The relationship between typical and maximum performance across levels of job complexity

Ken Jordan
University of Nebraska at Omaha
THE RELATIONSHIP BETWEEN TYPICAL AND MAXIMUM
PERFORMANCE ACROSS LEVELS OF JOB COMPLEXITY

A Thesis

Presented to the
Department of Psychology

and the
Faculty of the Graduate College

University of Nebraska

In Partial Fulfillment
of the Requirements for the Degree
Master of Arts

University of Nebraska at Omaha

by

Ken Jordan

August, 1990
THESIS ACCEPTANCE

Accepted for the faculty of the Graduate College, University of Nebraska, in partial fulfillment of the requirements for the degree Master of Arts, University of Nebraska at Omaha.

Thesis Committee

Name                   Department

Wayne Horin          Psychology
David A. Hinton       Public Administration
James M. Thomas       Psychology

Chairman: Lisa Scherer
Date: 7/30/90
ACKNOWLEDGEMENTS

I would like to acknowledge my committee members: Dr. Wayne Harrison, Dr Dave Hinton, Dr. Lisa Scherer, and Dr. Jim Thomas for their helpful comments and suggestions on this project. I would also like to thank Dr. Ray Millimet and Dr. Richard Wikoff for their patience and guidance during the analysis. I am also indebted to Teresa Cochran who helped with the data entry, Shar Whigham and Paul Pietig for their critiques of the manuscript and Vern Peterson for assisting with the manipulation (who else but Vern). Last, and most important, I thank Tom Kranda and Karen Gardner for making this project possible.
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Typical and Maximum Performance</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Job Complexity</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Summary of Job Complexity Literature</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Relevance of Typical/Maximum Distinction</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Application of Typical Measures with Utility</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Review of I/O Literature</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Summary Points of Literature</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Hypotheses</td>
<td>24</td>
</tr>
<tr>
<td>II</td>
<td>METHOD</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Subjects</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Measures</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Overview</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Job complexity pilot survey</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Maximum/typical performance</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Wrap time</td>
<td>33</td>
</tr>
</tbody>
</table>
## TABLE OF CONTENTS (Continued)

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relationship Between Wrap Time and Complexity</td>
<td>35</td>
</tr>
<tr>
<td>Distribution of Calls</td>
<td>36</td>
</tr>
<tr>
<td>Procedure</td>
<td>38</td>
</tr>
<tr>
<td><strong>III RESULTS</strong></td>
<td>43</td>
</tr>
<tr>
<td>ANOVA Results</td>
<td>44</td>
</tr>
<tr>
<td>Length of Maximum Test Period</td>
<td>53</td>
</tr>
<tr>
<td>Follow-Up Survey Results</td>
<td>54</td>
</tr>
<tr>
<td><strong>IV DISCUSSION</strong></td>
<td>57</td>
</tr>
<tr>
<td>Previous Research</td>
<td>59</td>
</tr>
<tr>
<td>Implications for Testing</td>
<td>61</td>
</tr>
<tr>
<td>Future Directions</td>
<td>63</td>
</tr>
<tr>
<td>Conclusion</td>
<td>69</td>
</tr>
<tr>
<td>References</td>
<td>70</td>
</tr>
<tr>
<td>Appendix A: Job Complexity Survey</td>
<td>78</td>
</tr>
<tr>
<td>Appendix B: Memo to Subjects</td>
<td>82</td>
</tr>
<tr>
<td>Appendix C: Consent Form</td>
<td>84</td>
</tr>
<tr>
<td>Appendix D: Follow-Up Survey</td>
<td>87</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Descriptive Statistics for Job Complexity Survey</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>Subjects’ Mean Number of Calls and Percentage of Calls by Category and Condition</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>Correlation Matrix by Condition</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>Summary Table: Repeated Measures ANOVA of Wrap Time by Complexity and Performance</td>
<td>47</td>
</tr>
<tr>
<td>5</td>
<td>Mean Wrap Times by Performance Condition and Job Complexity Level</td>
<td>48</td>
</tr>
<tr>
<td>6</td>
<td>Simple Effects Analyses on Mean Wrap Time for Complexity at Fixed Levels of Performance</td>
<td>52</td>
</tr>
<tr>
<td>7</td>
<td>Follow-Up Survey: Subjects’ Mean Ratings for Items Within Clusters</td>
<td>56</td>
</tr>
<tr>
<td>8</td>
<td>Stem and Leaf Plot: Self-Report of Effort During Maximum Performance Period</td>
<td>66</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mean Wrap Time (transformed) as a Function of Performance Condition and Level of Job Complexity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>51</td>
</tr>
</tbody>
</table>
Abstract

Previous research by Sackett, Zedeck, and Fogli (1988) found that measures of typical and maximum work performance correlate .36. The present study expands upon this performance distinction by examining differences in maximum and typical performance of 69 tele-services representatives across three levels of job complexity. Computerized performance data were collected for each subject, grouped according to job complexity, and analyzed using a repeated measures design. Results showed that the overall interaction between complexity and typical/maximum performance was not significant, but the correlation of subjects’ performance across all levels of complexity was moderately high, with performance measures showing significant improvement in the maximum condition.
Chapter I  
INTRODUCTION  

Historically, the field of industrial psychology has concerned itself with the application of measurement principles to personnel decisions (Landy, 1985). Central to this principle is the practice of performance appraisal (PA) and the use of different criterion measures. Despite years of experimentation and volumes of research, we are still left with what Smith (1976) calls the "criterion problem." That is, the question remains as to which method of measuring job performance (supervisor ratings, supervisor rankings, production output, work samples, peer assessment, quality of work, etc.) results in the most useful and psychometrically sound measures. Zedeck and Cascio (1984) concluded, "In essence we have developed procedures and conducted comparative studies of appraisal methods using indices that are unclear, ambiguous, and in part, wrong" (p. 471). Indeed, it appears that we still have not progressed beyond what Wherry (1957) refers to as "the dark ages of criteria" (p. 1).

Research efforts aimed at resolving this problem have perseverated in their attempts to find one performance measure (usually a subjective rating) that can be applied to a variety of formats. As Landy and Farr (1983) note, the search for the "Holy Format" has not remedied our performance appraisal problems, and other methods should be
investigated. Although this viewpoint is not new, it seems that very little has been accomplished regarding the development of new appraisal methods.

A popular suggestion regarding ways to handle the criterion problem is to use multiple criteria (Dunnette, 1963; Ghiselli, 1956; Guion, 1965; and Muchinsky, 1987). Cascio (1987) notes that most applied psychologists agree that job performance is multi-dimensional in nature. If we are to gain an understanding of the ingredients necessary for successful performance, it seems clear that multiple criteria (rather than composites) are the preferred method (Schmidt & Kaplan, 1971). A recent development that could be applied towards the goal of using multiple criteria is the distinction between typical and maximum performance measures (Sackett, Zedeck, & Fogli, 1988).

**Typical and Maximum Performance**

The primary difference between typical and maximum performance concerns the worker's objective and goals. Maximum performance is analogous to a test-like situation where the objective is to maximize work output, scores, productivity, etc. (depending on the relevant criterion). Achievement tests (e.g., Graduate Record Exam, Certified Public Accountant Exam), aptitude tests (e.g., Differential Aptitude Test, Bennett Test of Mechanical Comprehension, General Aptitude Test Battery) and specific ability tests (e.g., Purdue Pegboard, Snellen Eye Chart) are prime
examples of maximum performance measures. Data from these
tests are collected under controlled, short term
circumstances with the underlying assumption that test
takers will attempt to maximize their scores. These
c characteristics correspond closely to the criteria Sackett
et al. (1988) describe for a maximum performance measure:
"1) explicit awareness that one is being evaluated; 2)
awareness and acceptance of implicit or explicit
instructions to maximize effort; and 3) measured over a
short enough time duration that the performer’s attention
remains focused on the accepted goal..." (p. 482).
Contrasted to the notion of maximum performance is what
Sackett et al. (1988) call typical performance. In this
case, subjects are not aware that performance is being
monitored; they would not be consciously trying to maximize
performance, and measures would be aggregated over time.
Typical performance is the rate and/or quality of work that
one is capable of producing consistently. Examples of
typical performance would be a student’s homework grade for
a given class averaged over the semester, or the average
speed of a runner over a 10 K run.

The previously mentioned research by Sackett et al.
(1988) showed only a modest correlation between typical and
maximum performance. These authors examined the
performance of grocery store cashiers and found that
typical and maximum speed and typical and maximum accuracy
correlated .36 and .16 respectively, after correction for attenuation. Based on these low correlations, they concluded that typical and maximum measures are not equivalent. The implication is that typical and maximum measures tap different dimensions of performance and do not yield comparable information.

As an example, if a maximum performance test (knowledge test, work sample) were used as a selection device, the employer would be basing the decision on a maximum measure which may have little or no relationship to the applicant’s long-term typical performance. As Sackett et al. (1988) note, use of a maximum performance test in such situations implies that maximum performance measures and typical performance measures are interchangeable. The existing evidence does not support this assumption. Likewise, appraisal methods using maximum and typical indices that do not take this into account are also in danger of confounding the two types of measures. Failure to acknowledge these differences would result in measures that are less specific and not as meaningful, especially for purposes of employee development. Another benefit from research such as this is the recognition that performance cannot be measured on a common metric across different job settings. The realization that different jobs necessitate different types of performance, and thus different types of performance measures, would help to clarify what our
appraisal methods are measuring and how they can be used most effectively.

The present study examined the differences between typical and maximum performance measures across three levels of job complexity. For the purposes of this study, job complexity was conceptualized as the amount of information-processing and problem-solving required of the incumbent (Hunter, Schmidt, & Judiesch, 1990; Schmidt, Hunter, & Pearlman, 1981).

Specifically, the objectives of this study were to: (a) determine whether the results of Sackett et al. (1988) could be replicated using a different sample, and (b) examine the effects of job complexity as a moderator in the relationship between typical and maximum performance measures.

Job Complexity

Job complexity has been measured using a variety of methods (Ferris & Gilmore, 1985) and is difficult to operationalize. The notion of job complexity has been used extensively in job-related research as measured by such models as the Job Diagnostic Survey (JDS) (Hackman & Oldham, 1976) and the Job Characteristics Inventory (JCI) (Sims, Szilagyi, & Keller, 1976), and has recently been shown to have construct validity (Gerhart, 1988). The current literature discusses several ways that job complexity can be approached. Essentially, the measures
that are available can be classified in terms of: 1) measures based on classifications of work activities made by external observers (U.S. Department of Labor, 1977); 2) the incumbent’s perception of job complexity and work-oriented activities (Hackman & Oldham, 1976; Sims et al., 1976); 3) social and information processing cues from other co-workers (Salanzik & Pfeffer, 1978); and 4) individual differences such as field dependence (Stone, 1979).

