ABSTRACT

MURPHY, KRISTIN FREUND. An examination of factors affecting the accuracy, reliability and validity of job analysis task ratings. (Under the direction of Mark A. Wilson.)

This study examined several issues related to evaluating job analysis information. The first issue concerned the relationship between job analysis reliability and accuracy and the utilization of these variables to gain an estimation of job analysis validity. In this study, the performance on reliability and accuracy indices was compared to determine the extent to which such indices identify a consistent set of reliable and accurate raters. Similar to Green and Stutzman (1986), this study also compared the rating profile of the selected accurate and reliable raters with that of the entire rater population. The second issue concerned the impact of other individual differences concerning job tenure and experience on the validity of job analysis ratings. In addition, the effect of individual rater’s fatigue may have on job analysis reliability was explored. Finally, the study addressed whether the distinction between in-role and extra-role behavior may help explain some of the variance in job analysis ratings. Thus, the study examined whether the nature of the task has any impact on job analysis reliability or accuracy.

A significant relationship was found between reliability and accuracy. Highest mean reliabilities were found for the reliable raters as compared to accurate raters and the incumbent population. Thus, the impact of the reliability of raters on the validity of the instrument was demonstrated. A correlation analysis among reliability and accuracy scores and individual difference variables revealed a significant negative correlation between reliability and organization tenure. No significant relationships were found between education and reliability scores or between education and accuracy scores. An analysis of reliability over the course of a job analysis inventory showed that mean reliabilities trend downwards initially and slope slightly back upwards. The downwards trend may suggest that fatigue impacts the reliability of the rater over time. The sudden change in trend could indicate that point at which incumbents changed survey format.
(from computer to paper-pencil or vice-versa). Lastly, nature of the task appears have an impact on reliability of job analysis ratings as in-role task mean reliabilities were higher than extra-mean reliabilities.

In sum, this study sought to explain the potential impact of the factors of accuracy, reliability, individual differences and the nature of the task on the validity of job analysis ratings. This study expands the knowledge base concerning inter-relationship among these factors and the extent to which this knowledge could lead to a model of selecting accurate and reliable job analysis subject matter experts (SMEs).
AN EXAMINATION OF FACTORS AFFECTING THE ACCURACY, RELIABILITY AND VALIDITY OF JOB ANALYSIS TASK RATINGS

By

Kristin Freund Murphy

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

DEPARTMENT OF PSYCHOLOGY

Raleigh

2000

APPROVED BY:

________________________________              _____________________________
Chair of Advisory Committee
Dedication

This project is dedicated to my husband, Sean, my biggest fan and best friend, whose love and support helped keep me going during the whole graduate school process and who is cheering me on all the way to the finish line.
Biography

Kristin Freund Murphy was born on June 25, 1969 in Johnson City, NY. It was one of three exciting, historical events to occur that summer, next to the lunar walk and Woodstock. After doing what was the routine IBM moves at that time, her family settled down in the beautiful Hudson Valley. Kristin received her elementary and high school education in Wappingers Falls, NY, graduating from Roy C. Ketcham High School in 1987. She moved back up toward her cold and snowy birthplace in central New York to attend college at Colgate University in Hamilton. During the spring semester of her junior year, Kristin studied German language and history at the Albert Ludwigs Universität in Freiburg, Germany. She graduated magna cum laude with high honors from Colgate University in May 1991. She received a Bachelor of Arts degree in History and a minor in German and was inducted into Phi Beta Kappa.

It was during her junior year abroad to Germany that Kristin first became interested in Industrial-Organizational Psychology. Her decision to obtain a doctorate in this field led her to North Carolina State University and to establish a new home in the Raleigh-Durham area in the summer of 1994. Kristin earned a Master of Science degree from North Carolina State University in 1996 and currently works professionally for a global telecommunications and internet networking company in Research Triangle Park, NC.
Acknowledgments

I would like to take the opportunity to thank my committee members for their support and assistance throughout this project: Dr. Mark Wilson, Dr. Bert Westbrook, Dr. Bill Cunningham and Dr. Mike Wogalter.

I would like to express my gratitude to Johnny Weissmuller for supplying the ATSURVEY software package for use in this study. His generosity, hard work and dedication to research in this area are greatly appreciated.

I would especially like to thank my family: my Mom for her constant support; my brother John, for making me laugh during the times when I couldn’t find any humor in life, my dogs Riley and Portia who remind me to take pleasure in the simple things of life (food, mud puddles, and a good back scratch) and especially my husband, Sean, who is always there for me and who is my very best friend. Lastly, a very special thanks to my Dad, who is with me in spirit and who would be very proud of me!
**Table of Contents**

List of Tables vi

List of Figures vii

Introduction 1

Research Problem 7
  Reliability, Accuracy and Validity 8
  Individual Differences 14
  Nature of the Task 16
  Research Questions 18

Method 20

Results 23

Discussion 27

References 37

Tables 43

Figures 53
List of Tables

1. Repeat task items 41
2. Distractor task items 42
3. Descriptive statistics of reliability and accuracy indices 43
4. Correlations among accuracy and reliability index scores and individual difference variables 44
5. Mean reliabilities for task level and significance ratings of selected incumbent groups 45
6. Descriptive statistics of reliability profile analysis using task level ratings 46
7. Descriptive statistics of reliability profile analysis using task significance ratings 47
8. Mean reliabilities for task ratings by inventory position grouping 48
9. Mean correlation of task ratings by in-role and extra-role categorizations 49
List of Figures

1. A model of the potential factors impacting the accuracy, reliability and validity of job analysis ratings 53

2. A graph depicting changes in mean reliability of job analysis ratings throughout job analysis inventory 54
An Examination of Factors Impacting the Accuracy, Reliability, and Validity of Job Analysis Task Ratings

Introduction

As the world looks toward the twenty-first century with anxious anticipation, it comes as no surprise that the future marketplace will be driven by industries not backed by capital in the traditional sense of manufacturing plants, production equipment and financing. In contrast to the dominance of manufacturing giants of yesteryear, high market growth will be determined by the unique contribution of human capital. In this new era of knowledge work, it has become a common recognition that competitive advantage arises more from the strength of an organization’s intellectual capital rather than its business processes, bricks and mortar (Davis & Meyer, 1998). As an article on the importance of proper people management for productivity growth advises, “The answer [to the question of how to gain competitive advantage] is all that separates you from your competitors are the skills, knowledge, commitment and abilities of the people who work for you (Fast Company, 1998).”

Taking such advice seriously, organizations are increasingly realizing the importance of taking inventories on existing job activities and characteristics as well as the capabilities of the people who currently are performing or are needed to perform them. Organizations must have a clear understanding of what expertise each new project requires and where within and outside of the organization the knowledge, skills and abilities required for them lies. In order to achieve agility in the dynamic world marketplace, organizations must assemble the people with the required expertise, in the right project roles in the shortest amount of time (McGrath, 1996).
Job analysis is defined as the collection and analysis of any type of job related information (McCormick, 1979). Traditionally, job analysis has been the way to identify and describe expertise. Knowledge of the activities that a job entails and the personal characteristics of the people who perform the job serves several purposes. For the organization, it provides the information required for defining the activities and capabilities required for a particular work process and determining the criteria used to select or train appropriate candidates to perform the process. For the individual, knowledge about specific jobs is essential for making effective decisions about career counseling and vocational preparation.

