INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

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 copyright 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 original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 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.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600
RICE UNIVERSITY

TASK FREQUENCY RATING ACCURACY: OBJECTIVE AND PERCEPTUAL BASES OF AGREEMENT

by

WENDY RICHMAN

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

APPROVED, THESIS COMMITTEE

Miguel A. Quinones, Chair
Assistant Professor of Psychology

Robert L. Dipboy, Professor of Psychology

Jeffrey S. Bowers, Assistant Professor of Psychology

Houston, Texas
June, 1995
ABSTRACT

Task Frequency Rating Accuracy: Objective and Perceptual Bases of Agreement

by

Wendy Richman

This study examined the impact of task engagement and task familiarity on frequency rating accuracy using an objective measure of task frequency. By contrasting findings from the memory literature on frequency estimation with current task analysis practices, this research proposed that task performers and novel experienced respondents would generate more accurate frequency estimates compared to task observers and familiar experienced respondents. Participants were randomly assigned to a task engagement condition (performer vs. observer) as well as to a task familiarity condition (novel vs. familiar). In support of the hypotheses, performers and novel experienced subjects demonstrated greater accuracy in their frequency ratings. These findings were illustrated using several different measures of accuracy. The use of an objective measure of task frequency revealed that discrepancies in frequency ratings are due to respondents' varying job perceptions as opposed to differences in actual task frequency. Implications for task analysis research and practices are discussed.
Acknowledgments

First, I am extremely grateful to my advisor and committee chair, Mickey Quiñones, for his continual encouragement, support, and direction. His professional expertise and commitment to academic excellence helped me to produce a thesis of which we are both proud. Mickey's friendship and unwavering confidence played a significant role in my personal growth and my strides towards independence for which I will always be grateful.

Second, I am also very appreciative to Robert L. Dipboye and Jeffrey S. Bowers for their interest and intellectual contribution to the project. Their diverse viewpoints proved to be invaluable.

Third, I would like to thank Roman Longoria for his endless reassurance, guidance, and companionship. He was always able to help me gain perspective when I was feeling overwhelmed by the task before me.

Most importantly, I am grateful to my family, whose unconditional love and devotion guide my achievements. Their unceasing confidence and faith in my abilities constantly remind me that I will always achieve great feats.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Agreement: Perception or Reality?</td>
<td>3</td>
</tr>
<tr>
<td>Task Frequency Rating Accuracy</td>
<td>6</td>
</tr>
<tr>
<td>Memory Processes in Frequency Estimation</td>
<td>10</td>
</tr>
<tr>
<td>Task Engagement</td>
<td>12</td>
</tr>
<tr>
<td>Task Familiarity</td>
<td>15</td>
</tr>
<tr>
<td>Task Frequency Rating Format</td>
<td>18</td>
</tr>
<tr>
<td>Method</td>
<td>19</td>
</tr>
<tr>
<td>Participants</td>
<td>19</td>
</tr>
<tr>
<td>Design</td>
<td>19</td>
</tr>
<tr>
<td>Materials and Equipment</td>
<td>20</td>
</tr>
<tr>
<td>Manipulations</td>
<td>21</td>
</tr>
<tr>
<td>Procedure</td>
<td>22</td>
</tr>
<tr>
<td>Measures</td>
<td>25</td>
</tr>
<tr>
<td>Dependent Variables and Analytic Strategy</td>
<td>27</td>
</tr>
<tr>
<td>Results</td>
<td>32</td>
</tr>
<tr>
<td>Determining Actual Task Frequency</td>
<td>32</td>
</tr>
<tr>
<td>Descriptive Statistics</td>
<td>32</td>
</tr>
<tr>
<td>Test of Hypotheses</td>
<td>35</td>
</tr>
<tr>
<td>Absolute Deviations</td>
<td>35</td>
</tr>
<tr>
<td>Relative Deviations</td>
<td>38</td>
</tr>
<tr>
<td>Detection Accuracy</td>
<td>38</td>
</tr>
<tr>
<td>Rank Order Accuracy</td>
<td>51</td>
</tr>
<tr>
<td>Summary of Accuracy Measures</td>
<td>54</td>
</tr>
</tbody>
</table>
Table of Contents (continued)

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancillary Analyses</td>
<td>54</td>
</tr>
<tr>
<td>Consistency Across Rating Format</td>
<td>54</td>
</tr>
<tr>
<td>Individual Differences</td>
<td>55</td>
</tr>
<tr>
<td>Discussion</td>
<td>60</td>
</tr>
<tr>
<td>Memory Processes in Frequency Estimation</td>
<td>63</td>
</tr>
<tr>
<td>Implications</td>
<td>64</td>
</tr>
<tr>
<td>Limitations and Directions for Future Research</td>
<td>68</td>
</tr>
<tr>
<td>Conclusion</td>
<td>72</td>
</tr>
<tr>
<td>References</td>
<td>74</td>
</tr>
<tr>
<td>Appendix A Toy Company Simulation In-Basket Tasks</td>
<td>82</td>
</tr>
<tr>
<td>Appendix B Absolute and Relative Frequency Rating Questionnaires</td>
<td>88</td>
</tr>
<tr>
<td>Appendix C Pre-Experiment Questionnaires</td>
<td>97</td>
</tr>
<tr>
<td>Appendix D Importance and Interest Rating Questionnaires</td>
<td>102</td>
</tr>
</tbody>
</table>
List of Tables

Table 1: Descriptive Statistics and Intercorrelations of Study Variables 34
Table 2: ANOVA Results on Absolute Deviations 36
Table 3: Hierarchical Regression on Total Task Frequency, Task Engagement, and Task Familiarity on Absolute Deviations 37
Table 4: ANOVA Results on Relative Deviations 39
Table 5: ANOVA Results on the Hit Rate 40
Table 6: ANOVA Results on the False-Alarm Rate 43
Table 7: ANOVA Results on Detection Accuracy 46
Table 8: Repeated Measures MANOVA on the Hit Rate Proportions 49
Table 9: ANOVA Results on the Correlation between Absolute Frequency Ratings and Actual Task Frequency of Performance 52
Table 10: ANOVA Results on the Correlation between Relative Frequency Ratings and Actual Task Frequency of Performance 53
Table 11: ANOVA Results on the Consistency between Subjects' Absolute and Relative Frequency Ratings 56
Table 12: Intercorrelations between Individual Difference Variables and Accuracy Measures 58
List of Figures

Figure 1: Procedure Outline  
Figure 2: Interaction between Task Engagement and Task Familiarity on the Hit Rate  
Figure 3: Interaction between Task Engagement and Task Familiarity on the False-Alarm Rate  
Figure 4: Interaction between Task Engagement and Task Familiarity on Detection Accuracy  
Figure 5: Interaction between Hit Rate Proportions and Task Familiarity  
Figure 6: Interaction between Task Engagement and Task Familiarity on the Consistency Between Absolute and Relative Rating Scales  

page
23
42
44
47
50
57
Introduction

Knowledge about human work is obtained through job analysis procedures. Job analysis is defined as the method used to collect data on the job tasks, work procedures, and general worker-behaviors, such as decision making and supervision, that comprise a given position. Knowledge concerning the context in which the job is performed, including both physical and interpersonal characteristics of the environment, and descriptions of the necessary KSAs needed on the job are obtained through job analysis procedures (McCormick, 1976). The information gained from various job analysis approaches is necessary in order for organizations to develop processes of personnel selection, classification, placement, training, performance appraisal, job evaluation, job design, etc. (McCormick, Jeanneret, & Mecham, 1972). Thus, it can be seen that job analysis is a fundamental tool used by organizations in a variety of human resource management functions.

The majority of job analytic systems obtain job-relevant information from job incumbents and their supervisors. Some researchers speculate that it has become more common to use Subject Matter Experts (SMEs) in job analysis, rather than a single analyst, as a result of various well-publicized legal cases and decisions (e.g., Guardians v. New York City Civil Service Commission, 1980; Landy & Vasey, 1991). In addition, recent legislation has increased the need for accurate job analysis data. For example, the Americans with Disabilities Act (ADA) has put a stronger emphasis on job descriptions and the distinction between various job elements that comprise a given position. In order to make the appropriate accommodations for disabled employees, one has to use a job analytic system to determine the essential job elements. Hence, the ADA, among
other things, requires that researchers begin to examine the factors that may affect the accuracy of job analysis ratings.

When using SMEs, job incumbents and their supervisors are usually asked to complete a questionnaire in which they rate, for example, the relative importance or frequency of certain tasks that comprise the incumbents' jobs. Many researchers accept the accuracy of these ratings and consider any differences between multiple raters for the same job as random error or "noise" (Harvey, 1991; Landy & Vasey, 1991). Thus, low interrater reliabilities are interpreted as an indication of measurement error (Borman, Dorsey, & Ackerman, 1992). For example, in a study by Harvey & Lozada-Larsen (1988), a task inventory was developed based on ratings from job incumbents. The task descriptions given to the subjects in the study were based on the average incumbent relative-time-spent rating for the task items. However, the study did not question the accuracy of such ratings as indications of actual behavior.

Another example of how task frequency ratings have been used is in analyzing the successful transfer of trained tasks from a training program to the job. Recent approaches to examining transfer of training have focused on the extent to which trainees obtain the opportunities to perform trained skills on the job (Ford, Quiñones, Sego, & Sorra, 1992; Quiñones, Ford, Sego, & Smith, in press). In order to examine the various dimensions of the "opportunity to perform" construct (breath, activity level, and task type), trainees are asked, for example, to identify the number of times they have performed each of the sampled tasks within a given time. However, these studies have not examined the accuracy of these ratings.
In order to overcome any individual bias in job description, it has become common practice to aggregate the individual incumbent responses (Landy & Vasey, 1992). Researchers hope to eliminate the disagreement in job analysis ratings by computing job mean profiles for the data (Harvey, 1991). Sanchez and Levine (1989) assert that job analysis researchers aggregate data across subjects because it is presumed that the aggregation process will remove individual biases. Although it has become common practice to employ such methods, researchers have failed to examine the sources of the discrepancies among raters' responses. Before one can aggregate data, one must understand the processes involved in obtaining such data. Therefore, Harvey (1991, p. 118) states that "the solution to the problem of job analysis ratings disagreement is not to simply compute a mean and hope that errors will cancel out".

Agreement: Perception or Reality?

The completion of a job analysis typically involves making various task analysis ratings in which one must rate the frequency of performance of given tasks. Past job analysis research aimed at examining the accuracy of task frequency ratings has been unable to conclude whether discrepancies in task analysis responses are due to differences in respondents' perceptions of the job or due to actual differences in what incumbents do on the job (Taylor, 1978, in Mullins & Kimbrough, 1988; Schmitt & Cohen, 1989). Schmitt and Cohen (1989) recognized such a limitation in their research in stating that "there is no way of determining whether the members of some subgroups were giving different task ratings because they had been assigned different tasks or whether the response differences were perceptual differences common to all members of a particular subgroup" (p.103). Consequently, if researchers were to gather
information on actual task performance, one could determine if discrepancies in task frequency ratings are a reflection of real differences in task performance or due to varying perceptions of the job. Without such an examination, it will be difficult to fully examine the factors that impact the accuracy of task ratings and develop ways to ultimately improve the accuracy of job analysis data.

Researchers have noted that the results of empirical studies on the job analysis rating process are not optimistic (Harvey, 1991). Studies have found that incumbents and supervisors often provide different task ratings for the same job and that raters with different levels of experience provide varying task analysis responses (Harvey, 1991). Although past studies have examined these issues, one might come to different conclusions if the research made use of accuracy indices that compared task ratings to actual task performance.

This study attempts to utilize an objective measure of task frequency to reexamine the accuracy of task performer and observer ratings and reevaluate the impact of experience on task frequency rating accuracy. Consequently, the main goal of this research is to examine these past task analysis issues in a laboratory setting in order to obtain objective true score measures that help determine if discrepancies in task frequency ratings are due to differences in actual task frequency or differences in respondents' conceptualizations of the job.

Although this study examines the accuracy of task frequency ratings from a task analysis perspective, one should note that there are many uses of task frequency ratings beyond that of job analytic procedures. For example, in performance appraisal research using behavioral rating scales, raters are sometimes required to estimate the frequency of targets'
behavior. When examining the usefulness of behavioral expectation scales over other formats, Zedeck, Kafry, & Jacobs (1976) reported that some studies use frequency of occurrence as the frame of reference for responses. Likewise, Murphy, Garcia, Kerkar, Martin, & Balzer (1982) used task frequency ratings to assess observational accuracy and to examine the relationship between accuracy in observing behavior and accuracy in rating performance.

In addition to their use in job analysis and performance appraisal research, task frequency ratings serve additional purposes. As stated earlier, task frequency ratings are utilized when examining the opportunity to perform trained skills on the job (Ford et al., 1992; Quiñones et al., in press). Furthermore, to investigate the factors that affect task performance, Lance, Hedge, & Alley (1989) used "the number of times performed" measure as an indication of task experience to predict job performance. In social and psychological research, subjects are often asked to report the frequency with which they encounter a particular experience or engage in a particular behavior (Schwarz, Bless, Bohner, Harlacher, & Kellenbenz, 1991). Also, in clinical research and practice, practitioners often use self-report measures to assess the frequency of patients' behaviors (e.g. Kassielke & Hänsgen, 1982; von Zerssen & Koeller, 1975, 1976). Consequently, it can be seen that an examination of task frequency rating accuracy is not only relevant to industrial/organizational research on task analysis and performance appraisal, but it is also important in social and clinical investigations. Although this study maintains a task analysis perspective, the findings may be pertinent to improving research and practice in a variety of psychological disciplines.
Task Frequency Rating Accuracy

Although some investigators have examined the factors that impact the accuracy of task analysis ratings (Arvey, Davis, McGowen, & Dipboye, 1982; Arvey, Passino, Lounsbury, 1977; Borman et al., 1992; Conley & Sackett, 1987; Cornelius & Lyness, 1980; Green & Stutzman, 1986; Hahn & Dipboye, 1988; Harvey & Lozada-Larsen, 1988; Hazel, Madden, & Christal, 1964; Landy & Vasey, 1991; Meyer, 1959; Mullins & Kimbrough, 1988; O'Reilly, 1973; Schmitt & Cohen, 1989; Smith & Hakel, 1979), there are methodological and theoretical issues surrounding the accuracy measures used as criteria in such research. "Accuracy of measurement is a term used to describe both the strength and kind of relation between one set of measures and a corresponding set of measures (e.g., true scores) considered to be an accepted standard for comparison" (Sulsky & Balzer, 1988, p.497).

