi et al. (1984) model, many of the processes correspond to those identified by Ilgen & Feldman. In particular, attributional assessments using consensus, dis-
Demand from environment to conduct performance appraisal

→

Awareness of ratee and perceived situation

→

Recall information regarding:
1) Specific aspects of ratee's behavior
2) The situation, setting, and base rate
3) Additional relevant category prototypes

→

Seek information to verify hypotheses

→

Causal attributions
Affective responses
Predictions of behavior

→

Take into consideration:
1) Comparative standards
2) Extrinsic organizational factors

→

Rating of performance

→

Attempts to verify evaluations

Figure 2: Abstracted version of Ilgen & Feldman (1983) model of performance appraisal
Figure 3: Ilgen & Feldman (1983) model of performance appraisal

tinctiveness, and consistency information are represented in analogous forms in the two models. Base rate information, for example, can be considered identical to consensus information.

The DeNisi et al. model also makes two predictions that go beyond those included in the Ilgen & Feldman (1983) model that are relevant to this study. First, when raters make performance evaluations, those evaluations are expected to be preceded by internal attributions. If external attributions have been made, an evaluation regarding the ratee's performance is not expected to be formed. While this may be somewhat oversimpli-
fied, it is logical since an external attribution indicates that a factor outside of the ratee's control was responsible for the observed performance. Thus, no information is provided in terms of the ratee's performance.

Second, DeNisi et al. hypothesize that consensus, or base rate information, will be used in evaluating a ratee only to the extent that the rater does not already have clear (objective) standards against which to compare performance. In fairness, it should be noted that Ilgen & Feldman (1983) also make this assertion, but the point is made more casually and not included in their model. This hypothesis predicts that information will be used differentially depending upon the amount of knowledge the rater has regarding the standards of performance. It is quite possible that different rating sources will have different perceptions regarding their knowledge of the standards of performance for the ratee. This possibility is discussed later in this section.

The final model that will be discussed is provided by Larson (1984) in a discussion of the informal performance feedback process. Although the model does not address the formal performance appraisal process, it does provide additional support for the processes that occur during an evaluation of performance. Prior to providing individuals with feedback, there must be an evaluation of performance. This antecedent aspect of providing performance feedback is relevant to the purpose of
this research. Larson (1984) proposes that individuals do use consensus, distinctiveness and consistency information in order to form causal attributions in the process of evaluating performance.

The discussion of the above models was designed to demonstrate that several writers believe that raters do consider attributional information in evaluating performance and that they are expected to have awareness of this information. We would expect the processes outlined above to take place following a demand from the environment for an appraisal. Precisely what constitutes such a demand from the environment remains an empirical question. In the above discussion, Ilgen & Feldman imply that the demand is a formal request for an appraisal while the demand to begin the evaluation process is more informal in the case of Larson's model.

The primary demand for an evaluation in this investigation will consist of a specific, formal request for an open-ended evaluation without allowing the rater to access any supporting documentation or performance appraisal forms. There are a number of reasons for proceeding in this manner. First, this type of demand for an appraisal would not be significantly more familiar to one rating source than the other. For instance, the standard performance appraisal form is commonly completed for subordinates, but not for one's self. Second, it requires more information processing on the part of the rater and,
therefore, takes advantage of the decision to use verbal protocol analysis. Third, and perhaps most important, an unstructured demand is assumed to result in a more natural evaluation which is more consistent with the view that performance appraisal is an ongoing form of performance management. Finally, many practical personnel decisions are made in the absence of a completed performance appraisal rating form. This investigation would presumably be directly applicable to such decisions.

The primary purpose of this study was to explore differences in the formation of causal attributions and in the use of attributional information (consensus, consistency and distinctiveness) for two different rating targets (each manager was asked to evaluate him/herself and three subordinates). Attribution theory, an area of social psychology, was developed to describe how individuals determine the cause(s) for another person's behavior (see Ross & Fletcher, 1985 for a thorough review). Kelley (1967) suggested that in attempting to determine the causes of another's behavior, an observer will rely upon consensus information (Are others behaving the same way?), distinctiveness information (Does the person behave this way on everything he/she does?) and consistency information (Does the behavior continue to reoccur?). While such a rational approach to determining causality has been criticized, it has also provided the impetus for a great deal of fruitful research.
This study has been designed to address the attribution process for three reasons. First, it was seen as important to define a specific focus for the study since it entails an early investigation into the cognitive processing involved in performance appraisal. Second, there is clear evidence that the causal attributions ratees make can have a significant effect upon the ratings they provide (see, for example, Ilgen & Knowlton, 1980; Knowlton & Mitchell, 1980; Vance & Coover, 1986). Finally, verbal protocol analysis has been shown to be sensitive to attributional information during preliminary lab and field research.

With respect to specific predictions regarding differences in attributional processing for the two rating types, there was very little guidance for this investigation. One general finding in the attribution literature indicates that observers and actors tend to arrive at different causes for the actors performance. This has been termed the "fundamental attribution error" and refers to the tendency for observers of other's behavior to underestimate situational factors and to overestimate personal factors. It would be a contribution to demonstrate that such actor versus observer differences extend to self- versus supervisor ratings in an applied setting. Further, exploratory research was conducted to determine if any differences in attributional tendencies for the rating sources were moderated by performance level.
Another useful prediction derived from preliminary field work involved the differential use of consistency information for the two types of ratings. Based on an examination of verbal protocols collected during pretesting, it appeared that self-ratings were more likely to reflect consistency information in terms of an improvement or "gain score" than are supervisor ratings. That is, self-ratings were more apt to point out how much performance had improved over time.

Finally, the literature (Ilgen & Feldman, 1983; DeNisi et al., 1984) suggested that ratings of others might be more apt to reflect distinctiveness, consensus and other comparative information than self-ratings due to a clearer set of internalized standards for one's own performance. While such a prediction appeared legitimate, research on the use of consensus information has been inconclusive. Brown (1984) argued that consensus information is the most readily available and accurately assessed type of diagnostic information in the work setting. Williams, DeNisi, Blenscoe & Cafferty (1985) found that raters searched for consensus information when they had to select a ratee for a certain treatment. However, despite such findings, a great deal of research has indicated that consensus information is greatly underutilized (Borgida, 1978; McArthur, 1972; Kahneman & Tversky, 1973; Nisbett & Borgida, 1975). Thus, predictions with respect to the use of consensus information must be made with a degree of caution.
Before presenting hypotheses, it would appear to be valuable to relate the cognitive processes that occur during performance appraisal to some of the information presented in previous sections. While the information relevant to the cognitive processes discussed in this section is felt to be attended to contemporaneously by the rater, some of the information is being recalled from memory while other thought processes are being generated at that time. This is consistent with Ilgen & Feldman's (1983) claim that performance appraisal is best viewed as a memory-based task.

A framework developed for the purpose of this research is designed to illustrate the role memory plays in performance appraisal (see Figure 4). Time is made explicit in this presentation. A rater is expected to observe performance and think about that performance at that point and occasionally over time. Then, during the integration and judgment phase of performance appraisal, the rater must recall previous observations and thoughts as a major input in the process of thinking about making an appropriate assessment.

Thus, while forming a judgment for an annual performance appraisal may be memory-based, performance assessment is actually a more general, ongoing process and is best viewed as an on-line task (Lord, 1986). That is, judgments about performance are made "on-line" as performance information is encountered over time. Hastie & Park (1986) indicate that such judg-
ments are expected whenever individuals believe that a judgment is likely to be required at a later point in time. Precisely what factors prompt such on-line processing is another interesting research question that has implications for the different rating sources.

**Figure 4**: Framework of verbal protocol for the integration and judgment phase of performance appraisal
With respect to protocol analysis, as the rater makes an appraisal of performance upon request, some of the inputs and outputs of his/her cognitive processing will be attended to (stored in short-term memory). The rater will presumably attempt to provide labels and form sentences for a subset of this information passing through short-term memory. The result of this final encoding process will be the verbal protocol as recorded.

In summary, the attributional processes outlined in the above section will be used to determine if there are any differences in the manner in which information is processed for the two rating targets. Clearly, a number of design decisions needed to be made in order to insure that the specific research questions were appropriate to protocol analysis methodology. Unfortunately, while some researchers have suggested using protocol analysis for understanding performance appraisal (Ilgen & Favero, 1985; Schweiger, 1986), these comments have been made in passing and provide little guidance for this investigation. Ilgen & Favero (1985) merely suggest that using adaptations of verbal protocol analysis and information search methods will enable researchers to better generalize from social cognition research to performance appraisal processes. Schweiger's (1986) comments are similarly vague and are limited to an opening sentence and a concluding paragraph in the context of a three-page presentation that focused on the basic methodolog-
cal issues relevant to verbal protocol analysis. The discussion will now turn to the hypotheses that were used to guide this investigation.
Validity of Verbal Protocols

The primary purpose of this study was to provide insight into the processing of attributional information for different rating targets in the context of performance appraisal. However, it was essential for the study to demonstrate the validity of verbal protocol analysis. While the Literature Review presented evidence for the validity of verbal protocols, the evidence was derived entirely from the decision making and problem solving domains. It was necessary to provide an empirical test of the validity of verbal protocols in the performance appraisal domain. It was clear that no single study would be sufficient for establishing the construct validity of verbal protocols. However, this study included a set of measures which, taken collectively, were capable of providing significant support for the validity of protocol analysis in the performance appraisal domain.

This study also incorporated a number of internal indicators and precautions in an attempt to provide evidence for the validity of verbal protocols. Some of these measures (e.g.
number of words reported per minute during a protocol) and pre-
cautions (e.g. inducements for honest reporting of self-
relevant information) were discussed previously and a summary
table containing a complete list of such information is pre-
sented later.

The general design of this study required each manager to
evaluate him/herself and three subordinates. Thus, there was
one self-rating and three supervisor ratings (the manager eval-
uating three of his/her subordinates).

The first explicit test of the validity of the verbal proto-
cols involved the extent to which the data were consistent with
the prevailing literature on performance appraisal. As dis-
cussed in the beginning of the Literature Review, most research
has found self-ratings of performance to be more lenient than
supervisor ratings. While protocols reflect more information
processing than summary ratings, the expectation was that the
overall favorability of the evaluation would still correspond
to that found with ratings. Thus, the prediction regarding the
verbal protocol data was that:

Hypothesis 1. Verbal protocols for self-ratings will
be more lenient than those for supervisor
ratings.
The next indicator involved the extent to which the protocols reflected differences in ratee performance level. Each manager provided a verbal protocol and summary performance ratings using a performance appraisal form for the first two subordinates they were asked to evaluate. It was expected that if the supervisor's summary ratings for the first two subordinates were different, such differences in perceived performance level would also be detected in the verbal protocols. Thus, it was expected that:

Hypothesis 2. When differences in subordinate performance level are reflected in the supervisor's ratings of subordinates, they will also be reflected in the corresponding verbal protocol transcripts.

A third measure of the validity of verbal protocols involved the extent to which patterns of attributional processing were consistent across ratings for the different subordinates. Each manager provided a verbal protocol for three subordinates. The expectation was that:

Hypothesis 3. The attributional findings discovered for the supervisor evaluations of subordinates will be consistent regardless of which supervisor evaluation of a subordinate is selected for
The final test involved the ability to predict a change in the information processing exhibited in a verbal protocol following a specific demand from the environment. After managers provided verbal protocol evaluations for two subordinates and one's self, they were asked to complete performance appraisal forms for the two subordinates for which they previously provided protocols. Immediately following the completion of these performance appraisal forms, managers were asked to provide a verbal protocol for a third (different) subordinate. Since the managers completed two performance appraisal forms prior to this protocol and did not review any prior to the previous protocols, the prediction was that:

Hypothesis 4. The last verbal protocol will reflect the processing of more of the dimensions on the performance appraisal form than the first two protocols involving supervisor evaluations of subordinates.

Each of the four approaches presented above represents a different method of supporting the validity of measurements. While all of the hypotheses may not be confirmed, the approach was designed to provide some degree of converging evidence for the validity of verbal protocol analysis in this domain.
Another issue that should be addressed is the utility of applying verbal protocol analysis to the performance appraisal domain. As presented throughout the Literature Review, it was demonstrated that verbal protocol analysis is capable of capturing information that other methods (e.g. policy capturing, information boards, recording eye fixations, follow-up measures as would be used in surveys or interview procedures) can not. However, more specifically, does verbal protocol analysis offer more in terms of this investigation than can be derived using more traditional approaches in this area? This is an empirical question which goes beyond the scope of this study. The goal of this investigation was not to demonstrate that verbal protocol analysis is the only method that would arrive at similar conclusions.

However, until this question is fully resolved, it is felt that verbal protocol analysis will offer more information than a more traditional approach due to the increased confidence one can place in its results. This point is made quite clearly in support of verbal protocol analysis in general (Ericsson & Simon, 1984) and in a prescriptive discussion of an appropriate paradigm for studying attributional processes (Weiner, 1985). That is, data derived from more open-ended, less potentially reactive, concurrent measures (surveys or interviews would be considered retrospective) allows the researcher to rule out more alternative explanations and have more faith in the inferences drawn from one's results.
PRIMARY HYPOTHESES

The discussion now turns to the attributional processes of interest in this investigation. The six primary hypotheses follow directly from the final section of the Literature Review. The first two hypotheses were based on the "fundamental attribution error" and predicted that self-ratings would reflect different attributional processing than supervisor ratings. In particular:

Hypothesis 5. Supervisor evaluations of subordinates will reflect more internal attributions than self-evaluations.

Hypothesis 6. Self-evaluations will reflect more external attributions than supervisor evaluations.

Specific hypotheses regarding performance level were not made with respect to these two hypotheses, but were examined in an exploratory manner. Descriptive information on the nature of the internal attributions was also collected and analyzed.

The next hypothesis was based solely on findings from protocols collected during pretesting phases of the investigation. This was the only primary hypothesis to be formed in this man-
The remaining primary hypotheses were based on the previously discussed theoretical models of performance appraisal. This hypothesis involved the differential use of consistency information and predicted that:

**Hypothesis 7.** Self-evaluations will reflect more references to an improvement or gain in performance from previous levels of functioning than will supervisor evaluations.

The final hypothesis regarding differential patterns of processing by the rating sources involved distinctiveness and consensus information. It is possible that individuals have clearer internalized standards of performance for their own work than they do for the work of others. As a result, a plausible prediction was that:

**Hypothesis 8.** Supervisor evaluations will reflect more references to distinctiveness, consensus and other comparative information than self-evaluations.

The remaining two hypotheses addressed the order in which specific processes occur. There was no a priori reason to
believe that such processing would vary by rating focus so the hypotheses were expected to hold for self- and supervisor evaluations. However, the possibility of differences was explored. The first sequence hypothesis was based on the DeNisi et al. (1984) model which suggests that:

Hypothesis 9. Internal attributions will be followed by more evaluations than will external attributions.

The final hypothesis was based primarily on the Ilgen & Feldman (1983) model, but is implied by the DeNisi et al. and Larson models as well. This hypothesis addressed the relationship between attributional information (consensus, consistency and distinctiveness) and the formation of attributions. The prediction was that:

Hypothesis 10. Attributional information (consensus, consistency and distinctiveness) use will be more likely to precede the attributions it is associated with than it will be to follow them.
ANCILLARY HYPOTHESES

The hypotheses made in the previous section were drawn primarily from attribution theory and some of the existing models of performance appraisal. However, in addition to providing support for theoretical predictions, protocol analysis is capable of providing rich, descriptive information which can be used in an exploratory manner (Payne et al., 1978). It was considered desirable to take advantage of such an opportunity in this investigation. A more exploratory approach was particularly useful in providing more specific insights into the information processing underlying the more general hypotheses presented in the previous section.

In conducting this exploratory analysis, the entire integration and judgment phase of performance appraisal was addressed (rather than strictly the attributional processes) in order to gain additional insights which were provided by the context in which attributional processes appear. While such general models of performance appraisal exist (Ilgen & Feldman, 1983; DeNisi et al., 1984; Larson, 1984) and were discussed earlier, they are based primarily on deduction and logical propositions. The opportunity to inductively model the same process based on empirical research was exciting. Such empirically driven results provide an excellent response to Banks & Murphy (1985), Ilgen & Favero (1985) and Guion (1987), authors who have been critical of laboratory research on performance appraisal.
Exploratory research has been considered particularly useful during early phases of research on a phenomenon. A characteristic of such research is to provide the experimenter with the freedom to explore the data in search of hypotheses which can serve as the basis for future studies. However, while such freedom can be desirable, without any limiting parameters there is the very real concern of capitalizing on any idiosyncrasies of the particular data set.

This study was designed to reach a compromise between a strictly exploratory approach to data analysis and a hypothesis driven investigation. Thus, tentative hypotheses were presented a priori as derived from a framework of the integration phase of performance appraisal. This preliminary framework was developed from careful scrutiny of 15 verbal protocols obtained from managers during pretesting. It was generated prior to formal data collection, but this phase of the research must still be considered exploratory. However, the preliminary model based on pretesting did serve two purposes. First, it provided a guide for the researcher in examining future protocols and for discovering additional hypotheses. Second, it allowed for the possibility of cross-validation of portions of the preliminary model. That is, there was the potential for early confirmation or disconfirmation of the preliminary hypotheses.

The preliminary framework is presented in Figures 5 and 6 and outlines the pattern of processing discovered in the ini-
tial verbal protocols for supervisor and self-evaluations, respectively. These models were not meant to imply that every rater would follow the suggested flow of processing with probability 1.0. As was made clear in investigating the preliminary data, not unlike most models in the behavioral sciences, this is a probabilistic model. When appropriate, portions of the final results were analyzed statistically.

A brief discussion of this framework will now be presented. The model is generally consistent with the primary hypotheses presented in the previous section, but there are some minor discrepancies which require clarification. In examining these preliminary models for supervisor and self-evaluations, it is important to compare and contrast them with the theoretically based models of performance appraisal as presented in the Literature Review.

Figure 5 presents the flow of information processing for supervisor evaluations. Managers often began the evaluation process by reflecting on the subordinate's job responsibilities or one particularly salient dimension of performance. This initial component of the model is consistent with existing prescriptive models of performance appraisal.

The next component of the protocols often involved a general evaluation corresponding to the aspect(s) of performance considered in the previous step. This finding is not consistent with existing models of performance appraisal and would, if
EVALUATION OF SUBORDINATE

Consideration of job responsibilities or dimension of performance

Retrieve current overall evaluation or evaluation on performance dimension

Overall Positive Evaluation

Retrieve primarily positive episodes followed by negative episodes

Consensus info or other Comparative standards

External attribution or "consistency" information

Internal attribution

Overall Negative Evaluation

Retrieve primarily negative episodes followed by positive episodes

Consensus info or other Comparative standards

External attribution

Internal attribution

Figure 5: Preliminary empirical framework for supervisor ratings
replicated, be an interesting result. Prescriptive models imply that a great deal of information processing is required before a rater forms an evaluation. This does not appear to be the case and implies that authors of such models should be more specific in indicating what it is they are modeling.

The remaining sequence of processing was contingent upon whether the overall evaluation for the subordinate was generally positive or negative. If the overall evaluation was positive, rater's appeared to retrieve more positive episodes of behavior than negative episodes. Positive episodes also appeared to be recalled earlier than negative episodes. This finding was informative with regard to the controversy over whether information that is consistent or inconsistent with one's existing schema is most likely to be recalled. A number of researchers have found that information which is inconsistent with one's expectations is most likely to be retrieved.
EVALUATION OF SELF

Consideration of job responsibilities or dimension of performance

Retrieve current overall evaluation or evaluation on performance dimension (Primarily positive with some "could use improvement" as negative)

Retrieve positive episodes of behavior

Consistency information (Improvement in performance)

External attribution

Positive evaluation

Predictions to improve/learn more

Figure 6: Preliminary empirical framework for self-ratings

(Craik & Tulving, 1975; Hastie & Kumar, 1979; Lingle & Ostrom, 1979; Smith, 1973; Taylor, Crocker, Fiske, Sprinzen & Winkler, 1979). However, these preliminary results indicated that performance appraisal may represent a context that provides addi-
tional support for research which indicates that information confirming one's overall impressions is most apt to be retrieved (Howard & Rothbart, 1980; Rothbart, Evans & Fulero, 1979; Rothbart, Fulero, Jensen, Howard & Birrell, 1978; Sentis & Burnstein, 1979).

Review of positive episodes of behavior was likely to be followed by comparisons of the ratee's performance with the performance of other employees or with objective standards. The ratee's performance was expected to be viewed as better than the performance of other employees or as satisfactorily meeting high standards of performance. The final sequence of thoughts was consistent with existing theory and involves some type of internal attribution, predisposing a final positive evaluation and predictions of continued positive performance.