Complexity, as defined by the 1974 Dictionary of Occupational Titles (DOT), pertains to the three digit codes used to classify work activities according to skill level. This system was developed by Fine (1968) to categorize jobs according to the complexity of skills which incumbents use in dealing with data, people and things. The different skill levels within each category are used as the basis of the job classification scheme. Rousseau (1982) has found that job complexity, as defined by the data, people, and things (DPT) code, shows a stronger relation to only the data and people scales when measuring perceived job characteristics (autonomy, task significance and variety) using the JDS (Hackman & Oldham, 1976). The use of skills involving data and to a lesser extent people, may have the strongest impact in determining worker job perceptions. The things scale was least important in determining how an individual perceives job complexity. Complexity might, therefore, be more closely related to
skills involving people and data to the exclusion of skills required to interact with machines, equipment, tools, etc.

In a 3 year longitudinal study by Gerhart (1988), evidence was found to support the contention that incumbents' perceptions of job complexity (IPJC), as measured by a short version of the JCI (Sims et al., 1976) are related to the DOT measures of complexity. The results of Gerhart's study show convergent validity between incumbent-based measures of complexity and those measures used in the DOT derived through job analysis methods. In addition, discriminant validity was demonstrated by a lack of common variance between DOT-complexity and situational or individual variables such as education, tenure, wage, and physical demands. This indicates that IPJC is a function of measurable differences among jobs and not the variation attributed to the situational factors mentioned above. Gerhart (1988) further concludes that these results in conjunction with Gerhart (1987) constitute evidence of construct validation for the use of IPJC.

Previous research by Ferris and Gilmore (1985) and London and Klimoski (1975a) has also explored the relation between incumbent-based measures of job complexity and their relation to situational and individual factors. Their results concur with Gerhart (1988) in suggesting that overall job complexity is not consistently related to individual factors such as ratings of satisfaction and
effectiveness. London and Klimoski (1975b) also theorize that job complexity acts as a moderator variable between performance ratings and ratings of satisfaction and self-esteem. They suggest that performance and satisfaction are related to job complexity in the pattern of an inverse U (Landy & Farr 1983; London & Klimoski 1975b). For each worker there is an optimal level of job complexity under which that person functions best. In this case, job complexity was measured by taking the difference between incumbent's responses when asked "how much complexity is there in your job" and "how much should there be." If there was no difference between these measures, it was assumed that an optimal level of job complexity existed. According to London and Klimoski (1975b), if the level of complexity is increased, performance declines due to frustration. If complexity is lower than the optimal level, the incumbent may perceive him/herself as being underutilized and will again experience frustration.

Although the JCI will not be used in the present study, this information supports the practice of using job complexity information gathered from the employees. It also increases confidence that estimates of complexity are not biased by individual or situational variables such as education, tenure, or pay differences.
Summary of Job Complexity Literature

Several conclusions can be made regarding the literature on complexity. Although a variety of methods are available to measure complexity, it seems that surveying the incumbents is the preferred method (Landy & Farr 1983; London & Klimoski 1975a). Not only is this the most common, but the results correlate with findings derived through job analysis methods used by the DOT (Gerhart, 1988). Complexity has also been shown to be strongly related to the people and things code used by the DOT (Rousseau, 1982). This is particularly relevant to the present study because the subjects used in this experiment work mostly with customers and data. London and Klimoski (1975b) also found that complexity was related to several performance variables. In light of these findings, it was reasonable to suggest that complexity would be implicated in the relationship between typical and maximum performance.

Relevance of the Typical/Maximum Distinction

Although the differences between typical and maximum performance have been acknowledged previously (Weiner & Stewart, 1974), typical performance scores have not been recognized as useful measures of productivity. One of the few areas where typical performance was an acceptable criterion was in the use of personality assessment or interest inventories --almost all other tests were oriented
towards obtaining the maximum or highest possible score. However, in many work settings, especially those concerned with productivity or defect rate, typical rather than maximum measures should be the basis for performance appraisal. The reason is that typical measures have a more direct relationship to the organization's goals because they reflect what the worker actually does during the normal work day.

The Society for Industrial and Organizational Psychology (1987) has stated, "It must be possible to obtain or develop a relevant, reasonably reliable and uncontaminated criterion measure. Of these characteristics, the most important is relevance" (p.7). Hence, one rationale for the typical/maximum distinction can be based on relevance. That is, the appropriate fit (i.e., logical relationship) between the actual criterion used to measure performance and the conceptual criterion representing the organization's goals should be evident (Cascio, 1987; McCormick & Tiffin, 1974). It is argued here that typical measures used as the actual or operational criteria may in some cases have a more direct relationship to the conceptual criterion because the domain of job behaviors is better represented.

In addition, for jobs that are routine or repetitious (assembly line, data entry, etc.) the case for using a typical measure could be based simply on the deficiency or
contamination which accompanies the use of maximum performance measures. That is, for a very routine job such as a grounds keeper or stock person, use of a job knowledge test for promotions or merit raises may be deficient because it lacks a measure of physical stamina and manual dexterity. Strict reliance on the job knowledge test as a criterion would also be contaminated by measures such as reading comprehension and possibly test anxiety. This concern becomes more important if management is interested only in measuring the employee's proficiency for stocking shelves rather than promotion potential or some future-oriented criteria.

Similarly, if the job requires a high level of skill or ability, such as an emergency room physician or a research scientist, a maximum performance measure might be more relevant. The logical relationship is based on the similarity between the objectives and procedures of the maximum performance measure, and the actual job behaviors. With more complex jobs that periodically require high levels of performance, the maximum performance measure would be the favored method. It has the highest probability of providing information that is vital to determining one's ability to fulfill the job demands. Specifically, a maximum measure would demonstrate whether or not the incumbent has the resources and skills necessary for high achievement. Even though the emergency room
physician may not be required to function at a maximum ability level constantly, any measure that taps this ability would give useful information about the potential to do so. Similarly, police officers are not required to function under conditions of high stress everyday. However, it is important that they are capable of doing so in order to perform adequately in situations that are dangerous or life-threatening. For this type of occupation, performance is more likely dependent upon ability rather than motivation. The best indication of whether they can reach high levels of performance is to measure their capabilities using a maximum performance test (P. R. Sackett, personal communication, Nov. 29, 1988).

The relevance of the criterion measure to the conceptual criterion is vitally important for making valid personnel decisions. Any evaluative standard (criterion) used to measure performance should have a causal link to the conceptual criterion or the general purpose of the organization. For example, if the organizational goals are to reduce production costs per unit, average defect rate would directly influence this goal and therefore be a relevant measure. The 1978 Equal Employment Opportunity Commission (EEOC) Uniform Guidelines clearly support this viewpoint by stating, "Whatever criteria are used should represent important or critical work behavior(s) or work outcomes" (p.161).
Application of Typical Measures with Utility

Relevant criterion measures are not only more equitable to employees, they are also valuable to the employer. Typical measures can be used in a variety of applications associated with performance appraisal. As an example, typical measures can be used for calculating utility estimates of appraisal methods. Utility figures represent the amount of money a company can save by using a more valid selection procedure (Schmidt, Hunter, McKenzie & Muldrow, 1979) or improving their performance appraisal method (Landy, Farr, & Jacobs, 1982). When an organization uses utility estimates, they will be using an estimate based on means. A problem with this is that means are affected by extreme scores, which is precisely what one is likely to find using maximum measures. One solution would be to use mean scores based on typical performance because they are more representative of worker output. Typical performance data are derived from the employee’s day-in day-out work rather than the occasional peaks and valleys. Use of data such as these would be most accurate in jobs that are routine and operate in a stable environment.

In the Decision Theoretic Utility Model originated by Brogden (1949) and later refined by Schmidt and Hunter (1977) and Schmidt, Hunter, McKenzie, and Muldrow (1979), the role of "mean job performance" is crucial in figuring the validity of a selection instrument. This model has
been adapted for use with performance appraisal to determine the value of feedback on productivity. The equation:

$$\Delta U = t N_s d_t SD_y (C_1 - C_2)$$

is similar to the one used by Schmidt et al. (1979) to estimate the utility of a selection instrument. Here, \(\Delta U\) represents change in utility or the expected gain in dollars. The number of appraisals for a given employee over the anticipated employment period is represented by \(t\). \(N_s\) is the number of people affected by the new appraisal system and \(d_t\) represents the average difference in true scores between those who received feedback with the new system and those who did not. \(SD_y\) is an estimate of the pooled standard deviation in dollars, and \((C_1 - C_2)\) represents the per employee cost differences between the old and new evaluation system.

Although it is not specifically stated, the most accurate estimates for calculating \(d_t\) would come from typical work. Because \(d_t\) is based on the difference between means, and there is evidence suggesting that modal performance does not correspond to maximum performance (Kane, 1982), worker output should be measured by typical work estimates that are less likely to be biased by extreme scores.

More recently, Hunter, Schmidt, and Judiesch (1990) have shown that the standard deviation in Performance \((SD_p)\)
varies with job complexity. As complexity of the job increases, so does the amount of variation observed in productivity. Hunter et al. (1990) found that the differences between low, medium, and high complexity jobs correspond to an increase in SDp starting from 19% for low complexity, 32% for medium complexity and 48% for high complexity. They aptly point out that these performance differences may translate into considerable monetary gains (or losses) when selecting employees. If one standard deviation unit for a high complexity job equals a 48% difference in output as compared to 19% difference for a low complexity job, it is understandable that more emphasis should be place on the selection of applicants for high complexity jobs.

Given that Hunter et al. (1990) found differences in productivity among jobs of varying complexity, it was reasonable to expect complexity to interact with typical/maximum performance measures. Due to the increased variability of high complexity jobs, obtaining typical performance measures for purposes of selection becomes even more important. A point that Hunter et al. (1990) did not consider is that increases in performance variability, as noted for high complexity jobs, make it more probable that scores on a selection test will correspond to an increase in the overall range of maximum scores as compared to the range for typical scores. This would suggest that typical
performance measures would be of increasing importance as the complexity of the job increased.

For these reasons typical performance, in comparison to achievement or aptitude tests, may be more useful because it represents the amount of work that the employer can consistently expect from the employee. For purposes of performance appraisal, the typical/maximum dimension may be a partial solution to the criterion problem which has hindered appraisal efforts for many years. By opening new possibilities such as this, I/O psychologists may be able to develop measures that are more sensitive to different types of work behaviors instead of merely developing new methods that measure the same thing (e.g., BARS, BOS, BES).