Despite some recent calls for the end of job analysis (Carson & Stewart, 1996), it appears that a tangible business need for knowledge regarding the activities a job entails and the characteristics of the capability required to perform them still remains. As described above, job analysis is fundamental for building effective human resource systems and for helping the organization attain superior business performance. It provides the foundation from which organizations build selection, promotion and performance appraisal systems (Ash, 1988). More importantly, as previously stated, there is still a need for methods that identify the activities, knowledge skills and abilities required for a given job or role and that inventory human capability within an organization to meet those requirements. Organizations that set up the processes for profiling roles and inventorying capability will gain tremendous understanding of the capability needed to meet business requirements. In addition, it will provide a picture of what human capability exists within the organization and what is still needed. In short, organizations gain competitive advantage from knowing the expertise they currently have
and what they still need to learn. As former Hewlett-Packard CEO Lew Platt once said (echoing a former head of HP Labs), "If HP knew what HP knows, we would be three times as profitable." Platt expresses a belief in the potential value of knowledge, one shared by many other corporate executives (Davenport & Prusak, 1997).

Thus, organizations are now realizing that inventoring current job activities and the capability of employees as well as drawing predictions of what job activities and capability will be required for the future brings competitive advantage. Traditionally, job analysis has provided a means for describing such knowledge of the activities a job entails and personal characteristics of the people who perform the job.

The questionnaire format of the job description inventory is a common method used for obtaining job analysis information (Ash, 1988). When choosing the job analysis methods to employ, the importance of yielding similar results upon repetition has been identified as an important and practical issue (Levine, Thomas & Sistrunk, 1988). In the structured questionnaire method, a listing of various job-related items pertaining to a particular job is created and then given to those individuals in the job or subject matter experts (SME) to rate. Ratings of the items are made in response to such scales as level of involvement and relative time spent in (Christal & Weissmuller, 1988) and importance of certain job activities. Structured questionnaires traditionally require respondents to consider the extent to which certain task and behaviors are performed and often the listed knowledge, skills and abilities (KSA) are used in the job in question.

Data gathered from job analysis questionnaires provides a foundation of information from which selection and promotion processes are designed and training programs are developed. Typical applications of job analysis such as these have always
been central for human resource development in organization. However, in today’s fast-paced work environments, organizations are facing a greater challenge of deploying people with the right talents to different work opportunities routinely and in rapid timeframes (Mohrman, Galbraith & Lawler, 1998). As quickly as customer needs are identified, project teams are launched to develop solutions, and organizations must scramble to identify the project requirements and the talent needed for the projects. This increasing emphasis on inventorizing an organization’s “human capital” for more effective deployment of people on projects as well as other more traditional core human resources processes demonstrate the need for systematic development of reliable job-related information (McCormick, 1979).

The reliability of job analysis information has been and continues to be an important research issue. A large amount of research on the reliability of task inventories has been conducted in conjunction with the United States Air Force (Morsh, 1964). The inventory method is a quite common approach to job analysis and has been used extensively by the United States Air Force as its primary job analysis method (McCormick, 1979). As the Air Force research demonstrates, the test-retest methodology is one way to demonstrate the reliability of task inventories (McCormick, 1960; McCormick & Ammerman, 1960; McCormick & Tombrink, 1960; Cragun & McCormick, 1967). The Air Force research provides an excellent foundation for examining the role of reliability in job analysis ratings. Since that research was published, many practitioners recognize that the time and effort required to re-administer an often lengthy job description survey after a specified period of time is not feasible. In more recent research, Wilson, Harvey and Macy (1990) offered the repeat item technique
as one less time consuming alternative to estimate the reliability of task inventories. This study provided a new approach for calculating reliability of job analysis.

From the amount of research on reliability of task inventories and job description surveys (McCormick, 1960; McCormick & Ammerman, 1960; McCormick and Tombrink, 1960; Wilson, Harvey, Macy, 1990), a number of factors affecting reliability have been identified. For example, from their research with the Air Force and that of others, Christal and Weissmuller (1988) suggested that alphabetizing the task list under major duty headings may help respondents to pay close attention to tasks under their assigned duty headings. The increase in test-reliability of tasks' responses shown to be dependent on the amount of information (increasing the number of scales upon which to rate an item) requested of incumbents per item (McCormick, 1960) may be explained by greater attentiveness. Thus, the incumbent may be more attentive to the rating of a particular item when he or she is required to spend more time with it (Morsh, 1964). In addition, the variety of scales used in task inventories yield different average reliability coefficients (Birt, 1968). The scales for frequency of task performance and task duration have been found to have reliabilities ranging from the .50s to the .70s (McCormick & Ammerman, 1960; Morsh, 1964). Difficulty scales generally have lower reliabilities ranging from the .30s to the .50s. (McCormick & Ammerman, 1960; McCormick & Tombrink, 1960; Wilson, et. al., 1990).

The research on the reliability of job analysis methodology answers the needs for the systematic development of reliable job-related information (McCormick, 1979). However, outside the research on reliability of job analysis methodology, there is not a tremendous amount of job analysis literature devoted to the topic of job analysis
reliability, accuracy and validity overall. In his landmark chapter on job analysis, Harvey (1991) posed challenges that face the field in the area of job analysis. One challenge is to increase our knowledge of the job analysis rating process. The Air Force studies had set the stage for this work and more recent research has added interesting findings concerning job analysis methodology and ratings (McCormick, 1960; McCormick & Ammerman, 1960; McCormick & Tombrink, 1960; Cragun & McCormick, 1967; Wilson, Harvey & Macy, 1990). Although this research has expanded what is known about the job analysis rating process, there is much left to be addressed. In particular, the issue of job analysis ratings validity has remained uncharted territory until some very recent debates in the literature (Sanchez & Levine, 1999; Harvey & Wilson, 1999). It is hoped that current interest in this topic and the research that it sparks (such as this current study) will add to the ever growing domain of research on job analysis reliability, accuracy and validity.
Research Problem

Given the importance of job analysis information, few question that the reliability, accuracy and validity of such job analysis assessments should be of the utmost concern to industrial/organizational psychologists. Reliability is a precondition for validity and is necessary for developing any valid prediction process based on job analysis results (Nunnally, 1967). If reliability of job analysis data has not been established, one could question the validity or utility of the results (McCormick & Jeanneret, 1988). For this reason, researchers stress the importance of reliability when collecting data on human work behavior such as the knowledge, skills and abilities (KSAs) and work activities of employees (McCormick, 1979; Harvey, 1991; Wilson, Harvey & Macy, 1990). From their research, they have suggested a number of methods for assessing reliability and accuracy and have identified potential factors that can impact the validity of job analysis data.