In research on task analysis rating accuracy, the standard for comparison is not an actual (objective) true score. The true score measures used to compute an accuracy index tend to be either the average rating provided by all the task analysis respondents (e.g. Conley & Sackett, 1987; Cornelius & Lyness, 1980; Green & Stutzman, 1986; Harvey & Lozada-Larsen, 1988) or the average rating of a group of expert raters (e.g. Borman, 1977; Hahn & Dipboye, 1988; Murphy et al., 1982). In assessing the accuracy of task frequency ratings, for example, neither strategy would involve a comparison to an objective true score indicating the number of times a task was really performed. The validity of the conclusions drawn by task analysis accuracy studies depends on the "the degree to which the operationally defined true scores approximate the correct true scores" (Smither, Barry, & Reilly, 1989, p.143). Consequently, if such perceived
true scores are not accurate reflections of what occurs on the job, then research using such methodologies may be suspect.

Although researchers examining performance appraisal rating accuracy have utilized objective true score measures of performance as the standard of comparison (e.g. Smither et al., 1989), task analysis accuracy studies have neglected to so. Reexamination of the factors that are thought to impact task analysis frequency ratings may yield different results if an objective measure of frequency is employed in computing accuracy indices. Hence, this study attempts to analyze the accuracy of task frequency ratings with the use of objective task frequency as the standard of comparison. Such a comparison will allow one to determine if discrepancies in task frequency ratings are due to differences in actual task performance or differences in respondents' perceptions of the job.

The manner in which accuracy has been measured in task analysis research continues to be a problem. Although the operational definition of task analysis rating accuracy includes a comparison of the raters' ratings with some perceived true score, such accuracy measures are, in essence, measures of agreement, not accuracy. Research aimed at examining the factors that impact the accuracy of task analysis ratings tend to employ either Cronbach's (1955) ratings components (elevation, differential elevation, stereotype accuracy, and differential accuracy) (Conley & Sackett, 1987; Hahn & Dipboye, 1988; Harvey & Lozada-Larsen, 1988; Murphy et al., 1982) or reliability coefficients (Cornelius & Lyness, 1980; Shaffer, Saunders, & Owens, 1986; Smith & Hakel, 1979). Others utilize the ANOVA paradigm to examine group differences in task analysis responses (Arvey et al., 1977; Arvey et al., 1982; Landy & Vasey, 1991; Mullins & Kimbrough, 1988; Schmitt & Cohen, 1989). Regardless of
which approach is used, the true score to which raters' responses are compared is usually a mean or an average of all the respondents' ratings. Whether the approach employed utilizes the average rating score from either incumbents, supervisors, or expert raters as the perceived true score, such measures of accuracy only indicate the extent to which an individual's response deviates from the average. Consequently, an individual is deemed "accurate" if his/her rating is close to, or is in agreement with, the average rating.

**Cronbach Approach.** In determining the accuracy of task ratings, Cronbach's (1955) components of rating accuracy are based on deviations from average ratings. "The rater whose overall average is close to the (group's) overall average true score will tend to be more accurate than one whose average rating is far from the true score average" (Murphy et al., 1982, p. 321). Hence, accuracy per se is a measure of similarity or agreement between respondents' ratings and the corresponding average. For example, Green and Stutzman (1986) examined the accuracy of job analysis respondents by measuring the gap between an employee's judgment on a given task and his/her unit's centroid on that task. They conclude that the respondents are not equally accurate in their ratings, yet they state that their criteria "assessed similarity between an individual's ratings and the population's mean ratings" (p.543). One cannot determine if the mean aggregate rating, the centroid, is an accurate reflection of what occurs on the job. If the centroid is not representative of actual frequency of performance, for example, and one concludes that an individual's response is "accurate", according to Cronbach's method, a misleading conclusion may be reached. The individual may give a rating that is similar to the
mean, yet the response may not be comparable to the objective frequency of performance.

**Reliability Approach.** The use of reliability coefficients as an indication of accuracy poses the same problem as Cronbach's method. Those studies that compute reliabilities using correlation coefficients are essentially measuring agreement between raters' responses, not accuracy (Cornelius & Lyness, 1980; Shaffer et al., 1986; Smith & Hakel, 1979). For example, Shaffer et al. (1986) examined the *accuracy* of biographical data, yet they report that lower correlations are an indication that *agreement* is lower. Furthermore, using correlation coefficients, Smith and Hakel (1979) found that "naive job analysts provided with only a job title can rate *as well as* analysts using extensive information" (Harvey & Lozada-Larsen, 1988, p.457). Finding no significant differences in reliabilities between the different rater sources, however, does not necessarily mean that the different raters were equally accurate or that incumbents performed the same tasks the same number of times. The correlation between novice and expert raters may be high, yet both may be inaccurate when compared to an objective measure of task performance. Consequently, such *agreement* will not yield any information about the accuracy of the responses.

**ANOVA Approach.** Besides using Cronbach's method and reliability coefficients to assess rating accuracy, some studies use an ANOVA approach to examine group differences in task analysis responses (Arvey et al., 1977; Arvey et al., 1982; Landy & Vasey, 1991; Mullins & Kimbrough, 1988; Schmitt & Cohen, 1989). Schmitt and Cohen (1989) "used ANOVAs to determine whether there were mean differences in task ratings across different respondent groups" (p.97). Such a method,
however, does not allow one to examine the accuracy of the task ratings. Although it can be determined if the mean task rating for a given group is significantly different than the mean task rating for another group, one cannot make a conclusion concerning the extent to which the ratings are representative of actual behavior on the job.

When researchers utilize Cronbach's rating components, reliability coefficients, or ANOVAs to assess accuracy, one is unable tease apart such measures of agreement as either reflections of reality or perception. If researchers continue to estimate agreement, as opposed to measuring accuracy, and continue to employ perceived measures of frequency that may not be indicative of actual task behavior, one will not be able to determine if discrepancies in task frequency ratings are due to differences in respondents' perceptions of the job or due to actual differences in what incumbents do on the job.

Memory Processes in Frequency Estimation

Although task analytic approaches typically involve asking SMEs to rate the frequency with which they perform certain tasks, task analysis research has neglected to investigate how individuals estimate frequency of occurrence. The recollection of how many times a given task, event, or stimuli occurred is essentially a memory task. Consequently, the topic of frequency estimation and its representation in memory has been heavily documented in the memory literature (Begg, Maxwell, Mitterer, & Harris, 1986; Bradburn, Rips & Shevell, 1987; Bruce & Van Pelt, 1989; Greene, 1984; Greene, 1986; Greene, 1989; Hasher & Zacks, 1984; Hintzman & Block, 1971; Hintzman & Stern, 1978; Hockley, 1984; Howell, 1973; Johnson, Peterson, Chua Yap, & Rose, 1989; Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978; Maki & Ostby, 1987; Pitz, 1976;
Rowe, 1974; Rowe & Rose, 1977; Schwarz & Bienias, 1990; Watkins & LeCompte, 1991; Zacks, Hasher, & Sanft, 1982). Such research tends to focus on either the encoding of event frequency (Begg et al., 1986; Greene, 1984; Hasher & Zacks, 1984; Hintzman & Block, 1971; Hintzman & Stern, 1978; Hockley, 1984; Howell, 1973; Zacks et al., 1982) or the retrieval of frequency estimates (Begg et al., 1986; Bruce & Van Pelt, 1989; Bradburn et al., 1987; Huttenlocker, Hedges, & Bradburn, 1990). Since verbal stimuli (e.g. word lists), numbers, and digits tend to be the repeated event that subjects must estimate, memory researchers can determine the accuracy of subjects' frequency estimates by comparing their estimates to the actual frequency of presentation.

Although very few studies on frequency estimation have used natural events or behaviors as their independent variable (see Lichtenstein et al., 1978; Schwarz & Bienias, 1990 for exceptions), the methodologies and paradigms used in the memory literature can be applied to task analysis accuracy research. Consequently, the present research borrows from the memory literature and applies its various models of frequency estimation to a task analysis perspective.

When examining the factors that impact the accuracy of task frequency ratings, task analysis researchers might gain greater insight into the task frequency estimation process if they utilized the memory paradigms, along with actual (objective) true score measures of task performance. An exploration of both the task analysis research and the memory literature suggests that one might reconsider some of the ways in which task analysis approaches are currently conducted.
Task Engagement

In the industrial/organizational literature, several studies have examined the impact of task engagement on the accuracy of task ratings by comparing rating responses between individuals who perform the job (i.e. incumbents) and those who observe the job (i.e. supervisors and job analysts). Some research has investigated the extent to which incumbents and supervisors agree on the tasks that comprise the incumbents' job (Cornelius & Lyness, 1980; Hazel et al., 1964; Levine, Flory, & Ash, 1977; Meyer, 1959; O'Reilly, 1973; Smith & Hakel, 1979). Both Meyer (1959) and O'Reilly (1973) found high disagreement between incumbent and supervisors regarding the responsibilities of the incumbent.

Some researchers have suggested that individuals, such as supervisors, who only observe task performance may not have adequate knowledge of what occurs on the job compared to individuals, such as incumbents, who actually perform the tasks that comprise the job. For example, Hazel et al. (1964, p. 78) concluded that "the generally moderate agreement (or disagreement) found between supervisors and workers concerning the nature of the worker's job suggests that a supervisor may not know precisely what any one subordinate does task by task". In contrast, Smith & Hakel (1979) concluded that students, incumbents, supervisors, and job analysts are equally accurate in their job analysis ratings.

Additional research aimed at assessing the extent to which either incumbents or supervisors can provide accurate ratings also presents varying findings (Murphy et al., 1982; Repp, Nieminen, Olinger, & Brusca, 1988; Russell, Mattson, Devlin, & Atwater, 1990; Shaffer et al., 1986) Thus, the results of the above studies on task engagement present a
mixed picture concerning the convergence in task frequency ratings given by incumbents and supervisors (Smith & Hakel, 1979).

Since the recollection and estimation of how many times a given task was performed is essentially a memory task, task analysis researchers might gain insight into these mixed findings by utilizing paradigms suggested by the memory research. Some memory investigators have proposed that the encoding of frequency information is an automatic process (Zacks et al., 1982; Hasher & Zacks, 1984). They suggest that people automatically learn how often an event has occurred and that such automaticity is insensitive to various processes such as intentionality, practice, and strategy manipulations. However, several researchers have demonstrated that the encoding of event frequency is not automatic, but rather it is a product of controlled processes. Such investigators have found that intentionality of learning, practice, and semantic processing impact memory for frequency (Begg et al., 1986; Fisk & Schneider, 1984; Greene, 1984; Greene, 1986; Hockley, 1984; Maki & Ostby; 1987; Rowe, 1974; Rowe & Rose, 1977). Therefore, this suggests that various controlled processes can impact the accuracy of individuals' frequency estimates.

When examining the impact of strategy manipulation on event frequency, some memory researchers have found that individuals demonstrate greater accuracy in frequency estimation when they employ deeper levels of processing during encoding (Fisk & Schneider, 1984; Greene, 1986; Maki & Ostby; 1987; Rowe, 1974; Rowe & Rose, 1977). For example, Maki and Ostby (1987) manipulated level of processing by having subjects in the shallow processing condition indicate which presented words were relatively long or short and having subjects in the
semantic processing condition indicate which words were easy or difficult to imagine. They found that when subjects utilize a semantic processing strategy, they demonstrate improved performance on a frequency discrimination task. Their results also indicate that the level of processing manipulation impacts upon frequency estimates as a function of actual frequency presentation.

Although there is not much research that has found that individuals who engage in a task will employ deeper levels of processing, some cognitive research suggests that the cognitive processes involved in performing a task may be different than those associated with observing a task (Ghnogi, 1986; Kohl & Shea, 1992; Weir & Leavitt, 1990). For example, Ghnogi (1986) found that, on a memory task, subjects who perform the task demonstrate higher learning scores on a subsequent retention test than subjects who only observe the task. Furthermore, some researchers suggest that differences in actual practice and observational practice may be due to differences in levels of processing. MacKay (1981) argued that "whereas actual practice activates nodes throughout the hierarchy (as represented in the motor system), mental practice, because of its apparent lack of overt activity, only activates nodes at the mental level" (as cited in Kohl & Shea, 1992, p.258). Consequently, one might presume that performing a task, as opposed to observing a task, results in deeper levels of processing.

Since the above research suggests that performers of tasks may process activities at a deeper level than those who only observe the task and that deeper levels of processing leads to more accurate frequency estimates, one might submit that task performers, as opposed to observers, will generate more accurate frequency estimates. Consequently, this study
proposed that if incumbents engage in a deeper level of processing while performing a given task, they will subsequently generate more accurate frequency estimates than supervisors who only observe the task, employing a shallower processing strategy. In sum, this study made the following hypothesis concerning levels of processing and frequency estimation:

_Hypothesis #1:_ Performers who engage in task activities will generate more accurate frequency estimates than observers who only watch others perform the tasks.