In addition, typical performance measures are useful for utility estimates which have become increasingly important in figuring productivity gains using valid appraisal and selection procedures.

**Review of the I/O Literature**

It is surprising that the relatively simple distinction between typical and maximum performance has not been the subject of empirical research until recently (Sackett et al. 1988). A review of the I/O literature has shown that this distinction in performance measures has not been acknowledged, much less considered, as a viable source of performance data. Several authors have touched on the typical/maximum topic, but none have addressed it
specifically. Kingsbury (1933) noted that workers may perform at different levels of proficiency on various dimensions of work, but may still make comparable overall contributions to the organization. Kane (1982) developed the Performance Distribution Assessment (PDA) which incorporates the notion of typical and maximum performance within individual performance dimensions. This method of performance appraisal considers different levels of performance (e.g., exceptional, average, poor) along separate dimensions or tasks, weighted by any external factors that may constrain performance (Kane & Lawler, 1979). The basic tenet of PDA is that appraisal formats should take into consideration the opportunities that a ratee had to perform his/her work at the various levels of proficiency. Kane (1982) also noted that two individuals may have the same modal performance but show large differences in their maximum performance. Similarly, two individuals may show large differences in typical performance measures but have similar maximum performance scores. Implicit in Kane’s writing is the idea that performance measures taken under different conditions are not equal. Although the PDA is more concerned about opportunities to perform at different levels of proficiency, it also supports the typical/maximum distinction first noted by Sackett et al. (1988). In a
sense, Kane has used the variable of opportunity to scale differences in performance among employees.

Lee (1985) presents a slightly different approach by suggesting that appraisal formats be categorized according to the type of task for which they are best suited. In her research, Lee (1985) developed a classification system that matches task type (i.e., is the task observable and does it result in measurable outcomes) with a specific appraisal format. Although dimensions other than typical and maximum performance were used, her idea was conceptually similar because it emphasized the use of different appraisal formats suited to particular types of jobs. This type of framework may also be useful for organizing different criterion measures in relation to the tasks required by different jobs.

Differences in levels of performance have also been discussed in the literature. Landy and Farr (1983) suggest that performance is best measured in terms of deviation from an "optimal" level of performance. Optimal was defined as an arbitrary point where performance and employee satisfaction were high and turnover and accidents were kept to a minimum. The author’s point is that the best level of performance may not be the same as the maximum level. The optimal level may be somewhat lower in comparison to the maximum level in order to limit accidents and turnover. Again, this is not direct support for the
value of typical measures. However, there is evidence that researchers and managers have considered performance from a perspective similar to the typical/maximum dimension.

To date, only one published study (Sackett et al. 1988) has dealt specifically with the maximum/typical issue as applied to performance measures. Sackett et al. (1988) examined differences in maximum and typical performance measures for supermarket cashiers. Speed (number of items per minute) and accuracy (number of voids per shift) were measured under both average (typical) and test (maximum) conditions. Data were collected from the computerized cash registers because of the improved accuracy and reliability they offered. Maximum performance data (speed and accuracy) were collected using timed tests. The employees were asked to process several shopping carts filled with identical items. They were instructed to place equal emphasis on both speed and accuracy. Typical performance was assessed unobtrusively by using data collected by the cash register systems and averaged over a four-week period. Subjective ratings of each cashier's performance were also collected from the supervisors and later correlated with the typical and maximum measures.

Results of this study suggest that typical and maximum performance are not highly correlated. For experienced cashiers, measures of typical and maximum speed correlated .36 and accuracy correlated .16. From these relatively low
correlations, it seems likely that these performance measures are tapping independent criteria. Even though all other aspects of the work situation were consistent, large variances in behavior were observed between typical and maximum measures.

The implication of these results suggest that differences between typical and maximum work behavior should be considered when evaluating performance data. Appraisal methods should focus on the behavior(s) which most closely approximates work activities and conditions. At the least, typical/maximum influences should be separated from each other (when necessary) so that performance measures are not clouded by extraneous or unwanted influences. By disaggregating the dimensions that are involved in the rating process, more reliable information can be obtained from subjective rating methods (Sanchez & Levine, 1989).

Although very little data are available on the typical/maximum distinction, indirect support for its use can be found in other areas of research. One fact that is hard to ignore is the lack of convergence found among different performance measures reported in the literature. Numerous studies (Alexander & Wilkins, 1982; Heneman, 1980; and Meyer, 1987) have revealed low correlations between subjective and objective performance measures. Other research (Nathan & Alexander, 1988; Vance, MacCallum,
Coover, & Hedge, 1988) contradicts these findings. Clarification on these issues is difficult to find in light of the widespread psychometric problems (leniency, halo reliability, etc.). One possible reason for these inconclusive results may be that different dimensions of performance are being sampled during the appraisal process. While objective performance measures tend to be more reliable, subjective measures such as supervisor ratings may be based only on specific dimensions which are largely unknown. Sackett et al. (1988) noted that supervisor's ratings correlated most highly with maximum speed of the cashiers. The probable cause of such a correlation is that supervisors only observe the cashiers during peak work flow and are primarily concerned with whether or not cashiers can function efficiently under such conditions. Because supervisors typically help out during these times, their observations are based on instances of maximum performance and, thus, it is reflected in the ratings (Sackett et al., 1988). For these reasons, it seems clear that the typical/maximum distinction may help to explain the lack of convergence among performance measures gathered from different sources.

The implications of the Sackett et al. (1988) research suggest several noteworthy issues. The authors point out that maximum performance predictors are typically used as criteria in validation studies. In view of these findings,
however, this practice would be acceptable only if the employer was interested in predicting maximum performance capabilities. The drawback is that maximum performance measures do not guarantee the same performance in everyday work. Often the employer is more interested in long term, average performance. In this case, the typical measure would be more useful. Furthermore, typical performance measures appear to be more reliable because subjective ratings are often biased due to insignificant fluctuations in performance (DiNisi & Stevens, 1981). Performance measures that use data aggregated over time have also proved to be more reliable (Epstein, 1980; Rambo, Chomiak, & Price, 1983).

**Summary Points of Literature**

The usefulness of the typical/maximum distinction in performance measures can be shown in several ways. First, the work by Sackett et al. (1988) demonstrated a lack of convergence between typical and maximum measures supporting the notion of using this dimension as a source of multiple criteria. However, typical and maximum measures do not always have to be used together. The circumstances under which each type of measure would be appropriate will be indicated by the needs of the employee and employer as well as the organizational setting (Miles, 1980). Secondly, the logical relationship between actual criterion measures and conceptual criterion supports the usefulness of this method
based on relevance. If the job setting is a highly formalized, mechanistic, and routine setting where inputs and outputs are stable and predictable, typical performance measures may be more appropriate because the criterion measure and actual job performance have many of the same qualities in common.

Third, typical measures derived from aggregated data or actual work samples have proven to be more accurate and reliable as compared to test scores and supervisor ratings (Epstein, 1980; Rambo et al., 1983). Finally, job complexity has been noted to be a possible moderator variable in the relationship between typical and maximum measures (Muchinsky, 1987; Rambo et al., 1983; Sackett et al., 1988). Current research (Hunter et al. 1990) has also shown that increases in job complexity are positively related to increases in the standard deviation of mean output ($SD_p$). Further understanding of the relationship between complexity and the typical/maximum performance distinction may facilitate the correct use and application of appraisal methods and selection instruments.

The purpose of the present study was to explore the relation of typical and maximum performance across three levels of job complexity (high, medium, & low). Tele-services operators served as subjects and levels of task complexity were established by rating the various types of customer calls. A performance measure known as "wrap time"
(measure of time spent performing necessary tasks after a caller hangs-up) was used as the dependent measure. It was hypothesized that as complexity of the tasks increased, differences between typical and maximum scores would decrease. In other words, low complexity tasks should show considerable improvement when maximum scores are compared to typical scores. However, the difference between typical and maximum scores should be considerably less when high complexity tasks are examined. Sackett et al. (1988) hypothesized that at higher levels of complexity, ability (rather than motivation) would play an increasingly important role in determining maximum performance. Their research, however, did not offer any evidence to support that claim. The focus of this study was to test that relationship using a design that would control for many of the outside influences and look specifically at the impact of complexity on typical and maximum performance.

Hypotheses

Hypothesis I. Based on the results of a pilot survey (discussed in method section) showing a positive relationship between mean wrap scores (dependent variable) and complexity, a significant overall main effect for complexity levels was hypothesized.

Hypothesis II. It was hypothesized that the type of performance measure (maximum/typical) will show a significant overall main effect. When comparing typical
and maximum measures, it was expected that maximum measures would have the lowest average means.

**Hypothesis III.** It was hypothesized that task complexity would interact with the type of performance measure. A significant interaction would suggest that one or possibly both measures (typical/maximum) are not equivalent across levels of complexity. The consequences of such a finding were discussed earlier. Because this was the primary hypothesis, the simple main effects were also tested.

**Hypothesis IV.** Differences between typical and maximum performance were hypothesized to be the greatest in the low complexity condition compared to both the medium and high complexity conditions.
Chapter II

METHOD

Subjects

Sixty-nine inbound tele-services representatives (operators) working for a medium sized insurance company served as subjects in this experiment. Job demands for the operators included answering incoming calls from policy owners and potential customers, describing policy benefits, answering questions about claims and using the computer to call up information or update files. Subjects' ages ranged from 70 to 19 years with a mean of 35 years, SD = 12.8 years; average tenure in the current position ranged from 19.5 years to 3 months with a mean of 3.5 years and SD = 3.3 years. Approximately 86% of the operators were female, and only those operators who had completed training were used for data collection purposes. Training is necessary to ensure that operators possess the requisite skills to use the computer, answer basic questions about insurance products, and correctly code calls. Approximately 80 tele-services representatives were employed by the company during the data collection; all were asked to participate in the study, only three declined. The subject pool was also reduced due to incomplete data for some operators. Absences and terminations further reduced the available number of operators to 69.
Measures

Overview. A pilot survey was given to 20 job knowledge experts to determine the relative complexity for each type of call. Based on the results of this survey, calls were grouped into three levels of complexity for use as an independent variable. The second independent variable, typical/maximum performance measures, consisted of collecting wrap-up time data under the following conditions: 1) when the operators were unaware that their performance was being monitored (typical); and 2) when they were asked to maximize their output with the knowledge that their performance measures would be used for research purposes (maximum). The data were analyzed to examine the differences between maximum and typical performance across levels of complexity. The dependent variable is a performance measure known as "wrap-up time." Wrap-up time or "wrap" represents the amount of time an operator spends performing specific tasks after completing each customer call. Those tasks include, but are not limited to, documenting the customer's call, updating records, or contacting other company employees to resolve a question. All operators receive comparable proportions of high, medium, and low complexity calls due to random distribution of incoming calls.