Drawing upon this research, the present research examines several issues related to evaluating job analysis information. First, it addresses job analysis reliability and accuracy as a means to gain an estimation of job analysis validity. A description of the various methods for determining reliability and accuracy of job analysis ratings are included in this discussion. Next, it investigates the relationship between individual differences variables and job analysis validity. That is, it examines whether certain rater characteristics can help determine a profile of an accurate and consistent rater. Finally, the impact of the type of task being rated has on job analysis validity is explored. More specifically, this paper investigates the question of whether incumbent ratings of operational (in-role) tasks appear to be more reliable or accurate than ratings of citizenship (extra-role) work activities. Thus, by addressing these questions, it is hoped a more complete understanding of how to collect more reliable, accurate and valid job analysis data will be attained.
Figure 1 is a graphical depiction of how various factors addressed in the literature may be related to the validity of job analysis data. The model is meant to integrate all the factors known to have an impact on the validity of incumbent job analysis ratings. At the heart of the model is the accuracy and reliability of job analysis ratings. Taken together, reliability and accuracy are often cited as evidence for job analysis validity. One goal of this study was to examine various methods that operationalize each variable and the interrelationship between accuracy and reliability. Two factors that may have some impact on these variables are the individual differences of raters and the nature of the tasks they are rating. A second goal of this study was to examine what role these variables play in a more systematic manner. Subsequent sections of this proposal review current knowledge of the accuracy and reliability of job analysis information and the relationship between these two variables. The role of individual differences and the nature of the task being rated is also explored. Thus, by identifying the implications of these issues, a model of how to collect reliable and accurate job information has been developed.

Reliability, Accuracy and Validity

In order to expand current knowledge of reliability and accuracy of job analysis information, the study of factors related to the validity of job analysis data and the individual rater has to be taken further. What still remains unclear from the research is the relationship among reliability, accuracy and validity of job analysis data. The frequency with which these terms appear together in the literature may give the appearance that they are interchangeable terms. However, further investigation of their definitions demonstrates that they are quite distinct.

According to its general psychometric definition, reliability “refers to the consistency of scores obtained … when reexamined with the same test on different
occasions (Anastasi, 1988, p. 109).” For example, test-retest reliability is the consistency in responses across separate administrations of a test. Underlying the computation of error measurement, test reliability “indicates the extent to which individual differences in test scores are attributable to ‘true’ differences in the characteristics under consideration and the extent to which they are attributable to chance errors (Anastasi, 1988, p. 109).” A test’s statement of reliability provides practitioners with some assurance as to the consistency of scores when measuring a particular psychological phenomena.

Similarly, reliability in job analysis also refers to consistency in responses regarding descriptions of work activities. In job analysis, test-retest reliability refers to the stability over time of job profiles or individual responses to job analysis questionnaires rather than test scores. However, in contrast to psychological testing, job analysis results are expected to be not as stable over longer periods of time as most measures of psychological attributes. Job duties often change over time due to such things as seasonal duty cycles or proficiency increases (Wilson, Harvey & Macy, 1990). Although such an environmental impact on the testing situation may be considered by psychological testing principles as a source of error, these shifts in job profiles may actually reflect behavior changes in the true job situation rather than error (Harvey, 1991). Despite some differences in interpretation, researchers’ primary intent for examining reliability in job analysis ratings is to determine the extent to which the job analysis information collected is dependable and can serve as the foundation for developing valid prediction processes.

For the most part, the majority of research on reliability of job analysis has focused on methods and sources of data collection (Gael, 1988; Harvey, 1991). For example, the United States Air Force sponsored research on the reliability of task inventories during the 1960s. These studies, for the most part, assessed reliability of task inventories by using the test-retest method. For example, Cragun and McCormick (1967) assessed the test-retest reliability of task inventories by administering the surveys to the
same job incumbents twice within a four month period. They then excluded from the
reliability calculation any task that was not rated at both administrations. Mean test-retest
reliabilities for the five task rating scales ranged from .35 (difficulty scale) to .60 (point-
of-the-position scale). In more recent research, Wilson, Harvey and Macy (1990)
examined the use of the repeat-item technique to assess reliability. In this technique, task
items are repeated within a task analysis survey to simulate a test-re-test reliability
method but within the same survey administration. The reliabilities found among the
three inventories used ranged from .70s to .90s. Murphy and Wilson (1998) built upon
this research by comparing repeat-item and test-retest reliability calculation methods.
They demonstrated that the two reliability calculation methods reveal similar tendencies
but the repeat-item reliabilities tended to be consistently higher. They also found that as
items being rated become more abstract the reliability of ratings declined. Tasks were
found to have higher reliability estimates (mean reliabilities in .50s) than KSAs (mean
reliabilities of .30s).

Accuracy can be defined as being within or only deviating slightly from a
standard. From a psychometric perspective, rating accuracy generally refers to the degree
of conformity or closeness of a measure to a true value or “true score.” Cronbach (1955)
argued that global accuracy measures are not useful to interpret. Rather one should
examine accuracy by breaking it down into four components. Elevation (E) concerns
how closely a rater’s grand mean of ratings agrees with the grand mean of the target
scores. Differential elevation (DE) refers to accuracy in differentiating between different
ratees’ mean target scores. Stereotype accuracy (SA) relates to how correctly raters
differentiate between dimension target scores. Lastly, differential accuracy (DA) is the
degree to which raters correctly order ratees on each dimension when controlling for
ratee and dimension effects. This delineation of accuracy provided the standard model by
which most rating assessment research abides.
In the last decade, the term “accuracy” has been used quite often in reference to job analysis (Green & Stutzman, 1986; Green & Veres, 1990; Morgenson & Campion, 1997). In applying Cronbach’s model to the job analysis setting, Harvey and Lozada-Larsen (1988) suggest elevation and differential elevation as the most useful accuracy measures. Elevation indicates the overall leniency or severity of a job analyst. Differential elevation reflects how closely the average rating across all task dimensions of a given job compares to the expert average rating (Harvey & Lozada-Larsen, 1988). Therefore, accuracy in job analysis ratings can be considered as incumbent responses that are within acceptable limits around the standard profile of work behavior composed of tasks, generalized work behaviors and KSAs.

Green and Stutzman (1986) utilized Cronbach & Gleser’s (1953) $D$ statistic or a distance index to determine the similarity between an individual’s ratings and the population’s mean ratings. This index assesses the distance between an individual’s responses and that of average rater profile (Green & Stutzman, 1986). The $D$ statistic was used here to calculate profile similarity, the distance between an employee’s ratings of task statements and the centroid on these statements for all raters. They found that the average centroid correlations for the accurate raters was much higher ($r = .73$ to $.87$) than that found for the population of incumbents ($r = .35$ to $.79$). Thus in this study, accuracy was operationalized by comparing the psychometric distance between an individual’s responses and the average rater profile.

Carelessness and infrequency indices can also indicate accurate respondents (Green & Stutzman, 1986; Green & Veres, 1990; Harvey, 1991; Pine, 1995). The carelessness index includes two sets of tasks: those considered by job analysts to be related (true items) and those considered to be unrelated (distractor items) to the job in question on the basis on SME interviews, focus groups and other data sources (Green & Stutzman, 1986). Individuals are classified as inaccurate respondents if they rate tasks considered to be unrelated to the job as an important component to the work. Green and
Stutzman (1986) found that the interrater reliability was higher for incumbents selected as accurate ($r = .75$) as compared to interrater reliability across all incumbents ($r = .45$). This designation of accurate or inaccurate was based on the individual’s rater’s score on the carelessness index.