Even if a subordinate is clearly viewed in a positive manner, an occasional negative episode of behavior may still be retrieved. However, in the pilot study it was likely to be dismissed by attributing the poor performance to an external factor or by considering how much the subordinate has improved since the occurrence of the poor performance. This allowed the rater to continue to consider the subordinate in a positive manner.

If the overall evaluation for a subordinate were negative, the rater retrieved more negative episodes of behavior and these negative episodes of behavior were recalled earlier than
positive episodes. As a word of caution, very few subordinates were viewed as poor performers so the sequence of information processing for such performers is based on only a small number of protocols. However, it is comforting to recognize that the general flow of processing is analogous to that found with good performers. Incidents of poor performance by the subordinate were compared to the performance of other employees (consensus) and to other standards and were viewed as less effective. The supervisor was likely to then make an internal attribution and form a final negative evaluation. The supervisor considered what the subordinate must do (internal attribution) in order to improve performance and reflected on an action plan to verify the negative evaluation.

If the supervisor retrieved a positive episode of behavior for a poor performer it appeared that the behavior was dismissed by means of an external attribution. This enabled the supervisor to continue to provide a negative evaluation.

Figure 6 presents the flow of processing expected when a supervisor evaluates him/herself. This evaluation appeared to begin in the same manner in which a supervisor evaluates a subordinate. The manager was likely to consider his/her job responsibilities or one particularly salient dimension of performance. These initial thoughts were often followed immediately by a corresponding positive evaluation or a "could use improvement" assessment which served to replace direct negative
evaluations. Managers then appeared to retrieve positive episodes of behavior. Even if they provided a "could use improvement" or negative evaluation, it was unlikely that they would retrieve corresponding negative episodes of behavior.

With regard to attributional information, managers often relied on "consistency" information by considering how much their performance had improved over time. This corresponds directly to one of the primary hypotheses presented in the previous section.

Managers also appeared to use external attributions and to avoid making internal attributions for their own performance. However, the external attributions appeared to take on a slightly different form than those used to account for subordinate's performance. In the case of subordinate's performance, external attributions were used to explain away or dismiss performance that was not easily incorporated into the manager's schema for that subordinate. For the supervisor's self-evaluation, external attributions were used to enhance or improve their self-evaluations because they were able to achieve as much as they did despite various environmental constraints.

Self-evaluations were often completed with a positive evaluation and a prediction to improve or learn more in the future.

The entire framework, representing both supervisor and self-evaluations, is generally consistent with the primary hypoth-
eses presented in the previous section. This need not have been the case since all but one of the primary hypotheses were derived from the prescriptive models of performance appraisal. The framework was consistent with the first hypothesis. Supervisor evaluations did reflect more internal attributions than self-evaluations. In terms of the second hypothesis, self-evaluations also appeared to reflect more external attributions. Even though external attributions were present in the supervisory evaluations, they appeared within two paths of information processing which occurred with low probability. That is, external attributions for subordinates appeared to occur following the retrieval of negative episodes of behavior for good performers and positive episodes of behavior for poor performers. Both types of retrieval were expected to occur fairly infrequently.

The empirical framework was also consistent with the next two primary hypotheses regarding a greater use of consensus information in supervisor evaluations and more references to "improvements in performance" for self-evaluations. Of course this had to be the case for this later hypothesis since this was the one primary hypothesis based on the pretest data.

The final two primary hypotheses addressed the order in which specific processes occur. The first sequence hypothesis stated that internal attributions would be followed by evaluative statements while external attributions would not be fol-
lowed by evaluations. This was the only primary hypothesis that was not entirely consistent with the empirically based model of performance appraisal. The model in Figures 5 and 6 indicates that this hypothesis held for supervisor evaluations, but was not consistent with the processing found for self-evaluations. Positive evaluations did appear to follow external attributions in the context of self-evaluations. This may be due to the different form external attributions take when used in self-evaluations. This point was addressed earlier.

The empirical framework for the self- and supervisor evaluations was consistent with the final primary hypothesis. The final hypothesis stated that attributional information was more likely to precede the attributions it was associated with than it was to follow them.
CHAPTER IV
METHOD

SUBJECTS

Preliminary pilot work was conducted with 26 introductory psychology students and 15 managers employed by the City of Columbus. The formal study was conducted with 18 data processing managers and 18 office managers employed by Ohio State University. Within each of the job types, half of the sample consisted of first-line managers and the other half were second-line managers.

All managers participating in the formal study did so voluntarily. The Department of Personnel Services provided the researcher with the names of all data processing and office managers employed by the University. The managers were contacted by telephone and asked if they would be willing to participate in confidential performance appraisal research for a dissertation in the field of industrial psychology. Approximately 70% of the managers contacted were willing to participate. Interviews were conducted privately in each manager's office or in a nearby conference room.
It was considered important to document each manager's experience with: 1) performance appraisals in general; 2) the performance appraisal system presently used; 3) the ratee and; 4) the job (as manager and employee). The last time the manager completed a performance appraisal was also recorded. The context in which the manager and employee work was also documented.

PROCEDURE

The managers received a brief introduction to the study which included a discussion of performance appraisal and the "think aloud" methodology. Before managers provided the verbal protocols, a 20 minute training session was provided. This was required to insure that the verbalizations reflected the ongoing contents of short-term memory. Training is also useful in minimizing individual differences in managers' abilities to provide verbal reports (Ericsson & Simon, 1984) and in increasing their accuracy (Mabe & West, 1982).

The training session consisted of five practice exercises. The selection of practice exercises is not a trivial issue. The exercises were rather simple and were taken from the procedures prescribed by Ericsson & Simon (1984). The five tasks involved: 1) two simple addition problems; 2) generating as many words as possible given a set of letters; 3) determining the number of windows in your house or apartment and; 4) naming 15 different animals.
The instructions to managers as to how to "think aloud" were adopted from Ericsson & Simon (1984). Managers were told that the researcher can detect deviations from the correct procedure and will intervene during the practice exercises to assist them in reporting appropriately. The researcher also prompted the manager to continue verbal reporting whenever there was a period of silence lasting for more than approximately 10 seconds. Prompting is also a technique that is recognized in the self-disclosure literature as facilitating the completeness of verbal reports (Cozby, 1973).

Another approach for eliciting greater disclosure from managers is for the experimenter to disclose information about himself. In the context of this study, it would have been possible for the researcher to self-disclose on variations of the exercises discussed above. However, there might have been a trade-off in benefits as the researcher's verbal reports might have biased the managers to approach the exercises and provide reports in a manner consistent with the researcher's verbalizations. As a result, the experimenter provided self-disclosure on a brief task unrelated to performance appraisal or the practice exercises.

Following each of the five exercises, the researcher provided feedback on the manager's verbal reports. Appropriate reporting was reinforced by means of this social approval.
Once the researcher felt the managers were consistently reporting on the contents of short-term memory, they were seen as prepared to provide the verbal reports while completing the performance evaluations. As support to the managers' self-esteem, they were told that they are in an excellent position to evaluate themselves and subordinates. Managers were informed that complete confidentiality would be maintained and were encouraged to be honest in their verbal reports.

Each manager then completed a series of the performance evaluations for the different ratees while providing simultaneous verbal reports. A tape recorder was operating to record the entire reports. As is conventional (Ericsson & Simon, 1984), the researcher was present during data collection and sat behind or to the side of the manager. This practice allowed the experimenter to prompt the manager into providing a continuous verbal report on the information passing through short-term memory.

Each manager was asked to provide a "thorough, complete" open-ended evaluation for two subordinates and one's self. The time necessary to complete each evaluation was entirely the manager's decision. No documentation or other decision aids were used during these open-ended appraisals. The experimenter used a table of random digits to randomly select which of the manager's subordinates was to be evaluated to prevent any preliminary summary judgments from biasing the evaluations. The
managers completed the open evaluations for the ratees in the following sequence: subordinate, self and another subordinate. The manager stopped after each evaluation at which time the experimenter provided the instructions for the next exercise.

The ordering for the above series of evaluations was designed to provide a measure of the consistency of findings across different subordinates. It was also designed to serve as a partial control for any order effects. For instance, if results using the first subordinate were similar to those found using the second subordinate, this would provide some evidence against significant order effects. Order effects would have appeared as systematic differences between the first and second supervisor evaluations. Lack of consistency, on the other hand, would result in more random, unsystematic findings.

A complete control for order was considered somewhat problematic in this investigation. There are practical and logical reasons for not asking managers to begin with a self-evaluation. In addition, it is important that the two supervisor evaluations were not made consecutively so that a simple repetition of evaluative processing was not confounded with consistency of processing associated with the rating source.

Following completion of the three open evaluations, managers were asked to complete Ohio State University Employee Performance Evaluation forms for the same two subordinates previously evaluated using the open-ended format. The appraisal forms
consisted of seven specific dimensions (Job Knowledge, Quality of Results Obtained, Productivity, Cooperation with Others/Attitude, Dependability/Reliability, Reasoning/Judgment, and Supervision of Others) and an overall evaluation. Each dimension required a rating on a 5-point scale from unsatisfactory to exceptional. Verbal reports were not provided as the performance appraisal forms were completed. These data were useful in assessing the validity of verbal protocols and in providing supplementary information on the performance of the subordinates.

After managers completed the two performance appraisal forms, they were asked to "think aloud" as they provided a final open evaluation for a third randomly selected subordinate. This verbal protocol was analyzed to determine if it reflected more of the specific dimensions found on the performance appraisal form than the protocols collected from the first two supervisory evaluations of subordinates. One would expect that by completing two appraisal forms the dimensions of performance would become more salient to the rater and would be reflected in the protocol for a subsequent open evaluation. Such findings would serve to increase the researcher's confidence in the validity of protocol analysis.

In the process of collecting verbal protocols it is important to assess the quality of the data. Members of the dissertation committee had an opportunity to examine and react to
portions of the preliminary data and protocol data collected for the formal study. An evaluation regarding the extent to which the data were a reflection of the information stored in short-term memory was required. Ericsson & Simon (1984) suggest that the number of words reported as well as the nature of the data can indicate what type of information has been verbally reported.

Protocols have been found to generate approximately 100 words per minute. If protocols drop below 50 words per minute or produce more than 150 words per minute, the data may be of poor quality. A low word count may indicate that the subject is describing or explaining information that is being attended to. An extremely high word count may indicate a predominance of recognition processes with very little problem solving.

In conjunction with the word count, examining the nature of the data provides a more accurate indication of its quality. The verbal reports should not contain descriptions, explanations, extremely inconsistent utterances, or verbs in past tense (in certain contexts). More specifically, Ericsson & Simon (1984) claim that verbal protocols should include few of the following words: "I," "my head," "remember," "feel," and "know." Finally, the verbal reports should be relevant to the task at hand and should be provided with minimal prompting.
IDENTIFYING PHRASES

Once the protocols were recorded, the next step was to make a complete transcript of the verbal reports given by each subject. It was then necessary to break the protocols up into short phrases (Newell & Simon, 1972). Each phrase represented, as much as possible, a single thought, idea, or task assertion made by the subject. The logic for this procedure is that small phrases provide a detailed analysis of the protocol as well as allowing for a series of relatively unambiguous assessments of the information subject's attend to at a given time.

Before verbal protocols from the actual study were broken up, two undergraduate psychology students received training and had the opportunity to practice breaking up the protocols. These students, who had no knowledge of the hypotheses of interest, broke up segments of the protocols independently and compared results. This independence in forming phrases provided an indication of agreement. Once the coders were capable of consistently reaching agreement of 90% or better, they were considered prepared for breaking up the protocols for the formal study. The first 10 protocols of the formal study were also broken up into phrases by both assistants who worked independently. This allowed for computing an index of agreement for the data actually used in the study.
CODING PROCEDURES

After the verbal protocols were broken up into phrases, each phrase was assigned one or more codes based on different categories for the attributional processes involved in performance appraisal. While Ericsson & Simon (1984) discuss different task independent coding schemes, they recommend coding categories based on the task of interest (in this case it involves the process of conducting a performance appraisal). A task-based coding scheme was more appropriate for this research due to the nature of the hypotheses and the specificity of information provided by a task-based category system. An illustration of the coding scheme used for this research is provided in Table 1. These categories were derived directly from the attributional processes discussed in the performance appraisal models and from the coding of preliminary data.

Coding was conducted by reading each phrase individually and assigning it a code from one or more of the categories in Table 1 based on a similar process. It is preferable to code each phrase independently of the surrounding phrases and the context in which the phrase appears. Reducing the influence of such extraneous factors will tend to enhance the reliability of the coding (Ericsson & Simon, 1984). The coders attempted to code each phrase independently, but occasionally found it necessary to consider the surrounding phrases in order to make an appropriate decision.
Table 1

Preliminary verbal protocol coding scheme

1) Positive evaluation ("She does an outstanding job.")

2) Negative evaluation ("He is not a good supervisor.")

3) Could improve ("I could improve my interpersonal skills.")

4) Consensus information ("Better than anyone else who has held this job.")

5) Other standards ("I expect at least three inspections per year.")

6) Distinctiveness information ("This is the only project he has not completed on time.")

7) Dimension/Area of performance ("I am responsible for completing the monthly reports.")

8) Consistency information ("He has done this for the past two years.")

9) Consistency/Improvement information ("He has come a long way since he started here.")

10) External attributions ("I don't have enough time to do everything.")

11) Trait attribution ("She can be aggressive.")

12) Knowledge attribution ("She is quite knowledgeable.")

13) Skill attribution ("My interpersonal skills are strong.")
Table 1 (continued),

14) Experience attribution ("He needs more experience.")

15) Effort attribution ("He works very hard.")

16) Interest attribution ("He seems enthusiastic about the work.")

17) Other internal attributions ("He is reliable.")

18) Prediction of negative behavior ("This poor performance will probably continue.")

19) Prediction of positive behavior ("I expect her to be my best supervisor.")

20) Improving/Expectations to improve ("I hope to accomplish more.")

21) Plan to test evaluation ("I'll have to assign him some drafting work to see if he knows how to do it.")

At the outset of this coding process, the two assistants received training and coded segments of the preliminary protocols independently. This provided an independent check on the reliability across coders. The two assistants continued preliminary coding until they were consistently capable of achieving reliability estimates of .90 or above. Both students also independently coded the first 10 protocols of the formal study in order to provide an index of the reliability of the coding process for the study. After the first 10 protocols were cod-
ed, each student coded every other protocol. After all the data were coded, each student was asked to recode one of the protocols they had previously coded (either the 13th or 14th protocol) in order to provide an index of the intra-rater or "retest" reliability for the data used in the study.

The researcher was involved in preliminary coding as it would have been acceptable to modify the coding scheme based on the content of the data and the hypotheses of interest. The coding scheme might also have been modified on the basis of reliability assessments. While reliabilities have been reported to be in the .8 to .9 range (Ericsson & Simon, 1984), they may be reduced substantially by an overly detailed coding scheme. However, while definitions were clarified and decision rules were defined, the coding scheme was not altered during the practice sessions with the two coders.
CHAPTER V
RESULTS

VALIDITY OF VERBAL PROTOCOLS

The initial analyses were designed so that a conclusive statement could be made regarding the validity of using verbal protocol analysis in the performance appraisal domain. These initial findings provide the foundation for the remainder of the study.

The first step in evaluating the protocols was to examine them in terms of various internal indices. All of the data were examined qualitatively by the researcher and portions of the data were examined by two members of the dissertation committee. A transcript of the first protocol collected for the formal study is included in Appendix A.

The managers appeared generally comfortable with the task and took their responsibilities seriously. The data were clearly relevant to the performance appraisal task. Also, no direct inconsistencies in facts or information were detected. The data appeared to reflect on-line processing of raters recalling previously interpreted information and occasionally thinking about that information in order to provide an evalua-
tion. It was apparent from the tense of the verbs used that the protocols did not reflect an attempt to recall a previous evaluation.

There were a few cases in which a manager made an attempt to explain his/her thoughts. Each of these instances took place at the outset of the session. For these cases, the researcher stopped the manager immediately in order to clarify the nature of the task. In each situation, this intervention appeared to be sufficient in producing appropriate verbalizations.

There was very little need to prompt the managers to "think aloud" once the performance appraisal exercises began. In fact, the only time prompting was necessary was at the outset of the first supervisor evaluation or, more frequently, at the beginning of the self-evaluation. The self-evaluation appeared to be more difficult for the managers, at least initially, and to require an increase in information processing. This would have limited their ability to "think aloud" simultaneously.

The number of words verbalized per minute provides another index of the quality of the data. The number of words verbalized during a randomly selected minute of the self-evaluation and of the second supervisor evaluation were recorded for all 36 managers. The overall mean for both evaluations for all 36 managers was 99.8. This is particularly comforting since it falls precisely in the middle of the generally appropriate range of 50 to 150 words per minute. A repeated measures
t-test was performed to determine if there was a significant difference between the number of words verbalized during self-evaluations (mean=96.83) and supervisor evaluations (mean=102.75). The test statistic indicated that there was not a significant difference between the two means, t(35)=1.15, p=.26.

An attempt was also made to determine if the number of words verbalized per minute was a function of the rater's experience with the ratee, the job or the performance appraisal process. The number of words verbalized per minute during the self- and supervisor evaluations were correlated with the number of: 1) months the rater has been on his/her job; 2) months the rater may have held the ratee's job; 3) months the rater has supervised the ratee and; 4) performance appraisals of any type that the rater has completed throughout his/her career. The number of appraisals completed for a given subordinate and the length of time the subordinate has been supervised provide redundant information since an appraisal was conducted on an annual basis for each ratee in the study. Pearson, Spearman and Kendall Tau correlation coefficients were computed and are presented in Table 2. The Spearman and Kendall Tau correlation coefficients are nonparametric procedures and are, therefore, free from the assumptions of normal theory. While the Spearman and Kendall coefficients are related, disagreement between pairs appears as the squared difference between ranks in computing Spearman's
coefficient. The process of squaring differences places a different weight on different inversions of order. In Kendall's coefficient, an inversion in order for any pair of objects is treated in the same way as evidence of disagreement.

Table 2

Correlations between words per minute and experience

<table>
<thead>
<tr>
<th>WORDS PER MINUTE IN:</th>
<th>Supervisor Evaluation</th>
<th>Self-Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Months on job</td>
<td>- .42 (p=.01)</td>
<td>- .24 (p=.16)</td>
</tr>
<tr>
<td></td>
<td>-.37 (p=.02)</td>
<td>-.22 (p=.18)</td>
</tr>
<tr>
<td></td>
<td>-.27 (p=.02)</td>
<td>-.19 (p=.10)</td>
</tr>
<tr>
<td>Months on ratee's job</td>
<td>-.07 (p=.67)</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>-.00 (p=.99)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-.00 (p=.97)</td>
<td></td>
</tr>
<tr>
<td>Months supervised ratee</td>
<td>-.08 (p=.66)</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>-.07 (p=.68)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-.06 (p=.49)</td>
<td></td>
</tr>
<tr>
<td>Number appraisals completed</td>
<td>-.06 (p=.69)</td>
<td>-.07 (p=.67)</td>
</tr>
<tr>
<td></td>
<td>-.09 (p=.57)</td>
<td>-.26 (p=.12)</td>
</tr>
<tr>
<td></td>
<td>-.08 (p=.49)</td>
<td>-.20 (p=.10)</td>
</tr>
</tbody>
</table>

The first, second and third row in each cell contains Pearson, Spearman and Kendall coefficients, respectively.

The results in Table 2 indicate that the number of months the rater has been on his/her job was the only experience vari-
able to be related to the rate of verbalization. It appears that managers who were on the job longer produced fewer words per minute, particularly for the supervisor evaluations. This finding was contrary to what would have been predicted. While legitimate explanations could account for this result, it is entirely possible that it is a product of chance alone, particularly since none of the other experience variables were significant.

In addition to the above internal indices, four specific predictions were made in an attempt to establish the validity of verbal protocol analysis in the domain of performance evaluation. The first hypothesis stated that the verbal protocols for self-ratings would be more lenient than those for supervisor ratings. This hypothesis was tested by having a sample of introductory psychology students rate (on a scale from 1=poor to 5=excellent) and rank order the protocol transcripts for the first two supervisor evaluations and the self-evaluation simultaneously. This was done for all 36 managers. Each manager's data were evaluated by a different student. The mean rating for all supervisor evaluations was 3.5 and the mean for the self-evaluations was 3.6. While the means were in the predicted direction, a correlated t-test indicated that the two means were not significantly different, t(35)=1.08, p=.14. Thus, the first hypothesis was not confirmed; the mean ratings for the self- and supervisor evaluations were quite similar.
An examination of the student's rankings of the three protocol transcripts indicated that for the manager's self-evaluations, 7 received a rank of one, 20 received a rank of two and 9 were ranked third. The binomial distribution can be used to make inferential statements under the null hypothesis that \( p = 0.33 \) for any particular rank order. For rank = 1, \( B = 7 \) and the standardized value of \( B = -1.77, p = 0.038 \). This indicates that significantly less than \( 1/3 \) of the self-evaluations were ranked first. For rank = 2, \( B = 20 \) and the standardized value of \( B = 2.83, p = 0.002 \). Significantly more than \( 1/3 \) of the self-evaluations received a rank of two. Finally, for rank = 3, \( B = 9 \) with a standardized value of \( -1.06, p = 0.14 \). This indicates that the number of self-evaluations ranked third was not significantly less than \( 1/3 \). The above results indicate that self-evaluations were viewed as less positive than the best supervisor evaluations, but not as negative as the least favorable supervisor evaluations. Again, these results did not confirm the hypothesis that self-ratings would be more favorable, but they do indicate that self-evaluations are not less favorable than supervisor evaluations.