Job complexity pilot survey. Based on a review of the literature, task complexity for the various types of calls
is best measured using the perceptions of job incumbents (Gerhart, 1988; London & Klimoski, 1975a; Rousseau, 1982). A job complexity survey (see Appendix A) was developed using a company operations manual, information gathered from several interviews with the department manager, and an interview with an incumbent. The purpose of the survey was to determine the relative complexity of each call category. The survey was given to a total of 20 job knowledge experts; 10 were fully trained incumbents, and the other 10 were working supervisory staff. Before discussing the results of the survey, it is necessary to explain the procedure used to assign call codes to individual calls.

Each call handled by an operator is categorized using a 6-digit code. Each code is a combination of three 2-digit codes: primary, secondary and tertiary. The first two numbers (primary code) designate the type of insurance product about which the caller has a question (e.g., life insurance, claims, new products, etc). The secondary code indicates how the operator responded to the call (provided information, transferred the call, sent a letter to the caller, etc.) The final two digits (tertiary code) specify in greater detail what type of service was provided to the caller. For example, if a call was coded 50 20 58, the first two digits indicate that the caller was interested in "purchasing" some type of insurance. The secondary code (20), indicates that the operator "handled" the call. The
tertiary code (58), specifies in greater detail that the operator handled the call by "explaining the premiums."

Data collection for the job complexity survey utilized only the primary and secondary codes for each of the calls. The tertiary codes were omitted from the survey because they did not contribute any additional information about complexity. As noted above, the tertiary code specifies in greater detail the information already given by the secondary code. When developing the survey, the 13 primary codes were grouped into 8 existing categories. It was not necessary to have the respondents rate all 13 of the primary codes because in several cases the only difference between two codes was the media source used for advertising. Consequently, two primary codes may be different because one was associated with newspaper advertising and the other in a magazine. This situation occurred for several call categories (see groups 2 and 4 in Table 1), in which case, the data were aggregated into a single group. In addition, this type of grouping procedure was more meaningful to the operators because it corresponded to the standard operating procedures used by the company. Operators were very familiar with the categories and could easily answer questions concerning the relative complexity of each group. Prior to administering the survey, the format was pre-tested on one incumbent and the supervisor to ensure clarity of the instructions and
proper organization of the material. The survey asked respondents to: 1) rate the complexity for each of the eight call categories using an 11 point graphic rating scale; and 2) rate the complexity of the secondary codes within each call category by writing either 1, 2, or 3 (1-low, 3-high) in the space provided. Results of the survey are illustrated in Table 1.

The objective of using the complexity survey was to determine how the three levels of call complexity should be established. A one-way within subjects ANOVA was used to test for significant differences among the complexity ratings for the different call categories. As hypothesized, both the incumbents and supervisors were clearly able to distinguish different levels of complexity among the call categories, $F (7, 133) = 48.94, p<.0001$. These distinctions were then used to establish three levels of job complexity for use as an independent variable. Furthermore, differences in mean ratings showed fairly consistent separation among the groups (see Table 1).

Examination of the mean complexity ratings clearly point out that groups 1 and 8 (categories 50, 90) should be used for the low complexity condition. This decision was based on two factors: a) low mean complexity rating; and 2) a large enough percentage of calls to provide a substantive sample. Group 2 (call categories 51, 52) were used for the medium complexity condition and groups 4, 5 and 6 (call
categories coded 55, 57, 60, 70) were used for the high complexity condition. Groups 7 and 3 were not utilized because the means for these groups did not clearly fit into the high, medium or low complexity levels.

**Maximum and typical performance.** The second independent variable, typical and maximum performance, represents two points on a performance continuum. Typical performance data exemplify average performance that has been aggregated across time. Typical measures also represent the subject's normal level of output without any external constraints or special circumstances. Maximum performance occurs when subjects are instructed to maximize their effort or do the best they can for a limited period of time, with the knowledge that their work behavior is being watched or measured.

In the present study, daily performance data were tabulated by computer and collected at the end of each day. Three samples of data were used as measures of typical performance to provide reliable and accurate measures of effort. Average wrap times and the number of calls were recorded for each operator according to call category. The data were then aggregated into levels according to those cutoffs established by the job complexity survey (see Table 1). Because operators were not aware of which days the typical data were collected, there was little chance that this measure might be confounded by subject bias.
**Table 1**

**Descriptive Statistics for Job Complexity Survey**

<table>
<thead>
<tr>
<th>Group</th>
<th>Call Category</th>
<th>Complexity</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>1.670</td>
<td>a</td>
<td>1.156</td>
</tr>
<tr>
<td>2</td>
<td>51,52</td>
<td>3.730</td>
<td>b</td>
<td>1.581</td>
</tr>
<tr>
<td>3</td>
<td>54</td>
<td>5.090</td>
<td></td>
<td>2.268</td>
</tr>
<tr>
<td>4</td>
<td>55,57</td>
<td>6.610</td>
<td>c</td>
<td>2.016</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
<td>5.340</td>
<td>c</td>
<td>2.167</td>
</tr>
<tr>
<td>6</td>
<td>70</td>
<td>6.415</td>
<td>c</td>
<td>1.711</td>
</tr>
<tr>
<td>7</td>
<td>80</td>
<td>5.070</td>
<td></td>
<td>1.731</td>
</tr>
<tr>
<td>8</td>
<td>90</td>
<td>0.495</td>
<td>a</td>
<td>0.286</td>
</tr>
</tbody>
</table>

*Note.*

- **a** Denotes the call categories used to establish Complexity Level 1 (low).
- **b** Denotes the call category used to establish Complexity Level 2 (medium).
- **c** Denotes the call categories used to establish Complexity Level 3 (high). Categories 2 and 4 contain two call codes because the type of call is the same, the only difference being the advertising source used by the company.
Maximum performance data were collected during the following week. Operators were informed by memo on the day that they should maximize performance (see Appendix B). Notices were also posted in obvious places in each cubicle work space, and a message was put on the sign-on screen of the computer. Operators read the sign-on screen daily or whenever they begin work in order to be updated on work-related information, changes in policies, etc. To reinforce the idea of maximum performance, an assistant working with the experimenter was present in the work place on the day that maximum performance data were collected. The assistant's duties were to answer questions about the research and, in general, make his presence known as a researcher from the university. Although some operators commented negatively about being observed, the overall effect seemed to facilitate the manipulation.

Wrap time. A variety of statistics are tabulated for each call that an operator receives. Some of these measures may seem like good indicators of productivity and effort, but in actuality are not. In fact, a large percentage of call statistics are not used to monitor performance of the operators, but rather to indicate sales volume, customer satisfaction, and training needs for new operators. Consequently, few indices exist that are true measures of performance and effort. More importantly, any performance measure that involves effort must be under the
control of the operator and free from situational constraints. Wrap time was used as the dependent variable in this experiment because it is sensitive to variations in effort as well as the complexity of the call, but not influenced by the number of calls received or other outside influences. Following is a description of how wrap time is measured and how it relates to job complexity.

For every call that is handled, there are certain tasks that must be completed after the caller hangs-up. These tasks vary as a function of the type of call and typically reflect the amount of information that must be processed to adequately document the call. The operator may have to change an entry on the computer, add some information to a file, write a note to another department, or make other calls to solve a problem. Thus, the amount of time that elapses after the caller hangs-up, and before the operator pushes the "release" button to take the next call, is recorded by the phone system as wrap time. Because this is done automatically, wrap time data are very precise and not prone to subjective errors as one might find when using ratings or observational data.

The amount of wrap time an operator accumulates is influenced by several factors, one of these is undoubtedly effort. Some operators work continuously, taking one call after another until their shift is over or a scheduled break time occurs. Others may have a tendency to pause
after calls or to spend time in activities that are not work related. An operator could do this by leaving his or her machine in the wrap mode longer than necessary after a call. This may occasionally happen but is unlikely to be excessive because the operators know the supervisor examines average wrap times and total wrap time for the month. Also, departmental standards suggest that wrap time should be held to approximately 15 seconds for each call and no more than 4% of total time logged on the system. Naturally, some calls will exceed the 15 second estimate and others will certainly be below it.

**Relationship Between Wrap Time and Complexity**

It was hypothesized that wrap time would be influenced by the type of call that was handled. A preliminary analysis using a sample of 40 operators was used to test this relationship by analyzing wrap times according to complexity level of the call. This analysis, performed prior to the collection of experimental data, will not be discussed extensively because similar results were found in the final analysis and are discussed there.

However, the results of this preliminary analysis using a repeated measures one-factor ANOVA indicated that wrap time was related to the complexity level of the call, $F (2, 78) = 5.55, p<.005$. These results were intuitively understandable because calls of a more complex nature should require the operator to accumulate additional wrap
time. He or she may be required to make multiple entries on the computer, call other departments to get information, or write memos. High complexity calls do require the operator to integrate more information about various types of insurance, applicable regulations, claim procedures, hospitals, medicare, etc.

Based on these results, it appeared feasible to use wrap time as the dependent measure. The data were measured objectively and accurately through the use of a computer and should not be contaminated by unwanted influences. Preliminary results showed that a positive relationship existed between complexity level and wrap time. More importantly, wrap time is a valid performance measure because it is sensitive to variations in motivation. Each operator has the ability to influence his or her wrap time measures by increasing or decreasing effort and efficiency.

Distribution of Calls

All incoming customer service calls that are handled by the operators go through an automatic call distributor (ACD). The ACD acts as a channeling device which connects callers to an available operator. Calls come in through the trunk line and are then channeled to one of the available operators. Although the ACD was not designed to operate on a random basis, there is virtually no way to influence what calls an operator receives. The process by which the ACD works is analogous to the system where each
customer takes a number and waits in line to be served. Customer calls are served when help is available according to the order in which they were received. If callers have to be put on hold, the ACD systematically searches each line until it finds an available operator. Because calls are not screened, the process closely approximates a random process and thereby ensures that call distribution is equal among the operators and across time.