In a 1990 study, Green and Veres devised an infrequency index in order to assess the tendency of job analysis questionnaire respondents to rate job-irrelevant tasks as important. Similar to the carelessness indices used in previous work by Green and Stutzman (1986), Green and Veres (1990) adopted the name of the infrequency index from the personality literature because it more aptly described the intention of the index. Tasks are selected to be part of the infrequency index based on the infrequency that these activities were performed on the job. Results showed that the index was useful to detect respondents making constant error due to response sets and distortions but appeared less useful for detecting random errors. The carelessness and infrequency indices are computed from the entire sample of job analysis data and a subset that meet certain accuracy criteria are selected to use in the final analysis.

According to its psychometric definition, validity concerns “the degree to which the test measures what it purports to measure (Anastasi, 1988, p. 28).” The validity of job analysis generally refers to the extent to which a job analysis method provides all the relevant data that describes the job in question. One technique for demonstrating the validity of a job analysis questionnaire is the use of task coverage ratings (Harvey, 1991; Wilson, 1997). This approach asks incumbents to estimate the percentage of their job tasks included in the inventory, or the task coverage (Harvey, 1991; Wilson, 1997). Using a test-retest design, Wilson (1997) demonstrated that incumbents’ task coverage ratings did not drop substantially as expected when incumbents took a shortened questionnaire (reduction of 70% in task coverage) on the second administration. Other research suggests that the carelessness indices and profile comparison methods for
assessing accuracy actually provide ways to represent something close to an estimation of validity (Harvey, 1991; Wilson, 1997).

However, the fact that the literature does not provide a solid definition of validity of job analysis makes estimation of validity exceedingly difficult. Research does appear to present reliability and accuracy as necessary factors for building valid job analysis data (Harvey, 1991). Further study would need to examine whether the profile of work activities and KSAs for a particular job matches the closest picture to a “true depiction” of this job.

As the development of these indices indicate, research has begun to suggest ways to accommodate for the infallibility of human judgment in job analysis through evaluating methods for selecting accurate respondents. The repeat-item reliability method and carelessness indices provide ways to assess raters’ reliability and accuracy within the same administration of a job analysis inventory. As the work of Green and Stutzman (1986) and Green and Veres (1990) demonstrate, carelessness indices can be used to detect inaccurate respondents to structured job analysis questionnaires and select more accurate raters. However, the similar intent of the reliability and carelessness indices begs the question, “are the reliability and accuracy indices independent?” This study explored how to use these indices more effectively. As depicted in Figure 1, the first research question to be addressed in this study was “what is the relationship between reliability and accuracy?” This study utilized both a repeat-item reliability method as well as distractor items to determine the extent to which the two are related. This question was investigated in two ways: 1) at the item level using individual raters: rater performance on the repeat item reliability index will be compared to performance on the carelessness index; 2) at the profile level using rater and job profiles: the profile of reliable and accurate raters will be compared with that of the average rater.
Individual Differences

Other research has followed a path of examining individual and organizational factors that play a role in the job analysis ratings process (Landy & Vasey, 1991; Morgenson & Campion, 1997). Incumbent responses form the foundation for job analysis data. It follows that these data most often are derived from people’s personal insight and unique perspectives of their jobs (Goldstein, Zedeck & Schneider, 1993). However, as mentioned in the previous section, individuals may respond quite differently when compared with the average profile of responses. Is the reason people respond differently due to: 1) cognitive ability (education); 2) lack of familiarity with the job (experience); 3) fatigue. If fatigue is the reason, one would expect reliabilities to decline throughout the job analysis inventory however, no research has examined this issue to date. Thus, the fallibility of human judgment has strong implications on the accuracy of the job analysis data (Madden, 1964; Morgeson & Campion, 1997).

Morgenson and Campion (1997) concentrated on social and cognitive sources of potential inaccuracy in job analysis. Based upon an extensive review of the social, cognitive and industrial-organizational psychology literature, they identified a framework of 16 potential sources of inaccuracy in job analysis. The framework describes such social sources as self-presentation processes and social influence as well as cognitive sources such as limitation of and biases in information. Other individual factors such as race and level of education have also been examined. Green and Veres (1990) found a slight tendency for the workers who had achieved a higher educational level to be more accurate in their responses as indicated by their scores on a carelessness index. In a 1991 study, Landy and Vasey (1991) examined whether incumbents’ responses to a task inventory for the job of patrol officer were significantly different based on years of education or race. Results showed that education level and race had little or no effect on job analysis ratings.
The impact of other individual differences concerning job performance and experience to the validity of job analysis ratings has also been examined. Conley and Sackett (1987) examined whether using high- versus low- performing job incumbents as sources of job analysis information would have any influence on the data collected. They found that the lists of tasks, knowledges, skills and abilities generated by both performance groups were virtually the same. No differences in human judgments of ability requirements were found when responses were compared between experienced and less experienced firefighters (Bayer, Fleishman & Holt, 1996).

In contrast to the lack of differences found in these studies across rater characteristics, Landy and Vasey (1991) found that incumbent experience had a substantial influence on task ratings. There were uniform and reliable differences in responses to the task inventory when three experience groups of patrol officers (1-5 years, 6-10 years and 11 + years) were compared. Based upon these results, the authors suggested the need for a broad representation of experience levels in job analysis subject-matter-expert (SME) samples. They demonstrated the importance of using the stratification model (stratify by experience) to construct an SME sample for a job analysis.

As the above describes, the literature provides some interesting questions regarding the influence of various factors play on the validity of job analysis ratings. However, as the above description of the research indicates, the specific impact of individual factors on job analysis validity appears somewhat mixed. Further research can determine a more complete picture of the role such factors as experience, tenure and educational level can play in other samples of the same job or across other job categories. Does the significant impact of experience level on job analysis ratings found by Landy and Vasey (1991) appear in other studies of the job of patrol officer? More importantly, research has yet to utilize simultaneously specific reliability and accuracy indices to measure the impact of individual factors on job analysis validity.
The current study will utilize information regarding incumbents’ level of education and job experience (and/or organization tenure) to investigate whether there is a relationship among such factors and the reliability and accuracy of job analysis data. It will address the following research question: Can an individual characteristic profile for the consistent and accurate raters be identified? Is there a relationship between level of education and rater consistency and accuracy? Is there a relationship between amount of job experience (and/or organizational tenure) and rater consistency and accuracy? Does the reliability of a respondent’s ratings decrease as they move through a large task inventory? Based upon these results, suggestions can be made concerning what individual and organizational information can be useful in determining which subject matter experts (S.M.E s) should be used in the job analysis rating process.