Analysis of the coded protocol transcripts can also be used to provide insight into the favorability of the ratings. Single degree of freedom contrasts analyzing the difference between self- and supervisor ratings indicated that there was no difference in the proportion \( (F(1,105) = 0.05, p = 0.82) \) or fre-
quency \((F(1, 105) = 3.17, p = .08)\) of positive statements made by the two rating sources. However, similar analyses for negative evaluations indicated that the supervisor evaluations included a greater frequency \((F(1, 105) = 16.27, p < .0001)\) and proportion \((F(1, 105) = 13.15, p = .0004)\) of negative phrases than did self-evaluations. Thus, the coded data serve to provide indirect support for the first hypothesis in that self-evaluations included less negative statements than the supervisor evaluations.

The second hypothesis stated that differences in subordinate performance level would be reflected in ratings and verbal protocol transcripts. Before the introductory psychology students completed the previous evaluation exercise, they were given verbal protocol transcripts of two supervisor evaluations for subordinates who had received different summary performance ratings. These two protocols were produced by the same manager. The students were asked to read each transcript to determine which one of the two they felt was more positive, more favorable, or better. The number of "successes" would follow a binomial distribution with \(p = .5\) under the null hypothesis. Although the total \(n = 36\), three of the managers provided identical ratings for all three of their subordinates so these managers were excluded from this analysis with the resulting \(n = 33\). The students correctly identified the protocol that corresponded to the more favorable or positive summary rating for 30 out
of the 33 pairs. Standardizing $B=30$ for $p=.5$ and $n=33$ provides a value of 4.7, $p<.0001$. This easily exceeds conventional critical values and demonstrates that the protocols reflected the rating differences and these differences could be detected by the undergraduate students.

The third hypothesis predicted that any difference in attributional processing between the two rating sources would hold regardless of which supervisor evaluation (first, second or third) was selected for analysis. This hypothesis was assessed in the context of evaluating the four main hypotheses which referred to differences between the self- and supervisor evaluations. Each of the three supervisor evaluations was compared to the self-evaluation for each of the four hypotheses. The expectation was that the results would be similar regardless of which supervisor evaluation was compared to the self-evaluation. Although these results will be presented more formally later, the p-values obtained from single d.f. contrasts for the four hypotheses of interest using frequency and proportion data are presented in Table 3. The p-values within each cell are generally quite consistent. There appears to be only one cell (consistency/improvement hypothesis for proportion data) that would lead a researcher to arrive at a different conclusion depending upon which supervisor evaluation was selected for analysis.
Table 3
P-values for contrasts between self- and supervisor evaluations

<table>
<thead>
<tr>
<th></th>
<th>Frequency Data</th>
<th>Proportion Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>External</strong></td>
<td>.0001</td>
<td>.0001</td>
</tr>
<tr>
<td>attribution</td>
<td>.0001</td>
<td>.0001</td>
</tr>
<tr>
<td>hypothesis</td>
<td>.0001</td>
<td>.0001</td>
</tr>
<tr>
<td><strong>Internal</strong></td>
<td>.0029</td>
<td>.0134</td>
</tr>
<tr>
<td>attribution</td>
<td>.0011</td>
<td>.0009</td>
</tr>
<tr>
<td>hypothesis</td>
<td>.0003</td>
<td>.0222</td>
</tr>
<tr>
<td><strong>Consistency</strong></td>
<td>.8400</td>
<td>.8485</td>
</tr>
<tr>
<td>improvement</td>
<td>.2500</td>
<td>.0564</td>
</tr>
<tr>
<td>hypothesis</td>
<td>.4400</td>
<td>.4512</td>
</tr>
<tr>
<td><strong>Consensus</strong></td>
<td>.0088</td>
<td>.0727</td>
</tr>
<tr>
<td>standards</td>
<td>.0013</td>
<td>.0009</td>
</tr>
<tr>
<td>hypothesis</td>
<td>.0052</td>
<td>.0388</td>
</tr>
</tbody>
</table>

First, second and third row of each cell corresponds to the first, second and third supervisor evaluation, respectively.

The final hypothesis directed at establishing the validity of verbal protocols predicted that the open evaluation following the completion of the two appraisal forms would reflect more of the dimensions of performance on the forms than the previous open evaluations. That is, subjects would tend to recall information in terms of the Ohio State performance evaluation form. This hypothesis was tested by providing introductory psychology students with two protocol transcripts. One of
the transcripts was completed at the end of the session and the
other one was whichever protocol of the first two protocols
corresponded most closely to the final protocol in terms of
performance rating level. In addition to the two protocol
transcripts, each student received a blank copy of the Ohio
State performance evaluation form. The students' task was to
determine which of the two protocols corresponded most closely
to the evaluation form in terms of the dimensions of perform­
ance. Of the total n=36, 26 students correctly identified the
protocol that followed the completion of the evaluation forms.
Based on the binomial distribution with p=.5 under the null
hypothesis, B=26 was standardized using p=.5 and n=36 to obtain
a value of 2.67, p=.0038. This probability is significant and
indicates that the protocols were affected in a predictable
manner by making particular dimensions of performance more
salient to the rater.

The above findings appear to provide significant support for
the use of verbal protocol analysis in this domain. A summary
of the results and precautions that attempted to establish the
validity of verbal protocol analysis are presented in Table 4.
The following results addressed the validity of verbal protocol analysis:

1) Contrary to the prediction, self-evaluations were not significantly more lenient than the supervisor evaluations. However, the means were in the predicted direction.

2) The protocol "rating level" corresponded to rating level on traditional appraisal forms.

3) The protocols revealed consistent findings within a given rating source.

4) The verbal protocols were manipulated to reflect the processing of different performance dimensions.

5) Internal indices were evaluated by the researcher and a member of the dissertation committee considered proficient in collecting verbal protocols. Such indices included:
   a) The internal consistency of each verbal report.
   b) The nature of the data (e.g. types of verbs used, lack of explanations).
   c) Number of words verbalized per minute.
   d) Amount of prompting required by the researcher.

6) The verbal reports were relevant to the performance appraisal task.

The following precautions were designed to rule out alternative hypotheses:
Table 4 (continued),

1) Managers were given the name of the ratee at the time they were to provide the verbal protocol to minimize the potential for rehearsal effects.

2) A number of factors were incorporated into the procedures in an attempt to reduce subjects' motivation to distort their verbal reports for self-enhancement purposes.

CODING RELIABILITY

Prior to conducting the statistical analyses to examine the primary hypotheses, the data were divided into phrases and coded. The first 10 protocols were coded by the two coders independently. For the remaining 26 protocols, each assistant coded every other protocol. A standard approach for assessing the quality of the coding procedure is to determine the reliability of the coding practices. There are a variety of different approaches for assessing reliability in this context.

The first approach taken was to compute the reliability of the phrasing and coding between the two coders for each of the first 10 managers. This provides an index of reliability for the coding scheme considered in its entirety. Reliability for the phrasing was defined as the proportion of phrases coders agreed upon divided by the total number of different phrases formed by the two coders.
Reliability of the coding was determined by computing three different types of correlation indices. Pearson's correlation coefficient was computed, but this is only interpretable for $r$ between 0 and 1 when the two variables of interest satisfy a normal bivariate distribution. Due to this limitation, Spearman rank order correlation coefficients were also computed. Spearman $r$ is simply a Pearson correlation coefficient based on ranked data and has the benefit of being interpretable without assuming normality. All Spearman correlation coefficients presented have been corrected for ties. The final statistic used to assess the reliability of the coding was Cronbach's alpha. Cronbach's coefficient provides an estimate of internal consistency in which coders become analogous to test items in a test construction context.

The various estimates of agreement for the first 10 managers are presented in Table 5. All of the statistics in the table are equal or greater than .75. These estimates demonstrate that there was substantial agreement between the coders. The means across the first 10 managers were .95 for phrasing, .98 for Pearson's $r$, .87 for Spearman's $r$ and .99 for Cronbach's alpha.

There is the potential for these estimates to be artificially inflated since certain categories were used less frequently than others to the extent that a given manager might not have any phrases assigned to certain categories. Whether estimates
Table 5.
Reliability estimates for 10 managers

<table>
<thead>
<tr>
<th>Manager</th>
<th>Phrasing</th>
<th>Pearson</th>
<th>Spearman</th>
<th>Cronbach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.93</td>
<td>.99</td>
<td>.97</td>
<td>.99</td>
</tr>
<tr>
<td>2</td>
<td>.91</td>
<td>.96</td>
<td>.90</td>
<td>.98</td>
</tr>
<tr>
<td>3</td>
<td>.93</td>
<td>.99</td>
<td>.95</td>
<td>.99</td>
</tr>
<tr>
<td>4</td>
<td>.95</td>
<td>.98</td>
<td>.75</td>
<td>.99</td>
</tr>
<tr>
<td>5</td>
<td>.97</td>
<td>.98</td>
<td>.88</td>
<td>.99</td>
</tr>
<tr>
<td>6</td>
<td>.93</td>
<td>.99</td>
<td>.81</td>
<td>.99</td>
</tr>
<tr>
<td>7</td>
<td>.93</td>
<td>.97</td>
<td>.84</td>
<td>.98</td>
</tr>
<tr>
<td>8</td>
<td>.98</td>
<td>.99</td>
<td>.82</td>
<td>.99</td>
</tr>
<tr>
<td>9</td>
<td>.98</td>
<td>.97</td>
<td>.80</td>
<td>.99</td>
</tr>
<tr>
<td>10</td>
<td>.94</td>
<td>.93</td>
<td>.94</td>
<td>.96</td>
</tr>
<tr>
<td>Mean</td>
<td>.95</td>
<td>.98</td>
<td>.87</td>
<td>.99</td>
</tr>
</tbody>
</table>

derived from such data are entirely artificial can be debated since all categories were legitimate and potentially usable by each coder for each manager. However, as a precaution, the total set of frequency data (21 categories collapsed across the entire protocol for a manager) for all 10 managers was collapsed into a single vector for coder 1 and into a second vector for coder 2. These two vectors of cumulative frequencies were correlated to produce Pearson, Spearman and Cronbach coefficients of .99, .97 and .99, respectively. The results were consistent with those above in demonstrating significant inter-rater agreement.

A more specific method of assessing reliability is to index the extent of agreement for a given category or process in the
coding scheme. It was expected that some processes would be detected with greater reliability than others. Obtaining reliability information by category was useful in interpreting the results for the primary hypotheses which were related to the specific categories. This type of analysis required determining the extent of agreement between coder 1 and coder 2 for each of the 21 categories. The data consisted of frequency counts for each of the first 10 managers for coder 1 and for coder 2 for a given category. The Pearson, Spearman, and Cronbach estimates for each of the 21 categories are presented in Table 6.

The results in Table 6 indicate that the extent of agreement for approximately half of the categories was in the .8 to .9 range. This was encouraging since these results served as the basis for the remaining analyses. However, the results also indicate that distinctiveness information could not be detected reliably. One reason for this is that very few phrases were coded into this category so the estimates were more likely to be affected by a few disagreements. However, this does not change the fact that the coding for this category was unreliable and could not be interpreted properly for the primary analyses.

The "prediction of negative behavior" and "plan to test evaluation" categories were exploratory categories and apparently were not used by either coder (both agreed perfectly!) while coding the data for the first 10 managers.
Table 6
Reliability estimates for 21 categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Pearson</th>
<th>Spearman</th>
<th>Cronbach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Evaluation</td>
<td>.99</td>
<td>.96</td>
<td>.99</td>
</tr>
<tr>
<td>Negative Evaluation</td>
<td>.93</td>
<td>.88</td>
<td>.96</td>
</tr>
<tr>
<td>Could Improve</td>
<td>.98</td>
<td>.90</td>
<td>.98</td>
</tr>
<tr>
<td>Consensus Information</td>
<td>.92</td>
<td>.89</td>
<td>.92</td>
</tr>
<tr>
<td>Other Standards</td>
<td>.41</td>
<td>.41</td>
<td>.53</td>
</tr>
<tr>
<td>Distinctiveness</td>
<td>-.06</td>
<td>-.05</td>
<td>.00</td>
</tr>
<tr>
<td>Dimension Performance</td>
<td>.90</td>
<td>.85</td>
<td>.95</td>
</tr>
<tr>
<td>Consistency Information</td>
<td>.93</td>
<td>.85</td>
<td>.92</td>
</tr>
<tr>
<td>Consistency/Improvement</td>
<td>.55</td>
<td>.39</td>
<td>.60</td>
</tr>
<tr>
<td>External Attribution</td>
<td>.85</td>
<td>.80</td>
<td>.89</td>
</tr>
<tr>
<td>All Internal Attrib</td>
<td>.82</td>
<td>.70</td>
<td>.85</td>
</tr>
<tr>
<td>Trait Attribution</td>
<td>.76</td>
<td>.80</td>
<td>.83</td>
</tr>
<tr>
<td>Knowledge Attribution</td>
<td>.78</td>
<td>.70</td>
<td>.87</td>
</tr>
<tr>
<td>Skill Attribution</td>
<td>.41</td>
<td>.61</td>
<td>.58</td>
</tr>
<tr>
<td>Experience Attribution</td>
<td>.53</td>
<td>.63</td>
<td>.58</td>
</tr>
<tr>
<td>Effort Attribution</td>
<td>.37</td>
<td>.28</td>
<td>.53</td>
</tr>
<tr>
<td>Interest Attribution</td>
<td>.85</td>
<td>.85</td>
<td>.86</td>
</tr>
<tr>
<td>Other Internal Attrib</td>
<td>.44</td>
<td>.57</td>
<td>.56</td>
</tr>
<tr>
<td>Predict Negative</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Predict Positive</td>
<td>.92</td>
<td>.99</td>
<td>.93</td>
</tr>
<tr>
<td>Predict Improvement</td>
<td>.78</td>
<td>.50</td>
<td>.59</td>
</tr>
<tr>
<td>Test Evaluation</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
</tbody>
</table>

***No phrases were assigned to this category by either coder for any of the first 10 managers.

The remainder of the categories were characterized by reliabilities in the .4 to .6 range. While such estimates are less than ideal, they didn't necessarily pose a serious problem. Certain analyses were only exploratory and could be adjusted accordingly and other analyses were interpreted in light of these reliability estimates. For instance, analyses for inter-
nal attributions were based primarily on the composite of all internal attributions combined rather than each internal attribution considered separately. In addition, it is probably appropriate to view these reliability estimates as lower bounds for the reliability of the data used in the study for three reasons. First, the data used to compute these reliability estimates were not the data used to test the hypotheses. After the two assistants coded each manager's data independently, they convened in order to produce a final set of coded data which they could agree on. It can be expected, from either an intuitive or statistical perspective, that the final set of coded data would be more reliable than the original data. Second, the additional discussion required to reach agreement for the final codings served to clarify additional issues which would increase reliabilities for the data not yet coded. Finally, the principal researcher and the two coders convened to review the reliability results based on the first 10 protocols. Further discussion and training was conducted and appeared to clarify additional issues. This would presumably serve to increase reliability for data coded during the remainder of the study.

A final index of agreement computed in the study involved a single measure of intra-rater reliability for each coder. This provided a measure of the extent to which each coder was reliable or consistent over time. The intra-rater reliability coef-
Coefficients were obtained by having each assistant recode one of the protocols they had previously coded (either the 13th or 14th protocol) after all 36 protocols had been coded. This provided an index of agreement for the last 26 protocols which represented the phase of the study in which each coder worked independently.

The Pearson, Spearman and Cronbach coefficients for coder 1 were .96, .92 and .98, respectively, and for coder 2 the estimated coefficients were .97, .94 and .98, respectively. These results indicated that the two coders were capable of providing reliable data for the entire phase of the study when they coded independently. The findings are particularly reassuring since the coders began this second phase immediately following the first phase in which sufficient inter-rater reliability was demonstrated. Thus, there is some evidence indicating that all data used in the study were reliably coded.

GENERAL ANALYSES OF CODED DATA

This section examines the attributional processing of the two rating types based on the coded protocols. Prior to addressing the specific hypotheses of interest, the entire data set was examined in a broader framework. The data can be analyzed using a two factor repeated measures analysis of variance (ANOVA). Although the overall results from this analysis were not particularly relevant to the hypotheses of interest, this
approach has been conventional and did provide an overall framework in which to analyze the specific hypotheses.

The coded data could be entered into the analyses in one of two forms. The data could be represented in terms of frequencies or proportions. The frequency data consisted of the number of phrases coded into a particular category for one of the rating types by a given manager. A second approach would be to analyze the data in terms of the proportion of phrases out of a manager's total for that particular evaluation that were coded into a specific category. Both types of data were analyzed for this study.

The first ANOVA performed was a 4 (rating target) X 13 (coding categories) completely within subjects design. The four rating types or "sources" consisted of the three supervisor evaluations (evaluations of subordinates) and the self-evaluation. The categories were based on those found in Table 1. The 21 coding categories were reduced to 13 by collapsing the seven specific internal attribution categories and by collapsing the distinctiveness, consensus and standards categories. This second collapsed category was formed in this manner because the hypothesis made regarding these categories was made in terms of this entire composite.

The summary table of the ANOVA applied to the frequency data is presented in Table 7. The results indicate that there was a main effect for rating target, F(3,105)=4.11, p=.0078. The
mean number of phrases per category was 2.0 for subordinate 3, 1.9 for subordinate 2, 1.9 for subordinate 1, and 1.6 for the self-evaluation. Tukey's test for multiple comparisons indicated that the self-evaluation and the third supervisor evaluation were the only two to differ significantly at a .05 level of significance. These results were clarified by examining the findings for the proportion data which revealed no significant differences between the rating sources. Thus, it appeared that the rating focus differences were due to the fact that the supervisor evaluations were slightly longer than the self-evaluations. This main effect is not particularly relevant to any of the hypotheses in the study.

Table 7 also reveals a main effect for the different categories, $F(12,420)=131.54, p<.0001$. This finding was not unexpected since some categories (e.g., positive evaluation, internal attributions, negative evaluation, dimension of performance) will obviously be used more frequently than other categories (e.g., plan to test evaluation, prediction of negative behavior, expectations for improvement, prediction of positive behavior) regardless of rating type. This finding does not have any particular significance for this study, but the average number of times each of the 21 categories was used during a single rating is presented in Appendix B.

The final result presented in Table 7 is relevant to the focus of this study and involves the interaction between the
Table 7
Complete ANOVA for composite categories for frequency data

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3</td>
<td>43.7</td>
<td>14.6</td>
<td>4.11 (p=.0085)</td>
</tr>
<tr>
<td>Category</td>
<td>12</td>
<td>16015.2</td>
<td>1334.6</td>
<td>131.54 (p&lt;.0001)</td>
</tr>
<tr>
<td>Rater X Category</td>
<td>36</td>
<td>381.4</td>
<td>10.6</td>
<td>3.76 (p&lt;.0001)</td>
</tr>
<tr>
<td>Manager</td>
<td>35</td>
<td>1245.4</td>
<td>35.6</td>
<td></td>
</tr>
<tr>
<td>Rater X Manager</td>
<td>105</td>
<td>372.4</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>Category X manager</td>
<td>420</td>
<td>4261.4</td>
<td>10.1</td>
<td></td>
</tr>
<tr>
<td>Rater X Category X Manager</td>
<td>1260</td>
<td>3546.0</td>
<td>2.8</td>
<td></td>
</tr>
</tbody>
</table>

rating source and the use of the different categories. This interaction effect is significant, $F(36,1260)=3.76$, $p<.0001$, and indicates that at least two of the ratings differ in the use of the coding categories. Specific contrasts will be presented later to determine whether such differential processing corresponds to the specific hypotheses of interest. Such follow-up contrasts would have been conducted even if the overall $F$ test for the interaction had not been significant since such specific predictions do not require a significant overall $F$ value in order to be tested.
In interpreting the above ANOVA results, it is necessary to check the assumptions upon which such an analysis is based. The assumptions of this analysis of variance are that the errors are normally distributed with mean zero and constant variance. An additional assumption due to the presence of within subjects factors involves what is commonly referred to as homogeneity of covariance or compound symmetry. However, Huynh & Feldt (1970) have shown that the correct assumption is actually homogeneity of treatment difference variances. Each of these assumptions should be addressed when conducting ANOVA for inferential purposes.