**Task Familiarity.**

Although many studies have attempted to examine the effect of familiarity or experience on task analysis ratings, the findings are quite mixed (Borman et al., 1992; Conley & Sackett, 1987; Cornelius & Lyness, 1980; Green & Stutzman, 1986; Hahn & Dipboye, 1988; Landy & Vasey, 1991; Mullins & Kimbrough, 1988; Schmitt & Cohen, 1989; Smith & Hakel, 1979). Landy and Vasey (1991) and Borman et al. (1992) both conclude that incumbent experience impacts task ratings. They found that as levels of experience change, the frequency of the tasks conducted change. In contrast, several studies have discovered that job knowledge and experience in conducting job analyses does not affect subsequent task ratings (Arvey et al., 1982; Conley & Sackett, 1987; Cornelius & Lyness, 1980; Green & Stutzman, 1986; Mullins & Kimbrough, 1988; Schmitt & Cohen, 1989; Smith & Hakel, 1979). For example, when comparing incumbents, supervisors, job analysts, and college students, Smith and Hakel (1979, p. 691) conclude that "the source of responses of a job analysis inventory makes little practical difference."
Other investigators, however, have proposed that the source of the
task ratings is of primary importance (Cornelius, Denisi, & Blancoe, 1984; Harvey & Lozada-Larsen; 1988; Smither et al., 1989; Sulsky & Balzer, 1988). Harvey and Lozada-Larsen (1988) conclude that job-naive raters, who are unfamiliar with the job being analyzed, should not substitute job content experts because the amount of information available to raters significantly affects task analysis accuracy. Cornelius et al. (1984) also found that job experts and college students do not provide equivalent job analysis ratings. In the performance appraisal research, experts are found to be more accurate than nonexperts in providing performance ratings (Smither et al., 1989; Sulsky & Balzer, 1988). Consequently, expert ratings serve as true score estimates in appraisal accuracy research.

The above empirical studies aimed at examining the impact of familiarity or experience on task ratings may have yielded mixed findings due to either differences in respondents' perceptions of the job or differences in actual task performance. Consequently, the response differences found between job-naive raters and job content experts may be due to differing job conceptualizations or due to differences in actual frequency of performance between incumbents. Since the investigators failed to take actual task frequency of performance into account, one cannot make definitive conclusions concerning the effect of familiarity on the accuracy of task frequency ratings.

Task analysis research requires that SMEs have direct, immediate experience with the job for a period of time in which to become familiar with all of its tasks (Thompson & Thompson, 1982). Consequently, SMEs (incumbents, incumbent supervisors, and job analysts) are presumed to be in the best position to provide accurate and complete knowledge about the
tasks that comprise a given job (Landy & Vasey, 1991; Green & Stutzman, 1986). The memory literature, however, might suggest the opposite. Memory research on autobiographical events has shown that people often confuse and forget everyday events, especially if similar happenings are encountered frequently (Barclay & Wellman, 1986; Bruce & Van Pelt, 1989; Robinson & Swanson, 1990; Shaklee & Mims, 1982). Robinson and Swanson (1990) report that there is evidence to suggest that individuals have a more difficult time recalling specific events when the event is one of many similar incidents than when it is unique. Furthermore, Bruce and Van Pelt (1989) found that when subjects attempted to estimate the frequency of certain events that occurred on a 47-day bicycle tour, their estimates were more accurate for those events that occurred less frequently.

If SMEs, such as incumbents, are constantly engaging in the tasks that comprise their job, according to the memory literature, one would expect them to be highly inaccurate at remembering how many times a given task was performed. Since the memory literature suggests that routine aspects of the same event are less likely to be accurately recalled (Hudson, Fivush, & Kuebli, 1992), one might question the use of job content experts, who are highly familiar with the tasks being rated, in obtaining task frequency ratings. Contrary to what has been suggested in the industrial/organizational literature, this study proposed that job-naive raters may be more reliable in providing accurate task frequency ratings than SMEs, who are highly familiar with the tasks being rated.
Hypothesis #2: Raters with novel experience on a task will generate more accurate task frequency estimates compared to raters who are more familiar with a task.

Task Frequency Rating Format

Although it has become common practice to employ relative rating scales in task analysis research (Harvey, 1991), cognitive research on frequency estimation support the use of open-answer formats when assessing behavioral frequencies (Schwarz & Bienias, 1990; Schwarz et al., 1991). One reason why some performance appraisal researchers advocate the use of relative frequency rating scales is because raters' reports of the exact number of times each behavior occurs shows extremely low levels of interrater agreement (Murphy et al., 1982). Such discrepancies, however, may be representative of actual differences on the job. If the low interrater reliabilities are accurate reflections of task frequency, avoidance of absolute frequency estimates, in favor of relative frequency rating scales, may only hinder researchers attempts to directly examine rating accuracy.

Some cognitive investigations on frequency estimation support the use of absolute frequency assessments (Schwarz & Bienias, 1990; Schwarz et al., 1991). Schwarz & Bienias (1990) assert that the range of response alternatives (e.g. 'rarely', 'sometimes', and 'frequently') may serve as a frame of reference and ultimately affect respondents' judgments. Respondents may use the average frequency as a standard of comparison or may use the range of the response scale to anchor subsequent ratings (Schwarz et al., 1991). In addition, Schwarz et al. (1991) maintain that a large body of evidence illustrates that "the use of vague qualifiers is highly
problematic because respondents' understanding of terms such as 'rarely' or 'sometimes' shows considerable variation and different respondents use different terms for the same absolute frequency" (p.41). Consequently, such researchers advocate the use of open-ended frequency assessments.

Furthermore, although the use of relative rating scales is more common in task analysis research, such practices may reduce researchers' ability to make inferences beyond the job being investigated. Relative frequency scales only provide within-job-relative task data that can be used to make comparisons within a given job (Harvey, 1991). Harvey (1991) advocates the use of absolute frequency scales because they "provide a common metric that allows cross-job-meaningful comparisons to be drawn" (p. 98).

Since this study investigates the accuracy of task frequency ratings, it is pertinent to examine the extent to which absolute and relative rating scales affect individuals' ability to give accuracy information. By contrasting raters' responses on relative frequency ratings to responses on absolute frequency ratings, this study hopes to provide empirical evidence to illustrate whether certain individuals are better at giving consistent information across different rating formats.

Method

Participants

Sixty-four Rice undergraduates participated in the study in exchange for course credit. Fifty-two percent of the participants were women and 48% were men. The median age of the subject population was 19.

Design

The design was a 2 x 2 analysis of variance. The independent variables were task engagement (performer and observer) and task
familiarity (novel and familiar). The frequency rating questionnaires given to all the subjects included both absolute and relative rating scales.

**Materials and Equipment**

A Meccano® Erector® set, in which toy models are assembled by connecting metal pieces together with screws and nuts, was used for this research. Pieces used in the experiment included, among others, various sized screws, square nuts, tires, different sized perforated strips, wheels. An open-ended wrench, an allen wrench, and an instruction manual with an illustrated model were also provided.

Groups of four undergraduates participated in an organizational simulation in which the group members together represented a small toy manufacturing company (simulation adapted from Shaw's (1963) Group Task Inventory as described in Vanderslice, Rice, & Julian, 1987). The groups had to buy toy parts, construct toys according to model specifications, and sell the toys to the experimenter. The groups also had to complete a series of in-basket tasks such as creating a company name, developing a vision statement, filling out a supply request form, updating an inventory sheet, and recording sales on a financial record sheet (see Appendix A for the in-basket tasks). Each group that engaged in the simulation was videotaped so that an additional group of four subjects could observe their behavior.

After engaging in the simulation, group members were asked to complete a questionnaire to measure their perceptions of the frequency of performance of given task elements. Subjects who observed the group members performing the simulation also completed the same questionnaire. This questionnaire was used to measure observers' perceptions of the
frequency with which the group members engaged in the given job elements.

The tasks selected for examination in the questionnaires were 21 individual task elements that composed the simulation. The tasks were chosen to be unique behaviors that could be observed. In addition, bogus tasks that were not actually performed during the simulation were also included in the questionnaires in order to create an additional measure of accuracy.

**Manipulations**

**Task Engagement.** The task engagement factor was manipulated by having half the subjects engage in the organizational simulation while the other half watched a videotape of people engaged in the simulation. Consequently, each session of the experiment consisted of either four subjects assigned to a Performer condition or four subjects assigned to an Observer condition.

**Task Familiarity.** The task familiarity factor was manipulated by having subjects assigned to a Novel condition perform (or observe) the organizational simulation once before completing the task frequency questionnaires. Subjects assigned to a Familiar condition performed (or observed) the simulation three times before completing the task frequency questionnaires. It was believed that the toy company simulation would no longer be novel to those subjects observing or performing the simulation on more than one occasion. Subjects in the Familiar condition rated the behavior that occurred during the last simulation. Thus, the subjects in the Familiar condition were instructed to *only* rate the behavior that occurred in the third session of the simulation, not the first or the second session.
**Frequency Rating Format.** The questionnaires contained both an absolute frequency rating scale and a relative frequency rating scale. For the absolute frequency rating scale, participants were instructed to indicate the number of times they engaged in (or observed their target Performer engage in) each of the given 26 tasks. For the relative frequency rating format, participants were instructed to indicate how often they engaged in (or observed their target Performer engage in) each of the given 26 tasks relative to the other tasks. The absolute and relative frequency rating scales were counterbalanced on the rating forms so as not to create an effect of rating scale order.

Both the absolute and relative rating scales contained both real tasks and bogus tasks. Bogus tasks which were not actually performed during the simulation were dispersed among the "real" tasks in order to examine the extent to which individuals could identify those tasks that were and were not performed. See Appendix B for the absolute and relative rating scale questionnaires.

**Procedure**

Upon entering the experiment, half of the subjects were given instructions on the simulation while the other half received instructions on observing the simulation. During the course of the procedure, all subjects completed both a mechanical reasoning test and a cognitive ability test. Upon completion of their (last) simulation, subjects were instructed to complete frequency rating questionnaires. Participants were then debriefed on the true purpose of the experiment and thanked for their time. See Figure 1 for a general outline of the experimental procedure.
Figure 1
Procedure Outline

<table>
<thead>
<tr>
<th>Session 1</th>
<th>Test</th>
<th>Session 2</th>
<th>Test</th>
<th>Session 3</th>
<th>Questionnaires</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FAMILIAR:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>simulation</td>
<td>mechanical reasoning test</td>
<td>simulation</td>
<td>cognitive ability test</td>
<td>simulation</td>
<td>complete rating questionnaires (rate Session 3 only)</td>
</tr>
<tr>
<td>(observe or perform)</td>
<td></td>
<td>(observe or perform)</td>
<td></td>
<td>(observe or perform)</td>
<td></td>
</tr>
<tr>
<td><strong>NOVEL:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>mechanical reasoning test</td>
<td>-----</td>
<td>cognitive ability test</td>
<td>simulation</td>
<td>complete rating questionnaires</td>
</tr>
<tr>
<td></td>
<td>(observe or perform)</td>
<td></td>
<td>(observe or perform)</td>
<td>(observe or perform)</td>
<td></td>
</tr>
</tbody>
</table>
Simulation. Participants in the Performer condition were instructed to imagine themselves and their group members as employees in a toy manufacturing company. Their goal was to maximize profit for their company. Performers were informed that in order to accurately examine their actions, they had to be videotaped and that following their final simulation they would have to answer questions about how well their organization functioned.

After an explanation of the general nature of the simulation, performers were given additional instructions on what to accomplish during the two phases of the simulation. During the planning phase, performers were given 8 minutes to create a name for the company, determine the company's vision or goal, and fill out a supply request form. Upon completion of the planning phase, performers watched a brief videotape demonstrating how to put various pieces together using the tools provided. Subjects then received the supplies from the experimenter and began their 15 minute production phase. Performers had to assemble toy scooters, update an inventory sheet, check the quality of the completed scooters, sell the toys, and update a financial record sheet. Sales were made by giving completed scooters to the experimenter who in turn gave the group members $5 of play money for each toy that met required standards. Participants were informed that only completed toys could be sold and that any leftover supplies would be counted as inventory costs that would be subtracted from their overall profits. When the production phase was complete, the group's money, toy parts, partially completed toys, and financial records were collected.

Simulation Observation. Participants in the Observer condition were informed that the purpose of the experiment was to examine how the
videotaped participants performed as a group in an organizational simulation. Each observer was assigned to specifically watch one of the performers on the video. Observers were told that, after watching the video, they would have to evaluate their target person's behavior toward maximizing the company's profits and answer questions about how well the organization functioned.

After an explanation of the general nature of the simulation, observers were given additional instructions on what the performers had to accomplish during the two phases of the simulation. Observers were given copies of all the in-basket tasks, as well as a copy of the model building instruction sheet and an actual completed toy scooter. These were provided so that the observers could follow the actions of the performers. The observers also watched the same instruction video shown to the performers so that they could better understand how the performers were using the tools to construct the toys. Lastly, prior to viewing the actual simulation, observers watched a few seconds of both the planning and production phases of a demonstration simulation so that they could get an idea of what the simulation would look like.

**Measures.**

**Actual Frequency.** The actual frequency with which performers engaged in various task elements was determined by analyzing the videotaped sessions. Three trained raters, who were blind to the experimental conditions, watched the videos to objectively count the number of times the subjects performed each of the 21 tasks.

**Reported Absolute Frequency.** Participants' rated their reported absolute frequency by indicating the number of times they executed (or
observed their target Performer execute) certain tasks during the simulation.

**Reported Relative Frequency.** Subjects' rated their reported relative frequency by indicating how often they executed (or observed their target Performer execute) various task elements relative to how often other task elements were performed. Participants indicated the relative frequency of each task using a 5-point Likert scale ranging from 1 (*very infrequently*) to 5 (*very frequently*).

**Individual Difference Measures.** This study also examined the effects of various individual difference variables (e.g. cognitive ability, mechanical reasoning, self efficacy, and performance). The Wonderlic Personnel Test was used to assess cognitive ability while the mechanical reasoning section of the Differential Aptitude Test was used to assess mechanical reasoning ability. Of the 50 original items included in the mechanical reasoning section of the Differential Aptitude Test, only 35 items were randomly selected to be used in this study (α = .60). Item and factor analyses resulted in the elimination of 11 items. The resulting scale reliability was α = .70.

Furthermore, self efficacy was measured using a 10-item questionnaire on 5-point Likert scale (α = .85), and subjects' previous experience using an Erector® set was assessed by asking subjects whether or not they had played with or previously owned an Erector® set (α = .90). See Appendix C for theses measures.