Once the operator takes a call, he or she may answer questions, give information, complete forms or make referrals to another department. Immediately following each call, the operator enters a six-digit code into the Infoswitch Phone System to document specific information about the content of the call. The criteria for assigning call codes were discussed earlier. After a call is complete, the operator is available to take another call. If more callers are waiting the next caller is connected.

Each operator has two pieces of equipment at his or her work station; a computer terminal, which can be used to reference customer files, and an Infoswitch Automated Phone which serves as a terminal to the ACD system. The Infoswitch phone also has a built in L.E.D. display. Whenever the operator signs-on or off the system, dials a number or enters a code, the information registers on the display. Operators can use this feedback to correct mis-keyed codes or check system messages.
Procedure

Prior to data collection, each operator was asked to sign a consent form (see Appendix C) allowing the use of his or her performance data in the study. The operators who agreed to participate were given a brief explanation of the experiment.

Data for the experiment were collected at approximately weekly intervals for a period of one month. Typical data were collected by measuring wrap times on one day out of each week for three weeks. This portion of the data was collected first because it did not require any manipulation and could be done unobtrusively. Maximum data were collected during the following week in the manner described earlier. All data collection activities were conducted on Tuesdays. The decision to use Tuesdays was based on the recommendation of the department manager and her understanding of weekly customer calling patterns. Mondays and Fridays were excluded because these two days were quite variable in terms of calling patterns. Wednesdays and Thursdays were also eliminated because the company would occasionally use local advertising early in the week that would generate increases in calling towards midweek. It was obvious that a potential problem may exist if advertisements for a specific product biased the number and type of calls, particularly if the calls required lengthy follow-up paperwork or some other activity that
might bias wrap times. The proportional mix of calls could be considerably different compared to other days as a result of such advertising. Consequently, calling activity on Tuesdays seemed the most stable and was less likely to be affected from outside factors that may interact with the experimental treatment. The means, standard deviations, and percentages for each call category are listed in Table 2. The means in the upper part of the table indicate the average number of calls received by an operator for that call category (i.e., on the first day that typical data were collected (Typical 1) each operator received an average of 15.6 category 1 calls, 27.6 category 2 calls, and 4.9 category 4 calls).

Looking across the rows, it is obvious that the percentages of calls remained very stable during the data collection period. These results are consistent with earlier projections made from the pilot data regarding the grouping of the call categories into complexity levels. The goal was to split the call categories into three levels that were not only clearly different based on incumbent ratings but also sufficient in terms of the number of calls at each level. Tables 1 and 2 show that each complexity level was based on approximately equal numbers of calls for each of the four days.

As discussed earlier, the typical data were averaged across the first three days of data collection for each
Table 2

Subjects' Mean Number of Calls and Percentage of Calls by Category and Condition

<table>
<thead>
<tr>
<th>Call Category</th>
<th>Typical 1</th>
<th>Typical 2</th>
<th>Typical 3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>1a</td>
<td>15.6</td>
<td>8.7</td>
<td>18.4</td>
<td>13.1</td>
</tr>
<tr>
<td>2b</td>
<td>27.6</td>
<td>15.4</td>
<td>25.5</td>
<td>13.4</td>
</tr>
<tr>
<td>3</td>
<td>.1</td>
<td>.2</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>4c</td>
<td>4.9</td>
<td>4.7</td>
<td>4.0</td>
<td>4.9</td>
</tr>
<tr>
<td>5c</td>
<td>19.7</td>
<td>12.1</td>
<td>19.3</td>
<td>13.0</td>
</tr>
<tr>
<td>6c</td>
<td>.9</td>
<td>1.2</td>
<td>1.4</td>
<td>1.6</td>
</tr>
<tr>
<td>7</td>
<td>2.8</td>
<td>2.8</td>
<td>2.5</td>
<td>2.2</td>
</tr>
<tr>
<td>8a</td>
<td>1.3</td>
<td>1.4</td>
<td>1.7</td>
<td>1.5</td>
</tr>
<tr>
<td>Totals</td>
<td>5242</td>
<td>5548</td>
<td>5463</td>
<td>5629</td>
</tr>
</tbody>
</table>

Percentage of Calls by Call Category

<table>
<thead>
<tr>
<th>Category</th>
<th>Typical 1</th>
<th>Typical 2</th>
<th>Typical 3</th>
<th>Maximum</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>21.4</td>
<td>25.2</td>
<td>21.3</td>
<td>25.7</td>
<td>23.5</td>
</tr>
<tr>
<td>2b</td>
<td>38.0</td>
<td>35.0</td>
<td>35.4</td>
<td>35.2</td>
<td>35.9</td>
</tr>
<tr>
<td>3</td>
<td>.0</td>
<td>.1</td>
<td>.1</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>4c</td>
<td>6.8</td>
<td>5.5</td>
<td>7.6</td>
<td>6.0</td>
<td>6.5</td>
</tr>
<tr>
<td>5c</td>
<td>27.0</td>
<td>26.4</td>
<td>28.7</td>
<td>25.7</td>
<td>26.9</td>
</tr>
<tr>
<td>6c</td>
<td>1.2</td>
<td>1.9</td>
<td>2.2</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>7</td>
<td>3.8</td>
<td>3.5</td>
<td>3.1</td>
<td>3.5</td>
<td>3.4</td>
</tr>
<tr>
<td>8a</td>
<td>1.8</td>
<td>2.4</td>
<td>1.6</td>
<td>2.0</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Note. a = low complexity, b = medium complexity, c = high complexity.
operator. The averaging process was based on the number of calls received by that operator and the total amount of wrap time. For the high and low conditions which were composites of several call categories (see Table 1), a weighted average was used to calculate the mean wrap time. Wrap times were then averaged across the three days during which typical data were collected. Subjects were aware that typical data were being collected for research purposes but did not know the specific days. This procedure helped to reduce demand characteristics that may have influenced subjects' performance. The one week time lags were used so that both typical and maximum data were collected on the same day of the week. This procedure helped to eliminate any cyclical variation in call activity or other sources of bias due to the work week.

Maximum data were collected by instructing the operators to work at their highest level of output for the duration of their shift. It would have been desirable to collect maximum data for a shorter period of time, but this was not possible due to equipment limitations. A shorter period of time would have been more conducive to attaining a true measure of maximum performance (Sackett et al. 1988). The assistant who helped with the maximum performance manipulation did not collect any of the actual data, nor did he have knowledge of the specific hypotheses.
Only the wrap times were used as the dependent measure in this study. As a control measure, operators were instructed to maximize performance on all relevant dimensions of the job. This precaution was necessary so that performance in one area was not sacrificed for higher performance in another.

The other relevant areas of performance were "idle time" and "number of calls handled." As discussed previously, these criteria are not always related to effort, especially during slow calling periods. The number of calls an operator receives is dictated by the calling activity during the day. Idle time, which occurs when there are not any customer calls and the operator has little to do, is also beyond the operator's control and would be a poor measure of performance for that reason.
Chapter III

RESULTS

The total number of customer calls used in the analyses was approximately 20,705. Table 2 lists the descriptive statistics and percentages of calls by category. Initial inspection of the data indicated that one of the assumptions for a repeated measures design (sphericity) was violated. The repeated measures design rests on three assumptions: 1) independence of observations, 2) multivariate normality, and 3) sphericity (Stevens, 1986). Independence was not a concern in this setting because each subject works independently, and there is little chance that the performance of one operator may influence that of another.

To examine the normality of the data, Stevens (1986) and Norusis (1988) suggest using the graphical test of normality if \( N > 20 \). Results from the detrended normality plots indicated that the data in the individual cells were not normally distributed. Although the MANOVA procedure is robust with respect to normality, the same is not true regarding sphericity (Edwards, 1985; Stevens, 1986). The sphericity assumption states that the covariance matrix for the transformed variables should be a diagonal matrix with equal variances. In this instance the sphericity assumption was violated for one of the effects indicating that the F ratios may be positively biased (Box, 1954;
Edwards, 1985). On the basis of these results, a log transformation was used to normalize the data and satisfy the sphericity assumption. All reported F values are based on the Geenhouse and Geisser (1959) epsilon. Stevens (1986) suggests the use of this method so that the actual value of alpha is close to the nominal value. All subsequent analyses are based on the transformed data unless indicated otherwise. It should also be noted that an identical analysis was performed on the untransformed data and showed no differences in the significance of the main effects or interaction.

Table 3 contains a correlation matrix of subjects’ scores by condition. Although only three of the correlations will be used for discussion it is obvious that the degree of relationship is strong among all conditions. The significance of these correlations and their relevance to previous research are discussed later.

ANOVA Results

A 3 X 2 repeated measures analysis of variance was used to analyze the data. As mentioned earlier, because each subject responds to calls as assigned by the ACD, calls are randomly distributed, and the mixture of high, medium, and low complexity calls is proportional across subjects. The complexity level of the calls was previously established through the use of survey ratings in which employees and supervisors rated the complexity of calls.
Table 3

Correlation Matrix by Condition

<table>
<thead>
<tr>
<th>Condition</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typ-High</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max-High</td>
<td>2</td>
<td>.59</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typ-Med</td>
<td>3</td>
<td>.66</td>
<td>.64</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max-Med</td>
<td>4</td>
<td>.56</td>
<td>.60</td>
<td>.71</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Typ-Low</td>
<td>5</td>
<td>.48</td>
<td>.40</td>
<td>.56</td>
<td>.33</td>
<td>-</td>
</tr>
<tr>
<td>Max-Low</td>
<td>6</td>
<td>.56</td>
<td>.63</td>
<td>.62</td>
<td>.57</td>
<td>.51</td>
</tr>
</tbody>
</table>

Note. Typ = typical performance, Max = maximum performance. High, Med, and Low are levels of complexity. N=69.

All correlations are significant beyond p<.01.
The dependent measure was the wrap time for each call as recorded by a computer linked to the telephone lines. Subjects' scores were compiled for the three typical performance days, averaged according to complexity level and then matched with the data collected on the maximum performance day. Table 4 contains the summary table for the overall ANOVA and Table 5 shows the descriptive statistics for each condition with the transformed and untransformed data.

The F values are clearly significant for both the main effects of complexity (high, medium, and low) $F(2, 136) = 30.23, p < .001$ and performance (typical/maximum) $F(1, 68) = 36.73, p < .001$. The significant effect for complexity replicates the earlier finding of the pilot data and supports Hypothesis I by showing that complexity of the call is related to the wrap time measure. Ratings for the job complexity survey given earlier suggested that certain types of calls were more complex than others. Results here support the contention that calls which are more complex actually do require more time to process and therefore have longer wrap times. In addition, the significant effect for complexity also shows that it is possible to differentiate among the different types of calls as a way of establishing levels of complexity.