The normality assumption was examined for the above analysis using various approaches. Residual analysis is the primary diagnostic tool in checking the assumptions. The first check for the assumption of normality was to conduct the Kolmogorov-Smirnov test using the residuals. The resulting p-value was less than .01 which required rejecting the null hypothesis that the errors were normally distributed.

The second check for normality was to construct a normal probability plot of the residuals. A normal probability plot is a graph of the cumulative distribution of the residuals on the ordinate scale so that if the residuals are normally distributed, it will appear as a straight line on the main diagonal. The normal probability plot for the data increased from left to right, but not in a linear fashion throughout the middle of the range.
A third check of the normality assumption was made by plotting a histogram of the residuals. The histogram was symmetric about zero, but it was clearly leptokurtic.

The final check of the normality assumption was to construct a plot of the residuals versus the predicted values. If the model is appropriate and if the assumptions are satisfied, this plot should not reveal any obvious patterns. The plot did show a fairly random scatter about zero.

These diagnostic checks indicated that the residuals did not approximate a normal distribution, but it was also apparent that there were no gross deviations. In appealing to the robustness of the F test and in recognizing that ANOVA is being conducted for general, preliminary purposes in this investigation, the conclusions drawn from the above ANOVA are most likely appropriate.

The next assumption to be examined involved homogeneity of variance. One check of this assumption consisted of an inspection of the plot of the residuals versus the predicted values. Very often the variance of the observations increases as the magnitude of the observations increases. In such cases, the residuals get larger as the predicted values get larger. This problem does characterize some of the residuals in the data, but the major portion of the plot shows constant variance.

A test for homogeneity of variance was also conducted. Bartlett's test is commonly used in this regard, but the test
relies heavily on the assumption of normality, which has been shown to be questionable in this case. As a result, Burr-Foster's Q test, which has been found to be less susceptible to deviations from normality, was conducted. The result \(q^* = .08\) was greater than the critical value for \(q(.001) = .06\) which allowed for rejection of the homogeneity of variance assumption. Again, while the constant variance assumption was not fully satisfied, the departure did not appear to be significant enough to invalidate any of the conclusions drawn.

The final assumption involves equal variances of differences. Methods of checking and accounting for violations of this assumption are less clear cut. Some researchers (Keppel, 1973, pp. 466-467) have actually proposed doing nothing since the positive bias is expected to only be in the order of two or three percent. A more responsible correction has been provided by Geisser & Greenhouse (1958) which involves evaluating the obtained F value against a new critical value that assumes maximal heterogeneity. For the above results, all would have remained significant even when accounting for a positive bias of three percent. All results were also easily significant at a .10 alpha level after applying the Geisser & Greenhouse correction for the case of maximum heterogeneity. As a result, the original interpretations of the ANOVA results for the frequency data appear to hold with regard to the assumptions.
The next step was to perform an analogous 4 X 13 within subjects ANOVA based on the proportional data. The summary table for this analysis is presented in Table 8. The results indicate that there was no main effect for rating source, \(F(3,105)=1.28, p=.29\), as was discussed previously. This is reassuring since we did not expect one rating source to have more of their phrases coded into the categories than the other rating source.

Table 8

Complete ANOVA for composite categories with proportion data

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3</td>
<td>.018</td>
<td>.006</td>
<td>1.28 (p=.2865)</td>
</tr>
<tr>
<td>Category</td>
<td>12</td>
<td>82.0</td>
<td>6.8</td>
<td>230.2 (p&lt;.0001)</td>
</tr>
<tr>
<td>Rater X Category</td>
<td>36</td>
<td>1.31</td>
<td>.036</td>
<td>3.14 (p&lt;.0001)</td>
</tr>
<tr>
<td>Manager</td>
<td>35</td>
<td>.61</td>
<td>.036</td>
<td></td>
</tr>
<tr>
<td>Rater X Manager</td>
<td>105</td>
<td>.498</td>
<td>.005</td>
<td></td>
</tr>
<tr>
<td>Category X Manager</td>
<td>420</td>
<td>12.5</td>
<td>.030</td>
<td></td>
</tr>
<tr>
<td>Rater X Category X Manager</td>
<td>1260</td>
<td>14.6</td>
<td>.012</td>
<td></td>
</tr>
</tbody>
</table>
Table 8 also reveals a main effect for the categories factor, $F(12,420)=230.2, p<.0001$. Again, this result was expected and doesn't have any particular significance for this study. Collapsing across rating type, the average proportion of the protocols coded into each of the 21 categories is presented in Appendix B. Of more interest, however, is the rating source by category interaction which was significant, $F(36,1260)=3.14, p<.0001$. This differential use of the categories by the rating sources was the focus of this investigation.

With regard to the assumptions of this analysis of variance, the Kolmogorov-Smirnov test on the residuals was significant at a .01 level allowing for rejection of the null hypothesis that the errors were normally distributed. The normal probability plot was still "flat" toward the middle of the distribution, but it was an improvement over the plot for the frequency data. The histogram of the residuals was symmetric, but remained leptokurtic. The plot of the residuals versus the predicted values showed a fairly random scatter about zero. In sum, the assumption of normality was questionable, but it is unlikely that the results were affected.

With respect to the constant variance assumption, the plot of the residuals versus the predicted values did reveal a trend in which larger residuals were associated with larger predicted values, but the variance throughout the major portion of the plot appeared constant. The Burr-Foster Q test resulted in a
value of $q^* = .08$ which is slightly larger than the critical value of .06. This resulted in the decision to reject the homogeneity of variance assumption, but the problem was not viewed as serious enough to question the interpretation of the results.

In terms of the homogeneity of treatment difference variances assumption, the results would not have been affected even there was a positive bias in the order of three percent. In addition, the two significant F values would easily have been significant at alpha=.10 using the Geisser & Greenhouse correction.

The above ANOVA procedures were based on 13 categories used in the coding scheme. Some of these categories had direct relevance to the hypotheses of interest while others were purely exploratory and based on the coding of protocols collected during pretesting. As a way of further clarifying the results, it appeared important to know whether the above findings would hold if the categories factor was limited only to those coding categories directly relevant to the hypotheses in the study. In order to test this, similar ANOVA procedures were conducted for frequency and proportion data with a restricted categories factor. The new category factor was limited to the following categories: 1) could improve; 2) consensus, standards and distinctiveness information; 3) consistency/improvement; 4) external attributions and; 5) the composite of internal attributions.
The summary table for the ANOVA applied to the frequency data is presented in Table 9. The results corresponded directly to the earlier analyses. There was no main effect for rating source, $F(3,105)=2.21$, $p=.09$. There was, however, a significant main effect for category, $F(4,140)=202.46$, $p<.0001$, and a significant interaction between rating type (or source) and category, $F(12,420)=7.02$, $p<.0001$.

Table 9

Restricted ANOVA for composite categories for frequency data

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3</td>
<td>21.42</td>
<td>7.14</td>
<td>2.21 (p=.0915)</td>
</tr>
<tr>
<td>Category</td>
<td>4</td>
<td>6892.7</td>
<td>1723.2</td>
<td>202.5 (p&lt;.0001)</td>
</tr>
<tr>
<td>Rater X Category</td>
<td>12</td>
<td>237.3</td>
<td>19.8</td>
<td>7.02 (p&lt;.0001)</td>
</tr>
<tr>
<td>Manager</td>
<td>35</td>
<td>646.5</td>
<td>18.5</td>
<td></td>
</tr>
<tr>
<td>Rater X Manager</td>
<td>105</td>
<td>339.4</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>Category X manager</td>
<td>140</td>
<td>1191.6</td>
<td>8.5</td>
<td></td>
</tr>
<tr>
<td>Rater X Category X Manager</td>
<td>420</td>
<td>1183.6</td>
<td>2.8</td>
<td></td>
</tr>
</tbody>
</table>

The summary table for the ANOVA based on the proportion data is presented in Table 10. Again, the results were consistent
with those above even with the categories restricted only to those related to the hypotheses of interest. There was no main effect for rater, $F(3,105)=.88$, $p=.45$. There was a significant category effect, $F(4,140)=258.7$, $p<.0001$, and a significant interaction, $F(12,420)=6.38$, $p<.0001$.

Table 10

Restricted ANOVA for composite categories for proportion data

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3</td>
<td>.02</td>
<td>.007</td>
<td>.88 ($p=.4530$)</td>
</tr>
<tr>
<td>Category</td>
<td>4</td>
<td>37.4</td>
<td>9.35</td>
<td>258.7 ($p&lt;.0001$)</td>
</tr>
<tr>
<td>Rater X Category</td>
<td>12</td>
<td>.70</td>
<td>.058</td>
<td>6.38 ($p&lt;.0001$)</td>
</tr>
<tr>
<td>Manager</td>
<td>35</td>
<td>.75</td>
<td>.02</td>
<td></td>
</tr>
<tr>
<td>Rater X Manager</td>
<td>105</td>
<td>.72</td>
<td>.007</td>
<td></td>
</tr>
<tr>
<td>Category X Manager</td>
<td>140</td>
<td>5.1</td>
<td>.036</td>
<td></td>
</tr>
<tr>
<td>Rater X Category X Manager</td>
<td>420</td>
<td>3.8</td>
<td>.009</td>
<td></td>
</tr>
</tbody>
</table>

A check on the assumptions for the above analyses revealed results similar to those found previously. The residuals appeared to be symmetric about zero, but were leptokurtic. There was slight heterogeneity of variance in both analyses.
However, these departures from the assumptions and any heterogeneity of treatment difference variances would pose no serious problems due to the highly significant F values obtained in the analyses.

Additional ANOVA procedures were performed for the frequency and the proportion data without collapsing any of the categories. These analyses were conducted to be sure that the composites, which were formed by the researcher, were not entirely responsible for the obtained results. One set of analyses involved two 4 X 21 within subjects designs. These consisted of all of the coding categories considered separately. The other set of analyses involved only those categories directly relevant to the hypotheses of interest, but all such categories remained distinct. Thus, all the internal attributions were considered separately and the consensus, standards and distinctiveness categories remained separate.

The results from these analyses were virtually identical to those presented above. The rater main effect was insignificant while the category effect and the interaction effect between rater and category were highly significant.

Prior to addressing the specific hypotheses, there was one last set of analyses that was conducted using the overall ANOVA procedures. In collecting the data, managers were obtained from two different jobs (data processing and office managers) and from two levels (first and second line). It was considered
important to determine whether there were any differences in the data collected that would correspond to these two characteristics of the subjects. The overall ANOVA provides a simple, yet powerful method of detecting such differences. There were no specific predictions regarding job type or managerial level. The sampling strategy was designed to focus on a fairly homogeneous group of managers to insure interpretable results, but whether differences would actually emerge remained an empirical question.

The decision was made to explore the effect of job and level separately rather than simultaneously for two reasons. First, analyzing the data with job and level factors included would have reduced each cell size to nine. In this field study where results have not been overly robust such an analysis would have lacked sufficient power. And secondly, a four-way interaction was not expected and the conceptual difficulties of attempting to interpret such interactions provided a strong argument against conducting a four-way ANOVA.

The level factor was the first to be examined. Level was introduced as a between subjects factor into the original 4 (rating source) X 13 (categories with composites) completely within subjects design. The summary table for this ANOVA as applied to the frequency data is presented in Table 11. The results indicated that level did not have an effect upon the data. There was no main effect for level, $F(1,34)=.09$, $p=.76$. 
That is, there was no overall difference in the number of coded phrases by level. There was also no interaction of level with category, $F(12,408)=1.52$, $p=.11$, or with rating source, $F(3,102)=.18$, $p=.91$. Most importantly, there was no three-way interaction which indicates that the rater by category interaction, which is the focus of the study, was the same for both levels, $F(36,1224)=.42$, $p=.99$.

These analyses were repeated with the proportion data. They were also conducted by introducing the level factor into the full 4 (rating source) X 21 (category) within subjects design for both proportion and frequency data. The findings were consistent with those presented in Table 11. Thus, the evidence was conclusive in suggesting that the data could be analyzed without regard to level.

The job factor was examined in a similar manner. The summary table for the 2 (job) X 4 (rating source) X 13 (category) one between, two within ANOVA for the frequency data is presented in Table 12. As with level, the job factor did not appear to account for significant variance. There was no overall difference for the number of coded phrases between data processing and office managers, $F(1,34)=0.0$, $p=.98$. The job factor did not interact with category, $F(12,408)=.35$, $p=.98$, or with rating source, $F(3,102)=1.44$, $p=.24$. There was also no three-way interaction between job, rating source and category, $F(36,1224)=1.15$, $p=.25$. A similar set of analyses was per-
Table 11

Complete ANOVA for composites with level for frequency data

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>1</td>
<td>3.4</td>
<td>3.4</td>
<td>.09 (p=.7615)</td>
</tr>
<tr>
<td>Manager in level</td>
<td>34</td>
<td>1242.0</td>
<td>36.5</td>
<td></td>
</tr>
<tr>
<td>Rater</td>
<td>3</td>
<td>43.7</td>
<td>14.6</td>
<td>4.01 (p=.0096)</td>
</tr>
<tr>
<td>Category</td>
<td>12</td>
<td>16015.2</td>
<td>1334.6</td>
<td>133.5 (p&lt;.0001)</td>
</tr>
<tr>
<td>Level X Rater</td>
<td>3</td>
<td>2.0</td>
<td>.67</td>
<td>.18 (p=.9087)</td>
</tr>
<tr>
<td>Level X Category</td>
<td>12</td>
<td>182.5</td>
<td>15.2</td>
<td>1.52 (p=.1135)</td>
</tr>
<tr>
<td>Rater X Category</td>
<td>36</td>
<td>381.4</td>
<td>10.6</td>
<td>3.70 (p=.0001)</td>
</tr>
<tr>
<td>Level X Rater X Category</td>
<td>36</td>
<td>43.0</td>
<td>1.2</td>
<td>.42 (p=.9991)</td>
</tr>
<tr>
<td>Manager X Rater in Level</td>
<td>102</td>
<td>374.0</td>
<td>3.7</td>
<td></td>
</tr>
<tr>
<td>Manager X category in Level</td>
<td>408</td>
<td>4078.9</td>
<td>10.0</td>
<td></td>
</tr>
<tr>
<td>Manager X Rater X Category in Level</td>
<td>1224</td>
<td>3503.0</td>
<td>2.86</td>
<td></td>
</tr>
</tbody>
</table>

formed for the proportion data using the 4 X 13 design and for the frequency and proportion data using the full 4 X 21 design.
All findings were consistent with those presented above. As a result, additional analyses in the study were conducted without differentiating between data processing and office managers.

Table 12

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job</td>
<td>1</td>
<td>.03</td>
<td>.03</td>
<td>0.0 (p=.9758)</td>
</tr>
<tr>
<td>Manager in Job</td>
<td>34</td>
<td>1245.4</td>
<td>36.6</td>
<td></td>
</tr>
<tr>
<td>Rater</td>
<td>3</td>
<td>43.7</td>
<td>14.6</td>
<td>4.16 (p=.0080)</td>
</tr>
<tr>
<td>Category</td>
<td>12</td>
<td>16015.2</td>
<td>1334.6</td>
<td>129.1 (p&lt;.0001)</td>
</tr>
<tr>
<td>Job X Rater</td>
<td>3</td>
<td>15.1</td>
<td>5.0</td>
<td>1.44 (p=.2353)</td>
</tr>
<tr>
<td>Job X Category</td>
<td>12</td>
<td>43.5</td>
<td>3.6</td>
<td>.35 (p=.9787)</td>
</tr>
<tr>
<td>Rater X Category</td>
<td>36</td>
<td>381.4</td>
<td>10.6</td>
<td>3.78 (p=.0001)</td>
</tr>
<tr>
<td>Job X Rater X Category</td>
<td>36</td>
<td>116.0</td>
<td>3.2</td>
<td>1.15 (p=.2513)</td>
</tr>
<tr>
<td>Manager X Rater in Job</td>
<td>102</td>
<td>357.2</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>Manager X category in Job</td>
<td>408</td>
<td>4217.9</td>
<td>10.3</td>
<td></td>
</tr>
<tr>
<td>Manager X Rater X Category in Job</td>
<td>1224</td>
<td>3430.1</td>
<td>2.8</td>
<td></td>
</tr>
</tbody>
</table>
The discussion will now focus on the specific attributional hypotheses of interest. The first of these hypotheses stated that supervisor evaluations would reflect more internal attributions than self-evaluations. This hypothesis was initially evaluated by examining the differential use of the composite of internal attributions by the two rating sources. This was performed by conducting a one-way ANOVA for the rating source factor with the composite of internal attributions as the dependent variable. A single degree of freedom contrast was then conducted comparing the self-evaluation to the three supervisor evaluations. The results of these analyses for the frequency and proportion data are presented in Table 13. Analysis of the frequency data revealed a main effect for rater, $F(3,105)=5.89$, $p=.0009$, and a significant difference between the self- and the average of the supervisor ratings, $F(1,105)=17.18$ ($p<.0001$). The self-evaluations had an average of 6.4 internal attributions while the supervisor evaluations had an average of 9.0 internal attributions.

An examination of the proportion data reveals similar results with a main effect for rater, $F(3,105)=4.22$, $p=.0009$, and an effect for the difference between the two rating sources, $F(1,105)=11.32$, $p=.0011$. Self-evaluations made reference to internal attributions for 5% of the phrases while supervisor evaluations made internal attributions for 10% of the protocol.
Table 13  
ANOVA and contrast for use of internal attributions

I. For Frequency Data:

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3</td>
<td>177.6</td>
<td>59.2</td>
<td>5.89 (p=.0009)</td>
</tr>
<tr>
<td>Self vs. 3</td>
<td>1</td>
<td>172.5</td>
<td>172.5</td>
<td>17.18 (p&lt;.0001)</td>
</tr>
<tr>
<td>Supervisor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manager</td>
<td>35</td>
<td>1466.1</td>
<td>41.9</td>
<td></td>
</tr>
<tr>
<td>Rater X</td>
<td>105</td>
<td>1054.7</td>
<td>10.0</td>
<td></td>
</tr>
<tr>
<td>Manager</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

II. For Proportion Data:

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3</td>
<td>.326</td>
<td>.109</td>
<td>4.22 (p=.0074)</td>
</tr>
<tr>
<td>Self vs. 3</td>
<td>1</td>
<td>.292</td>
<td>.292</td>
<td>11.32 (p=.0011)</td>
</tr>
<tr>
<td>Supervisor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manager</td>
<td>35</td>
<td>4.81</td>
<td>.137</td>
<td></td>
</tr>
<tr>
<td>Rater X</td>
<td>105</td>
<td>.825</td>
<td>.008</td>
<td></td>
</tr>
<tr>
<td>Manager</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A check on the assumptions for the above analyses indicated that there were no serious violations of the assumptions. While the Kolmogorov-Smirnov test rejected normality for both analyses at alpha=.01, the histograms and normal probability plots looked reasonable. The plot of the residuals versus the
predicted values revealed a constant, random scatter about zero. Finally, based on the Geisser & Greenhouse correction, results would have remained significant even under the most extreme case of heterogeneity of difference variances.

A more qualitative, global assessment regarding the differential use of internal attributions by the rating sources was also made. For a given manager, one of the research assistants read the four evaluations provided by that manager and selected the one that he felt placed the most emphasis on internal attributions. Of the 36 protocols, the research assistants selected a supervisor evaluation 33 times. The binomial distribution was used to determine whether $B=33$ out of $n=36$ is significantly greater than $.75$, which one would expect under the null hypothesis. The standardized value of $B=2.3$ ($p=.0107$), which is significantly greater than chance alone would suggest. Thus, it appears that managers did make more internal attributions in the context of evaluating their subordinates than in evaluating themselves.

In an attempt to further explore the differential use of internal attributions, analyses of variance and contrasts were performed for each of the internal attribution categories separately. These follow-up analyses did not prove particularly fruitful. Based on the proportion and frequency data, there were no significant differences between self- and supervisor evaluations in knowledge, experience, effort or interest attri-
butions. Trait attributions were made more frequently, $F(1,105)=5.45$, $p=.02$, in supervisor evaluations, but this difference disappeared when accounting for the fact that supervisor evaluations were longer, $F(1,105)=1.48$, $p=.23$. Only skill attributions (e.g. interpersonal skills) were made more frequently in the supervisor evaluations, $F(1,105)=4.24$ ($p=.04$), and used a greater proportion of the time, $F(1,105)=3.96$ ($p=.04$). However, $p$-values of .04 in these analyses should not be considered highly significant due to the exploratory nature of these contrasts.