A measure of group performance was created for each group of performers who engaged in the simulation. Each leftover supply was considered to be worth 2¢ while each completed scooter was worth $5. The total number of leftover supplies was counted and then multiplied by
.02. This value was then subtracted from the total amount of money the group had received for each successfully completed toy scooter.

Lastly, in addition to the frequency rating questionnaires, subjects' rated the importance of each of the tasks as well as their interest in performing each of the tasks. Subjects indicated the importance of each of the tasks toward achieving the group's overall goal or vision using a 5-point Likert scale ranging from 1 (very minor importance) to 5 (extreme importance). In addition, subjects indicated their interest in performing each of the tasks using a 5-point Likert scale ranging from 1 (not interesting at all) to 5 (very interesting). See Appendix D for both rating formats.

Dependent Variables and Analytic Strategy

Analyses were performed in order to examine the extent to which subjects were accurate at generating task frequency estimates. The hypotheses were tested by creating various measures of accuracy such as (a) absolute deviations, (b) relative deviations, (c) detection accuracy, and (d) rank order accuracy. Additional analyses were performed to examine the extent to which subjects gave consistent information using various rating formats and to examine the impact of various individual differences on subjects' task frequency responses.

Absolute Deviations. To examine the accuracy of subjects' task frequency ratings, the dependent measure for several ANOVAs was a measure of deviation. Subjects were considered "inaccurate" in their frequency ratings to the extent that their reported absolute frequency ratings deviated from the actual task frequency. The following equation was used to determine the extent to which subjects' reported absolute frequency ratings deviated from actual frequency of performance:
\[ AD = \sum_{i=1}^{21} |A_i - R_i| \]

where,

- \( AD = \) Absolute Deviations
- \( A_i = \) Actual number of times a performer engaged in task \( i \)
- \( R_i = \) Subjects' reported number of times performing task \( i \)

The differences between subjects' reported absolute frequency rating and their actual task frequency of performance were summed across all 21 real tasks. This measure was then used as one of the dependent variables in the study.

In order to take into account the number of times the given tasks were actually performed, hierarchical regression was performed on the absolute deviations with the total number of times the 21 tasks were performed entered first. The use of hierarchical regression allowed an examination of the effects of task engagement and task familiarity on the accuracy of the frequency ratings after removing the variance due to the actual number of times tasks were performed.

**Relative Deviations.** To further examine the extent to which subjects' frequency estimates departed from the number of times the tasks were actually performed, a measure of relative deviations was computed. This measure was used to determine the accuracy of the participants' reported task frequency responses relative to the number of times the tasks were performed. As stated previously, subjects were considered "inaccurate" in their frequency ratings to the extent that their reported absolute frequency ratings deviated from the actual task frequency. The following equation was utilized to determine the degree of relative deviation in the subjects' ratings:
\[
RD = \frac{\sum_{i=1}^{21} |A_i - R_i|}{\sum_{i=1}^{21} A_i}
\]

where,

RD = Relative Deviations

\(A_i\) = Actual number of times a performer engaged in task \(i\)

\(R_i\) = Subjects' reported number of times performing task \(i\)

The absolute deviations between subjects' reported absolute frequency and their actual task frequency of performance were summed across all 21 tasks and then divided by the total number of times the 21 tasks were performed. The variable ranged from zero to infinity, where larger numbers indicated greater deviations.

**Detection Accuracy.** An additional measure of accuracy that was utilized involved the application of signal detection theory (Swets & Pickett, 1982; Lord, 1985). Having included both real tasks and bogus tasks on the rating questionnaires, an estimate of detection accuracy was computed from both hit rates and false-alarm rates. The hit rate referred to the proportion of real tasks correctly identified as having been performed or not performed as well as the proportion of bogus tasks correctly identified as having not been performed. The false-alarm rate referred to the proportion of real tasks falsely recognized as having been performed or not performed as well as the proportion of bogus tasks falsely recognized as having been performed. Detection accuracy was determined by taking the difference between the hit rate and the false-alarm rate.
\[ DA = \frac{N_c}{N} - \frac{N_i}{N} \]

where,

\( DA = \) Detection Accuracy
\( N_c = \) Number of tasks correctly identified
\( N_i = \) Number of tasks incorrectly identified
\( N = \) Total number of tasks
\( \frac{N_c}{N} = \) Hit Rate
\( \frac{N_i}{N} = \) False-Alarm Rate

This measure of detection accuracy was used as the dependent variable in an ANOVA to determine if subjects in the various experimental conditions differed in their ability to detect those tasks that were and were not performed. This helped to identify those subjects that could detect frequency of occurrence above the level of noise that was present from the inclusion of bogus tasks on the frequency questionnaire.

**Rank Order Accuracy.** Analyses were performed to examine how well subjects were able to provide frequency estimates that reflected the true rank order of the number of times the tasks were actually performed. The rank order of tasks reflected by subjects' absolute frequency ratings and relative frequency ratings were matched against the actual rank order of task frequency. Correlations were computed between subjects' actual task frequency of performance and both their absolute frequency ratings and their relative frequency ratings. Larger correlations indicated that tasks that were actually performed more often were subsequently rated as having been performed more often.
Since correlations tend to have skewed distributions, the two rank order correlations were transformed to Fisher's z equivalents. The Fisher's z equivalents were then used in subsequent ANOVAs to determine if subjects' ability to give frequency estimates that reflected the actual rank order of task frequency was differentially affected by experimental condition. Furthermore, these correlations were used to determine if subjects were better able to capture the actual rank order of tasks using an absolute rating scale or a relative rating scale.

**Ancillary Analyses.** A correlation between subjects' reported absolute frequency ratings and their reported relative frequency ratings was also computed. Using this correlation as the dependent variable in an ANOVA, subjects' ability to give consistent information on both types of rating scales across experimental conditions was examined.

To examine the impact of various individual difference variables on the accuracy of subjects' frequency judgments, the different accuracy measures were correlated with individual difference measures (i.e. cognitive ability, mechanical reasoning, self efficacy, and performance). Regressions were also performed on several accuracy measures to determine the amount of variance in accuracy scores explained by the individual difference variables.

Lastly, to determine if the importance or the level of interest assigned to the tasks impacted the accuracy of the frequency estimates, the importance and interest ratings were correlated with the various accuracy measures.
Results

Determining Actual Task Frequency

To determine the actual frequency of task performance, three trained raters watched the videotaped simulations to objectively count the number of times each task was performed. Prior to coding, the three raters spent 2 one hour-long sessions discussing what behaviors constituted the execution of a given task. An additional one hour session was spent practicing coding a demonstration simulation and comparing raters' responses.

After coding the actual experimental data, the average reliability across all 21 tasks for the three raters was found to be $\alpha = .91$, while the average correlation between any pair of raters across the 21 tasks was .83. The average amount by which raters' responses deviated from one another was less than 1 ($M = .68; SD = .25$). Due to the high correspondence between the raters' judgments, the three raters' responses were averaged for each of the 21 tasks in order to create an objective measure of the frequency of task performance.

Descriptive Statistics

Prior to analysis, variables were investigated for accuracy of data entry, missing values, and outliers. Examination of the distributions of subjects' reported absolute frequency ratings on the 21 real tasks revealed several outliers. Closer examination indicated that 3 individuals' responses were more than 5 standard deviations away from the mean on 24% of the items. It appears that these individuals did not take the task seriously as they gave disproportionately large responses. For example, when reporting how many times they had used a tool, two of these subjects gave estimates of 50 and 60 while the mean response was only 10.17 ($SD = \ldots$
Due to their extreme ratings on a large percentage of the items, these three individuals were removed from all subsequent analyses.

Table 1 presents the means, standard deviations, and intercorrelations of the variables in the study. With the exception of detection accuracy, only ratings on the "real" tasks were included in the accuracy measures. Examination of Table 1 reveals that subjects performed tasks an average of 3.19 times (SD = .108), yet subjects reported performing the tasks an average of 2.24 times (SD = .94). In general, subjects' reported frequency ratings deviated from actual task frequency of performance by an average of 67% (SD = .19).

Further examination of Table 1 indicates that there were no significant correlations between the actual number of times tasks were performed and both task engagement and task familiarity (r = -.05, p > .05 and r = -.01, p > .05, respectively). In addition there were no significant correlations between the total number of tasks performed and both task engagement and task familiarity (r = -.02, p > .05 and r = .12, p > .05). This suggests that there were no differences between performers and observers and between novel and familiar subjects in the total number of tasks performed and the total number of times tasks were performed.

The total amount of absolute deviations, however, was significantly correlated with the actual number of times subjects performed the tasks (r = .71, p < .01). This suggests that as subjects performed tasks more often, their frequency estimates began to deviate more from the actual frequency with which they performed the tasks. Therefore, relative deviations were also computed in order to determine the accuracy of respondents' frequency ratings relative to the number of times the tasks were performed.
<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>MEAN</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Task Engagement&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.48</td>
<td>.50</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Task Familiarity&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.52</td>
<td>.50</td>
<td>.05</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Reported Number of Times Performed Real Tasks</td>
<td>2.24</td>
<td>.94</td>
<td>.34**</td>
<td>.28*</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Actual Number of Times Performed Real Tasks</td>
<td>3.19</td>
<td>.11</td>
<td>-.05</td>
<td>-.01</td>
<td>.24</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Total Number of Real Tasks Performed</td>
<td>12.92</td>
<td>2.72</td>
<td>-.02</td>
<td>.12</td>
<td>.01</td>
<td>.53**</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Absolute Deviations</td>
<td>44.47</td>
<td>17.13</td>
<td>-.34**</td>
<td>.28*</td>
<td>-.03</td>
<td>.71**</td>
<td>.32*</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Relative Deviations</td>
<td>.67</td>
<td>.19</td>
<td>-.41**</td>
<td>.45**</td>
<td>-.22</td>
<td>-.02</td>
<td>-.07</td>
<td>.67**</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Correlation b/w Actual Absolute and Reported Absolute</td>
<td>.54</td>
<td>.24</td>
<td>.39**</td>
<td>-.04</td>
<td>.40**</td>
<td>.16</td>
<td>.19</td>
<td>-.23</td>
<td>-.47**</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Correlation b/w Actual Absolute and Reported Relative</td>
<td>.55</td>
<td>.31</td>
<td>.42**</td>
<td>-.20</td>
<td>.17</td>
<td>-.28*</td>
<td>-.33**</td>
<td>-.56**</td>
<td>-.51**</td>
<td>.41**</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Correlation b/w Reported Absolute and Reported Relative</td>
<td>.29</td>
<td>.14</td>
<td>.25</td>
<td>.05</td>
<td>.31*</td>
<td>-.01</td>
<td>.08</td>
<td>.05</td>
<td>.11</td>
<td>.31*</td>
<td>.28*</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>11. Detection Accuracy</td>
<td>1.62</td>
<td>.81</td>
<td>.30*</td>
<td>-.38**</td>
<td>-.13</td>
<td>-.16</td>
<td>-.06</td>
<td>-.46**</td>
<td>-.49**</td>
<td>-.04</td>
<td>.39**</td>
<td>-.44**</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**NOTES:**

<sup>a</sup> Coded 0=observer 1=performer

<sup>b</sup> Coded 0=novel 1=familiar

n=61

* p < .05

** p < .01
Test of Hypotheses

This study hypothesized that performers who engaged in the simulation would generate more accurate frequency estimates than observers. In addition, this study proposed that novel participants would generate more accurate frequency estimates than familiar participants. These hypotheses were tested using each measure of accuracy.

Absolute Deviations. The differences between subjects' reported absolute frequency ratings and their actual task frequency of performance were summed across all 21 tasks. Table 2 presents the results of an ANOVA on total deviations. The results yielded a main effect for both task engagement, $F(1,57) = 7.50$, $p = .01$, and task familiarity, $F(1,57) = 4.72$, $p = .03$. Across all tasks, observers' frequency estimates deviated from actual task frequency by an average of 2.38, whereas performers' estimates only deviated by 1.28. In addition, familiar subjects' frequency estimates deviated from actual task frequency by an average of 2.35, whereas novel subjects' estimates only deviated by 1.90.

Given the significant correlation between absolute deviations and actual number of times tasks were performed ($r = .71$, $p < .01$), an analysis taking the actual number of times tasks were performed into account was performed. Table 3 presents the results of a hierarchical regression on subjects' absolute deviations with the total number of times the 21 tasks were performed entered first. After removing the variance explained by actual task frequency, both task engagement and task familiarity still predicted 17% additional variance in absolute deviations ($\beta = -.27$, $p < .01$ and $\beta = .29$, $p < .01$, respectively). This indicates that, regardless of how many times the tasks were performed, performers were more accurate than observers and novel subjects were more accurate than familiar subjects.
Table 2

ANOVA Results on Absolute Deviations

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Engagement (E)</td>
<td>1887.17</td>
<td>1</td>
<td>1887.17</td>
<td>7.50**</td>
</tr>
<tr>
<td>Task Familiarity (F)</td>
<td>1187.63</td>
<td>1</td>
<td>1187.63</td>
<td>4.72*</td>
</tr>
<tr>
<td>E x F</td>
<td>35.66</td>
<td>1</td>
<td>35.66</td>
<td>.14</td>
</tr>
<tr>
<td>Within</td>
<td>14338.32</td>
<td>57</td>
<td>251.55</td>
<td></td>
</tr>
</tbody>
</table>

* p < .05
** p < .01
Table 3
Hierarchical Regression of Total Task Frequency, Task Engagement, and Task Familiarity on Absolute Deviations (N = 61)

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\beta_a$</th>
<th>$R^2$</th>
<th>$\Delta R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Task Frequency</td>
<td>.50**</td>
<td>.50**</td>
<td></td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Engagement$^b$</td>
<td>.67**</td>
<td>.17**</td>
<td></td>
</tr>
<tr>
<td>Task Familiarity$^c$</td>
<td>-.27**</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes:
a = regression weight after all variables have been entered in the equation.
$R^2$ = variance accounted for by the predictor set upon entry into the regression equation.
$\Delta R^2$ = increment in variance accounted for when the predictor set is entered into the regression equation.