Hypothesis II is also supported as evidenced by the overall decrease in wrap times for the maximum condition as
Table 4

Summary Table: Repeated Measures ANOVA of Wrap Time by Complexity and Performance

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>( \omega^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>413</td>
<td>43.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between</td>
<td>68</td>
<td>25.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td>345</td>
<td>17.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>2</td>
<td>2.54</td>
<td>1.27</td>
<td>30.23*</td>
<td>4.1</td>
</tr>
<tr>
<td>C X Subj</td>
<td>136</td>
<td>5.70</td>
<td>.042</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td>1</td>
<td>1.69</td>
<td>1.69</td>
<td>36.73*</td>
<td>2.8</td>
</tr>
<tr>
<td>P X Subj</td>
<td>68</td>
<td>3.14</td>
<td>.046</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C X P</td>
<td>2</td>
<td>.12</td>
<td>.06</td>
<td>1.76</td>
<td>.1</td>
</tr>
<tr>
<td>Error (C X P)</td>
<td>136</td>
<td>4.60</td>
<td>.034</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The two levels of the performance variable are typical and maximum. N=69.

* p< .001.
Table 5
Mean Wrap Times by Performance Condition and Job Complexity

<table>
<thead>
<tr>
<th>Level</th>
<th>Transformed</th>
<th>Untransformed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Typical Performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Complexity</td>
<td>.510</td>
<td>.267</td>
</tr>
<tr>
<td>Medium Complexity</td>
<td>.438</td>
<td>.309</td>
</tr>
<tr>
<td>Low Complexity</td>
<td>.325</td>
<td>.255</td>
</tr>
<tr>
<td>Maximum Performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Complexity</td>
<td>.354</td>
<td>.348</td>
</tr>
<tr>
<td>Medium Complexity</td>
<td>.358</td>
<td>.355</td>
</tr>
<tr>
<td>Low Complexity</td>
<td>.177</td>
<td>.316</td>
</tr>
</tbody>
</table>

Note. Transformed values were derived using a log transformation. The untransformed values are in seconds.
compared to the typical condition. In this case, the decrease in wrap times represents improved performance because the operator took less time to perform the task. This effect was important for several reasons. First, it serves as a manipulation check in the experiment. That subjects were able to show a significant difference in their performance is evidence that: 1) they have the flexibility in their job to alter their performance level; and 2) the subjects were compliant with the researcher's request to increase their performance when requested to do so. Both conditions were necessary for making tests of Hypotheses III & IV. However, even with these results there is no guarantee that the subjects worked to the very best of their abilities (such as in a true maximum performance test) but, it does show a statistically significant increase in performance.

Although the overall interaction was not significant, certain relationships within the interaction were hypothesized a priori. To examine these relationships, analyses of the simple effects using cell totals were performed at the alpha = .01 level. These analyses examined the change between typical and maximum performance at each separate level of complexity. Results of the F tests show that typical and maximum performance levels were significantly different in both the high complexity condition $F(1, 68) = 20.00$, $p<.01$ and the low
complexity condition $F(1, 68) = 18.00, p<.01$. Typical and maximum performance were not significantly different in the medium complexity condition. Figure 1 illustrates the relationship among the three levels of complexity. Visual inspection of Figure 1 also suggests that the medium complexity calls were primarily responsible for the weak interaction effect. Possible explanations for these differences will be discussed later.

Simple effects analyses were again used to examine differences among individual cells (high, medium, and low) at fixed levels of performance (typical and maximum). These analyses showed that there were differences among the high, medium, and low conditions within both the typical $F(2, 136) = 28.57, p<.01$, and maximum data $F(2, 136) = 35.11, p<.01$. The Tukey A procedure for post hoc tests was used to test for differences in the cell totals. Table 6 shows the results of the post hoc tests.

The Tukey A procedure was chosen for the post hoc comparisons because it offers better protection against Type II errors when sample size is adequate and is appropriate to use if the sphericity assumption is met (Stevens, 1986). In addition, all tests were made using alpha = .01 in order to control the error rate per contrast. The results indicate that for both the typical and maximum performance data, there was no difference between the medium and high complexity levels, but that the
Figure 1. Mean wrap time (transformed) as a function of performance condition and level of job complexity.
Table 6

**Simple Effects Analyses on Mean Wrap Time for Complexity at Fixed Levels of Performance**

<table>
<thead>
<tr>
<th>Complexity Level</th>
<th>Performance Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Typical</td>
</tr>
</tbody>
</table>
| High             | .510  
   \(^a\) | .354  
   \(^b\) |
| Medium           | .438  
   \(^a\) | .358  
   \(^{a,b}\) |
| Low              | .325  
   \(^{b,c}\) | .177  
   \(^d\) |

*Note.* Analyses were performed on transformed data. Values not sharing the same superscript are significantly different at \(p < .01\).
low complexity condition was significantly different from the other two. Even though the high and medium levels were not different, the slight interaction effect in the medium level was sufficient justification for not aggregating the data into only two groups. Other explanations are also possible, and these will be discussed later.

Considering the overall outcome of the analysis, there is little statistical support for the existence of an interaction between task complexity and the typical/maximum performance distinction. Thus, Hypothesis III concerning the interaction between complexity and performance level is not supported; and Hypothesis IV, which specified the direction of the interaction, is also not supported.

**Length of Maximum Test Period**

The average duration of an operator’s shift was 5.7 hours, SD = 1.6. This is an important factor in regard to the original criteria described by Sackett et al. (1988); namely, that the maximum test session be short enough so that the subject’s attention remains focused on the task. At this time, guidelines have not been established to determine how much time constitutes a typical or maximum measure. It is quite possible that these time intervals will vary according to the task being performed.

The obtained mean shift time of 5.7 hours was longer than the original average of 4.4 hours that was taken from the pilot data. This variation is likely due to the small
sample size and the normal variation in work schedules. Although the maximum test period was longer than it should have been, it is still shorter than a full eight hour workday. The problem this may have caused is explored more fully in the discussion section.

**Follow-Up Survey Results**

Immediately following the last day of data collection, a follow-up survey was given to all participants (see Appendix D). The survey contained 24 questions and was designed to measure five different aspects of the operators' performance: (a) self-efficacy at work; (b) the amount of effort applied to work on the maximum performance day; (c) perceived constraints on performance; (d) consistency of work rate; and (e) perceptions of research project (suspicious or not). Due to the small sample size and the exploratory nature of the questionnaire, a cluster analysis was used to analyze the data.

Correlations were used as the distance measures, and the Average Linkage Within Groups method (Norusis, 1988) was used for the clustering procedure. The agglomeration schedule suggested that 5 clusters were present in the data. Results further showed that for clusters 3, 4, and 5, the questions loaded exactly as intended. For dimensions 1 and 2, the majority of questions also loaded as intended thereby substantiating the decision to interpret the clusters as they were originally designed.
The means for each question were averaged to arrive at the mean score for each cluster. Table 7 shows the descriptive statistics for each of the 5 clusters.

The most notable results of the questionnaire concern the operators' average level of self-efficacy ($M = 7.72$), and their seeming lack of concern regarding the purpose of the research ($M = 4.06$). The cluster labeled "consistency of work rate" was included to assess the operators' ability to maintain a high level of performance throughout their shift. The mean level of reported consistency ($M = 6.89$) indicates that performance output was relatively consistent or that the operators made only marginal efforts to comply with the maximum performance instructions. Examination of the cluster for "amount of effort" would suggest the latter alternative is correct.
Table 7

Follow-Up Survey: Subjects' Mean Ratings for Items Within Clusters

<table>
<thead>
<tr>
<th>Cluster</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Self-Efficacy</td>
<td>7.72</td>
<td>1.93</td>
</tr>
<tr>
<td>2. Amount of Effort on Maximum Performance Day</td>
<td>4.89</td>
<td>2.44</td>
</tr>
<tr>
<td>3. Constraints on Performance</td>
<td>3.82</td>
<td>2.60</td>
</tr>
<tr>
<td>4. Consistency of Work Rate</td>
<td>6.89</td>
<td>2.45</td>
</tr>
<tr>
<td>5. Perception of Research</td>
<td>4.06</td>
<td>2.90</td>
</tr>
</tbody>
</table>

*Note.* Tabled values were derived from the average of each question's mean within the particular cluster. Response scale for all items on follow-up questionnaire was: 1 - 10, (1=low, 10=high).

Clustering of questions: 1 = (1,3,6,7,9); 2 = (5,10,13,17,18,11,15,1-4,1-5); 3 = (1-6,14); 4 = (12,16); 5 = (2,4,8).
Chapter IV
DISCUSSION

The findings of this experiment do not show direct support for an interaction effect between task complexity and the typical/maximum performance criteria. Although the typical to maximum performance increases were not consistent across levels of complexity, there is only limited grounds on which to speculate about the true relationship of these variables. Of primary interest are the results of the medium complexity call group. Examining Figure 1, it is clear that performance on medium complexity tasks was much more consistent across the typical/maximum distinction in comparison to the other two levels of complexity. The correlation between typical and maximum performance in this case was .71 which is considerably larger as compared to .59 for high complexity and .51 for low complexity.

There are multiple explanations that may explain this finding. First, the difference may be due to the types of calls that were combined to form the three complexity groups (see Table 1). In this case, only one call category was used for the medium condition whereas three different call categories were combined to make the high complexity condition and two different call categories were combined to make the low complexity condition. Due to the increased variability of aggregating multiple call categories, the
observed differences in the correlations may be a function of the grouping rather than the complexity level.

Another possible explanation for the high correlation of the medium complexity group may be related to the type of call. Referring to Table 1, it may be that call group two (medium complexity) was much more homogeneous in terms of the tasks and demands placed on the operators. That is, despite any effort to increase performance, built in limitations of the system or a high degree of routine behavior may be the reason why performance did not increase compared to the other two complexity levels. This type of floor effect could easily account for the intersecting lines in Figure 1.

The third and probably least supportable explanation is that complexity and performance do interact. The fact that the interaction is limited to one level of complexity is not in accordance with the hypothesis but still suggests that some type of effect is present. It is possible that the relationship between complexity and typical/maximum performance is shaped similar to an inverted U rather than linearly. In other words, there may be differences between typical and maximum performance measures, but only with the most and least complex of job tasks. Further research would be necessary in order to test such an hypothesis.