The above results regarding internal attributions are somewhat unexpected. While internal attributions are made more often in the context of supervisor evaluations, there doesn't appear to be any specific type of internal attribution that can account for this difference. One explanation for not identifying any specific differences could be the fact that the skill, experience and effort coding categories were detected less reliably. However, the reliability estimates for the trait, knowledge and interest categories were sufficient for making meaningful comparisons between the means for the rating sources.

The final set of analyses involving internal attributions involved an exploration of the effect of the ratee's performance level and the supervisor's job/appraisal experience on the number of internal attributions made. A number of performance
level measures were collected for each manager's three subordinates. There were three overall measures that were examined in relationship to the formation of internal attributions. These measures consisted of: 1) an overall evaluation on the Ohio State performance evaluation form; 2) an overall rating collected by the researcher based on a five-point scale (1=poor to 5=excellent) and; 3) an overall rank ordering of the performance of all the subordinates reporting to a given manager.

The proportion and frequency of internal attributions were correlated (using Pearson, Spearman and Kendall coefficients) with the three measures of performance for all three subordinates. The resulting coefficients fluctuated about zero without any of the p-values approaching a .05 level of significance. There appeared to be no relationship between the formation of internal attributions and the performance level of the ratee.

A similar set of correlations were computed between the formation of internal attributions and four potentially relevant experience variables. These measures of experience included: 1) the number of months the manager was on his/her present job; 2) the number of months the manager held the subordinate's job; 3) the number of months the manager supervised the subordinate and; 4) the number of performance appraisals the manager had completed in his/her career. Again, with an alpha level of .05, there were no significant correlations between the fre-
quency or proportion of internal attributions and any of the experience variables.

ANALYSES OF EXTERNAL ATTRIBUTIONS

The next specific hypothesis stated that self-evaluations would include more external attributions than supervisor evaluations. This hypothesis was examined by conducting a one-way ANOVA on the frequency and proportion of external attributions made by the different rating sources. A summary table of these analyses and the contrasts between self- and supervisor evaluations are presented in Table 14.

The results for the frequency data indicate a main effect for rater, $F(3,105)=8.45$, $p<.0001$, with a significant difference between the self- and supervisor evaluations, $F(1,105)=24.3$, $p<.0001$. The self-evaluations had an average of 1.86 external attributions while the supervisor evaluations averaged .49 external attributions.

Analysis of the proportion data revealed consistent results with a general effect for rating source, $F(3,105)=12.87$, $p<.0001$, and a significant contrast effect between the two types of ratings, $F(1,105)=38.32$, $p<.0001$. Twelve percent of the phrases in the self-evaluations involved external attributions while only two percent of the phrases in the supervisor evaluations were coded as external attributions.
Table 14
ANOVA and contrast for use of external attributions

I. For Frequency Data:

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3</td>
<td>53</td>
<td>17.7</td>
<td>8.45 (p&lt;.0001)</td>
</tr>
<tr>
<td>Self vs. 3</td>
<td>1</td>
<td>50.7</td>
<td>50.7</td>
<td>24.3 (p&lt;.0001)</td>
</tr>
<tr>
<td>Supervisor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manager</td>
<td>35</td>
<td>129.5</td>
<td>3.7</td>
<td></td>
</tr>
<tr>
<td>Rater X</td>
<td>105</td>
<td>219.5</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>Manager</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

II. For Proportion Data:

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3</td>
<td>.267</td>
<td>.089</td>
<td>12.87 (p&lt;.0001)</td>
</tr>
<tr>
<td>Self vs. 3</td>
<td>1</td>
<td>.265</td>
<td>.265</td>
<td>38.32 (p&lt;.0001)</td>
</tr>
<tr>
<td>Supervisor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manager</td>
<td>35</td>
<td>.275</td>
<td>.008</td>
<td></td>
</tr>
<tr>
<td>Rater X</td>
<td>105</td>
<td>.726</td>
<td>.007</td>
<td></td>
</tr>
<tr>
<td>Manager</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A diagnostic check on the satisfaction of the assumptions for the above analyses revealed no general problems, but there were two outlying residuals. The Kolmogorov-Smirnov test was significant at an alpha=.01 level, indicating that it was unlikely that the data were drawn from a normal distribution.
The normal probability plots and histograms reflected the outliers, but otherwise revealed no serious problems. The plot of the residuals versus the predicted values showed a fairly random scatter about zero. The residuals did increase as the magnitude of the predicted values increased, but with the size of the F values obtained in these analyses there appeared to be no problem. The results were also such that any degree of heterogeneity of treatment difference variances would not have affected the interpretations.

The overall, qualitative assessment by the research assistants was conducted for the use of external attributions as well. The total n in this case was 31 since the students were not required to select one of the four evaluations if they could not find differences between the protocols that would enable them to make a decision. The results indicated that out of n=31 choices made, 21 self-evaluations were selected as placing the greatest emphasis on external attributions. The null hypothesis in this case is p=.25 since there was only one self-evaluation within each set of four evaluations. The standardized value was 5.52, p<.0001, which indicates that significantly more than one-quarter of the self-evaluations were selected as placing the most emphasis on external attributions.

There is an additional, somewhat subtle, issue that must be addressed when comparing self-evaluations to supervisor evaluations. This comparison involves an inherent confound between
rating source and job level. That is, the self-evaluations were made for first- and second-line managerial jobs while the supervisor evaluations were made for first-line managerial jobs and nonmanagerial jobs. It is possible that more external attributions are made in higher level jobs. In order to be sure the differences attributed to rating source were not confounded by level, an independent t-test was performed between self-evaluations for first-line managers and the average of the three supervisor evaluations made by second-line managers. This type of analysis serves to analyze rating source while holding the job level constant at first-line manager. For the frequency data, the mean number of external attributions for self-evaluations was 2.1 and for supervisor evaluations the mean was .67. The independent t-test between these means was significant, \( t(23.6) = 2.24, p = .03 \), after adjusting for unequal variances.

The proportion data produced similar results, \( t(18.9) = 2.96, p = .008 \). The self-evaluations consisted of external attributions for 15% of the appraisal while supervisors made external attributions only 3% of the time when evaluating their subordinates. These results provide compelling evidence for being able to attribute the findings in this study to differences in rating source as opposed to job level.

The final set of analyses made with respect to external attributions involved an examination of the correlation between
external attributions and the performance level and relevant experience variables. Investigation of the role of ratee performance level was restricted to supervisor evaluations since this information was not obtained for use in the self-evaluations. In terms of the performance level variables, there was no significant relationship between the number or proportion of external attributions made and ratings or rankings of job performance.

For the experience variables in the context of the supervisor evaluations, the frequency and proportion of external attributions were significantly negatively correlated with the number of months the supervisor worked at his/her present job. That is, the longer the manager occupied his/her position the less apt he/she was to make external attributions for his/her subordinate's performance. None of the other experience variables (months on the subordinate's job, months supervised the subordinate and number of times a performance evaluation had been completed) were related to the formation of external attributions. It is important to point out that the lack of a relationship between the formation of external attributions and time on the subordinate's job contradicts findings by Mitchell & Kalb (1982) who did find a positive relationship between the two variables.

For the self-evaluations, there was no relationship between the frequency or proportion of external attributions made and
the manager's tenure on his present job or the number of times a manager completed a formal performance appraisal in his/her career.

**ANALYSES OF CONSISTENCY/IMPROVEMENT INFORMATION**

An additional hypothesis stated that self-evaluations would reflect more references to an improvement or gain in performance from previous levels of functioning than would supervisor evaluations. A summary of the one-way ANOVA for the use of consistency/improvement information by the different rating sources is presented in Table 15. While the reliabilities for the consistency/improvement category were only in the .39 to .60 range, this did not appear low enough to necessarily preclude detecting differences between the two rating sources.

The results of the ANOVA for the frequency data revealed no overall differences for rating source, $F(3,105)=1.31$, $p=.27$. The specific contrast between the two sources was also insignificant, $F(1,105)<1$, $p=.94$. Analyses of the proportion data produced similar results with no overall rater main effect, $F(3,105)=1.85$, $p=.14$, and an insignificant contrast effect between the two rating types, $F(1,105)=1.04$, $p=.31$. In sum, there appeared to be absolutely no difference in how the two rating sources used consistency/improvement information in their evaluations. While the various assumptions were not fully satisfied, there were no problems severe enough to invalidate the above results.
Table 15

ANOVA and contrast for the use of consistency/improvement

I. For Frequency Data:

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3</td>
<td>1.5</td>
<td>.5</td>
<td>1.31 (p=.27)</td>
</tr>
<tr>
<td>Self vs. 3 Supervisor</td>
<td>1</td>
<td>.002</td>
<td>.002</td>
<td>.01 (p=.94)</td>
</tr>
<tr>
<td>Manager</td>
<td>35</td>
<td>25.7</td>
<td>.734</td>
<td></td>
</tr>
<tr>
<td>Rater X Manager</td>
<td>105</td>
<td>39.3</td>
<td>.37</td>
<td></td>
</tr>
</tbody>
</table>

II. For Proportion Data:

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3</td>
<td>.005</td>
<td>.002</td>
<td>1.85 (p=.1425)</td>
</tr>
<tr>
<td>Self vs. 3 Supervisor</td>
<td>1</td>
<td>.001</td>
<td>.001</td>
<td>1.04 (p=.3109)</td>
</tr>
<tr>
<td>Manager</td>
<td>35</td>
<td>.037</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>Rater X Manager</td>
<td>105</td>
<td>.095</td>
<td>.001</td>
<td></td>
</tr>
</tbody>
</table>

The overall, qualitative assessment was also conducted for this hypothesis. The research assistants read the four evaluations for each manager and selected the one that relied most heavily upon consistency/improvement information. The results of this analysis agree with those presented above. Of a total
n=24 selections, only 6 of these corresponded to self-evaluations. Six out of 24 or .25 is precisely what one would expect under the null hypothesis that there was no difference between the rating sources in the use of consistency/improvement information.

An examination of the use of consistency/improvement information was also conducted by controlling for job level. References to consistency/improvement information for self-evaluations by first-line managers were compared to those made during supervisor evaluations by second-line managers. In each evaluation, the job level of the ratee involves a first-line manager. No significant differences were detected using the frequency data, $t(34)=.86$, $p=.39$, or the proportion data, $t(19.4)=1.69$, $p=.11$. Thus, there were no differences in the use of consistency/improvement information even while holding job level constant.

An attempt was made to explore the relationship between the use of consistency/improvement information and the measures of experience and performance levels. These analyses were conducted using Pearson, Spearman and Kendall correlation coefficients for each of the four evaluations. No significant relationships were discovered for the experience or performance level variables.
The final hypothesis regarding differential patterns of processing by the two rating sources involved consensus, standards and distinctiveness information. This hypothesis predicted that supervisor evaluations would reflect more references to distinctiveness, consensus, and other comparative information than self-evaluations. A one-way ANOVA was performed with the four raters (self and three supervisor) serving as the four levels of the factor. The analysis of variance summary table for a composite of the three types of information is presented in Table 16.

With respect to the frequency data, the mean was 1.5 phrases for the supervisor evaluations and .61 for the self-evaluations. The single d.f. contrast between the two rating sources was highly significant, $F(1,105)=13.03$, $p=.0005$. The average percentage of references made to consensus, standards or distinctiveness information was 10% for the supervisor evaluations and 5% for the self-evaluations. The contrast between the means for the proportional data was also significant, $F(1,105)=8.98$, $p=.0034$. Thus, the hypothesis regarding consensus, standards and distinctiveness information was clearly confirmed with significantly more references to this information when a supervisor evaluates a subordinate.
Table 16
ANOVA and contrast for use of consensus, standards composite

I. For Frequency Data:

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3</td>
<td>22.1</td>
<td>7.4</td>
<td>4.49 (p=.0053)</td>
</tr>
<tr>
<td>Self vs. 3 Supervisor</td>
<td>1</td>
<td>21.3</td>
<td>21.3</td>
<td>13.03 (p=.0005)</td>
</tr>
<tr>
<td>Manager</td>
<td>35</td>
<td>190.9</td>
<td>5.45</td>
<td></td>
</tr>
<tr>
<td>Rater X Manager</td>
<td>105</td>
<td>171.9</td>
<td>1.64</td>
<td></td>
</tr>
</tbody>
</table>

II. For Proportion Data:

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3</td>
<td>.094</td>
<td>.031</td>
<td>4.0 (p=.0097)</td>
</tr>
<tr>
<td>Self vs. 3 Supervisor</td>
<td>1</td>
<td>.071</td>
<td>.071</td>
<td>8.94 (p=.0034)</td>
</tr>
<tr>
<td>Manager</td>
<td>35</td>
<td>.542</td>
<td>.015</td>
<td></td>
</tr>
<tr>
<td>Rater X Manager</td>
<td>105</td>
<td>.825</td>
<td>.008</td>
<td></td>
</tr>
</tbody>
</table>

As a post hoc procedure, each component of the above composite was analyzed separately. Recall, however, that the reliability estimates for the consensus, standards and distinctiveness categories were in the range of .90, .45, and 0, respectively. As a result, any findings regarding distinctiveness
information would be uninterpretable and will not be presented.
The reliability estimate for the standards category was viewed
as sufficient since it should probably be viewed as the lower
bound for the actual estimate.

Consensus information was used more frequently
\( (F(1,105)=5.22, \ p=.0244) \) in the supervisor evaluations
\( (\text{mean}=.63) \) than in the self-evaluations \( (\text{mean}=.25) \). It was
also used a greater percentage of the time, \( F(1,105)=4.62, \ p=.034 \). Consensus information was used 4% of the time in
supervisor evaluations and only 1% of the time in the self-
evaluations.

The same general findings appeared for the use of other
standards. Standards of performance were referred to more fre­
quently \( (F(1,105)=14.74, \ p=.0002) \) during supervisor evaluations
\( (\text{mean}=.5) \) than during self-evaluations \( (\text{mean}=.06) \). Standards
of performance were used a higher percentage of the time
\( (F(1,105)=13.57, \ p=.0004) \) during supervisor evaluations (4%) than during self-evaluations (0.2%).

A check on the assumptions required for interpreting ANOVA
revealed some minor violations, but no significant problems.
The F values for the contrasts presented above were significant
enough that slight deviations from the assumptions would have
been unlikely to have affected the results.

The qualitative judgment based on the overall assessment
involved two separate decisions. The first decision involved
determining which of the four protocols relied most heavily on consensus information or other comparative standards. A second, separate decision was then made regarding the use of distinctiveness information. In each case, the null hypothesis was $p = .75$ since there were three out of four supervisor evaluations and selection of a supervisor evaluation was defined as a "success."

For the use of consensus/standards information, out of $n = 35$, 33 supervisor evaluations were selected as placing a greater emphasis on consensus information or other standards. The standardized value was $2.64$ ($p = .0041$), which means the proportion selected was significantly greater than $p = .75$. The results for distinctiveness information are probably still highly suspect in this context as well since those individuals making the decisions for the coding were the same as those making the overall assessments. However, out of $n = 29$ decisions, 24 supervisor evaluations were selected. With the null hypothesis of $p = .75$, the standardized value of $B$ was .97 ($p = .16$), which is approximately what one would expect under the null hypothesis. Whether this result is due to the low reliability or to a similar use of distinctiveness information by the two rating sources cannot be answered with the data obtained.

Again, to be sure the above differences were due to rating source, $t$-tests were performed between self-evaluations for
first-line managers and supervisor evaluations for second-line managers. Based on the composite of consensus, standards and distinctiveness information, supervisor evaluations referenced this information more frequently, $t(24.7) = -3.32, p = .0028$ and with greater proportion, $t(34) = -3.47, p = .0014$, than did those making self-evaluations.

Finally, the relationships between the different types of information discussed above and the performance level and experience variables were examined. No consistent relationships were discovered using the overall composite or consensus and standards information considered separately.

RELATIONSHIP BETWEEN ATTRIBUTION FORMATION AND EVALUATIONS

The last two hypotheses referred to the sequence or order in which thought processes related to attribution formation are thought to occur. It is this type of analysis that takes advantage of verbal protocol analysis as a method of data collection. These hypotheses were predicted to hold for both rating sources. However, the data were analyzed separately in order to be sure the predicted patterns occurred for both rating sources.

The first of these hypotheses stated that internal attributions would be followed by more evaluations (positive or negative) than external attributions. The contingency nature of this hypothesis suggested arranging a $2 \times 2$ frequency table and
using a chi-square test for association. That is, the vertical dimension of the table represented whether a given attribution was internal or external while the horizontal dimension represented whether or not an evaluation was associated with the attribution. If an evaluation occurred in the phrase immediately following an attribution or in the same phrase it was considered associated with the attribution. Thus, four frequency counts were obtained for a given table by examining the data for all 36 managers. The frequency counts and the chi-square values for the self-evaluations and the combined supervisor evaluations are presented in Table 17.

One should refrain from attempting to interpret the findings by simply examining the frequency counts in each of the four cells due to extreme differences in the marginal totals. The chi-square test statistic compares the obtained frequencies to expected frequencies. The expected frequency for a given cell is the product of the two corresponding marginal totals divided by the total frequency.

The chi-square values for the self- (97.36, p<.001) and supervisor (70.93, p<.001) evaluations were clearly significant. These results indicate that, regardless of rating source, there is a clear relationship between the type of attribution and the formation of an evaluation. As hypothesized, internal attributions were more likely to result in an evaluation than were external attributions.
Table 17
Relationship between attributions and evaluations

I. For Self-Evaluations:

<table>
<thead>
<tr>
<th></th>
<th>Not Associated w/ Evaluation</th>
<th>Associated w/ Evaluation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal Attribution</td>
<td>8 (34.3)</td>
<td>176 (149.7)</td>
<td>184</td>
</tr>
<tr>
<td>External Attribution</td>
<td>38 (11.7)</td>
<td>25 (51.26)</td>
<td>63</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
<td>201</td>
<td>247</td>
</tr>
</tbody>
</table>

Expected values in parentheses.
Chi-square (1)=97.36, p<.001.

II. For the Combined Supervisor Evaluations:

<table>
<thead>
<tr>
<th></th>
<th>Not Associated w/ Evaluation</th>
<th>Associated w/ Evaluation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal Attribution</td>
<td>18 (49.6)</td>
<td>848 (816.4)</td>
<td>866</td>
</tr>
<tr>
<td>External Attribution</td>
<td>35 (3.38)</td>
<td>24 (55.6)</td>
<td>59</td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>872</td>
<td>925</td>
</tr>
</tbody>
</table>

Expected values in parentheses.
Chi-square (1)=70.93, p<.001.
As an illustration of this type of processing, consider the following excerpt from the first manager's evaluation of one of her subordinates (see Appendix A for the entire transcript): "Jean is probably the most problem person I have. Not because she does not do her work or know her work, but because she resents being asked to do anything outside her own job description and that is something that I don't like for somebody to say to me...." It appears that the manager is considering the subordinate's personal or internal attributes (i.e. "resents being asked") and following this with an evaluation (i.e. "don't like").

An example of an external attribution and a reluctance to form an evaluation is provided by the following excerpt from one of the second manager's protocols: "Considering all the staffing problems she has, I think they produce a reasonable amount of work. I won't say that it's outstanding, but I also don't believe that's her fault. Given a full staff I would think that her productivity would be a lot higher."

RELATIONSHIP BETWEEN ATTRIBUTIONAL INFORMATION AND ATTRIBUTIONS

The final hypothesis stated that attributional information was more likely to precede the attributions it was associated with than it was to follow them. This hypothesis was not tested quite this literally. The hypothesis was designed to address the role of attributional information (consensus, con-
sistency and distinctiveness) in the formation of attributions. In its most general form, the prediction was that attributional information would be used in the formation of attributions and therefore would precede the attributions it was associated with. Limiting the analysis to attributions preceded and/or followed by attributional information would not address the broader issue. Such an analysis would not take into consideration the formation of attributions in the absence of any attributional information. That is, the general relationship between attributional information and attributions would not have been examined.

As a result, the analyses corresponding to this hypothesis were designed to examine the overall relationship between attributions and the attributional information thought to be useful in the formation of attributions. Again, the chi-square test for association was felt to be particularly appropriate for evaluating this relationship. A 2 X 2 frequency table was arranged with one dimension representing the use of attributional information and the other indicating whether an attribution was made. Every pair of phrases across all managers was examined in generating the frequency counts for the following four cells of the table: 1) use of attributional information (consensus or other standards, distinctiveness and consistency) in the first phrase of a pair along with an attribution in that phrase or in the second phrase of the pair; 2) use of attribu-
tional information without an accompanying attribution; 3) no attributional information in the first phrase, but evidence of an attribution in the second phrase and; 4) no attributional information in the first phrase and no coded attributions in the second phrase.

These frequency data are presented in Table 18 for the self-evaluations and for the combined supervisor evaluations. The chi-square values for the self- (11.89, p<.001) and supervisor (52.73, p<.001) evaluations were significant, but the deviations from the expected values were not as extreme as those obtained for the previous hypothesis. It is clear that there is a relationship between the use of attributional information and the forming of attributions, but post-hoc analyses of the frequency data appeared to be warranted.