Nature of the Task

The differences in the types of work activities one is rating may contribute to variance in job analysis ratings. The tasks or work activities required for a particular job can be categorized in more than one dimension of performance. Drawing upon research supporting a multi-dimensional perspective of performance, recent attention has been given to the role performance dimensionality plays in rating of job activity (Werner, 1994; Borman, White & Dorsey, 1995). There is considerable support for making a distinction between task and contextual performance. Task performance, or “in-role” behavior, refers to the operational activity one does to achieve role-prescribed duties or work objectives defined specifically by one’s job description (Werner, 1994). Contextual performance, or “extra-role” behavior, refers to citizenship activity one does that cannot be explicitly required in advance for a given job and that contributes toward the organizational, social and psychological work context (Borman & Motowidlo, 1997). Research often uses the term Organizational Citizenship Behavior to describe these
“extra-role,” discretionary, voluntary contributions for which one often receives no
formal credit (Organ, 1990). It defines the work that is outside the realm of required
operational activity. OCB is most commonly described and understood by the following
characteristics: altruism, courtesy, sportsmanship, conscientiousness and civic virtue.

Although the concept of organizational citizenship behavior (Organ, 1988) has
been around for more than a decade, there has been a recent surge of interest to explore
the role the nature of the task may play in rating behavior. For example, Werner (1994)
found that dimensions capturing both in-role and extra-role behaviors strongly influenced
various measures of rater search strategies as well as supervisory ratings. In addition,
high levels of extra-role behaviors were found to result in significantly greater halo error.
Motowidlo and Van Scotter (1994) compared ratings of task, contextual and overall
performance among Air Force personnel. Very similar correlations were found between
task and overall performance (.43) and contextual and overall performance (.41). In
addition, these two dimensions of performance contribute independently to an
individual’s overall performance. In addition, personality constructs were also found to
be good predictors of contextual performance.

Thus, the literature displays support for a multi-dimensional view of performance,
one that consists of both task and contextual performance. In addition, supervisor ratings
appear to demonstrate that individuals seek out and appreciate different types of work
behavior evidence when making performance assessments. Does this mentality also
apply to job analysis ratings? Do incumbents make distinctions between activities within
either task or contextual components of performance by rating them differently? As
Figure 1 shows, the third research question addressed in the present study will be “do
people rate more consistently in-role behaviors or extra-role behaviors?” Incumbent task
ratings concerning the level of activity and significance will be compared between in-role
and extra-role work activities. Therefore, the study will determine more specifically: 1)
do people more consistently rate their performance of in-role behaviors or their
performance of extra-role behaviors? 2) do people more consistently rate the significance of extra-role behaviors or the significance of in-role behaviors? These findings will provide a better understanding of the potential role the nature of the task may play as a source of unintended variance (systematic error) in job analysis ratings.

Research Questions

A review of the literature provides some interesting questions regarding the influence of various individual and organizational factors upon the validity of job analysis ratings. The current study will seek to determine whether an individual characteristic profile for the consistent and accurate rater can be identified. To achieve this, this study will begin by addressing the first research question: what is the relationship between reliability and accuracy? The second set of research questions will explore whether an individual characteristic profile for the consistent and accurate raters be identified. More specifically it will seek to answer the following research questions: Is there a relationship between level of education and rater consistency and accuracy? Is there a relationship between amount of job experience (and/or organizational tenure) and rater consistency and accuracy? Does the reliability of a respondent’s ratings decrease as they move through a large task inventory? Lastly, the third research question to be addressed in the study will be “do people who rate more consistently and accurately rate in-role behaviors differently from extra-role behaviors?” Incumbent task ratings concerning the level of activity and significance will be compared between in-role and extra-role work activities. Therefore, the study will determine more specifically 1) do people more consistently rate their performance of in-role behaviors or their performance of extra-role behaviors? 2) do people more consistently rate the significance of extra-role behaviors or the significance of in-role behaviors? Based upon these results, suggestions can be made concerning what individual and organizational information could be useful
in determining which subject matter experts (S.M.E.’s) should be used in the job analysis rating process.
Method

Participants

Two data sets were used in this study. Each set contains the results of a job analysis inventory given to state police officers of four different ranks: lieutenant, first sergeant, sergeant, trooper that was administered twice. Incumbents were told that the survey would be used for the development of a promotion system. The data set is from the first administration involved nearly all incumbents (n=1181) in the organization in the specified ranks. The second data set is includes data from a second administration (or retest of the original inventory) as well as data from the original administration for a sample of incumbents. In the second administration, a stratified random sampling technique using a random number table was employed to select a second group of incumbents (n = 125), approximately, 11% of the original sample (n = 1181). The second administration took place approximately three and a half months later.

Instrument Design and Procedure

The Job Description Survey had five sections. The first section of the Job Description Survey used in this study contains 345 tasks, grouped under 18 duty categories and one Miscellaneous category. Incumbents rated tasks on a Level of Involvement scale ("I ___ this item"), with the anchors do not do, do, fill-in for someone doing, and supervise for someone doing. Incumbents who rate the task "do not do" on the Involvement scale were instructed that they do not have to rate the item according to the Significance scale. Tasks items were also rated according to a Significance scale ("This item is ___ significant to my job"), with the anchors not, slightly, moderately, very, and extremely.

The second section of the survey contains a listing of forms and reports. Items were also rated on an Involvement scale ("I ____ this form or report.") with anchors: does
Incumbents rated items in this section using the same Significance scale described above.

In the third section, incumbents rated knowledge and skills requirements of nineteen duty categories which will be categorized by whether they pertain to "Patrol," "People" or "Paperwork." Incumbents rated knowledge ("Knowledge in this area is ____ to my job.") and skill ("Skill in this area is ____ to my job.") on an Importance scale with the following anchors: unimportant, somewhat important, moderately important, and very important.

In the fourth section, the inventory required incumbents to rate seven abilities in relation to their importance to the job Overall or to the "Patrol," "People," and "Paperwork" areas of their work. The seven abilities used were drawn from recent factor analysis work (Powell, Cunningham, Wimpee & Wilson, 1994) and included equipment-control sensorimotor ability, general physical ability, manual ability, numerical ability, reasoning & problem solving, verbal ability, and visual perception. Incumbents were then asked to rank the top three abilities in relation to their job overall, and with regard to the categories of "Patrol", "People", "Paperwork".

In the fifth section of the survey, there were demographic and reaction questions. Demographic questions concerned tenure and level of education and included such items as "My ethnicity is ___." or "I have been in this position for ___ years." Reaction questions required incumbents to provide feedback regarding such issues as instruction difficulty, item difficulty, time to complete survey and job coverage.

The first administration of the job description survey involved only the traditional paper-pencil format. Three and a half months later the survey was administered again to a random sample of the original participants. All incumbents on the second administration took half of the task section (approximately 180 items) in a paper-pencil format and half in a computer format. The order in which the two formats was administered (computer format first or paper-pencil format first) was counterbalanced to
minimize any confounding effect from administration sequence. The knowledge, skills and abilities items and the demographic and reaction questions were administered in a paper-pencil format for both administrations of the survey. The ATSURVEY software package for DOS, a job analysis questionnaire program, was used for the computer portion of the survey administration.