$^b$ Coded 0 = observer 1 = performer
$^c$ Coded 0 = novel 1 = familiar

* $p < .05$
** $p < .01$
These results parallel those found when only task engagement and task familiarity were used to predict absolute deviations.

Relative Deviations. To further determine the accuracy of respondents' frequency ratings relative to the number of times tasks were performed, a measure of relative deviations was computed. To create the measure of relative deviations, the absolute deviations between subjects' reported absolute frequency and their actual task frequency of performance were summed across all 21 tasks and then divided by the total number of times the 21 tasks were performed. Table 4 presents the results of an ANOVA on the relative deviations. The results yielded a main effect for both task engagement, $F(1,57) = 13.29, p < .01$, and task familiarity, $F(1,57) = 15.95, p < .01$. More specifically, observers' frequency estimates deviated from actual task frequency by an average of 74%, whereas performers' frequency estimates deviated by only 58%. Furthermore, familiar subjects' frequency estimates deviated from actual task frequency by an average of 75%, whereas novel subjects' frequency estimates deviated by only 59%. This suggests that individuals who performed the task and individuals who were relatively new at the task were more accurate in their frequency estimates.

Detection Accuracy. Using signal detection theory, an alternative estimate of accuracy was computed. In order to assess the accuracy of detection, the proportion of tasks correctly identified (the hit rate) and the proportion of tasks incorrectly identified (the false-alarm rate) were computed. Table 5 presents the results of an ANOVA on the hit rate. The results revealed a main effect for both task engagement, $F(1,54) = 6.55, p = .01$, and task familiarity, $F(1,54) = 10.38, p < .01$. Performers demonstrated a greater hit rate compared to observers ($M = 2.44$ vs. 2.19,
Table 4

ANOVA Results on Relative Deviations

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Engagement (E)</td>
<td>.32</td>
<td>1</td>
<td>.32</td>
<td>13.29**</td>
</tr>
<tr>
<td>Task Familiarity (F)</td>
<td>.39</td>
<td>1</td>
<td>.39</td>
<td>15.95**</td>
</tr>
<tr>
<td>E x F</td>
<td>.001</td>
<td>1</td>
<td>.001</td>
<td>.89</td>
</tr>
<tr>
<td>Within</td>
<td>1.39</td>
<td>57</td>
<td>.02</td>
<td></td>
</tr>
</tbody>
</table>

* p < .05
** p < .01
Table 5

ANOVA Results on the Hit Rate

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Engagement (E)</td>
<td>.81</td>
<td>1</td>
<td>.81</td>
<td>6.55*</td>
</tr>
<tr>
<td>Task Familiarity (F)</td>
<td>1.29</td>
<td>1</td>
<td>1.29</td>
<td>10.38**</td>
</tr>
<tr>
<td>E x F</td>
<td>.57</td>
<td>1</td>
<td>.57</td>
<td>4.60*</td>
</tr>
<tr>
<td>Within</td>
<td>7.10</td>
<td>57</td>
<td>.12</td>
<td></td>
</tr>
</tbody>
</table>

* $p < .05$
** $p < .01$
respectively), and novel subjects demonstrated a greater hit rate compared to familiar subjects ($M = 2.45$ vs. $2.15$, respectively).

In addition, a significant two-way interaction was found between task engagement and task familiarity, $F(1,57) = 4.60$, $p = .04$. Examination of Figure 2 indicated that the effect of task familiarity differed for observers and performers. T-tests of the means revealed that, for performers, novel experience lead to a greater proportion of tasks correctly identified, $t(27) = -3.41$, $p < .01$. For observers, however, familiarity with the task had no effect on the proportion of tasks correctly identified, $t(30) = -.97$, $p > .05$. Additional t-tests revealed that, for novel subjects, performing the task lead to a higher hit rate, $t(30) = -3.98$, $p < .01$. For familiar subjects, however, engagement with the task had no effect on the hit rate, $t(27) = -.18$, $p > .05$. In sum, these results suggest that performers with novel experience were the most accurate at correctly identifying whether or not a task was or was not performed.

Table 6 presents the results of an ANOVA on the false-alarm rate. The results revealed a main effect for both task engagement, $F(1,57) = 6.47$, $p = .01$, and task familiarity, $F(1,57) = 11.37$, $p < .01$. Specifically, on average, observers demonstrated a greater false-alarm rate than performers ($M = .85$ vs. $.62$, respectively), and familiar subjects demonstrated a greater false-alarm rate than novel subjects ($M = .85$ vs. $.54$, respectively).

Furthermore, a significant two-way interaction between task engagement and task familiarity was found, $F(1,57) = 4.78$, $p = .03$. Examination of the means in Figure 3 indicated that the effect of task familiarity differed for observers and performers. T-tests of the means revealed that, for performers, familiarity with the task lead to a greater
Figure 2

Interaction Between Task Engagement and Task Familiarity on the Hit Rate
Table 6

ANOVA Results on the False-Alarm Rate

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Engagement (E)</td>
<td>.78</td>
<td>1</td>
<td>.78</td>
<td>6.47*</td>
</tr>
<tr>
<td>Task Familiarity (F)</td>
<td>1.37</td>
<td>1</td>
<td>1.37</td>
<td>11.37**</td>
</tr>
<tr>
<td>E x F</td>
<td>.58</td>
<td>1</td>
<td>.58</td>
<td>4.78*</td>
</tr>
<tr>
<td>Within</td>
<td>6.86</td>
<td>57</td>
<td>.12</td>
<td></td>
</tr>
</tbody>
</table>

* $p < .05$

** $p < .01$
Figure 3

Interaction Between Task Engagement and Task Familiarity on the False-Alarm Rate
proportion of tasks incorrectly identified, $t(27) = 3.49, p < .01$. For observers, however, familiarity with the task had no effect on the proportion of tasks incorrectly identified, $t(30) = 1.08, p > .05$. Additional t-tests revealed that, for novel subjects, observing the task lead to a higher false-alarm rate, $t(30) = 4.05, p < .01$. For familiar subjects, however, engagement with the task had no impact on the false-alarm rate, $t(27) = .14, p > .05$. Combined with the analyses on the hit rate, these results suggest that performers with novel experience not only demonstrated the highest hit rate, but also demonstrated the lowest false-alarm rate.

Since it was possible for subjects to receive a relatively high hit rate by identifying every task as having been performed, a correction was applied to the hit rate in order to determine subjects' ability to accurately distinguish between those tasks that were and were not performed. Therefore, detection accuracy was computed by taking the difference between the hit rate and the false-alarm rate. As can be seen in Table 7, analysis of variance on detection accuracy revealed a main effect for both task engagement, $F(1,57) = 6.52, p = .01$, and task familiarity, $F(1,57) = 10.88, p < .01$. Specifically, performers were found to have higher detection accuracy compared to observers ($M = 1.76$ vs. $1.32$, respectively), and novel subjects were found to have a higher detection accuracy compared to familiar subjects ($M = 1.91$ vs. $1.30$, respectively).

In addition to the significant main effects, a significant two-way interaction between task engagement and task familiarity was found, $F(1,57) = 4.69, p = .03$. Examination of Figure 4 indicated that the effect of task familiarity differed for observers and performers. T-tests of the means revealed that, for performers, novel experience with the task lead to greater detection accuracy, $t(27) = -3.45, p < .01$. For observers,
Table 7

ANOVA Results on Detection Accuracy

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Engagement</td>
<td>3.19</td>
<td>1</td>
<td>3.19</td>
<td>6.52*</td>
</tr>
<tr>
<td>Task Familiarity</td>
<td>5.32</td>
<td>1</td>
<td>5.32</td>
<td>10.88**</td>
</tr>
<tr>
<td>E x F</td>
<td>2.30</td>
<td>1</td>
<td>2.30</td>
<td>4.69*</td>
</tr>
<tr>
<td>Within</td>
<td>27.89</td>
<td>57</td>
<td>.49</td>
<td></td>
</tr>
</tbody>
</table>

* p < .05
** p < .01
however, familiarity with the task had no effect on detection accuracy, 
$t(30) = -1.02, p > .05$. Additional t-tests demonstrated that, for novel subjects, performing the task lead to greater detection accuracy, $t(30) = -4.01, p < .01$. For familiar subjects, however, engagement with the task had no impact on detection accuracy, $t(27) = -.16, p > .05$. In sum, these results suggest that performers with novel experience were the most accurate at detecting those tasks that were and were not performed.

Additional analyses were performed on the various elements that compose the hit rate to determine if subjects were better at detecting tasks that were performed versus tasks that were not performed. The hit rate was composed of two elements; the proportion of tasks that were correctly identified as having been performed and the proportion of tasks correctly identified as having *not* been performed. As shown in Table 8, a repeated measures MANOVA on these two proportions indicates that, on average, subjects were better at correctly identifying tasks that were not performed ($M = 1.48$) than tasks that were performed ($M = .82$), $F(1,57) = 158.75, p < .01$.

A significant interaction was also found between task familiarity and the two hit rate proportions, $F(1,57) = 8.35, p < .01$. Examination of the means in Figure 5 indicated that although novel and familiar subjects demonstrated the same hit rate on the tasks that were performed, $t(59) = -.13, p > .05$, novel subjects demonstrated a greater hit rate on those tasks that were *not* performed, $t(59) = -3.18, p < .01$. This suggests that, in general, all subjects were better at determining whether a task was not performed, compared to a task that was performed, yet this effect was even stronger for novel subjects.
Table 8

Repeated Measures MANOVA Results on the Hit Rate Proportions

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-Subjects Effects:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Engagement (E)</td>
<td>.41</td>
<td>1</td>
<td>.41</td>
<td>6.55*</td>
</tr>
<tr>
<td>Task Familiarity (F)</td>
<td>.65</td>
<td>1</td>
<td>.65</td>
<td>10.38**</td>
</tr>
<tr>
<td>Error</td>
<td>3.55</td>
<td>57</td>
<td>.06</td>
<td></td>
</tr>
</tbody>
</table>

| Within-Subjects Effect:       |      |    |     |       |
| Hit Rate Proportions          | 12.90| 1  | 12.90| 158.75**|
| Hit Rate Props. x Engagement  | .10  | 1  | .10 | 1.27  |
| Hit Rate Props. x Familiarity | .68  | 1  | .68 | 8.35**|
| Error                         | 4.63 | 57 | .08 |       |

* p < .05
** p < .01
Figure 5

Interaction Between Hit Rate Proportions and Task Familiarity

Performance of a Task

Hit Rate

NOVEL
FAMILIAR

PERFORMED
NOT PERFORMED
Rank Order Accuracy. The rank order of tasks reflected by subjects' absolute frequency ratings and relative frequency ratings were matched against the actual rank order of task frequency. A correlation was computed for each subject between subjects' absolute frequency ratings and the actual task frequency of performance. In addition, a correlation between subjects' relative frequency ratings and the actual frequency of performance was computed for each subject. These two correlations were then used as dependent variables in subsequent ANOVAs. On average, subjects were equally good at capturing the rank order of tasks with the absolute rating scale (mean $r = .54$; $SD = .15$) and the relative rating scale (mean $r = .55$; $SD = .31$). This suggests that, regardless of which rating scale format was used, a task that was actually performed more often was rated as having been performed more often.

Since correlations tend to have skewed distributions, the two rank order correlations were transformed to Fisher's $z$ equivalents. Results of an ANOVA on the Fisher's $z$ correlation between subjects' absolute frequency ratings and actual task frequency yielded a main effect for task engagement, $F(1,57) = 10.57$, $p < .01$ (see Table 9). In addition, results of an ANOVA on the Fisher's $z$ correlation between subjects' relative frequency ratings and actual task frequency also yielded a main effect for task engagement, $F(1,57) = 11.62$, $p < .01$ (see Table 10). Specifically, performers were better than observers at representing the actual rank order of task frequency whether they used the absolute rating scale ($M = .74$ vs. $.55$, respectively) or the relative rating scale ($M = .79$ vs. $.53$, respectively). For example, when using the absolute rating format, performers' frequency estimates reflected a rank order of tasks in which 33% of the tasks were ranked similarly to the actual rank order of task
Table 9

ANOVA Results on the Correlation between Absolute Frequency Ratings and Actual Task Frequency of Performance

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Engagement (E)</td>
<td>.52</td>
<td>1</td>
<td>.52</td>
<td>10.57**</td>
</tr>
<tr>
<td>Task Familiarity (F)</td>
<td>.001</td>
<td>1</td>
<td>.001</td>
<td>.03</td>
</tr>
<tr>
<td>E x F</td>
<td>.06</td>
<td>1</td>
<td>.06</td>
<td>.28</td>
</tr>
<tr>
<td>Within</td>
<td>2.78</td>
<td>57</td>
<td>.05</td>
<td></td>
</tr>
</tbody>
</table>

* $p < .05$

** $p < .01$
Table 10

ANOVA Results on the Correlation between Relative Frequency Ratings and Actual Task Frequency of Performance

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Engagement (E)</td>
<td>.94</td>
<td>1</td>
<td>.94</td>
<td>11.62**</td>
</tr>
<tr>
<td>Task Familiarity (F)</td>
<td>.17</td>
<td>1</td>
<td>.17</td>
<td>2.12</td>
</tr>
<tr>
<td>E x F</td>
<td>.03</td>
<td>1</td>
<td>.03</td>
<td>.31</td>
</tr>
<tr>
<td>Within</td>
<td>4.54</td>
<td>57</td>
<td>.08</td>
<td></td>
</tr>
</tbody>
</table>

* $p < .05$

** $p < .01$
frequency. Conversely, the observers' frequency estimates reflected a rank order of tasks in which only 19% of the tasks were ranked similarly to the actual rank order of task frequency.

After averaging across the two rating formats, the effect of task familiarity appeared to be in the hypothesized direction. The average correlations between novel subjects' absolute ratings and relative ratings with actual task frequency of performance were larger than familiar subjects' ratings with actual task frequency (M = .73 vs. .63, respectively).