Finally, it is possible that this experiment was not a fair test of the hypothesis given the range restriction of
using only one job. Compared to the variability found across other occupations, the data in this study represent a narrowly constricted range of complexity. If different positions with a wider range of complexity were used (e.g., surgeon, computer technician, laborer), the results could easily be different. This alternative was considered but seemed unfeasible due to the confounds that occur when comparing performance across different positions. As a result, the present study employed a more powerful design to examine the differences within only one position. The repeated measures design was powerful enough to detect significant differences among the three levels of complexity, but the total proportion of variance accounted for by the entire experiment was still small (see Table 4) and may not have been sufficient for an adequate test of the interaction.

Previous Research

In comparison to the previous work by Sackett et al. (1988), there is support for the notion that typical and maximum performance are in fact two different criteria. The results of the current study show a powerful effect between conditions (typical/maximum) and serve as a replication of Sackett et al.'s (1988) work. One difference worth noting is the comparison of the correlation coefficients between the two studies. Sackett et al. (1988) report a correlation of .36 between typical
and maximum speed measures for cash register operators. A test for the difference between independent correlations (Bruning & Kintz, 1968) shows that a correlation of .36 is significantly different from 2 of the 3 correlations in the present study. Specifically, the correlations between typical and maximum performance for medium (r = .71) and high complexity tasks (r = .59) are significantly different from Sackett et al.'s (1988) finding of .36. The correlation for low complexity tasks (r = .51) as found in the present study was not different from Sackett's finding.

The discrepancies between these correlations may be related to several factors. First, the differences may be a function of using two different samples. Although Sackett et al. (1988) used 453 cashiers in a between subjects design, the present experiment generated a comparable number of observations by using 69 subjects and a within subjects design. Secondly, the differences may be related to the nature of the tasks that are involved with the jobs. The cashier and the telephone operator positions are obviously not equivalent, but, when compared on a macro level, they do share some similar characteristics. Both jobs involve inconsistent work rates, use of computer and electronic equipment, daily interactions with others, and routinized behaviors.

Third, it may be that the range of variation between the typical and maximum condition is comparable across the
two jobs. That is, employees in both positions may have equal latitude as far as increasing or decreasing their productivity. If this is true, it follows that the relationship between typical and maximum performance changes as the nature of the task changes. Using the correlations from these two studies, human performance may become more variable as task complexity approaches one of the two extremes (high and low). This would support the general notion of Hypothesis III (i.e. an interaction between complexity and performance dimension) and suggest that the form of the relationship between the two variables is an inverted U. However, the results of this study cannot be taken as direct empirical support for such a claim.

**Implications for Testing**

One of the most important issues that surrounds the typical/maximum performance distinction is the impact that it may have on test-related activities and performance appraisals. The main effect for performance in this experiment demonstrates the mean differences (see Table 5) between typical and maximum performance and supports the notion that these measures are different types of criteria. Furthermore, typical and maximum measures should not be substituted or interchanged for one another even though they do show significant correlations. To the extent that dimensions such as this are ignored, evaluation criteria
will continue to be unreliable. Nagle (1953) suggests that the lack of stability in performance is the main limitation in achieving reliable criteria. Research by O’Leary (1972) further illustrates the importance of considering as many different dimensions of criteria as possible when evaluating performance.

Another aspect of the analysis concerns the correlations presented in Table 3. The correlations between typical and maximum performance for each level of complexity are moderately high compared to other predictor relationships normally found in psychological data. If we assume that motivation was consistent across the three levels of complexity (corresponding to the repeated measures), the relative performance increase can be attributed to ability factors of the subjects. As a result, maximum performance measures, used as a selection instrument, may be used to predict typical performance. Although these predictor criterion relationships are not perfect, they do rival the validity coefficients typically found with cognitive ability tests (Hunter & Hunter, 1984). If applicants do not possess the necessary work skills at the time of hiring, the typical/maximum relationship may also be used for promotion purposes.

An application of these findings to performance appraisal might include the use of weighted formats that are tailor-made for different types of positions (Lee,
Jobs that require steady, consistent performance would be typical-performance intensive and would therefore weigh that criterion more heavily. Jobs that regularly require maximum performance on critical tasks (police officer, fireman, etc.) would place greater emphasis on maximum performance tests (e.g., situational testing, work sample). The advantage to using weighted formats is that different dimensions would be evaluated using the most appropriate method and then weighted according to importance.

**Future Directions**

There are three important points that should be discussed in regard to future research in this area. As mentioned earlier, the question of range restriction on the job complexity variable should be examined to determine the extent to which this is a threat to internal validity. The possibility of using different jobs was also discussed as a way of increasing the amount of variance. Unfortunately, such methods may cast an unfavorable light on the results due to the lack of experimental equivalence among the different jobs.

Second, measures of work quality should be considered for use as an additional dependent variable. Although quality of work was held constant in the present experiment, it is possible that this variable may interact with job complexity and/or the typical/maximum performance
condition. In addition, quality of work may be unaffected by the performance condition when the task is highly routinized or well learned. However, tasks that are not well learned may be susceptible to performance decrements when measured under maximum performance conditions (Hebb, 1955). No research is yet available concerning the interaction between complexity, typical/maximum performance, and the degree of prior learning on the task.

Third, this study is only the second to deal with the emerging topic of typical and maximum performance. More work is needed to establish the interrelatedness of this criterion with other variables, as well as exploring the contexts in which typical or maximum performance are appropriate for use in selection or performance appraisal activities. Research is also needed to study these effects with management positions and other types of higher level jobs.

Attention should also be focused on the duration of the maximum test period. Equipment limitations in the present study did not allow for tabulation of wrap times at specific intervals. This would have been desirable given the average duration of the operators’ shift was longer than originally estimated (approx. 4.4 hr). In the present case, the maximum data were collected over a 24 hr period but the average length of an operator’s shift was only 5.7 hr. In retrospect, this is most likely too long to expect
maximum performance from a group of employees who are volunteering their time. On several occasions during the data collection, subjects complained about the expectations of the experiment. As a result, it is possible that the 5.7 hr shift over-extended the abilities of the operators. Even if they had planned to maximize their performance, the long day may caused fatigue and ultimately resulted in a performance decrement. Although the typical/maximum manipulation still proved to be significant, the interaction with complexity may have been altered due to the lengthy test period.

As a way to measure the subjects' motivational level, a survey was administered to the participants following the data collection period. The survey (see Appendix D) asked the respondents to estimate how much extra effort they put into their work on the maximum performance day. The survey results showed that approximately half of the subjects reported that they exerted 50% or more effort than usual during the maximum test period. Table 8 shows a stem and leaf plot illustrating subjects' self report of their motivational levels. Each stem represents an interval on a graphic rating scale. The leaves, located above the stems, are the fractional units within that interval (e.g.,) there were three responses scored 1.5 and two responses scored 1.3. It is important to remember that Table 8 is taken
Table 8

**Stem and Leaf Plot: Self-Report of Effort During Maximum Performance Period**

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
<th>Count</th>
<th>9</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>7 9</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7 7</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5 5 7</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>5 1 5</td>
<td>3</td>
<td>1 5</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>3 0 5 0 5 0 0 3</td>
<td>0</td>
<td>0 0</td>
<td></td>
</tr>
</tbody>
</table>

Note. The data were derived from subjects' scores using an 11 point graphic rating scale.

Response scale: 10 = large amount of extra effort, 0 = no extra effort. N = 52.
from self-report data and, as a result, may be unreliable due to a partial sample (N = 52) or a response bias.

Another point that is implicitly illustrated by Table 8 is that maximum performance cannot always be expected without the proper reward structure. Although a reward structure was in place (normal pay and incentive system), it may be naive to believe that the 80 employees would voluntarily work their hardest without any type of extra compensation. Furthermore, employees may be wary of showing any drastic increases because of the repercussions from management. In the present experiment, motivation to perform may also have been fueled by the suspicion that the company was interested in the performance data or that there may be sanctions taken against those who failed to cooperate or show improvement. Although considerable effort was made to assure the subjects that the outcomes of this research could in no way impact their jobs, the possibility of suspicion always remains.

The implementation of an incentive system prior to the collection of data may have also had an effect. This incentive system allowed the operators to earn paid time off for the reduction and maintenance of their idle time, talk time, wrap time, etc. It is possible that the incentive system had an impact on the results of the present study. Due to the newness of the incentive plan, the operators may have already increased their efforts at
work in order to earn the time off. This would suggest that a ceiling effect might have been introduced during the experiment. The question that is difficult to answer is whether the current data were influenced by this effect or some other interaction of variables that would threaten the internal validity of the study. Had there not been a strong effect for the typical/maximum manipulation this would have been a serious methodological problem. However, since the operators were able to significantly increase their performance, it is safe to assume that any ceiling effect that may have been present was not severe enough to jeopardize the main effects. The possible impact of the incentive system on the interaction is unclear. This limitation in conjunction with the prolonged duration of the maximum test period represent the only drawbacks in the methodology.

Future research in this area should attempt to provide some type of positive reward system as a way to encourage and ensure maximum performance. Therefore, in addition to the three necessary conditions for maximum performance as stated by Sackett et al. (1988), i.e., explicit awareness of evaluation, awareness and acceptance of implicit or explicit instructions to maximize effort, and short enough time duration; a fourth one is proposed that pertains to incentives. That is, the subject must have some incentive, commensurate to the task, that is offered in exchange for
effort; intrinsic motivators would be an exception. It is clear that although people possess the necessary skills and abilities, a reward structure is necessary to provide motivation when testing the upper limits of performance capability (Bandura, 1982).

**Conclusion**

Despite any outside factors that may have influenced the conditions in this experiment, there are several general statements that can be made. First, reliable and significant differences do exist between typical and maximum performance measures. These differences should not be ignored when considering any type of evaluative results, regardless of the application. Performance appraisal measures and other types of tests are the most immediate concerns to the human resource practitioner. It also appears that while substantial mean differences exist between typical and maximum measures, performance among subjects is fairly consistent as evidenced by the significant correlations. This should not be taken as a general statement that the workers are capable of significantly increasing their performance on a daily basis. There is the possibility that an improvement in one area may cause a performance decrement in other areas (i.e., sacrifice quality for quantity). Secondly, although the hypothesized relationship between performance and complexity was not supported, it would be short-sighted to
assume that no differences exist across jobs that involve a wider range of task complexity.
References


APPENDIX A

Job Complexity Survey
JOB COMPLEXITY SURVEY

The following survey is part of a research project being conducted for a master's thesis at the University of Nebraska at Omaha. We are interested in your opinion regarding the complexity of the different types of calls the inbound operators handle. Your participation in this survey is strictly voluntary. If you choose not to participate there is no penalty involved and it will not affect your job with Physicians in any way. However, your opinion is valued and we hope that you will choose to complete this survey. The information you supply will be used primarily for research purposes.