Reliability and Accuracy Indices

The survey contained checks to estimate the reliability and accuracy of incumbents’ ratings. A repeat-item reliability index was included in the survey. Twenty-seven task items were randomly selected and repeat in this section of the survey. Eight task items were repeated within the their same duty category (repeated within context). The remaining 19 items were contained in final category labeled "Miscellaneous" (repeated out-of-context). Table 1 shows the items repeated within task categories and within the Miscellaneous category. Incumbents were cautioned that tasks may be listed under more than one group or may appear twice. They were told that not all tasks apply to their jobs and that they should rate all items that apply to their jobs. Responses to all 27 repeat task items were pulled together as an assessment of rater consistency.

Incumbents received a score from 1 to 27 depending on the number of repeat items they rate the same as the corresponding original item.

In addition to repeat-item reliability index, the inventory also included an accuracy index. Table 2 provides a list of five “distractor” task items that were included in the inventory. The “distractor” task items were created by an advisory council made up of representatives of the police organization. They were included in the “appropriate” task category. Incumbents received a score of one to five depending on the number of distractor items that they correctly answer as not applicable to their job. The five “distractor” task items were taken together as an estimate of validity of ratings.
Results

Scores on the reliability and accuracy index were computed as previously described. Table 3 provides descriptive statistics of the reliability and accuracy indices. As Table 3 demonstrates, dispersion of scores was greater on the reliability index with a mean of 24.146 and a standard deviation of 3.34 as compared to a mean of 4.569 and a standard deviation of .74 on the accuracy index. For the purposes of this study, reliable raters were considered those who scored a twenty-six or better on the reliability index (within the top quartile) and accurate raters were considered those who scored a five on the accuracy index.

Research question one concerned the relationship between reliability and accuracy. Are they similar variables? More specifically, this question refers to whether reliability and accuracy indices will identify a similar set of consistent and accurate raters. Two dimensions to this research questions were investigated. First, the relationship between reliability and accuracy was examined at the item level. Rater performance of the repeat item reliability index was compared to performance on the carelessness index. As shown in the correlation matrix in Table 4, the relationship between reliability and accuracy scores is significant ($r = .355, p < .0001$).

Similar to previous research, the relationship between reliability and accuracy was also examined at the profile level using item and rater profiles (Green & Stutzman, 1986). Table 5 presents the mean reliabilities for mean task ratings of different groups of raters: all incumbents, reliable raters (persons who scored of twenty-seven or better on the reliability index), accurate raters (persons who scored of five on the accuracy index) and reliable and accurate raters (persons who met both reliability and accuracy criteria). Mean reliabilities) for the task items were computed using Fischer's $z$ transformation and finding the weighted $Z$ mean of the correlations and converting it back to a mean correlation (Steel & Torrie, 1980, Cohen & Cohen, 1983). The weighted $z$ mean
computation takes into account the different variances of each correlation as opposed to the typical z mean procedure, which was also done for comparison purposes. By taking into account variances differences among the correlations included in the mean procedure, the weighted z mean calculation provides a more accurate assessment of the overall mean correlation for each type of item. Thus, the weighted z mean was the procedure chosen to calculate all the mean correlations shown in the data tables. As shown in Table 5, mean reliabilities were higher for all selected groups than that found for the population of incumbents. Mean reliabilities were highest for the reliable raters \( (r = .86 \text{ and } r = .85) \) and reliable and accurate raters \( (r = .87 \text{ and } r = .84) \). Although the mean reliabilities for the accurate incumbents were higher than that for the entire population, the reliability index appears to have a greater impact on the overall reliability of the inventory than the accuracy index.

In addition to examining the overall mean task ratings of different rater groups, comparisons of the reliability of individual rater profiles were also made. A dataset of the twenty-seven initial and repeat task items was transposed to provide overall repeat item reliabilities for each rater (Wilson, Harvey & Macy, 1990). Table 6 and 7 provide descriptive statistics from this profile analysis by the level of involvement and significance scales, respectively. The mean reliability of the rater profiles was quite high \( (r = .90, r = .89) \). In view of the profile distance between the average and reliable rater, there was a wide range of reliabilities (from a high of \( r = .99 \text{ and } r = 1.00 \) to a low of \( r = -.57 \text{ and } r = -.50 \)). Seventy-five percent of incumbents had a reliability of \( r = .69 \text{ and } r = .73 \) or greater.

The second set of research questions concern whether an individual characteristic profile for the consistent and accurate raters can be identified. Is there a relationship between level of education and rater consistency and accuracy? Is there a relationship between amount of job experience (and/or organizational tenure) and rater consistency and accuracy? In order to determine the extent of these relationships, a correlation
analysis among reliability and accuracy index scores and tenure and education variables was performed. Table 4 presents the results of these analyses. As expected, significant relationships were found between position tenure and organization tenure ($r = .605$, $p < .0001$), and prior education and post education ($r = .844$, $p < .0001$). A significant negative relationship was found between reliability scores and organization tenure ($r = -.190$, $p < .0001$). A significant negative relationship was also found between prior education and position tenure ($r = -.099$, $p < .0001$).

The study also examined the impact individual’s level of fatigue may have had on the reliability and accuracy of ratings. More specifically, this study examined whether the reliability of a respondent’s ratings decreases as they move through a large task inventory. In order to answer this question, task item reliabilities were computed based upon their position in the inventory. The data set from the test-retest situation (n =125) was be used for these analyses. The 345 task items were chunked in groups of 50 (and one group of 45) in order of their appearance in the inventory. Mean task rating test-retest reliabilities for each group of task items were compared with one another in order to identify any trends. As shown in Table 8, test-retest reliabilities for task level and task significant ratings were highest in the first task grouping ($r = .729$ and $r = .755$, respectively) and lowest at the end of the third task grouping ($r = .542$ and $r = .501$, respectively). Using the test for significance of the difference between independent correlation coefficients (Cohen & Cohen, 1975), these differences between the task item grouping with highest mean test-retest reliabilities and the task items grouping with the lowest mean test-retest reliabilities was not significant. Figure 2 portrays how the changes in reliabilities trend downward at the end of the third grouping and then increase over the remaining task groupings.

For this last set of research questions, incumbent task ratings concerning the level of activity and significance were compared between in-role and extra-role work activities. In order to determine the differences in rating behavior according to the nature of the
task, the following analyses was performed. The 345 task items were first divided into two categories depending on the nature of performance required for the task: 1) in-role task items; 2) extra-role task items. Decisions as to which category the task item was placed depended on whether the task item requires operational activity (in-role) or citizenship activity (extra-role) by the individual performing the task. Five industrial and organizational psychologists and subject matter experts on the job represented in this study categorized all 315 tasks. Thirty tasks were defined as extra-role task items and 345 tasks were defined as in-role task items. Once the task items were categorized, comparisons were made as to whether test-retest reliability changes when an individual is rating in-role tasks items as compared with extra-role task items. In order to make fair comparisons between the in-role and extra-role task item categories, three random samples of thirty in-role tasks were used. Similar to the previous research question, the data set from the test-retest situation (n =125) was used for these analyses as well. Mean test-retest reliabilities were computed for in-role task samples and the extra-role task group. As Table 10 shows, mean test-retest reliabilities were higher for the in-role task level and significance item samples ($r = .60, r = .56; r = .57, r = .54; r = .52, r = .54$) than for the extra-role task level and significance item samples ($r = .46$ and $r = .45$.) Using the test for significance of the difference between independent correlation coefficients (Cohen & Cohen, 1975), these differences between in-role mean reliabilities and extra-role mean reliabilities was however, not significant.
Discussion

Through its focus on a number of factors impacting the evaluation of job information, this study has shed some new light on job analysis validity, a topic that was once uncharted territory in the literature. In addition, the results of this study have implications for industrial/organizational psychologists in both research and practice alike. As depicted by the model in Figure 1, this study has examined several factors that impact the validity of job analysis ratings. It has delved further into the relationship between job analysis reliability and accuracy, investigated methods for selecting accurate and reliable raters, considered the role individual differences may play, and lastly, explored the influence of the nature of task. These issues are fundamental to the understanding of the reliability and accuracy of job analysis ratings and are issues at the heart of current debates in the literature concerning the lesser known concept of job analysis validity. With respect to the research questions, what can be concluded about the validity of job analysis ratings and where it leaves the importance of the evaluation of job information today?