**Summary of Accuracy Measures.** In sum, the analyses on the various accuracy measures (absolute deviations, relative deviations, detection accuracy, and rank order accuracy) suggest that although subjects, on average, engaged in the same number of tasks the same number of times, respondents were not equally accurate at generating the frequency with which tasks were performed. In testing the hypotheses, the analyses on the different accuracy measures indicate that individuals who actually performed the task and individuals who had novel experience with the task generated more accurate frequency estimates. More specifically, performers and novel subjects gave frequency responses that were closer to the actual frequency of performance, detected those tasks that were and were not performed with greater precision, and generated frequency responses that more accurately captured the actual rank order of task frequency. Consequently, these results support the hypotheses proposed by this study.

**Ancillary Analyses**

**Consistency Across Rating Format.** To examine the extent to which subjects were able to give consistent information using both an absolute and a relative rating scale, a correlation between subjects' absolute and relative
ratings was computed. Table 11 presents the results of an ANOVA on the Fisher's z equivalent correlation representing the consistency between subjects' absolute and relative ratings. The findings revealed a main effect for task engagement, $F(1,57) = 5.44$, $p = .02$. Specifically, performers gave more similar responses on the two rating scales compared to observers ($M = .42$ vs. .31, respectively). There was no main effect for task familiarity, $F(1,57) = .84$, $p > .05$, yet a significant two-way interaction between task engagement and task familiarity was found, $F(1,57) = 8.09$, $p = .01$ (see Figure 6).

T-tests of the means revealed that the effect of task familiarity did not significantly impact observers or performers, $t = -2.02$, $p = .05$ and $t = 1.79$, $p > .05$, respectively. Additional analyses, however, demonstrated that although the effect of task engagement did not significantly impact novel subjects, $t = .32$, $p > .05$, it did significantly impact familiar subjects, $t = -3.10$, $p < .01$. Specifically, familiar performers demonstrated greater consistency between their absolute and relative ratings compared to familiar observers ($M = .52$ vs. .27 respectively).

**Individual Differences.** To examine the impact of individual difference variables on the accuracy of subjects' frequency judgments, the various accuracy measures were correlated with cognitive ability, mechanical reasoning, self efficacy, performance, importance ratings, interesting ratings, etc. As can be seen in Table 12, only subjects' performance in the simulation was related to any of the accuracy measures. Group performance was significantly correlated with both absolute deviations ($r = .41$, $p = .03$) and detection accuracy ($r = -.51$, $p < .01$). This indicates that the higher the group's performance, the more group
Table 11

ANOVA Results on the Consistency between Subjects' Absolute and Relative Frequency Ratings

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Engagement (E)</td>
<td>.17</td>
<td>1</td>
<td>.17</td>
<td>5.44*</td>
</tr>
<tr>
<td>Task Familiarity (F)</td>
<td>.03</td>
<td>1</td>
<td>.03</td>
<td>.84</td>
</tr>
<tr>
<td>E x F</td>
<td>.25</td>
<td>1</td>
<td>.25</td>
<td>8.09**</td>
</tr>
<tr>
<td>Within</td>
<td>1.78</td>
<td>57</td>
<td>.03</td>
<td></td>
</tr>
</tbody>
</table>

* p < .05  
** p < .01
Figure 6

Interaction Between Task Engagement and Familiarity on the Consistency Between Absolute and Relative Ratings
Table 12
Intercorrelations between Individual Difference Variables and Accuracy Measures

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Absolute Deviations</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Relative Deviations</td>
<td>.67**</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Detection Accuracy</td>
<td>-.46**</td>
<td>-.49**</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Correlation b/w Actual Absolute and Reported Absolute</td>
<td>.23</td>
<td>-.47**</td>
<td>-.04</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Correlation b/w Actual Absolute and Reported Relative</td>
<td>-.56**</td>
<td>-.51**</td>
<td>.38**</td>
<td>.41**</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Self Efficacy</td>
<td>.12</td>
<td>.07</td>
<td>-.03</td>
<td>-.10</td>
<td>-.04</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Cognitive Ability</td>
<td>.01</td>
<td>.04</td>
<td>.25</td>
<td>-.26</td>
<td>-.06</td>
<td>-.06</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Mechanical Reasoning Ability</td>
<td>-.18</td>
<td>.00</td>
<td>-.12</td>
<td>-.02</td>
<td>-.01</td>
<td>-.08</td>
<td>-.29*</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Grade Point Average</td>
<td>-.05</td>
<td>.07</td>
<td>.08</td>
<td>-.15</td>
<td>.02</td>
<td>-.06</td>
<td>.39**</td>
<td>.07</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Group Performance</td>
<td>.41*</td>
<td>.11</td>
<td>-.51**</td>
<td>.20</td>
<td>-.19</td>
<td>.22</td>
<td>-.30</td>
<td>-.13</td>
<td>-.37</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Age</td>
<td>.02</td>
<td>.00</td>
<td>.02</td>
<td>-.20</td>
<td>.11</td>
<td>.20</td>
<td>.00</td>
<td>.18</td>
<td>.20</td>
<td>-.08</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Gender&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.11</td>
<td>.13</td>
<td>-.01</td>
<td>-.18</td>
<td>-.08</td>
<td>.12</td>
<td>-.03</td>
<td>-.18</td>
<td>-.32*</td>
<td>.40*</td>
<td>.10</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Previous Experience with an Erector® set</td>
<td>-.03</td>
<td>-.14</td>
<td>.13</td>
<td>.14</td>
<td>.14</td>
<td>.27*</td>
<td>-.08</td>
<td>.01</td>
<td>-.14</td>
<td>.40*</td>
<td>-.03</td>
<td>.10</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Importance Ratings</td>
<td>.10</td>
<td>-.03</td>
<td>.12</td>
<td>.12</td>
<td>-.14</td>
<td>.13</td>
<td>-.06</td>
<td>.02</td>
<td>.09</td>
<td>-.32</td>
<td>-.06</td>
<td>-.19</td>
<td>-.13</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>15. Interesting Ratings</td>
<td>.22</td>
<td>.23</td>
<td>-.15</td>
<td>-.07</td>
<td>-.21</td>
<td>.18</td>
<td>-.01</td>
<td>-.05</td>
<td>.11</td>
<td>.16</td>
<td>.10</td>
<td>.00</td>
<td>-.16</td>
<td>.50**</td>
<td>1.0</td>
</tr>
</tbody>
</table>

NOTES: <sup>a</sup> Coded 0=female 1=male
n=61
* p < .05
** p < .01
members' frequency responses deviated from the actual number of times tasks were performed. In addition, the higher the group's performance, the less group members' demonstrated precision in identifying whether a task was or was not performed. Together, these correlations suggest that subjects were less accurate in certain aspects of their frequency responses when they performed at a higher rate.

However, after taking into account the number of times the tasks were performed, group performance did not explain any additional variance in absolute deviations beyond actual task frequency of performance. In addition, after taking into account the total number of tasks actually performed by the subjects, group performance did not predict any additional variance in absolute deviations.

To determine if group performance explained any additional variance in detection accuracy after taking into account the number of tasks performed and the number of times tasks were performed, subsequent regressions were performed. The results demonstrated that group performance explained 28% of the variance in detection accuracy beyond that explained by the number of times tasks were performed ($\beta = -.60, p < .01$). In addition, group performance predicted 23% additional variance beyond that explained by the total number of tasks performed ($\beta = -.49, p < .01$). These two analyses indicate that individuals with high performance were less able to detect those tasks that were and were not exhibited regardless of the frequency of task performance and the number of tasks performed.

Therefore, there appears to be some aspect of group performance that significantly influenced the precision with which subjects identified tasks as having been performed or not performed. Whether or not
subjects' gave frequency ratings that were close to the actual number of
times tasks were performed, however, was not predicted by performance
level, but rather was partially explained by how many and how often tasks
were executed.

Discussion

The purpose of this study was to utilize an objective measure of task
frequency to determine if discrepancies in task frequency ratings were due
to actual differences in task frequency or differences in job perceptions.
This research examined the accuracy of task performer and observer
ratings and evaluated the impact of task familiarity on task frequency
rating accuracy by comparing respondents' frequency ratings to an
objective measure of task frequency. It was hypothesized that individuals
who engaged in the task activities would generate more accurate frequency
estimates compared to individuals who only observed others perform the
tasks. In addition, it was hypothesized that individuals with novel
experience on a task would generate more accurate frequency estimates
compared to those who were more familiar with the task.

The results of this study indicate that, although there were no
differences between respondents in different conditions in how many and
how often tasks were executed, in general, respondents were not equally
accurate in generating frequency ratings. Overall, participants gave
inaccurate frequency estimates more than 50% of the time.

Using several measures of accuracy, support was found for the
hypothesized relationship between task engagement and frequency rating
accuracy. Performers generated more accurate frequency estimates than
observers. More specifically, individuals who performed the task
activities, compared to those who only observed the tasks, produced
frequency estimates that were closer to the actual number of times the tasks were performed. Even after taking actual task frequency of performance into account, performers' frequency ratings were still more accurate.

Performers were also more accurate at detecting those tasks that were and were not performed compared to observers. They demonstrated less carelessness and greater precision in correctly identifying the execution of a task. Furthermore, performers, compared to observers, exhibited greater rank order accuracy in that their frequency estimates more accurately reflected the actual rank order of task frequency. This was found whether the performers used an absolute rating scale or a relative rating scale.

Lastly, performers gave more similar ratings on both the absolute and the relative rating scales compared to observers. Although the consistency in performers' responses between various rating formats is essentially not a measure of accuracy, one might consider this as evidence toward the reliability of performers' responses.

Using the same measures of accuracy, support was also found for the hypothesized relationship between task familiarity and frequency rating accuracy. Novel respondents generated more accurate frequency estimates compared to familiar respondents. More specifically, individuals with novel experience on the task, compared to those with familiar experience, generated frequency estimates that were closer to the actual number of times the tasks were performed. This effect was still found even after actual task frequency of performance was taken into account.

Furthermore, novel respondents were also more accurate at detecting those tasks that were and were not performed compared to familiar respondents. They demonstrated less carelessness and greater precision in
correctly identifying the execution of a task. Lastly, although there was no significant difference in the rank order accuracy between individuals with novel and familiar experience whether the absolute or relative rating format was used, when averaging across both formats, novel subjects demonstrated superior rank order accuracy.

Together, the results of the task engagement and task familiarity factors suggest that individuals who actually perform the task and individuals with novel experience at the task are better able to provide accurate frequency estimates. These effects were found using several different measures of accuracy (i.e. absolute deviations, relative deviations, detection accuracy, and rank order accuracy). In addition, the findings suggest that respondents were not equally accurate in their ratings although there were no differences between performers and observers and between novel and familiar subjects in how many and how often tasks were performed.

The results of this research further suggest that task analysis respondents have misconceptions of the frequency of task occurrence. Although an aspect of the respondents' inaccurate frequency estimates was explained by what the respondents actually performed, after taking the actual number of times tasks were performed into account, response deviations still remained. Therefore, since the discrepancies in frequency ratings were not due to differences in task frequency of performance, one might speculate that the discrepancies were due to respondents' differing conceptualizations of the tasks at hand. Respondents' varying perceptions of the tasks seem to have been driven by familiarity with the tasks and engagement with the tasks.
Memory Processes in Task Frequency Estimation

This study is the first attempt at applying memory research to the issues in task analysis rating accuracy. By incorporating various memory processing models of frequency estimation into a task analysis perspective, and by using an actual (objective) true score measure of task frequency, this study has helped provide insight into the task frequency estimation process.

The memory literature suggests that performers of tasks may process activities at a deeper level than those who only observe the tasks and that deeper levels of processing leads to more accurate frequency estimates. Since the results of this study indicate that some aspect of task engagement caused the performers to generate more accurate frequency estimates compared to the observers, one might speculate that the act of performing the tasks influenced the performers to encode the information at a deeper level. This processing of information at a deeper level, in turn, helped the performers to generate better frequency ratings.

The memory literature also submits that individuals have a more difficult time recalling specific events when the event is one of many similar incidents than when it is unique (Robinson & Swanson, 1990) and that routine aspects of the same event are less likely to be accurately recalled (Hudson et al., 1992). In this study, although there were no differences in task frequency during the simulation that was rated, familiar subjects engaged in the simulation three times while the novel subjects engaged in the simulation only once. Since the results of this study indicate that familiar respondents were less accurate at generating frequency estimates compared to novel respondents, one might speculate that this occurred because the familiar respondents were unable to make accurate
distinctions between similar task occurrences that were encountered in all
three simulations. The respondents with novel experience, on the other
hand, were more accurate in their frequency estimates because the tasks
that they rated were only encountered during the one simulation.

Implications

The results of this research have wide implications for the manner in
which task analysis practices are currently executed. More specifically, the
findings may have consequences for the sources of task analysis data, the
methods of collecting task analysis data, and the subsequent inferences
made from such data. First, it has become common practice to use Subject
Matter Experts (SMEs) in task analyses. SMEs are typically job
incumbents, supervisors, or job analysts who are considered to be in the
best position to provide complete and accurate knowledge about the tasks
that comprise a given job because they are familiar with all of its tasks
(Thompson & Thompson, 1982). This study, however, provides empirical
evidence to suggest that job incumbents who are relatively new at the job
may be in the best position to provide accurate task frequency ratings.
Therefore, one might question the current use of supervisors and job
analysts and the use of respondents who have a lot of experience with a job
when completing a task analysis.

Past personnel research on task analysis rating accuracy has been
unable to determine who is the most qualified to provide task ratings. For
example, the findings of several job analysis studies present a mixed picture
concerning the convergence in task frequency ratings given by respondents
with varying levels of familiarity or experience with a job. Past studies
may have found differences in task ratings by experience (i.e. Landy &
Vasey, 1991; Borman et al., 1992; Hahn & Dipboye, 1988), however, these
researchers were unable to make definitive conclusions about who was more accurate in their ratings because they did not utilize an *objective* true score measure from which to compare the raters' responses. In contrast, this study was able to compare the frequency ratings of novel and familiar respondents to an actual (objective) measure of task frequency. It was discovered that novel experienced individuals are more accurate in their task frequency ratings. Consequently, the results of this research may bring previous findings into question.