It was apparent in computing the chi-square value that the two cells which did not involve the use of attributional information (i.e. No attributional information--Associated with an attribution, No attributional information--Not associated with an attribution) contributed very little to the test statistic. That is, the deviations of the observed values from the expected values for these two cells were not particularly large. While the deviations were in the predicted direction for these two cells, the magnitude of these two deviations contributed very little to the overall test statistic.
Table 18

Relationship between attributional information and attributions

I. For Self-Evaluations:

<table>
<thead>
<tr>
<th>Use Attributional Information</th>
<th>Associated w/ Attribution</th>
<th>Not Associated w/ Attribution</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Associated</td>
<td>26</td>
<td>10</td>
<td>36</td>
</tr>
<tr>
<td>(18.16)</td>
<td>(17.84)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did Not Use</td>
<td>198</td>
<td>210</td>
<td>408</td>
</tr>
<tr>
<td>(205.84)</td>
<td>(202.16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>224</td>
<td>220</td>
<td>444</td>
</tr>
</tbody>
</table>

Expected values in parentheses.
Chi-square (1)=11.89, p<.001.

II. For the Combined Supervisor Evaluations:

<table>
<thead>
<tr>
<th>Use Attributional Information</th>
<th>Associated w/ Attribution</th>
<th>Not Associated w/ Attribution</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Associated</td>
<td>153</td>
<td>37</td>
<td>190</td>
</tr>
<tr>
<td>(106.55)</td>
<td>(83.45)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did Not Use</td>
<td>696</td>
<td>628</td>
<td>1324</td>
</tr>
<tr>
<td>(742.45)</td>
<td>(581.55)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>849</td>
<td>666</td>
<td>1514</td>
</tr>
</tbody>
</table>

Expected values in parentheses.
Chi-square (1)=52.73, p<.001.
As a result, it was necessary to decompose the 2 X 2 frequency table in order to better understand the relationship between attributional information and attributions. While an entire literature exists on the appropriate partitioning of chi-square tables into orthogonal components, that research is not relevant to this situation since this analysis only involved one degree of freedom. Thus, the only support for the post-hoc analyses was the researcher's desire to better understand the data. The only particular concern involved the probability of type I error, but this was not of major importance since only two follow-up tests were conducted for each table and the comparison that was expected to be insignificant was in fact insignificant.

The post-hoc analyses involved conducting two chi-square goodness-of-fit tests for each table (the table with the self-evaluation data and the table for the supervisor evaluation data). The first test addressed the two cells corresponding to the use of attributional information. This test evaluated whether the frequency of phrases making attributions and phrases not making attributions was the same when the preceding phrase involved a reference to attributional information. The results of this chi-square test for the self- and supervisor evaluations are presented in Table 19.
Table 19
Post-hoc chi-square for attributional information

I. For Self-Evaluations:

<table>
<thead>
<tr>
<th>Use of Attributional Information</th>
<th>Associated w/ Attribution</th>
<th>Not Associated w/ Attribution</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>26</td>
<td>10 (18)</td>
<td>36</td>
</tr>
</tbody>
</table>

Expected values in parentheses.
Chi-square (1)=7.12, p<.01.

II. For Supervisor Evaluations:

<table>
<thead>
<tr>
<th>Use of Attributional Information</th>
<th>Associated w/ Attribution</th>
<th>Not Associated w/ Attribution</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>153 (95)</td>
<td>37 (95)</td>
<td>190</td>
</tr>
</tbody>
</table>

Expected values in parentheses.
Chi-square (1)=70.82, p<.001.

The chi-square values for the self- (7.12, p<.01) and supervisor (70.82, p<.001) evaluations were significant at conventional alpha levels. These findings indicate that, regardless of rating source, when attributional information is considered it is much more likely that the rater will form an attribution than not form an attribution.
The second test involved the remaining two cells of the table and evaluated whether the frequency in the formation of attributions was dependent upon the preceding phrase not referencing attributional information. Thus, the two cells of the table represent the frequency of phrases referencing attributions and the frequency of phrases not referencing attributions when the preceding phrase did not indicate the use of any type of attributional information. The frequencies and the obtained chi-square statistics are presented in Table 20 for the self- and supervisor evaluations.

Although the results were in the predicted direction, the chi-square values for the self- (.36, p>.5) and supervisor (3.5, p<.10) evaluations were not significant for this portion of the table. These results indicated that a phrase was just as likely to reference an attribution as it was to not reference an attribution even when the preceding phrase did not reference attributional information. Attributions were still likely to occur without an explicit reference to attributional information. Combined with the previous post-hoc analyses, it appeared that when attributional information was referenced the formation of an attribution became a likely occurrence. However, the formation of attributions clearly did not depend upon an explicit reference to attributional information. Many attributions occurred (with nearly a .5 probability per phrase) without accompanying attributional information.
Table 20

Post-hoc chi-square without attributional information

I. For Self-Evaluations:

<table>
<thead>
<tr>
<th></th>
<th>Associated w/ Attribution</th>
<th>Not Associated w/ Attribution</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did Not Use Attributonal Information</td>
<td>198 (204)</td>
<td>210 (204)</td>
<td>408</td>
</tr>
</tbody>
</table>

Expected values in parentheses.
Chi-square (1)=.36, p>.5.

II. For the Combined Supervisor Evaluations:

<table>
<thead>
<tr>
<th></th>
<th>Associated w/ Attribution</th>
<th>Not Associated w/ Attribution</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did Not Use Attributonal Information</td>
<td>696 (662)</td>
<td>628 (662)</td>
<td>1324</td>
</tr>
</tbody>
</table>

Expected values in parentheses.
Chi-square (1)=3.5, p<.10.

It is difficult to provide a clear interpretation of these results. However, it is possible that attributional information is often used in the formation of attributions when behavior is initially observed. But, since the appraisal process involves a recall task, the rater has already done the diagnos-
tic work and, for the sake of cognitive economy, can simply recall the final attribution.

An example of a manager using attributional information to form an attribution was actually provided in the previous section. In this example, the manager preceded an internal attribution with the statement: "Jean is probably the most problem person I have." This type of comparison clearly makes use of consensus information. Examples can also be found where attributions are not preceded by attributional information. To illustrate, the following set of statements were selected from the second manager's protocols: "He produces a lot of work. He has a lot of job knowledge. He knows what he's doing." These statements demonstrate the manager's ability to form a knowledge attribution without an explicit reference to corresponding attributional information.

This completes the results for the primary hypotheses of interest. A summary of the results for all six primary hypotheses is presented in Table 21. The next section will focus on the results from the exploratory analyses.
1. Hypothesis: Supervisor evaluations will reflect more internal attributions than self-evaluations.

Results: Clearly supported based on coded data and global assessments.

2. Hypothesis: Self-evaluations will reflect more external attributions than supervisor evaluations.

Results: Clearly supported based on coded data and global assessments.

3. Hypothesis: Self-evaluations will reflect more references to an improvement or gain in performance from previous levels of functioning than will supervisor evaluations.

Results: No support using coded data or global assessments.

4. Hypothesis: Supervisor evaluations will reflect more references to distinctiveness, consensus and other comparative information than self-evaluations.

Results: Overall hypothesis was confirmed using coded data. Hypothesis restricted to consensus and standards information was confirmed based on coded data and global assessments. Coded data and overall assessments revealed no support for the differential use of distinctiveness information.

5. Hypothesis: Internal attributions will be followed by more evaluations than external attributions.

Results: Clearly supported for self- and supervisor evaluations.
Table 21 (continued),

6. Hypothesis: Attributional information will be used in the formation of attributions.

Results: Generally supported. When attributional information was referenced, an attribution was likely to be formed. However, many attributions were made without an explicit reference to attributional information.

EXPLORATORY ANALYSES

In addition to examining the primary hypotheses of interest, the study also included exploratory analyses related to the empirically based framework of performance appraisal developed during pretesting. The findings are presented primarily in terms of simple descriptive statistics, but inferential statistics are considered for those aspects of the model which are consistent with the primary hypotheses discussed earlier.

While a great deal of additional exploration was conducted with the data for the formal study, the framework for the investigation was provided by the sequence of processing described in the preliminary exploratory model. This model was developed using data collected during the pretesting phases of this investigation.

The first component of the preliminary model for supervisor and self-evaluations involved the rater's consideration of the ratee's job responsibilities. This prediction was not con-
firmed for the supervisor or self-evaluations. For the first and second supervisor evaluations and the self-evaluation, only 5 out of the 36 protocols included a consideration of the ratee's job responsibilities within the first 6 phrases of the protocol. The third supervisor evaluations had only four protocols that began in this manner. Thus, very few managers made an explicit reference to job responsibilities and this pattern was consistent for both supervisor and self-evaluations.

Although it is disappointing when one's hypotheses are not confirmed, there appears to be an explanation for the obtained findings which will be presented in the Discussion Section. An additional question involved the impact of this initial modification in processing on the remainder of the protocol. While an attempt was made to address this issue, nothing of importance could be discovered due to insufficient data and to the fact that many factors varied between the pretesting and formal phases of data collection. However, one possibility is mentioned later in this section.

The above findings suggested altering the preliminary exploratory model by removing "consideration of job responsibilities" from the model for supervisor and self-evaluations. The revised tentative model based on the primary data set is presented in Figures 7 and 8 and includes the processing discovered for the supervisor and self-evaluations, respectively.
EVALUATION OF SUBORDINATE

Consideration of ratee's experience/time on job

Retrieve current overall or dimensional evaluation

Initial positive evaluation

Retrieve primarily positive information followed by negative information

Consensus info or other comparative standards

Negative behavioral episodes

External attribution

Initial negative evaluation

Retrieve primarily negative information followed by positive information

Negative behavioral episodes

Consensus info or comparative standards

External attribution

Figure 7: Final empirical framework for supervisor ratings
Figure 7 (continued),

Internal attribution

Positive evaluation

Internal attribution

Negative evaluation

EVALUATION OF SELF

Retrieve current overall or dimensional evaluation (Primarily positive)

Use significantly more dimensions of performance

More "could use improvement" evaluations

External attribution

Positive, less negative evaluations

Figure 8: Final empirical framework for self-ratings

In terms of the supervisor evaluations, the protocols often began with the manager considering the ratee's experience or
tenure on his/her present job. Although this finding was not predicted, it is a particularly interesting discovery since it implies that managers do consider some type of standard against which they can evaluate employees. A consideration of job experience was made within the first 6 phrases for 17 out of 36 (17/36) first supervisor evaluations, 10/36 second supervisor evaluations and 15/36 third supervisor evaluations. In contrast, self-evaluations included similar references to experience or time on the job for only 3 of the 36 protocols. As a result, consideration of experience was represented in the model for supervisor evaluations, but was excluded from the description of self-evaluations.

The next component for the supervisor evaluations involved an overall or dimensional evaluation. Of the 36 first supervisor evaluations, 34 included evaluations within the first 3 phrases. Out of the 17 protocols that also referenced the subordinate’s experience or tenure, 10 of these phrases preceded the evaluation. From the set of second supervisor evaluations, 35 included an evaluation within the first 3 phrases. Of the 10 protocols that involved a reference to the ratee’s experience on the job, 8 of the 10 phrases preceded the evaluation. For the third supervisor evaluations, 31 of the protocols began with an evaluation with 10 of the 15 references to ratee experience preceding the evaluations.
An excerpt from the protocols may be useful in clarifying the nature of the first two components for the supervisor evaluations. The first supervisor evaluation provided by the first manager began as follows: "She has been here the longest. Rosemary, very intelligent. She has gone far beyond what you would expect from a person with her schooling because she dropped out of high school." It is clear that the manager made an initial reference to the subordinate's tenure and immediately followed that reference with evaluative statements. As another illustration, the third manager's evaluation of the first subordinate began as follows: "Alan is a very hard working individual. He's very ambitious. He's relatively new in terms of what we're doing over here." In this case, an early reference to the subordinate's tenure was made, but it followed initial evaluative statements.

Thus, there was evidence that evaluations consistently appeared during the initial phase of the protocols and that the evaluations were generally preceded by references to experience when these statements occurred. These findings regarding initial evaluations are not consistent with the existing models of performance appraisal. It may be that such models were designed to characterize evaluations made in a different context. However, the results of this study indicate that managers are capable of providing evaluations for research purposes without a great deal of information processing.
The self-evaluations provided similar results with 34 of the 36 protocols including an evaluation within the first 3 phrases. Evaluation is presented as the first component of the self-evaluation model since very few references to experience were made during these evaluations.

The remaining sequence of processing for the supervisor evaluations was generally contingent upon whether the initial evaluation for the subordinate was positive or negative. The favorability of the initial evaluation was generally consistent with the supervisor's final overall rating of the subordinate's performance. Of the 108 (3 x 36) total supervisor evaluations, 90 consisted of positive initial evaluations. Only three of these 90 subordinates received follow-up overall ratings as low as "Fair" on a five-point scale (Excellent, Very Good, Good, Fair, and Poor). None of these 90 subordinates received "Poor" ratings. This implies that a few managers may have modified the favorability of their statements throughout the protocol, but that for most managers the initial evaluation was consistent with their overall perception of the subordinate's performance.

For the remaining 18 subordinates who received initial negative evaluations, only one received an "Excellent" final overall rating and four received "Very Good" ratings. There were only two "Poor" overall evaluations and the initial evaluation for each of these subordinates was negative in the protocols.
Again, these results imply that the favorability of some protocols may have changed, but that generally the first evaluation was consistent with the manager's view of the ratee's job performance.

The above results indicate that the supervisor's initial evaluation is predictive of his/her general perception of the subordinate. As indicated in Figure 7, the type of processing discovered is contingent upon whether the initial evaluation is positive or negative. One general difference between the two types of evaluations involved the proportion of additional positive versus negative evaluations.

Following either type of initial evaluation (positive or negative), it was apparent that additional evaluations could be either positive or negative. However, as expected, if the rater's overall impression of the ratee was positive there was less apt to be follow-up negative evaluations. Likewise, if the rater had a negative impression of the ratee, follow-up positive evaluations were less likely than negative evaluations. The correlations between positive and negative phrases for the different ratings are presented in Table 22. Since the protocols varied in length, the most meaningful correlations in Table 22 involve those presented in terms of proportion data. Consistent with the above discussion, it is clear that the correlation between positive and negative evaluations is negative. As the proportion of positive evaluations increased, the proportion of negative evaluations decreased.
Table 22

Correlations of positive with negative evaluations

<table>
<thead>
<tr>
<th></th>
<th>Frequency Data:</th>
<th>Proportion Data:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First Supervisor</td>
<td>First Evaluation</td>
</tr>
<tr>
<td></td>
<td>-.28(p=.09)</td>
<td>-.80(p&lt;.0001)</td>
</tr>
<tr>
<td></td>
<td>-.41(p=.01)</td>
<td>-.81(p&lt;.0001)</td>
</tr>
<tr>
<td></td>
<td>-.30(p=.01)</td>
<td>-.63(p&lt;.0001)</td>
</tr>
<tr>
<td></td>
<td>Self-Evaluation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.26(p=.11)</td>
<td>-.39(p=.0200)</td>
</tr>
<tr>
<td></td>
<td>.13(p=.42)</td>
<td>-.60(p&lt;.0001)</td>
</tr>
<tr>
<td></td>
<td>.10(p=.43)</td>
<td>-.44(p=.0003)</td>
</tr>
<tr>
<td></td>
<td>Second Supervisor</td>
<td>Second Evaluation</td>
</tr>
<tr>
<td></td>
<td>-.09(p=.61)</td>
<td>-.67(p&lt;.0001)</td>
</tr>
<tr>
<td></td>
<td>-.06(p=.72)</td>
<td>-.63(p&lt;.0001)</td>
</tr>
<tr>
<td></td>
<td>-.04(p=.73)</td>
<td>-.48(p&lt;.0001)</td>
</tr>
<tr>
<td></td>
<td>Third Supervisor</td>
<td>Third Evaluation</td>
</tr>
<tr>
<td></td>
<td>-.12(p=.49)</td>
<td>-.80(p&lt;.0001)</td>
</tr>
<tr>
<td></td>
<td>-.21(p=.23)</td>
<td>-.82(p&lt;.0001)</td>
</tr>
<tr>
<td></td>
<td>-.15(p=.23)</td>
<td>-.66(p&lt;.0001)</td>
</tr>
</tbody>
</table>

The first, second and third row of each cell corresponds to Pearson, Spearman and Kendall coefficients, respectively.

Additional positive evaluations were likely to be provided for subordinates who were initially evaluated positively. The pattern of processing in the formation of these positive evaluations was presented in the context of the final two primary hypotheses. The manager compared the ratee’s performance to the performance of other employees (providing consensus information) or with objective standards. The ratee’s performance was typically viewed as better than the performance of other
employees or as satisfactorily meeting high standards of performance. The likely result of this comparison involved an internal attribution. As discussed in earlier sections, many more internal attributions were made than external attributions, particularly in the context of the supervisor evaluations. This internal attribution was then often followed by a positive evaluation. An example of this type of processing was included earlier. Statistical support for these final two associations is provided by the significant chi-square values presented for these specific hypotheses. Although these chi-square statistics were based on all 108 supervisor evaluations, 102 of these involved subordinates who received "Good," "Very Good" or "Excellent" overall performance ratings.

The final predicted component for "positive" employees receiving positive final evaluations involved a statement by the rater expecting additional positive behavior. This hypothesis was not confirmed. Supervisor evaluations included such statements infrequently (mean=.13) and only .6% of the time. This type of processing was not predicted to occur for the self-evaluations. However, for comparative purposes, self-evaluations included "positive prediction" phrases with a mean frequency of .06 and only .2% of the time. While the differences between the rating types were in the predicted direction, single d.f. contrasts using frequency (F(1,105)=.90, p=.34) and proportion (F(1,105)=1.40, p=.24) data revealed that there were
no significant differences between the two rating types. As a result, this component was removed from the model in the interest of parsimony.

Even if a subordinate was clearly viewed in a positive manner, an occasional negative episode of behavior was still retrieved. An episode of behavior refers to a single, specific event which involved the subordinate. This term was used too loosely in the preliminary model in that it failed to distinguish between behavioral generalizations and behavioral episodes. Very few behavioral episodes were retrieved in the formal study. This general finding provides an interesting insight into the lack of specificity in performance appraisal as it is actually practiced. Only seven behavioral episodes were retrieved for subordinates generally viewed as performing favorably.

However, as Figure 7 is meant to imply, 5 of the 7 behavioral episodes involved negative behavior. That is, most of the specific episodes of behavior were associated with negative activities.

A clear statement regarding the role, purpose or significance of these behavioral episodes is made later in this discussion after we examine their use in the context of generally poorly performing subordinates.

While five negative episodes of behavior were retrieved for generally positive employees, four of the five were dismissed
by attributing the poor behavior to internal or external factors beyond the subordinates control (e.g. "just thoughtlessness"). If we focus on the use of external attributions for "better" employees, 23 were associated with a positive or negative evaluation. Of these 23, 15 were used to dismiss negative information. As Figure 7 indicates, this ability to dismiss negative information allowed the rater to continue providing positive evaluations for employees they generally perceive as positive.

An excerpt from the first manager's protocol transcript should be useful in further illustrating the type of processing described above. The following quote was selected from an evaluation for a subordinate who received an "Excellent" final overall rating: "So, I cautioned her that the next time she took vacation I wanted her to watch that schedule. And darned if she didn't do it again this year. So when I asked her why she did it--she said 'Well, those are the days Bob wanted off, I guess I can change it.' And I noticed that she has changed it so I guess maybe it was just thoughtlessness that she put down the time that she did. So, I would say Rosemary is an excellent worker."

The above excerpt involves the recall of a negative incident of behavior or behavioral episode for a superior employee. It appears that the negative behavior was dismissed by attributing it to "just thoughtlessness." Immediately following this dis-
missal of negative behavior, the manager proceeded to provide an extremely positive evaluation which was consistent with her overall perception of the subordinate's performance.

The preliminary model also predicted that raters would dismiss negative information for better employees by considering how much the subordinate had improved since the occurrence of the poor performance. This hypothesis was not confirmed. While improvement information was used in 3% of the phrases (mean frequency was .24), it was not used to dismiss negative information as was predicted.

As mentioned, the type of information processing was different for those subordinates who were generally perceived as poor performers. As a word of caution, only six subordinates (out of 108) received "Poor" or "Fair" overall performance ratings so the sequence of information processing presented for such performers was based on little data. However, the first interesting finding for this path of processing involved the use of behavioral episodes. For employees viewed as poor performers, 13 negative episodes of behavior were retrieved and zero positive episodes of behavior were recalled. This finding is particularly significant since there were only six subordinates that elicited these 13 negative episodes of behavior.