First, this study examined two variables that are central for understanding job analysis validity: reliability and accuracy. A first look into the results of the study would show that there is a significant and positive relationship between job analysis reliability and accuracy. Given this relationship, it can be seen why these terms may often be used interchangeably in the research literature. However, when the results of the study are examined more closely, one sees some differences in the influences of reliability and accuracy on the validity of job analysis ratings and their relationship with other variables.
For example, in the comparison of reliabilities for selected incumbent groups, the mean reliabilities for reliable incumbents were remarkably higher than that of the population. In contrast, the mean reliabilities for accurate incumbents were no different from the population. One reason for this difference could be range restriction. Greater than seventy-five percent of incumbents had a perfect score on the accuracy index compared to less than twenty-five percent on the reliability index. Thus, the larger set of items on the reliability index may have helped its impact appear more pronounced at the item level.

In addition, from the analysis of individual raters responses across a set of items (i.e., profile analysis), one can see that mean individual rater reliabilities are also quite high. However, the individual profile reliabilities begin to trend rapidly downwards for the bottom half of raters. Therefore, there can be a dramatic difference in the reliability of individual raters. This finding taken with the large difference in the reliability of the inventory between reliable raters and the population suggests that unreliable raters can have a tremendous impact on the validity of job analysis results.

Slight differences in the relationship reliability and accuracy have with individual difference variables were also found. A correlation analysis that was performed revealed a significant and negative correlation between reliability and organization tenure. A negative correlation was also found between accuracy and organization tenure but it was not significant. No significant relationships were found between reliability and education or between accuracy and education. This is consistent with previous research with a similar population that found that educational level had minimal effect of job analysis ratings (Landy & Vasey, 1991). This could be due to the fact that individual’s education
may not have provided them with job specific knowledge or other skills that would positively influence rating reliability or accuracy. Perhaps what is most interesting about this comparison of reliability and accuracy with individual difference variables is the negative relationship that was found with organizational tenure. This would suggest that individuals that have spent the most time with the organization appear to be less reliable and less accurate in their ratings about their jobs. How could this be? Landy and Vasey (1991) also identified experience to be the most important demographic variable in their analysis. In addition, they found that the direction of the relationship between tenure and task ratings varied by task categories such as traffic activities (negative relationship) and non-criminal interventions (positive relationship). With regard to the current study, it is possible that as incumbents experience level increased, they took on slightly different responsibilities and in the organization and became less familiar with specific tasks in the job. The fact there was not as strong of a negative relationship between reliability and position tenure and accuracy and position tenure provides support for this observation. Another explanation concerns the distinctions among declarative, procedural and strategic knowledge. It is possible that over time, specific knowledge of operational activities and key job information becomes more strategic, almost second nature to more the experienced individuals. As expertise and knowledge is built up over time, one has to purposefully recall it to use it. The knowledge becomes less declarative and harder to readily describe. Thus, shifts in job duties and one’s developing expertise over time in an organization may have an impact on one’s accuracy and reliability in job analysis ratings.

Individual difference variables such as experience and education can be considered to be characteristics of the rater, or rater demographics. Rater demographics
have been examined extensively in the literature for its potential contribution as a source of unintended variance (systematic error) in job analysis ratings (e.g., Conley & Sackett, 1987; Landy & Vasey, 1991). As discussed above, the literature has pointed to such issues as cognitive ability (education) and lack of familiarity with the job (experience) as potential reasons for why people respond differently to job analysis inventories. In contrast to individual differences as characteristics of the rater, one can also examine variables that impact the rater specifically in the rating situation such as anxiety and fatigue (Anastasi, 1988). Job analysis inventories are often long and can be quite tiring and tedious situation for the rater. The current study looked at the impact of fatigue on the reliability of the job analysis reliability. This analysis of reliability over the course of a job analysis inventory showed that mean reliabilities trend downwards initially and slope slightly back upwards. The downward trend may suggest that fatigue does impact the reliability of the rater over time. However, why did the mean reliabilities increase slightly halfway through the inventory before trending down once more? One explanation could be the differences in rater’s response to particular task categories.

Another explanation could be related to the unique aspect to the rating situation of the smaller dataset that was used in this analysis. As described previously, the incumbents in this dataset participated in two administrations of the job analysis inventory (test re-test situation). On the second administration of the inventory, incumbents took the first half of the inventory in paper-pencil format and the second half in a computerized format (or vice-versa). The sudden change in trend could indicate that point at which incumbents changed survey format (from computer to paper-pencil or vice-versa). The change in format could have given the rater a slight “mental break” or
influenced their attentiveness enough to have a positive effect on reliability. Others have made similar connections between reliability and attentiveness. Suggestions have been made that alphabetizing the task list under duty headings or increasing the amount of information requested of incumbents for each task item may help respondents play closer attention to the task they are rating and therefore, increase reliability (McCormick, 1960; Christal & Weissmuller, 1988). Thus, fatigue and rater attentiveness may have an impact on rater reliability and accuracy and may also help to explain some of the unintended variance (systematic error) that occurs with job analysis ratings.

Although the characteristics of the rater or the rating situation are traditionally often identified as contributors to error in ratings, this study looked at a relatively new “culprit” in the area of job analysis research. The current study examined the potential role the nature of the task may play as a source of unintended variance (systematic error) in job analysis ratings. The literature on organizational citizenship behavior first emphasized the concept of performance dimensionality and the distinction between task performance (“in-role”) and contextual performance (“extra-role”) (Organ, 1988). There now is considerable support for the assertion that the types of activities one is rating may contribute to variances in the rating of job activity (Werner, 1994; Borman, White & Dorey, 1995).

This current study extended this research by demonstrating the potential impact of performance dimensionality on reliability of job analysis task ratings. Mean correlations of task ratings for random samples of in-role task were much higher than that of extra-role task ratings. What does this tell us? One explanation may be that this indicates that there is something else going on at a more cognitive level during the rating process.
Incumbents may be better able to get a picture in their mind of the day-to-day operational ("in-role") activities and can rate them more reliably. It may be harder to recall details of extra-role tasks, which are less straightforward, discretionary and outside the realm of required activity. A more simple explanation for the difference in reliabilities between in-role and extra-role tasks may be task frequency. Incumbents may perform in-role tasks more frequently than extra-role tasks, are more familiar with them and therefore, can rate them more reliably. Thus, similar to the previous discussion on fatigue, cognitive issues concerning the inventory content may also have an impact on reliability of job analysis ratings.