Industrial/organizational research on the impact of task engagement on task analysis ratings have also demonstrated mixed findings. Past studies have been unable to conclusively determine whether incumbents or supervisors are better at generating task analysis ratings because they did not employ an objective measure from which to compare raters' responses. For example, studies that have found high agreement or disagreement between different rating sources (e.g. Meyer, 1959; O'Reilly, 1973; Smith & Hakel, 1979) typically assessed the *agreement* or convergence in respondents' ratings. Therefore, they could not conclude who generated more *accurate* ratings. This study, however, compared performers' and observers' frequency estimates to the actual number of times tasks were performed and discovered that individuals who actually engage in the task activities are more accurate. Consequently, one might question previous findings that concluded that different rater sources are equally accurate in their task frequency ratings (e.g. Smith & Hakel, 1979).

The results of this study also have implications on the reliability of task analysis ratings. The findings indicate that individuals who perform the task activities are not only more accurate in their frequency responses, but they also demonstrate greater consistency in their frequency ratings.
across absolute and relative rating scale formats. One might speculate that such consistency in performers' frequency ratings illustrates greater reliability in their responses. Therefore, future researchers might consider using task performers in order to obtain the most accurate and reliable task frequency ratings.

Furthermore, although this study examined the accuracy of task frequency ratings from a task analysis perspective, the findings have implications on the sources of frequency ratings in a variety of psychological disciplines. For example, social and psychological researchers often ask subjects to report the frequency with which they encounter a particular experience or engage in a particular behavior (Schwarz et al., 1991). The results of this study suggest that, if the subjects are highly familiar with the experience or behavior being rated, one might question the accuracy of their frequency ratings. Furthermore, when examining the opportunity to perform (OTP) trained skills on the job (Ford et al., Quiñones et al., in press), the activity level dimension of the OTP construct is created by asking trainees to identify the number of times they have performed each of the sampled tasks within a given time. As suggested by the present research, respondents with many opportunities to perform the trained skills may provide less accurate frequency estimates compared to respondents with fewer opportunities. Consequently, one might question the accuracy of the activity level dimension of the OTP construct in that it may be confounded with respondents' degree of familiarity with the trained tasks.

In addition to having consequences on the sources of task analysis data, this study also has implications on the methods of collecting task analysis data. First, the results of this study impact issues concerning the
redundancy between task rating scales (Harvey, 1991). The nature and the number of task rating scales impacts both the ease of conducting task analyses and the usefulness of task analyses in making cross-job comparisons. This study suggests that a task analysis researcher may only need to employ either an absolute rating scale or a relative rating scale, not both. The findings demonstrate that individuals are equally good at capturing the actual rank order of task frequency whether an absolute or a relative rating scale is employed. In addition, the findings suggest that task performers are capable of providing similar information using both the absolute and relative rating formats. If future research substantiates such findings, one might consider improving the manageability of the task analysis process by utilizing only one of the two scales.

The choice of whether to use an absolute frequency scale or a relative frequency scale subsequently impacts the usefulness of task analyses in making cross-job comparisons. Although the use of relative rating scales is more common (Harvey, 1991), the within-job-relative task data collected from such scales only allows comparisons to be made within a given job. The use of an absolute frequency format, however, provides a common metric from which comparisons between jobs can be made. Consequently, if future researchers substantiate the findings of this study that task performers provide similar ratings on both an absolute and a relative rating scale and demonstrate equal ability in capturing the rank order of task frequency using both rating formats, one might consider using the absolute frequency scale with job incumbents to generate meaningful cross-job comparisons.

In addition to having consequences on the collection of task analysis data using different rating scales, the findings of this study have
implications on the inferences made from such data. Typically, the
criticality of individual tasks is important when making personnel decisions
from task analysis data. Criticality indexes are created in which "task
ratings are combined to form a single critical value" (e.g. by multiplying
Importance x Frequency) (Harvey, 1991, p. 107). This study demonstrates
that respondents are not equally accurate a generating frequency ratings
and that task observers are not as accurate as performers at generating
frequency estimates that reflect the actual rank order of task frequency. If
task analysis researchers unknowingly use inaccurate frequency estimates to
compute criticality indexes, a misleading ranking of critical tasks may be
obtained. Although documents such as the Uniform Guidelines do not
specify exactly how researchers should identify critical tasks (Harvey,
1991), the results of this study indicate that researchers need to examine the
accuracy of respondents' frequency estimates using an objective measure of
task frequency in order to avoid inaccurate computations of criticality.

Limitations and Directions for Future Research

Based on the results of this study, there are a number of issues that
future research could address, several of which deal with the
generalizability of the present findings. First, future research should
examine whether similar results are found when the investigation occurs
over a longer time span. This research found effects for both task
engagement and task familiarity, yet several past studies on these factors
did not find such effects. This discrepancy may be due to the fact that this
research was executed during a very short period of time. Future
investigators might consider using an objective measure of task frequency
and reexamine the influence of engagement and familiarity on task rating
accuracy over several weeks or months. This would help determine if the
findings of this study are generalizable to work settings in which individuals engage in task activities far longer than the length of a laboratory experiment.

The second issue concerns the generalizability of the results from a laboratory experiment to the actual workplace. Some might assume that objective true score measures, such as the actual number of times tasks are performed, can only be obtained in a laboratory setting. However, with the rise of computers into the workforce, it may become possible to obtain objective measures from which to compare raters' responses. For example, in certain occupations, such as a grocery store check-out clerk, essentially every action is recorded by computer. Each swipe of an item across the grocery store scanner, each digital entry into the cash register, each credit card authorization, and each receipt generation can be tabulated. Therefore, since some aspects of the job can be objectively measured, future researchers can determine the reliability of raters' responses. If respondents are found to generate ratings that are similar to the objective true score, one can feel confident in using these respondents to rate those aspects of the job that are not directly observable.

A third issue also concerns the generalizability of the results. One might consider the television viewing by the observers in this experiment as a limitation of this research. However, such intense, direct viewing of performers' behavior is superior to what actually occurs on the job. Supervisors, for example, do not spend all their time directly viewing incumbents' behavior because they have other tasks which demand their attention. In this experiment, however, observers were only instructed to watch the videotaped behavior. Despite this controlled setting, observers still generated less accurate frequency estimates compared to performers.
Future research should consider replicating this research and having observers watch the performers while they are actually engaging in the simulation. This would help to eliminate the possibility that observers generated inaccurate frequency responses because their viewing of the performers' behavior was videotaped and was limited to a television screen. In addition, having observers physically present in the room while performers engage in the task activities would help determine if the results of this study are generalizable to a more realistic setting.

The fourth issue concerns the role of respondents' performance levels on the accuracy of task ratings. Some past studies has found minimal differences in task analysis responses between effective and ineffective performers (e.g. Wexley & Silverman, 1978; Conley & Sackett, 1987). Such research may not have found performance effects on rating accuracy because the accuracy of rating responses could not be assessed without an objective true score as a standard of comparison. Using an objective measure of task frequency, this study found that, regardless of how many and how often tasks were performed, performance level influences the precision with which respondents are able to identify those tasks that have and have not been exhibited. Although performance levels did not significantly impact the accuracy of raters' responses, high performers were less able to detect those tasks that were and were not performed. Therefore, future research should examine the impact of performance on task analysis accuracy by comparing task responses to an objective measure. In addition, they should investigate what cognitive processes occur in high versus low performers to examine their influence on respondents' ability to identify the exhibition of particular tasks.
An additional issue that future investigators should examine concerns the possible motivational differences between task performers and observers and its impact on the accuracy of task ratings. This study found that there were no differences between observers and performers in their level of interest with each of the task elements. Although this finding hints at the notion that observers and performers may have been equally motivated, one might attempt to specifically measure respondents' level of motivation. For example, investigators might ask task raters to report what they were motivated to do during the experiment and have respondents rate how motivated they were toward providing accurate frequency estimates. Research using such methodology will help to determine if motivation differentially impacts the frequency ratings of task performers and observers.

Future researchers should also examine the role of memory processes in the generation of frequency estimates. Since the act of providing task frequency estimates is essentially a memory task in which SMEs have to recall the number of times given tasks are performed on the job, future research might consider examining the possible information biases that may occur. Some investigators mention the role of memory and recall in providing task ratings (e.g. Murphy et al., 1982; Shaffer et al., 1986; Smith & Hakel, 1979), yet most job analysis studies presume that only personal biases, such as boasting and self-preservation, impact job analysis ratings (Smith & Hakel; 1979; Anderson, Warner, & Spencer, 1984; Green & Stutzman, 1986; Harvey, 1991; Hazel et al., 1964; O'Reilly, 1973; Pannone, 1984; Shaffer et al., 1986). Harvey (1991) suggests that the process of collecting job analysis data may be subject to the same cognitive biases as performance appraisal research. "For
example, if the duty categories rated in a job analysis do not correspond to the cognitive schemata employed by the job analyst when the relevant job information was initially observed and stored, undesirable judgment heuristics may influence the duty ratings" (p.115). Therefore, future investigators should further the attempts of this research and apply additional findings from the memory literature to a task analysis perspective. Such attempts may help to determine the possible memory information biases that occur during the frequency rating process.

**Conclusion**

Despite the limitations in the generalizability of the findings, this study makes several contributions to the task analysis literature. First, it suggests that task analysis raters are not equally capable at providing frequency estimates. Familiarity with the task and engagement with the task can influence the accuracy of respondents' frequency ratings. Second, current task analysis practices are called into question. If the findings of this study are substantiated in future research, one might speculate that the use of SMEs may potentially invalidate the task analysis ratings and the computation of criticality. Job incumbents who are relatively new at a job may be more capable at providing accurate information about the tasks that comprise a job. In addition, if future research replicates the finding that task performers can provide accurate ratings using either an absolute or a relative frequency scale, one might consider using only the absolute frequency format with job incumbents in order to improve the manageability of the task analysis process and the usefulness of task analysis in making cross-job comparisons.

Third, this study highlights the need to use objective measures when determining the accuracy of raters' responses. The use of this
methodology illustrates that variations in raters' frequency responses are
not due to differences in task frequency, but instead due to differences in
raters conceptualizations of the job. Finally, this study highlights the need
to incorporate various findings from the memory literature to
investigations on frequency rating accuracy. Consequently, job analysis
and task analysis researchers, along with investigators from several other
psychological disciplines, should attempt to broaden their focus and
consider the memory processes that ultimately influence the generation of
frequency estimates.
References


frequency range on diagnostic judgments. *Applied Cognitive Psychology, 5*, 37-49.


Appendix A
Toy Company Simulation In-Basket Tasks
Write down the name of your toy manufacturing company below:
COMPANY VISION STATEMENT

A vision statement is the beginning of organizational leadership. It provides overall direction and clarifies your purpose and meaning. By referring to it, and internalizing its meaning, you will be more likely to choose behavior that serves your values and reject behavior that opposes them.

The strength of writing an organizational vision statement is three-fold:

1) It will encourage you to clarify the purpose of your company and identify what is really important to the group.
2) It will allow you to clarify and express succinctly your company's values and aspirations.
3) The integration of your vision statement into your planning will give you a tangible method of keeping that vision before you.

Sample Organizational Vision Statement:
For an organization that works with clients:

We are committed to providing our clients with high quality developments and responsive management services and creating enjoyable and productive living and working environments. We are fair, honest, courteous, and professional. We are sensitive to our clients' needs and dedicated to their satisfaction.

Your Company's Vision Statement:
# SUPPLY REQUEST FORM

You will have **15 minutes** for the production session in which you must make the toys, update a supply inventory sheet, sell the toys, and update a financial record sheet.

<table>
<thead>
<tr>
<th>SCOOTER</th>
<th># PER MODEL</th>
<th>X</th>
<th># MODELS EXPECT TO MAKE</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>square nuts</td>
<td>8</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>screws (4 small + 2 large)</td>
<td>6</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>driving band (tire)</td>
<td>2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pulley (yellow wheel)</td>
<td>2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>perforated strip (7 holes)</td>
<td>2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>perforated strip (3 holes)</td>
<td>1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>washers</td>
<td>2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>axle rod (5cm 2&quot;) (handlebar)</td>
<td>1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>reversed angle bracket</td>
<td>1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spring clip</td>
<td>2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>angle brackets (135 degrees) (kickstand)</td>
<td>1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Total = 28**

---

85
| Square Nuts | Screws (all) | Driving Bands (Tires) | Pulleys (Wheels) | Perforated Strips (7 holes) | Perforated Strips (3 holes) | Washers | Axle Rods (Handlebars) | Reversed Angle Brackets | Spring Clips | Angle Brackets (Kickstands) |

**STARTING INVENTORY:**

|   |   |   |   |   |   |   |   |   |   |   |

**SUBTRACT SUPPLIES USED:**

|   |   |   |   |   |   |   |   |   |   |   |

|   |   |   |   |   |   |   |   |   |   |   |

|   |   |   |   |   |   |   |   |   |   |   |

|   |   |   |   |   |   |   |   |   |   |   |

|   |   |   |   |   |   |   |   |   |   |   |

|   |   |   |   |   |   |   |   |   |   |   |

|   |   |   |   |   |   |   |   |   |   |   |

|   |   |   |   |   |   |   |   |   |   |   |

|   |   |   |   |   |   |   |   |   |   |   |

|   |   |   |   |   |   |   |   |   |   |   |

|   |   |   |   |   |   |   |   |   |   |   |

|   |   |   |   |   |   |   |   |   |   |   |

**CLOSING INVENTORY:**

|   |   |   |   |   |   |   |   |   |   |   |
## FINANCIAL RECORD SHEET

<table>
<thead>
<tr>
<th>DESCRIPTION OF TRANSACTION (i.e. sold 2 scooters)</th>
<th>MARKET PRICE (current selling price)</th>
<th>DEPOSIT (# sold X price)</th>
<th>BALANCE (starting balance = $0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix B

Absolute and Relative Frequency Rating Questionnaires
POST-EXPERIMENT QUESTIONNAIRE FOR PERFORMERS

PARTICIPANT#:__________________________

Using the response scale provided, indicate how often you engaged in the following tasks relative to the other tasks. Be sure to read over all the tasks listed below before attempting to answer the first item.