Each item on the survey asks you to do two things:

1) Rate the overall complexity for each category of calls (i.e., New Business, POS, Life, etc.). Place a checkmark (✓) anywhere along the line labeled 1-10 to indicate your answer. A "10" would mean that the category of calls is the most complex, a "1" would indicate that the calls are not very complex. Do not hesitate to score the easiest category with a "1" or a "0" and the most complex category with a "9" or "10".

2) Rate the complexity of each of the Level 2 codes within each call category. To do this, you will rate the different Level 2 codes by writing a 0, 1, 2, or 3 in the space provided. In this case, a 1 means low complexity, 2 means average and 3 means very complex. Use the "0" if you don't know or haven't had enough experience with that particular Level 2 code. Remember you are rating the complexity of the level 2 codes within a call category.

When rating the complexity of the different types of calls and the level 2 codes please consider the following guidelines.

* the amount of knowledge an operator must have to handle the calls adequately
* experience required
* the amount of time and effort you spend solving or analyzing the caller's problem rather than just supplying information

Example:
Old Business, code: 50

[ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
In this example, the Old Business calls were checked as being the least complex of the call categories overall. Following the overall complexity rating, the Level 2 codes within the Old Business category were rated. Calls that are coded 50-20 and 50-23, were scored with a "2" indicating medium complexity in comparison to the other Old Business calls. Calls coded 50-21 and 50-25 were scored with a "3" indicating high complexity for an Old Business call. The 50-24 call was scored with a "1" meaning it was the least complex of the Old Business calls. A "0" means the operator did not have enough experience with that type of call to make an accurate judgement. Please do not confer with other operators when making your decisions.

1. New Business, code: 50

2. POS, (NRECA, Reader's Digest, Fingerhut) codes: 51, 52, 53, 59
3. CHS, code: 54

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

**LEVEL 2 Codes:**
- 20 handled
- 21 call back
- 22 special
- 23 letter
- 24 transfer
- 25 call back later
- 60 claim handled

---

4. Medicare Supplement, codes: 55, 56, 57

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

**LEVEL 2 Codes:**
- 20 handled
- 21 call back
- 22 special
- 23 letter
- 24 transfer
- 25 call back later
- 60 claim handled
- 61 claim call back
- 62 claim special
- 63 claim letter
- 64 claim transfer
- 65 claim call back later

---

5. Mass Media Claims, code: 60

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

**LEVEL 2 Codes:**
- 60 claim handled
- 61 claim call back
- 62 claim special
- 63 claim letter
- 64 claim transfer
- 65 claim call back later

---
6. Agency, code: 70

<table>
<thead>
<tr>
<th>0</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>
</tr>
</thead>
</table>

**LEVEL 2 Codes:**
- 20 handled  60 claim handled
- 21 call back  61 claim call back
- 22 special  62 claim special
- 23 letter  63 claim letter
- 24 transfer  64 claim transfer
- 25 call back later  65 claim call back later

7. Life, code: 80

<table>
<thead>
<tr>
<th>0</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>
</tr>
</thead>
</table>

**LEVEL 2 Codes:**
- 20 handled  60 claim handled
- 21 call back  61 claim call back
- 22 special  62 claim special
- 23 letter  63 claim letter
- 24 transfer  64 claim transfer
- 25 call back later  65 claim call back later

8. Other, code: 90

<table>
<thead>
<tr>
<th>0</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>
</tr>
</thead>
</table>

**LEVEL 2 Codes**
- 11 no answer
- 13 misdial
- 24 transfer
- 90 general

Please return your survey to Karen Gardner when you are finished.
Thank you for your cooperation.
APPENDIX B

Memo Distributed to Operators
ATTENTION RESEARCH PARTICIPANTS

For those of you who are participating in the University of Nebraska research study, please take note of the following:

On Tuesday, May 2 we will be collecting statistical data on the calls handled by the inbound area. During that day, we ask that you work your hardest to improve your daily stats (e.g., idle time, wrap time, etc.) However, please do not sacrifice quality for quantity.

This phase of the research will only last one day and your participation is vital to the outcome of the project. Also, please remember that the purpose of this study is not to test or evaluate you or your fellow workers. The information being gathered will not be used for company purposes and cannot affect you in anyway.

Results of the study will be made available to you later this summer when the project is completed.

If you have any questions, concerns, or complaints, please call me at one of the numbers listed below.

Thank You

Ken Jordan, M.A.

UNO (daytime) 554-4811
Home (evening) 393-2858
APPENDIX C

Informed Consent
Consent Form

THE RELATIONSHIP BETWEEN TYPICAL AND MAXIMUM MEASURES OF PERFORMANCE ACROSS LEVELS OF JOB COMPLEXITY.

Invitation
You are invited to participate in a research study that compares measures of work performance.

Purpose of the Study
The purpose of this study is to examine differences between typical and maximal work performance. We are interested in the differences between average or everyday performance and short-term, high output performance.

Basis for Subject Selection
All tele-services representatives in the inbound area (except those in training) who are interested in participating will be accepted.

Explanation of Procedure
As part of the procedure, you will be asked to perform your regular work duties to the best of your ability during one shift. We are also asking your consent to: 1) use daily performance data (call rate, average talk & wrap time, idle, etc.) collected by the company and; 2) use background information and appraisal data. You will also be asked to complete a short survey following the data collection.

Potential Risks
There are no risks involved with your participation.

Potential Benefits
The direct benefits to you as a participant are minimal. However, your contribution as a subject may help to improve the appraisal and selection procedures used in the work place. By volunteering, you are not only improving the system of which you are a part but also making a contribution to science.

Financial Obligations
As a research participant you are not under any financial obligation.
Assurance of Confidentiality
Any information that is obtained in connection with this study will remain confidential and will be disclosed only with the subject's permission. Performance data will not be used for appraisal or promotion purposes and will not jeopardize your employment with Physicians in anyway. Your supervisor may see the results of this study but all information will be summed into groups so that no individual comparisons are possible.

Withdrawal from the Study
Participation is voluntary. Your decision to participate will not affect your present or future relationship with Physicians. If you decide to participate, you are free to withdraw your consent and to discontinue participation at any time.

Offer to Answer Questions
If you have any questions regarding this study, ask one of the experimenters prior to signing the consent form. If you have any questions later on, feel free to contact one of the researchers listed at the bottom of the page.

If you have any additional questions concerning the rights of research subjects you may contact the University of Nebraska Institutional Review Board (IRB), telephone 402/559-6463.

Concluding Consent Statement
YOU ARE VOLUNTARILY MAKING A DECISION WHETHER OR NOT TO PARTICIPATE. YOUR SIGNATURE INDICATES THAT YOU HAVE DECIDED TO PARTICIPATE HAVING READ THE INFORMATION PROVIDED ABOVE. YOU WILL BE GIVEN A COPY OF THIS CONSENT FORM TO KEEP.

Signature of Subject Date

Signature of Investigator Date

INVESTIGATORS: Ken Jordan 114 Arts & Sciences Hall 554-4811
Lisa Scherer 347 Arts & Sciences Hall 554-2698
APPENDIX D

Follow-up Survey
Follow-up Survey For Research Participants

If you participated in the University of Nebraska research project please complete the following survey. Your responses will be anonymous so feel free to be as candid as possible. The results of this survey will be used for research purposes only. When you are finished, put the survey in the envelope provided and drop it in the box by Freda’s desk before leaving work. Also, there will be a list of all the participants’ names attached to the box. Find your name on the list and put a check mark by it to indicate that you’ve finished the survey. Do not put your name or any identifying information on the survey.

PART I

1) Did you work on Tuesday, May 2? _____ yes _____ no
   (If you answered "no" to question 1 skip to Part II on the second page).

2) How many hours did you work on Tuesday, May 2? ____ hrs.

3) How long have you worked at PMIC as a Tele-Services Rep?
   _____ years _____ months

INSTRUCTIONS:
Read the questions that follow and mark your answers (using a check mark) on the scale provided. Please answer all the questions.

4) How much EXTRA effort (if any) did you put into your work on the day we asked you to work up to the best of your abilities (May 2)?
   |----|----|----|----|----|----|----|----|----|
   0%  50%  100%

5) How certain were you regarding your ability to improve your performance (idle, wrap, etc.) on that day (May 2)?
   |----|----|----|----|----|----|----|----|----|
   not very certain at all
6) On May 2, did anything get in your way or interfere with your work performance (e.g., large number of difficult calls, an argument with someone, bad mood, etc.)?

<table>
<thead>
<tr>
<th>nothing</th>
<th>several things</th>
</tr>
</thead>
<tbody>
<tr>
<td>interfered with</td>
<td>interfered that</td>
</tr>
</tbody>
</table>

PART II

INSTRUCTIONS:
Use the agreement scale below to answer the following questions. Write the number (1-10) in the blank provided to indicate your response. Each question should be answered in reference to your performance as a tele-services rep.

AGREEMENT SCALE

1----2----3----4----5----6----7----8----9----10
strongly disagree strongly agree

1. __ The departmental work standards are difficult to attain.

2. __ I was suspicious about the purpose of this research.

3. __ I am capable of a high level of performance on this job.

4. __ I enjoyed the opportunity to participate in this research project.

5. __ I believe that I have the capability to perform my daily job duties better than my co-workers.

6. __ I have more difficulty fulfilling my work responsibilities than other Tele-Services reps.

7. __ If I wanted to, I could get a job with any other company doing tele-services work.

8. __ Participation in this research project made me uneasy.

9. __ This job is well within the scope of my abilities.

10. __ I actually perform my daily job duties better than my co-workers.
11. __ I do not have much difficulty in reaching my work objectives.

NOTE:
The following questions all refer to Tuesday, May 2 when you were asked to work up to the best of your abilities. If you weren’t here on that day skip these questions and go to the last page.

12. __ The amount of effort I put into my work was consistent throughout the work day.

13. __ I was more tired than usual after working my hardest.

14. __ I was in a good mood and felt fine on that day.

15. __ I really exerted myself in order to try and improve my level of performance.

16. __ I was not able to maintain a high level of performance all day.

17. __ I was surprised to find out how hard I could work.

18. __ The amount of effort I put into my work wasn’t any different than any other day.

COMMENTS:
If you have comments regarding any aspect of this project please write them in the space provided. Your responses will be kept confidential. Once again, thank you for your cooperation!

Sincerely,
Ken Jordan, M.A.