Combined with the findings for better performers, it appeared that behavioral episodes generally involved negative behavior. Since there was not a differential recall of episod-
ic behavior for good and poor performers, it implies that there was no relationship between a type of worker schema and the type of behavioral episode retrieved. The results imply that supervisors have a general job schema for a properly functioning work environment. When behavior deviates from this schema it is salient to the supervisor, requires corrective action and is therefore retrieved for good and poor performers. As a result, the behavioral episodes were probably not retrieved for any particular purpose during the evaluations. That is, they were not retrieved as evidence to support the rater's evaluation since negative episodes were also recalled for better employees. Behavioral episodes were generally negative and it appeared that they were retrieved simply because they were more easily remembered.

In the context of generally positive evaluations, the rater felt compelled to dismiss negative episodes of behavior. However, for poor performers, these negative episodes of behavior played an integral role in the formation of an evaluation. We must speculate on the precise placement of negative behavioral episodes in the sequence of information processing due to insufficient data. However, it appeared generally consistent with the data and logical to view negative episodes as preceding the use of consensus information. Specific incidents of poor performance by the subordinate were compared to the performance of other employees and to other standards and were viewed as less effective.
As an illustration, the following excerpt was selected from the first manager's evaluation of her "most problem" subordinate: "...if I ask [Jean] to type [a letter] for me, she will say 'I'm the one that failed typing, you remember?' However, [Jean] can type. The only part of typing she failed was the speed part and probably the next time I want to let [Jean] type I'll say 'Well, Bonnie will you do it?' Because Bonnie's more agreeable and will say 'sure.'" This quote illustrates the recall of a negative incident of behavior and its use in confirming that the subordinate is less effective than other employees.

At this point the processing appeared to be similar to that discovered for better performers. Following a comparison to other employees and objective standards, an internal attribution and a negative evaluation were made. While these final two associations were supported statistically for all supervisor evaluations, the sample size was too small to conduct a similar analysis restricted to poor performers. However, a qualitative assessment of the data did indicate that this type of processing could be expected for poor performers as well.

The final step in the preliminary model predicted that the supervisor would consider what the subordinate must do in order to improve performance. The model also hypothesized that the supervisor would reflect on an action plan to verify the negative evaluation. Neither one of these predictions was con-
firmed. Out of the 108 supervisor evaluations only one phrase was coded into each of the two categories. It is difficult to account for these highly disconfirming results. Although it is mere speculation, it is interesting to note the correlation between the occurrence of these two types of information and the managers' consideration of job responsibilities. During pretesting with the use of more elaborate instructions, job responsibilities were considered along with what the subordinate should do to improve performance and a reflection on how to verify one's negative evaluations. It is possible that a request of the rater to consider the ratee's job responsibilities results in a more thorough evaluation in a variety of ways.

The final component of the model for the supervisor evaluations involved what occurred when raters made favorable references for generally poorly performing subordinates. As discussed, no specific behavioral episodes were referenced in this context. The lack of data precluded discovering any interesting patterns, but the role of external attributions was explored. There were only two external attributions associated with a positive or negative evaluation in the context of a poorly performing subordinate. Of these two external attributions, one was used to dismiss a positive statement while the other was used to dismiss a negative evaluation. An important point to make is that no external attributions were used to
form a more negative or more positive (as was the case for self-evaluations) evaluation. The amount of data used above was not sufficient to make any conclusive statements. However, the findings certainly did not contradict the use of external attributions in dismissing inconsistent behavior. As a result, the role of external attributions in dismissing positive behavior for generally poorly performing subordinates was included in the model.

Figure 8 presents the flow of processing discovered when a supervisor evaluates him/herself. As discussed, an overall or dimensional evaluation was the first component for the self-evaluations. The prediction was that these initial self-evaluations would be primarily positive. This hypothesis was confirmed to some extent since only two out of the 36 (.06) initial self-evaluations were negative while 18 out of the 108 (.17) initial supervisor evaluations were negative.

One of the more interesting findings with regard to the self-evaluations involved the explicit use of more dimensions of performance. In terms of frequency data, self-evaluations averaged 2.42 references to specific dimensions of performance while supervisor evaluations averaged only 1.82 such references, F(1,105)=3.51, p=.06. Analysis of the proportion data revealed similar results, F(1,105)=8.13, p=.0052. Self-evaluations referred to different dimensions of performance for 20% of the phrases, while supervisor evaluations involved specific dimensions of performance for only 12% of the phrases.
The preliminary model predicted that self-evaluations would retrieve primarily positive episodes of behavior. The data indicated that self-evaluations did not include positive or negative episodes of behavior. Based on the 36 self-evaluations, only one specific episode of behavior was detected. This episode involved a description of a situational factor. Thus, it was clear that self-evaluations did not involve retrieval of specific episodes of behavior. This certainly has significant implications for the use of different rating sources in the practice of performance appraisal. The self-evaluations in this study used information that was far less concrete or specific.

The preliminary empirical model also predicted that consistency/improvement information would be used more frequently during self-evaluations. This prediction was one of the primary hypotheses and was not confirmed. Supervisor evaluations used consistency/improvement information in 3% of its phrases while self-evaluations relied on this information only 2% of the time. Since this information was used rather infrequently and was not used differentially by the two rating sources it was removed from the model.

The results regarding general consistency information were similar to those found for consistency/improvement information. Analysis of the frequency ($F(1,105)<1, p=.70$) and proportion ($F(1,105)<, p=.98$) data revealed no significant differences.
Self- and supervisor evaluations included references to consistency information for 1% of the total phrases.

The final empirically based model is consistent with the preliminary model by excluding consensus/standards information and internal attributions in the model for self-evaluations. This is not meant to imply that this information was not used during the self-evaluations. It does, however, reflect the results of the primary hypotheses which indicated that self-evaluations included significantly fewer references to consensus/standards information and internal attributions. Since the models were designed to illustrate differential patterns of processing between the two rating types, consensus/standards information and internal attributions were not included in the model for self-evaluations.

As discovered in analyzing the data relevant to the primary hypotheses, self-evaluations did appear to make use of significantly more external attributions. However, the external attributions did take on a slightly different form than those used to account for subordinate's performance. In the context of a supervisor evaluation, external attributions were primarily used to explain away or dismiss behavior that could not easily be incorporated into the manager's schema for that subordinate. During the self-evaluations, external attributions were used to enhance or improve self-appraisals in that managers were able to perform well despite situational or environmental
constraints. Out of 26 external attributions associated with an evaluation, 17 were used in the formation of a positive evaluation while the other 9 were used to dismiss a negative evaluation. It is clear that the primary purpose of external attributions was to produce a more positive evaluation.

As an example, the following excerpt has been selected from the first manager's evaluation of herself: "We take care of the benefits here and that's quite a job. I try to disperse the work among the people under me as evenly and as fairly as I can. I try not to show partiality. I probably do a lot of things myself that I could pass down, because if I see that everybody is busy, I hate to burden them with more than they can do." This quote begins with a reference to the difficulty of the task (external attribution), but is followed by generally favorable comments about the manager's own performance. While these comments are not extremely positive summary statements, the information considered could certainly contribute to an overall positive evaluation.

Compared to the supervisor evaluations, self-evaluations included approximately the same percentage of positive evaluations (63% for self-evaluations and 60% for supervisor evaluations). However, supervisor appraisals included more negative evaluations than self-appraisals. In terms of frequency data, for self-appraisals the mean was 1.19 while the mean was 3.02 for supervisor evaluations, F(1,105)=16.27, p<.0001. For the
proportion data, the self-appraisals included negative evaluations for only 9% of its phrases; the supervisor appraisals made negative evaluations for 20% of the protocol, \( F(1,105)=13.15, p=.0004 \).

As predicted, self-evaluations did make more "could use improvement" statements. An example of this type of comment selected from one of the self-evaluations would be: "I think I need to learn microcomputer skills a little better than I do now." Such statements serve as a more desirable form of expressing a negative evaluation. Analysis of the frequency data revealed significant differences (\( F(1,105)=8.33, p=.0037 \)) between the self- (mean=.56) and supervisor (mean=.40) evaluations. The proportion data also yielded significant differences (\( F(1,105)=8.33, p=.0047 \)) between the self- (mean=4%) and supervisor (mean=1.5%) evaluations. The "could use improvement" statements were often made with respect to a specific dimension of performance, but did not appear to be preceded by any other type of information processing.

The final component to be discussed in Figure 8 involves the hypothesis in the preliminary model that self-evaluations would be terminated with the rater's prediction that he/she would improve or learn more. While this finding appeared robust during preliminary phases of data collection, it was not supported by the formal data set. In terms of frequency data, there was no significant difference between the two rating sources,
179

$F(1, 105) = 1.22$, $p = .27$. Predictions to improve or learn more were made infrequently during self- (mean=.06) and supervisor (mean=.13) evaluations. The proportion data produced similar results ($F(1, 105) < 1$, $p = .46$), with self- and supervisor evaluations making such predictions .5% and .9% of the time, respectively. As a result, the "prediction to improve or learn more" component was eliminated from the model characterizing self-evaluations.
VALIDITY OF VERBAL PROTOCOL ANALYSIS

One of the initial objectives of this study was to make a defensible statement regarding the use of verbal protocol analysis for the study of performance appraisal. Taken collectively, the various internal indices and the four specific "validity checks" provide fairly convincing evidence in favor of using verbal protocol analysis in this domain.

The general favorability of the self- versus the supervisor protocols was not inconsistent with the performance appraisal literature (Kraiger, 1986). While the self-evaluations in the study were not significantly more favorable, they were in the predicted direction. In addition, there was evidence that the self-evaluations included less negative statements than the supervisor evaluations. Such results are encouraging as they indicate that the verbal protocol method did not result in a more critical review of one's own performance than would be expected by the literature.

There was also evidence that perceived differences in subordinate performance were reflected similarly in both ratings and
verbal protocols. Although this provides support for the validity of the data used in this study, a broader statement is implied. In conducting verbal protocol analysis, one does not necessarily sacrifice the information typically collected in the form of ratings. Verbal protocol analysis provides the "rating level" information along with additional information (to be discussed in the next section).

The results also indicated that the processing discovered for the supervisor evaluations was consistent across the three subordinates. Such consistency in the protocols is reassuring in establishing the validity of the method. In addition, these findings imply that the order in which the ratings were completed in this study did not affect the nature of the protocols for the supervisor evaluations. These findings are encouraging since order was an undeniable potential confound in the study.

The final set of results directed at establishing the validity of protocol analysis indicated that the protocols could be affected in a predictable manner by having raters review relevant dimensions of performance prior to providing the protocol. These results have significant implications for this study and for the study and practice of performance appraisal in general. In terms of this study, it has been argued that we can have faith in the validity of the method of obtaining traditional ratings because we are capable of manipulating them in a predictable manner. Likewise, one should have increased faith in
the validity of verbal protocols when they can be affected in a way that is consistent with the researcher's a priori predictions.

However, the observed effect of the above manipulation has broader implications. These results serve as direct support for a DeNisi et al. (1984, p. 373) hypothesis that was not tested. The results indicate that an appraisal form is capable of affecting or organizing the consideration and recall of performance relevant information. The DeNisi et al. hypothesis stated that the types of rating scales used will influence the way performance is perceived by the rater. From an applied perspective, it is useful to recognize that an appraisal form can influence how a manager considers an employee's performance beyond the period of time in which the manager is actually completing the form. As a suggestion for future research, a logical extension of the manipulation conducted here would be to compare the effect of two different types of rating forms on a manager's perception of subordinates' performance.

In general, the study provides converging evidence regarding the validity of verbal protocols for research on performance appraisal. While researcher's have suggested using verbal protocol analysis to study the evaluation process (Ilgen & Favero, 1985; Feldman, 1987), this study begins to provide some of the necessary empirical support for using protocol analysis in this domain.
Another issue related to the validity of protocol analysis involves the nature of the data collected. More specifically, this involves the time at which the data is gathered. Although performance appraisal does involve the recall of previous information and assessments, protocol analysis theory suggests that the recorded thoughts should consist of an independent, concurrent evaluation. That is, the task should not require an attempt to retrieve a previous performance evaluation. Such retrieval would cause the verbal reports to be retrospective which, as discussed throughout the Literature Review, would not be particularly appropriate. Empirical evidence supporting the claim that this requirement was successfully met was presented in the Results section. However, three additional logical arguments support the concurrent nature of the performance appraisal data.

First, managers have easy access to the information they need to provide an evaluation. Even in a more general context, when an individual is asked how he/she feels about someone else, the individual is usually capable of forming a response immediately. It is much more difficult for the individual to attempt to retrieve the results of a previous assessment made at some point in the past. Thus, there would be no reason for managers to attempt to retrieve the results of previous evaluations unless they were specifically asked to do so.
Second, the manager encounters performance related information continuously over time. Events constantly arise in the workplace and managers recognize this. As a result, a prompt for an evaluation at different times is likely to result in evaluations based on different information.

Third, and finally, appraisals completed in the past are not likely to be perceived as significant events that should be recalled. Performance appraisals are simply a reflection of the manager's perception of the subordinate at a given point. For instance, it is unlikely that managers will attempt to recall the information they provided for this study when they are asked again for an evaluation. It is unlikely that they would associate any particular significance with the information they provided for the protocols. In retrospect, it appears that initial doubts regarding the concurrent nature of performance evaluations may not have been warranted.

There is one final issue to be discussed regarding the validity of the verbal protocol data. The above results address the validity of the complete verbal reports taken in their entirety. However, the hypotheses and discoveries made in this study are based on more specific aspects of the protocols. As a result, an argument might be that the results based on the "validity checks" do not apply to the specific research questions examined in this study. Such a criticism is legitimate, but additional procedures were conducted to insure the
validity of the data actually used in testing the specific hypotheses.

Before discussing these procedures, it is important to present the rationale for attempting to establish the validity of the verbal protocol data in general. Recall, one of the initial objectives of this study was to establish the validity of verbal protocol analysis in the context of performance appraisal. As support for the validity of the method, an attempt was made to establish the validity of the data obtained using the method. This involved a fairly general objective. In order to address such a general goal, it was essential that the "validity checks" be conducted at the most generalizable level of analysis rather than with aspects of the study that were specific to this investigation.

Despite the above decision, a great deal of evidence exists in support of the validity of the data used in testing the specific hypotheses. First, the overall "validity checks" are certainly relevant and should be perceived as comforting by the reader. While such findings are not sufficient conditions for establishing the validity of the data used for analysis, they are necessary.

Secondly, the only difference between the original protocol transcripts and the data used in the final analysis involved the phrasing and coding procedures. The inter-rater reliability estimates of the phrasing and coding procedures for the
entire coding scheme were consistently high. The intra-rater reliability estimates demonstrated that the raters were consistent over time. Although there were a few exceptions, the reliability estimates by category were also sufficient. In general, the phrasing and coding procedures were performed in a reliable manner in this study. These results also indicate that reliable coding is possible in the domain of performance appraisal given such factors as appropriate pretesting, training, motivation, etc.

One final aspect of the study provides fairly compelling evidence for the validity of the data used in the final analyses. Recall, four of the primary hypotheses were analyzed based on the coded data and the global, qualitative assessments by undergraduate psychology students. In all cases, the analyses based on these two types of data produced similar results (i.e. either both outcomes were significant or both were not).

UTILITY OF VERBAL PROTOCOL ANALYSIS

Another issue raised in the Hypotheses section involved the benefit of using verbal protocol analysis rather than more traditional methods such as questionnaires or interview schedules. Although specific hypotheses were not made in this regard, a number of findings in this study should serve to demonstrate the benefits of using protocol analysis over alternative methods. Such evidence is provided by the unexpected findings and
the sequence or order in which certain processes were found to occur. Daft (1983) suggests that the hallmark of good research is to allow for such uncertainty and surprise in conducting a study.

These findings are presented below and would not have been discovered or would have been found with less confidence in the case of specific prompting using structured rating scales or questionnaire items. In examining each of these findings, the critical question is whether you could have made the discovery using structured items designed prior to the investigation, not whether you could "find" such results after you already know they exist.

One of the most interesting findings of the study involved manager's tendency to consider subordinates' experience or time on the job. This finding was not expected and would not have been discovered using more structured approaches. Another such finding was the tendency for managers to make more "could use improvement" statements in terms of their own performance. The differential use of behavioral episodes for good and poor performers was a finding that probably would not have been discovered based on traditional techniques.

Another unexpected finding involved the effect of the instructions on the protocols. The final data revealed that few managers made an explicit reference to the ratee's job responsibilities even though this was done frequently during
pretesting phases of the study. During the preliminary phases
of this investigation, a variety of factors were being modified
and evaluated. One of these factors involved altering the
instructions given to managers. A set of instructions used for
most of the pretest data included a request of the managers to
"...consider the ratee's job responsibilities..." This request
was deliberately excluded from the instructions used in col­
lecting data for the formal study.

It is very possible that the preliminary instructions were
responsible for producing the initial processing found in the
pretest data. If this was the case, it serves to make several
points. First, the verbal protocols were sensitive to and
could be manipulated by subtle modifications in the instruc­
tions. Second, raters do not explicitly consider the ratee's
job responsibilities unless they are asked to do so. Further­
more, when they are asked to consider this information, they
are capable of doing so.

Analysis of the coded data also produced results which sup­
port the usefulness of verbal protocols. While the two rating
types involved a similar use of positive evaluations, supervi­
sor evaluations included more negative evaluations. This find­
ing would not have been discovered using standard ratings of
performance. The tendency for self-evaluations to reference
more specific dimensions of performance was another interesting
and unexpected finding.
Finally, the results based on the order or sequence in which processes were found to occur can be made with a great deal of confidence since they were produced concurrently and without providing subjects with specific prompts regarding the sequence of information processing. Results regarding the sequence of information processing were produced for the primary and exploratory analyses. These findings were presented earlier and will not be repeated to avoid redundancy. However, it should be clear that verbal protocol analysis has enhanced utility in exploratory work since, without guiding hypotheses, the demand on the respondent to seriously consider all possible sequences of events would be overwhelming.

COMMON PROCESSING FOR BOTH RATING TARGETS

This study was designed to discover and characterize differences between self- and supervisor evaluations. However, before highlighting these differences, it is important to review those findings or patterns which were found to characterize both self- and supervisor ratings.

Both types of ratings included a similar number and proportion of positive evaluations during the protocols. Also, of all 21 coding categories, the positive evaluation category was clearly used most often for the two rating targets. In addition, as the exploratory analyses revealed, managers were capable of providing an evaluation immediately following a demand
from the environment in the context of self- and supervisor assessments.

The verbal protocols for both rating targets revealed little use of consistency or distinctiveness information. A plausible explanation for neglecting to use this information is presented later. The primary analyses also indicated that there was a relationship between the use of attributional information and the formation of attributions for both types of ratings.

The two types of ratings used internal attributions much more often than they used external attributions. As expected, the emphasis during the evaluations was clearly on the rating target rather than the surrounding environment. The primary analyses also indicated that internal attributions were followed by more evaluations than external attributions. This finding was consistent regardless of rating target.

Finally, both types of ratings were similar in that they made relatively few predictions of performance. That is, there were very few statements indicating that performance would be positive, negative or improving in the future. One possible explanation for this finding was presented in the Results section. The two types of ratings were also similar in that there were few statements indicating a plan or strategy to test the validity of one's evaluation of the ratee. Such statements did appear occasionally during pretesting.
Processes Related to Attribution Theory

The primary purpose of this study was to use attribution theory to explore the potential for differences in information processing for the two types of ratings. The study has been successful in discovering such differences based on primary and exploratory analyses.

The study generalizes the actor-observer differences commonly found in attribution and laboratory research to an applied performance appraisal context. The findings in terms of internal and external attributions were consistent with the predictions based on attribution theory. In accord with the fundamental attribution error, supervisors made fewer external attributions and more internal attributions for the performance of their subordinates.

Performance level of the subordinate was not found to moderate this finding in terms of internal or external attributions. While seemingly inconsistent with the literature, this result can be accounted for. In terms of internal attributions, while research has found that more internal attributions are made for poor performers (Kipnis, 1972; Mitchell & Kalb, 1982; Ilgen, Mitchell & Fredrickson, 1981; Pence, Pendleton, Dobbins & Sgro, 1980), other research (Mitchell & Wood, 1980) and the fundamental attribution error suggest that performance level is not a factor. In addition, the restriction in range of the perform-
ance level of the subordinates could have contributed to the lack of a significant relationship. That is, there may not have been enough "poor" performers in this study to reveal a performance level effect even if one does exist.

In terms of external attributions, there was no prior prediction regarding differential use according to subordinate performance level. This study made a more interesting discovery. External attributions were used for both high and low performing subordinates. However, they were used differentially depending upon the performance level of the subordinate. That is, they were used to dismiss negative behavior for good performers and to dismiss positive behavior for poor performers.