(For example: I used a tool ("very infrequently", "infrequently", "sometimes", "frequently", or "very frequently") compared to the other tasks listed.)

1 = Very infrequently
2 = Infrequently
3 = Sometimes
4 = Frequently
5 = Very frequently

___ 1. used a tool (includes both the allen wrench and the open-ended wrench)

___ 2. voiced a suggestion concerning what toy to produce

___ 3. took individual pieces out of the supply box

___ 4. voiced a suggestion concerning the company's name

___ 5. connected pieces together with screws and nuts

___ 6. helped a fellow group member with model construction

___ 7. voiced a suggestion concerning the vision statement

___ 8. disassembled parts of a toy model

___ 9. filled out a task division form

___ 10. recorded the company's vision statement

___ 11. updated the inventory sheet
1 = Very infrequently
2 = Infrequently
3 = Sometimes
4 = Frequently
5 = Very frequently

12. assembled a propeller

13. used a writing instrument (for uses other than filling out this questionnaire)

14. filled out the supply request form

15. assembled a toy motor with the tools provided

16. voiced a suggestion concerning how many parts to request

17. tested a final product

18. recorded the company's name

19. used a calculator

20. passed a needed part to a co-worker

21. discussed ways to divide up the tasks among group members

22. counted money

23. hammered a nail

24. put pieces back into the supply box

25. recorded financial information

26. used the instruction sheet (flipped the sheet from one side to the other)
Please indicate **the number of times** you engaged each of the following tasks.

___ 1. used a tool (includes both the allen wrench and the open-ended wrench)

___ 2. voiced a suggestion concerning what toy to produce

___ 3. took individual pieces out of the supply box

___ 4. voiced a suggestion concerning the company's name

___ 5. connected pieces together with screws and nuts

___ 6. helped a fellow group member with model construction

___ 7. voiced a suggestion concerning the vision statement

___ 8. disassembled parts of a toy model

___ 9. filled out a task division form

___ 10. recorded the company's vision statement

___ 11. updated the inventory sheet

___ 12. assembled a propeller

___ 13. used a writing instrument (for uses other than filling out this questionnaire)

___ 14. filled out the supply request form

___ 15. assembled a toy motor with the tools provided

___ 16. voiced a suggestion concerning how many parts to request

___ 17. tested a final product

___ 18. recorded the company's name

___ 19. used a calculator
20. passed a needed part to a co-worker
21. discussed ways to divide up the tasks among group members
22. counted money
23. hammered a nail
24. put pieces back into the supply box
25. recorded financial information
26. used the instruction sheet (flipped the sheet from one side to the other)
POST-EXPERIMENT QUESTIONNAIRE FOR OBSERVERS

PARTICIPANT#: __________ TARGET GROUP MEMBER#: __________

Using the response scale provided, indicate how often your target engaged in the following tasks relative to the other tasks. Be sure to read over all the tasks listed below before attempting to answer the first item.

(For example: She used a tool ("very infrequently", "infrequently", "sometimes", "frequently", or "very frequently") compared to the other tasks listed.)

1  =  Very infrequently
2  =  Infrequently
3  =  Sometimes
4  =  Frequently
5  =  Very frequently

____  1. used a tool (includes both the allen wrench and the open-ended wrench)
____  2. voiced a suggestion concerning what toy to produce
____  3. took individual pieces out of the supply box
____  4. voiced a suggestion concerning the company's name
____  5. connected pieces together with screws and nuts
____  6. helped a fellow group member with model construction
____  7. voiced a suggestion concerning the vision statement
____  8. disassembled parts of a toy model
____  9. filled out a task division form
____ 10. recorded the company's vision statement
____ 11. updated the inventory sheet
1 = Very infrequently
2 = Infrequently
3 = Sometimes
4 = Frequently
5 = Very frequently

12. assembled a propeller
13. used a writing instrument (for uses other than filling out this questionnaire)
14. filled out the supply request form
15. assembled a toy motor with the tools provided
16. voiced a suggestion concerning how many parts to request
17. tested a final product
18. recorded the company's name
19. used a calculator
20. passed a needed part to a co-worker
21. discussed ways to divide up the tasks among group members
22. counted money
23. hammered a nail
24. put pieces back into the supply box
25. recorded financial information
26. used the instruction sheet (flipped the sheet from one side to the other)
Please indicate the number of times your target engaged each of the following tasks.

1. used a tool (includes both the allen wrench and the open-ended wrench)
2. voiced a suggestion concerning what toy to produce
3. took individual pieces out of the supply box
4. voiced a suggestion concerning the company's name
5. connected pieces together with screws and nuts
6. helped a fellow group member with model construction
7. voiced a suggestion concerning the vision statement
8. disassembled parts of a toy model
9. filled out a task division form
10. recorded the company's vision statement
11. updated the inventory sheet
12. assembled a propeller
13. used a writing instrument (for uses other than filling out this questionnaire)
14. filled out the supply request form
15. assembled a toy motor with the tools provided
16. voiced a suggestion concerning how many parts to request
17. tested a final product
18. recorded the company's name
19. used a calculator
20. passed a needed part to a co-worker
21. discussed ways to divide up the tasks among group members
22. counted money
23. hammered a nail
24. put pieces back into the supply box
25. recorded financial information
26. used the instruction sheet (flipped the sheet from one side to the other)
Appendix C
Pre-Experiment Questionnaires
PRE-EXPERIMENT QUESTIONNAIRE FOR PERFORMERS

PARTICIPANT #:__________________

Age__________

Gender (please circle)  M    F

Grade Point Average (overall)__________

Have you ever owned an Erector Set? (Please place a check on the appropriate response.)

(YES)  (NO)

Have you ever played with an Erector Set before? (Please place a check on the appropriate response.)

(YES)  (NO)

Using the scale below, please indicate how much previous experience you have had in using an Erector Set. Place your response in the blank space below.

1 = No experience  
2 = Some experience  
3 = A lot of experience

(Place your response here.)
Using the scale below indicate your level of agreement with the following items. Do not answer how you think you are expected to answer. Answer in an honest fashion.

1 = Strongly Disagree
2 = Disagree
3 = Neither agree nor disagree
4 = Agree
5 = Strongly Agree

___ 1. I feel confident in my ability to assemble the Erector Set models effectively.

___ 2. I think I can eventually reach a high level of performance on this task.

___ 3. I am sure I can learn how to perform this task effectively in a relatively short period of time.

___ 4. I don't feel that I am as capable of assembling the models as other people.

___ 5. On the average, other people are probably much more capable of performing this task than I am.

___ 6. I am a fast learner for these types of tasks, in comparison to other people.

___ 7. I am not sure I can ever reach a high level of performance on this task, no matter how much practice and training I get.

___ 8. It would take me a long time to learn how to perform this task effectively.

___ 9. I am not confident that I can perform this task successfully.

___ 10. I doubt that my performance will be very adequate on this task.
PRE-EXPERIMENT QUESTIONNAIRE FOR OBSERVERS

PARTICIPANT #:__________________________

Age__________

Gender (please circle)  M    F

Grade Point Average (overall)___________

Have you ever owned an Erector Set? (Please place a check on the appropriate response.)

(YES)    (NO)

Have you ever played with an Erector Set before? (Please place a check on the appropriate response.)

(YES)    (NO)

Using the scale below, please indicate how much previous experience you have had in using an Erector Set. Place your response in the blank space below.

1 = No experience
2 = Some experience
3 = A lot of experience

(Place your response here.)
Using the scale below indicate your **level of agreement** with the following items. Do not answer how you think you are expected to answer. Answer in an honest fashion.

1 = *Strongly Disagree*  
2 = *Disagree*  
3 = *Neither agree nor disagree*  
4 = *Agree*  
5 = *Strongly Agree*

___ 1. I feel confident in my ability to determine how the videotaped organization runs.

___ 2. I think I can eventually reach a high level of performance on this task.

___ 3. I am sure I can learn how to perform this task effectively in a relatively short period of time.

___ 4. I don't feel that I am as capable of evaluating group members' behavior as other people.

___ 5. On the average, other people are probably much more capable of performing this task than I am.

___ 6. I am a fast learner for these types of tasks, in comparison to other people.

___ 7. I am not sure I can ever reach a high level of performance on this task, no matter how much practice and training I get.

___ 8. It would take me a long time to learn how to perform this task effectively.

___ 9. I am not confident that I can perform this task successfully.

___ 10. I doubt that my performance will be very adequate on this task.
Appendix D
Importance and Interest Rating Questionnaires
For Performers:

Using the response scale provided, indicate the importance of each of the following tasks in achieving your group's overall goal or vision.

1 = Very minor importance  
2 = Low importance  
3 = Intermediate importance  
4 = High importance  
5 = Extreme importance

___ 1. using a tool (includes both the allen wrench and the open-ended wrench)

___ 2. voicing a suggestion concerning what toy to produce

___ 3. taking individual pieces out of the supply box

___ 4. voicing a suggestion concerning the company's name

___ 5. connecting pieces together with screws and nuts

___ 6. helping a fellow group member with model construction

___ 7. voicing a suggestion concerning the vision statement

___ 8. disassembling parts of a toy model

___ 9. filling out a task division form

___ 10. recording the company's vision statement

___ 11. updating the inventory sheet

___ 12. assembling a propeller

___ 13. using a writing instrument (for uses other than filling out this questionnaire)

___ 14. filling out the supply request form
1 = Very minor importance
2 = Low importance
3 = Intermediate importance
4 = High importance
5 = Extreme importance

15. assembling a toy motor with the tools provided
16. voicing a suggestion concerning how many parts to request
17. testing a final product
18. recording the company's name
19. using a calculator
20. passing a needed part to a co-worker
21. discussing ways to divide up the tasks among group members
22. counting money
23. hammering a nail
24. putting pieces back into the supply box
25. recording financial information
26. using the instruction sheet (flipped the sheet from one side to the other)
For Performers:

Using the response scale provided, indicate how interesting you found each of the following tasks.

1 = Not interesting at all
2 = Slightly interesting
3 = Neutral
4 = Somewhat interesting
5 = Very interesting

___ 1. using a tool (includes both the allen wrench and the open-ended wrench)
___ 2. voicing a suggestion concerning what toy to produce
___ 3. taking individual pieces out of the supply box
___ 4. voicing a suggestion concerning the company's name
___ 5. connecting pieces together with screws and nuts
___ 6. helping a fellow group member with model construction
___ 7. voicing a suggestion concerning the vision statement
___ 8. disassembling parts of a toy model
___ 9. filling out a task division form
___ 10. recording the company's vision statement
___ 11. updating the inventory sheet
___ 12. assembling a propeller
___ 13. using a writing instrument (for uses other than filling out this questionnaire)
___ 14. filling out the supply request form
1 = Not interesting at all
2 = Slightly interesting
3 = Neutral
4 = Somewhat interesting
5 = Very interesting

___ 15. assembling a toy motor with the tools provided
___ 16. voicing a suggestion concerning how many parts to request
___ 17. testing a final product
___ 18. recording the company's name
___ 19. using a calculator
___ 20. passing a needed part to a co-worker
___ 21. discussing ways to divide up the tasks among group members
___ 22. counting money
___ 23. hammering a nail
___ 24. putting pieces back into the supply box
___ 25. recording financial information
___ 26. using the instruction sheet (flipped the sheet from one side to the other)
For Observers:

Using the response scale provided, indicate the importance of each of the following tasks in achieving the group's overall goal or vision.

1 = Very minor importance
2 = Low importance
3 = Intermediate importance
4 = High importance
5 = Extreme importance

___ 1. using a tool (includes both the allen wrench and the open-ended wrench)
___ 2. voicing a suggestion concerning what toy to produce
___ 3. taking individual pieces out of the supply box
___ 4. voicing a suggestion concerning the company's name
___ 5. connecting pieces together with screws and nuts
___ 6. helping a fellow group member with model construction
___ 7. voicing a suggestion concerning the vision statement
___ 8. disassembling parts of a toy model
___ 9. filling out a task division form
___ 10. recording the company's vision statement
___ 11. updating the inventory sheet
___ 12. assembling a propeller
___ 13. using a writing instrument (for uses other than filling out this questionnaire)
___ 14. filling out the supply request form
1 = Very minor importance
2 = Low importance
3 = Intermediate importance
4 = High importance
5 = Extreme importance

15. assembling a toy motor with the tools provided
16. voicing a suggestion concerning how many parts to request
17. testing a final product
18. recording the company's name
19. using a calculator
20. passing a needed part to a co-worker
21. discussing ways to divide up the tasks among group members
22. counting money
23. hammering a nail
24. putting pieces back into the supply box
25. recording financial information
26. using the instruction sheet (flipped the sheet from one side to the other)
For Observers:

Using the response scale provided, indicate how interesting you believe each of the following tasks are.

1 = Not interesting at all
2 = Slightly interesting
3 = Neutral
4 = Somewhat interesting
5 = Very interesting

1. using a tool (includes both the allen wrench and the open-ended wrench)
2. voicing a suggestion concerning what toy to produce
3. taking individual pieces out of the supply box
4. voicing a suggestion concerning the company's name
5. connecting pieces together with screws and nuts
6. helping a fellow group member with model construction
7. voicing a suggestion concerning the vision statement
8. disassembling parts of a toy model
9. filling out a task division form
10. recording the company's vision statement
11. updating the inventory sheet
12. assembling a propeller
13. using a writing instrument (for uses other than filling out this questionnaire)
14. filling out the supply request form
1 = Not interesting at all
2 = Slightly interesting
3 = Neutral
4 = Somewhat interesting
5 = Very interesting

15. assembling a toy motor with the tools provided
16. voicing a suggestion concerning how many parts to request
17. testing a final product
18. recording the company’s name
19. using a calculator
20. passing a needed part to a co-worker
21. discussing ways to divide up the tasks among group members
22. counting money
23. hammering a nail
24. putting pieces back into the supply box
25. recording financial information
26. using the instruction sheet (flipped the sheet from one side to the other)