The general findings with respect to external attributions contribute to our overall understanding of ratings of different targets. In terms of self-evaluations, there were many more external attributions and these attributions were used to dismiss negative behavior and further enhance positive performance. Such findings serve to assist in explaining the tendency to find inflated self-ratings.

These results have implications for the nature of ratings as well as performance appraisal interviews and feedback processes. Supervisors are less apt to recognize the effect of external attributions on subordinate performance than are the subordinates themselves. In addition, when external factors are
recognized, they are used to confirm the supervisor's overall perception of the subordinate. Internal attributions will also have a similar effect. Internal attributions are used in a consistent manner so that good performers are perceived as extremely good and poor performers are perceived as extremely poor. The significance of these results for performance appraisal and feedback contexts should be clear. If a subordinate is perceived in a positive manner, performance feedback should be easily accepted. However, if a subordinate is generally viewed in a negative fashion, he/she could have some legitimate problems with the performance feedback received.

In addition to the type of attributions formed, use of the three types of attributional information was also examined. No differences were found in the use of consistency or distinctiveness information. Both types of information were used infrequently. Although there were reliability problems in coding distinctiveness information, there is a plausible explanation for why this information was not used in either type of rating.

The managers were asked to provide overall ratings for given subordinates. Distinctiveness information involves within subordinate (or between task) distinctions. As a result, distinctiveness or within subordinate information is not necessarily useful in providing overall ratings which are often made with respect to other subordinates' performance. This provides a
suggestion for follow-up research on the types of demands that would initiate the use of distinctiveness information. For instance, a prompt requesting the manager to determine which training programs would be appropriate for certain subordinates would be more apt to require "within subordinate" processing.

Consensus and other standards information was found to be used differentially by the two rating types. Supervisor evaluations included more references to this information than self-evaluations. This reduced use of comparative information in the context of the self-evaluations is consistent with comments made by Mabe & West (1982, p. 290) on this topic. Mabe & West suggest that individuals are not likely to compare their performance with a particular standard or the performance of others unless such comparative information is made salient. Since the managers were not asked to consider specific standards or reference groups in making their self-evaluations, the results obtained are consistent with the comments by Mabe & West.

This failure to consider evaluative standards in the self-evaluations provides another plausible explanation for why self-ratings have been found, at times, to be more favorable than other types of ratings in the literature. Such findings have clear implications for the performance appraisal interview. It appears that a supervisor may make ratings in terms of particular standards that the subordinate isn't even taking into consideration when evaluating his/her own performance.
Furthermore, exploratory analyses indicated that such comparative information is apt to be used differentially by the supervisor for good and poor performers. That is, for subordinates who were perceived favorably overall, comparative standards were found to be used in confirming that performance was very good. However, for subordinates who were perceived negatively, comparative information was used to confirm that the performance was below standard. These findings are consistent with literature which suggests that people tend to seek information which confirms their initial expectations (Snyder, 1981; Snyder & Cantor, 1979; Swann, Stephenson & Pittman, 1981; Wong & Weiner, 1981). Thus, once again, the subordinate who is perceived positively will be viewed even more positively while the subordinate who is perceived negatively overall will be viewed even more negatively.

The final aspect of attribution theory to be discussed involves the formation of attributions and evaluations. It appeared that for both rating types, attributitional information was used to form attributions and internal attributions were related to the formation of evaluations. However, as discussed, these patterns were more pronounced for the supervisor evaluations. The self-evaluations used more external attributions in forming evaluations and were less apt to rely on attributitional information (consensus and other comparative standards). As discussed, the tendency for self-evaluations to
include more external attributions is clearly consistent with attribution theory. However, the results were a discovery in that external attributions were used to further inflate positive evaluations rather than dismiss negative evaluations. The findings with respect to consensus information appear to be the first of their type. These findings provide additional confirmation of the significance of attribution theory for the practice of performance appraisal. The results also provide insight into the cognitive processing models of performance appraisal. This issue will be discussed further in a later section.

Processes Unrelated to Attribution Theory

Based on the coding categories not related to attribution theory and on the exploratory analyses, additional processes were found that differentiated between the two rating types. These findings are generally quite consistent with those presented in the previous section. The supervisor evaluations included more references to the ratee's experience or time on the job. This type of information serves as a comparative standard and its differential use is consistent with the Mabe & West (1982) suggestions which were discussed in the previous section. Individuals appear to be less apt to consider external standards in evaluating their own performance. However, in evaluating others, it appears that such standards become relevant in attempting to provide an appropriate evaluation. This
type of processing allows the rater to apply more or less stringent standards depending upon the ratee's experience on the job.

Significantly more negative episodes of behavior were also recalled in the context of the supervisor evaluations. Such differential recall may be a function of "selective attention" by the manager or by the manager's tendency to attribute negative aspects of his/her performance to external factors. Whatever the reason, this highly specific, negative information provides another explanation for the tendency for supervisor ratings to be less lenient than self-ratings. Such findings also imply that a supervisor may provide feedback in terms of specific episodes of behavior that the subordinate would not have retrieved independently. As a result, the subordinate could be expected to view such feedback as trivial and overly critical.

It was also discovered that more specific dimensions of performance were used for the self-evaluations. This implies, perhaps, a more thorough evaluation of one's self. However, when considered along with the fact that there were fewer standards and negative episodes referenced in the self-evaluations, it allows the manager to view him/herself as positive on many dimensions of performance. This is obviously less likely to lead to an overall negative evaluation.
Finally, the self- and supervisor did include a similar number of positive evaluations based on the coded data. However, the self-evaluations included less negative evaluations and more "could use improvement" evaluations. Such "rephrasing" clearly allows the manager to evaluate him/herself a little less stringently. Whether such statements are the result of other differential processing discussed in this section (e.g. fewer standards and negative behavioral episodes) is clearly possible, but is an empirical question worthy of additional study.

**CONSIDERATION OF EXISTING PERFORMANCE APPRAISAL MODELS**

The results obtained in this study provide some significant insights into the cognitive processing models of performance appraisal. As mentioned in the Literature Review, these models are typically designed to characterize supervisor evaluations. Based on findings from the previous section, it appears clear that generalizing such models to other types of ratings should be discouraged. It appears that the nature of the model is in fact contingent upon the type of rating being made.

Two of the major findings in this study do provide empirical support for the models representing supervisor evaluations. First, as characterized in the performance models, there did appear to be a relationship between the use of attributional information and the formation of attributions. Second, a rela-
tionship between the formation of internal attributions and evaluations was also discovered. While such results appear somewhat straightforward and logical, they seem to have a great deal of significance. They provide empirical support for the relevance of attribution theory to performance appraisal practices. They also provide the essential empirical basis for specific aspects of the cognitive processing models of performance appraisal.

In addition to the supporting evidence above, significant results which are absent from or inconsistent with the existing performance appraisal models were also discovered. The first of these discoveries involved the robust finding that supervisors were capable of providing an evaluation immediately following a demand from the environment. While these results may not appear surprising, they are clearly inconsistent with the performance appraisal models. The existing models suggest that a great deal of information processing is essential prior to forming an evaluation.

It may be that such models provide a more accurate representation of what occurs when performance is initially observed rather than the processes expected during a formal performance review. Although there is nothing undesirable about such models, it seems obvious that the authors of these conceptual frameworks should be clear in specifying what it is they are modeling. That is, the models may better characterize the ini-
tial formation of an evaluation as would occur in an assessment center context or with early assessments of a new employee.

The discussion above reflects a distinction between on-line (i.e. initial) and "follow-up" evaluations that has not really been made in the literature. It also suggests the need for additional research on the topic. One could envision a series of laboratory studies which systematically vary the length of time between the observation of performance and the evaluation. This would presumably serve to clarify the types of distinctions discussed above. Similar laboratory studies could also be conducted by varying such factors as rater familiarity with the ratee, evaluation experience of the rater, etc. While additional modeling will certainly be difficult, it appears that this type of combination between laboratory and field research will be essential if we are to continue to make interesting discoveries in this area.

Another finding in this study which is not specifically represented in the performance appraisal models involved the supervisor's tendency to consider the ratee's experience or time on the job. This finding is not necessarily inconsistent with the cognitive models. As mentioned, the ratee's experience or time on the job can be viewed as a type of standard which is represented in most of the models of performance appraisal. However, awareness of this specific standard is certainly informative and demonstrates the need to support the
"theoretical" models with empirical research. While such models may serve as a general framework, there is clearly much more we can learn in this domain.

The final result that has significant implications for our models of appraisal processes involved the differential processing discovered for subordinates perceived as performing at different levels. Existing models of performance appraisal do not suggest that cognitive processes are contingent upon the supervisor's overall perception of the subordinate's performance. As presented earlier, processes involving episodes of behavior, external attributions and consensus/standards information varied according to the supervisor's overall evaluation for the subordinate. Again, additional lab studies (restricted field studies might be even better) would be useful in further examining the effect of subordinate performance level on the supervisor's evaluation of the subordinate.

LIMITATIONS OF THE STUDY

There are a number of general issues that should be recognized in interpreting the results from this study. In following my suggestions from the previous section, it is important to specify what it is one is attempting to model. The managers in this study understood that they were providing performance appraisals for research purposes only. As a result, this was a study of assessments made without any real consequences for
either the manager or the subordinate. This investigation was less likely to capture "reporting" processes or the political nature of performance appraisal (see Longenecker, Gioia, & Sims, 1987, for an excellent and long overdue discussion of the political factors that obviously factor into the appraisal process). This should not be viewed as a particular problem. It is simply important to specify what processes one is attempting to capture. In fact, it might be argued that limiting the study to assessment (as opposed to reporting) processes is beneficial at this point in that it provides clearer suggestions for the interventions necessary to improve upon the formation of appraisals. Longenecker, Gioia, & Sims (1987) provide such suggestions in response to the many political factors that operate in the appraisal process.

Another potential limitation of the study was that managers were only asked to provide overall evaluations. The type of processing discovered for evaluations made in terms of specific dimensions could be different. For example, one might find a greater use of distinctiveness information during such evaluations. Because most organizational appraisals involve the use of specific dimensions, it seems to be an area worthy of further investigation.

As discussed previously, another critical limitation of this study involved the range restriction on the overall performance ratings. Only six subordinates received "Fair" or "Poor" over-
all performance ratings. While this type of sample may be ecologically valid in that managers may encounter relatively few poor performers, this does not reduce the need to understand the evaluation process for poor performers. A more focused field study targeted at different ratee performance levels would be useful.

Caution should also be taken in attempting to generalize from the job type and level findings obtained in this study. The two job types and levels were selected with the expectation that they were somewhat homogeneous. While job and level differences were not found in this study, they could easily arise in other contexts.

Finally, the sample size used to draw inferences is always a critical factor. The sample size in this study was as large or larger than what is typical for the type of in-depth analysis conducted in this investigation. However, in terms of external validity, there is a monotonic relationship between one's ability to generalize and the size of the sample drawn. Given such considerations, this sample places limitations on one's ability to make broad generalizations. Additional research and cross-validation is essential, particularly for the exploratory aspects of this study.
CHAPTER VII
CONCLUSION

Verbal protocol analysis appears to be a flexible method which can be used to access cognitive processes in an organizational context. Protocol analysis can be used to provide valid and useful information in the study of performance appraisal processes.

Empirical support has been found for the application of attribution theory to the appraisal context. However, a number of findings suggest ways in which our current models of performance appraisal should be revised.

Attribution theory has clearly been useful in enabling us to detect rating target differences. However, such differences are better understood as a result of using verbal protocol analysis.
FIRST SUPERVISOR EVALUATION

She has been here the longest. Rosemary, very intelligent. She has gone far beyond what you would expect from a person with her schooling because she dropped out of high school. Her math ability is uncanny and her ability to work with computers or with computer printouts is very good. She seems to enjoy finding, resolving problems like when there is a bug in a computer program. She is excellent at finding what causes the problem. So in the position she's in now she is really good. However, Rosemary does have some faults, one being she is, she thinks along the line of more of a blue-collar worker than manager or executive type. She believes in being at work right on the minute, but she also believes in leaving right on the moment whether the work is done or not. She's doing a lot better now with her ill time but her philosophy seemed to be in the past that if I don't use the ill time I'll lose it. So, everybody else is using theirs, I might as well use mine too. Rosemary's knowledge of the work around this place is better than anybody else's and she probably deserves more money than
what she gets. I would be hard put to try to find a replacement for Rosemary. However, it is strange about Rosemary. She'd come to work on the day that the payroll has to be in whether she is half dead. Then she decides that...you can almost tell when Rosemary decides that she is going to take the day off because her desk will be...everything will be straightened up and pretty much in order and all these things done that need to be done. She takes care of the job pretty good. There is only one thing that I have against Rosemary. She put down for vacation the other day. She wanted three weeks. Several years ago she took three weeks at one time and she...well, how our schedule goes is that there is one day that is really hard to be without her...hard to have someone to replace her. And that day comes every other week and she took two days off that she...that we were hard pressed to be without her. Where she could take three weeks and have only one of those days off. So, I cautioned her that the next time she took vacation I wanted her to watch that schedule. And darned if she didn't do it again this year. So when I asked her why she did it she said well, those are the days Bob wanted off, I guess I can change it. And, I noticed that she has changed it so I guess maybe that it was just thoughtlessness that she put down the time that she did. So, I would say Rosemary is an excellent worker. And, in the job she is in, she would be hard to replace.
SELF-EVALUATION

That's going to be rough because, you can always see the good you do. You can see the bad, but you sort of hesitate to hope nobody else sees it. Well, I would say that as far as dependability...well, I don't think anyone can find any fault there because I work longer hours than I ever put down. Of course, I do...I can take advantage if I am off an hour and a half at noon, well, nobody can say anything because I have already worked the eight hours. It's just different times. As far as getting the work done, there are a few things that are a little bit...that I tend to maybe push back. And evaluation is one of them because I dislike doing it so much I guess. When somebody does something wrong, I like to let them know immediately rather than to wait to the end of a period and then to put down "well, you're not so good because you did this way back then." And I figure if I tell the person about it immediately and it does not happen again much, it should be forgotten and it should not be in an evaluation. So probably, I would say my weak point is in evaluating. As far as getting the jobs done I figure that I meet deadlines real well. And if we can't meet the deadline, well of course we have to get an extension and I know how to go about that. Also, I do have to take care of a lot of corrections and that is something you just sort of pick up as you go along. We take care of the benefits here and that's quite a job. I try to disperse the work among the people
under me as evenly and as fairly as I can. I try not to show partiality. I probably do a lot of things myself that I could pass down, because if I see that everybody is busy, I hate to burden them with more than they can do. Self-evaluation is really hard because how do you try to see yourself like others see you. It is sometimes difficult to do.

SECOND SUPERVISOR EVALUATION

Jean is probably the most problem person I have. Not because she does not do her work or know her work, but because she resents being asked to do anything outside her own job description and that is something that I don't like for somebody to say to me, that's not my job, because there are certain things that are not particularly anybody's job. They are just whoever has the time to do them. Jean does one thing that I don't like and I probably tend to ask her to do it last because she will always come up with a remark like if I ask her to write a letter...we don't have a regular typist but all of our girls can type and if I ask her to type one for me she will say, "I'm the one that failed typing, you remember?" However, she can type, the only part of typing she failed was the speed part and probably the next time I want to let her type I'll say well, Bonnie will you do it, because Bonnie's more agreeable and she will say sure. She might not have the time although she'll fit it in. So, but as far as...Jean does have good
points. Like if there's a project to be done, she's probably the best at organizing the thing and setting it up of any of the people that I have. So, I guess maybe the thing to do is to keep Jean doing the things she does better. However, I know that we sort of had a little run in a couple years ago when we had a position vacant and Jean's work was way down and I asked her to cover that position. That time she got sort of perturbed and even was looking at the greensheets and secretly I wished she would find another job. Because I don't think there's any place she could work and make personal phone calls and do things like that she does here. And still I guess I would miss Jean if she went because she knows her job and if I'm not around her work will still get out without anybody pushing her or getting after her. As far as attendance goes I will never forget the time it was bad out and she arrived twenty minutes early. She sat there drinking coffee and reading the paper till it was time to start and at the end of the day she came around and said can I leave, I was here twenty minutes early. and I said, but you didn't go to work twenty minutes early and she said I'll be darned if I come in twenty minutes early again. And I said, it's all right with me, nobody asked you to come in twenty minutes early. So Jean does come on time and she leaves on time. But as far as sick leave, she takes it as she gets it. Now, she is plagued by problems. I know that she's got allergies and she's got things. However, I'll never forget
the day that she asked for a vacation day and I said that's the
day you're supposed to do that payroll. And she said well her
daughter had moved back in with her and she left and she wanted
to clean her house and I said, "well the house has been this
way for several weeks, now will one more day make a differ­
ence?" And she says "that's just what my husband says, that my
job has to govern and everything has to work around that job."And she took off and I had to do the payroll so that really
endeared me to Jean. So, you can't say that she's an excellent
employee, she's a good employee, but she does have faults.

THIRD SUPERVISOR EVALUATION

Anne is an excellent worker and in fact I think I hit pay
dirt when I hired Anne. Because, in the position she's in now,
Anne is an excellent receptionist. She's very thorough. She
takes care of the benefits as well or better than anybody can.
She is very courteous on the phone and with people who come in.
Anne is one of the most organized people I have ever seen, in
fact with all the forms that she has people sign up with, she
has everything in its place and she sends them over. When she
goes to send them over, she doesn't fool around having to hunt,
they're all together. She's more organized than I am. I'm real­
ly amazed when I look at Anne's organization. Anne is off one
day a week and she gets disturbed when somebody else, who is
put in her job, doesn't keep things as organized as she does.
Of course, that's one thing she has to put up with because she wanted to drop down to 80 percent for personal reasons. However, as organized as Anne is and as good as she is it's real surprising she doesn't want to move up. She's happy where she is. And she doesn't want the responsibility of any of the other jobs here. And that sort of hampers us on teaching our cross-training because of course with our deadlines and things we have to have more than one person trained for every job. It sort of puts Anne out of our cross-training program. But, still I wouldn't want to say anything against Anne because I would hate to have to replace her. As far as punctuality goes, Anne's here every day and she's here on time and she does a good days work so I would give her top credit there. In her job, Anne does an exceptional job, it's just that she doesn't want to move upward and she doesn't even apply for another job.
## Appendix B

CODING CATEGORY USE REGARDLESS OF RATING TYPE

<table>
<thead>
<tr>
<th>Category</th>
<th>Mean Frequency</th>
<th>Mean Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Evaluation</td>
<td>8.7</td>
<td>.61</td>
</tr>
<tr>
<td>Other Intern Attrib</td>
<td>3.4</td>
<td>.24</td>
</tr>
<tr>
<td>Negative Evaluation</td>
<td>2.6</td>
<td>.17</td>
</tr>
<tr>
<td>Dimension Performance</td>
<td>2.0</td>
<td>.14</td>
</tr>
<tr>
<td>Trait Attribution</td>
<td>1.5</td>
<td>.11</td>
</tr>
<tr>
<td>Skill Attribution</td>
<td>1.3</td>
<td>.11</td>
</tr>
<tr>
<td>External Attribution</td>
<td>0.8</td>
<td>.05</td>
</tr>
<tr>
<td>Effort Attribution</td>
<td>0.8</td>
<td>.06</td>
</tr>
<tr>
<td>Knowledge Attribution</td>
<td>0.5</td>
<td>.04</td>
</tr>
<tr>
<td>Interest Attribution</td>
<td>0.5</td>
<td>.04</td>
</tr>
<tr>
<td>Consensus Information</td>
<td>0.5</td>
<td>.03</td>
</tr>
<tr>
<td>Other Standards</td>
<td>0.4</td>
<td>.03</td>
</tr>
<tr>
<td>Distinctiveness</td>
<td>0.4</td>
<td>.02</td>
</tr>
<tr>
<td>Could Improve</td>
<td>0.3</td>
<td>.02</td>
</tr>
<tr>
<td>Consistency/Improvement</td>
<td>0.2</td>
<td>.01</td>
</tr>
<tr>
<td>Experience Attribution</td>
<td>0.2</td>
<td>.02</td>
</tr>
<tr>
<td>Consistency Information</td>
<td>0.2</td>
<td>.01</td>
</tr>
<tr>
<td>Predict Positive</td>
<td>0.1</td>
<td>.01</td>
</tr>
<tr>
<td>Predict Improvement</td>
<td>0.1</td>
<td>.01</td>
</tr>
<tr>
<td>Predict Negative</td>
<td>0.0</td>
<td>.00</td>
</tr>
<tr>
<td>Test Evaluation</td>
<td>0.0</td>
<td>.00</td>
</tr>
</tbody>
</table>
LIST OF REFERENCES


Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. Psychological Review, 93, 258-268.


Sargent, S. S. (1940). Thinking processes at various levels of difficulty. *Archives of Psychology*, 249.