In sum, what does this study tell us that is new? What also are the limitations? The fact that reliability of job analysis inventories does matter and can vary by groups of individual raters is arguably the biggest contribution of this study to research. It showed how the reliability of an inventory changes drastically by selecting only reliable raters to use in the analysis. A further analysis of individual rater profiles demonstrated the very existence of raters with very low reliability, the rater group that skewed the overall reliability of the rating instrument. Thus, if the job analysis inventory did not include this repeat-item reliability index, one would have never known the impact of a handful of unreliable raters nor would have been able to isolate them from the others. Thus, this study provides tremendous support for the inclusion of reliability indices in job analysis inventories as one step toward understanding the validity of the instrument.

A second goal of this study was to examine methods that operationalize accuracy and reliability (i.e., the accuracy and reliability indices) and the interrelationship between these variables. These variables sit at the heart of much debate in the recent literature
regarding the validity of job analysis. Some research uses the terms of reliability and accuracy interchangeably. In contrast, other research considers these variables to be independent. The current study shows that these variables are not independent but significantly related. In contrast, there were some differences in how these variables impacted the overall validity of the inventory and how they related to individual differences variables. The range restriction of the accuracy index may have inhibited the full impact of the accuracy variable from being observed. If the accuracy index was composed of a larger set of items, it may have been easier to draw comparisons with the reliability index. However, if there are too many items in the distractor index it can reduce the credibility of the instrument and would also make the inventory that much longer. The number of items in the accuracy index in this study was similar to that of previous research (Green & Stutzman, 1986). In addition to demonstrating further the use of an accuracy index, this study provided further evidence to the feasibility of repeating items on a job analysis inventory as an easy means of assessing reliability. Overall, the study did provide a clearer picture of how reliability and accuracy relate and how they impact validity.

Another goal of the current study was to investigate what role individual differences and the nature of the task may play in the reliability and accuracy of job analysis inventories. Although previous research has examined such demographic variables as job experience and education, this study was able to look at their specific relationship to both reliability and accuracy. As previously discussed, it identified a significant and negative relationship between organization tenure and reliability and was able to provide greater insight into the factors that may be driving this finding. Once
again, the range restriction in the accuracy index may have prevented the finding of a significant relationship with the individual differences variables, specifically tenure.

The current study also looked into some issues that are fairly new to research into job analysis validity: fatigue and the nature of the task. Fatigue potentially played a role in the changes observed in the mean reliabilities over the course of the inventory. The fact that there was a change in inventory format with the dataset that was used in this analysis may have confounded the results. Lastly, the research study focused on the impact of performance dimensionality on reliability of job analysis ratings examined a very new component to job analysis research. The considerable difference in reliability between in-role and extra-role tasks suggests that this distinction between task and contextual performance has a tremendous impact on incumbents’ rating. This distinction has been shown in performance appraisal ratings in previous research (Motowidlo & Van Scotter, 1994). Now, we are seeing the impact on job analysis ratings. One limitation in this analysis was that the small sample size of extra-role tasks restricted the analysis to a sample of sixty task items (thirty of each). A larger set of items for comparison would have provided a level of power at which significance of differences between correlations could be more appropriately assessed.

Future research could examine the impact of performance dimensionality more closely as a potential source of systematic error. Was the finding of this current study that the nature of the task can impact reliability merely an artifact of the inventory used or a realistic depiction of the distinction between task and contextual performance? A future study could examine this by including a larger set of tasks with a even number of in-role and extra-role job activities. By identifying potential sources of unintended
variance as this, one can potentially increase the reliability of methods for collecting job analysis ratings.

A second research area to explore as a potential source of systematic error concerns the individual rater. Should raters be selected based on such characteristics as organization tenure? As stated previously, the negative relationship between reliability and tenure may be explained by shifts in duties that occur over an individual’s time with an organization. However, future research could look into this relationship further. Does attitude also play a role here? Does the skepticism or disinterest in one’s job that often arises over time in an organization have an impact on reliability of job analysis ratings? There could be a number of issues underlying this relationship.

Fatigue and rater attentiveness could be another area to examine, particularly if the research focused on the structure of job analysis inventories. Shorter and faster approaches for collecting job analysis ratings would likely be more amenable to incumbents and logically, could lead to more attentive raters and better job analysis results. Lastly, future research could examine different ways for structuring reliability and accuracy indices. The current study demonstrated the feasibility for including such indices within a survey and tremendous value they provided as measures of the accuracy and reliability (and potentially the validity) of a job analysis inventory. In addition, it showed how these indices could be used to make distinctions between raters and the possible benefit from selecting inventory respondents based on their reliability and validity.

Research that focuses on easy methods for assessing reliability and accuracy and for selecting reliable raters would be of particular interest to practitioners. In view of the
growing importance of inventorying human capital within a company and the accelerating pace at which decisions about people must be made, practitioners often have a small window of opportunity to make assessments of existing job activities and characteristics as well as individuals’ skills and capabilities. Therefore, by developing faster and more effective ways for collecting reliable, accurate and consequently, valid job analysis information is highly desirable. The time to make these assessments and appropriate human resource decisions may be limited. However, the need to make the best decision possible is paramount. One cannot miss any opportunity to get the best people with the right skill sets in the shortest amount of time on the most critical projects in the company. Time to market is key to creating the most profit and time to talent, particularly in today’s competitive talent market, is even more fundamental. Neither are metrics that a company can afford to miss.
References


Harvey, R. J. & Wilson, M. A. (1999). Yes Virginia, there is an objective reality in job analysis. (To be published)


Table 1

**Repeat task items**

<table>
<thead>
<tr>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Repeat Items Within Task Categories</strong></td>
</tr>
<tr>
<td>Assign Troopers to various tasks</td>
</tr>
<tr>
<td>Attend SHP functions</td>
</tr>
<tr>
<td>Coordinate training schedule with the training center</td>
</tr>
<tr>
<td>Direct placement of allied agencies’ vehicles and equipment</td>
</tr>
<tr>
<td>Document all public complaints about Troopers</td>
</tr>
<tr>
<td>Give commendations</td>
</tr>
<tr>
<td>Maintain files of area/staff records and documents</td>
</tr>
<tr>
<td>Stop violators</td>
</tr>
<tr>
<td><strong>Repeat Items Within Miscellaneous Category</strong></td>
</tr>
<tr>
<td>Apply safety awareness techniques</td>
</tr>
<tr>
<td>Arrest individual for bookable violations and warrants</td>
</tr>
<tr>
<td>Attend press conferences/ media events on special projects</td>
</tr>
<tr>
<td>Chart trainees’ daily activities</td>
</tr>
<tr>
<td>Check schedules for accuracy</td>
</tr>
<tr>
<td>Correspond regularly with court officials to discuss cases, court performance, &amp; maintain good working relationship</td>
</tr>
<tr>
<td>Enforce laws</td>
</tr>
<tr>
<td>Enter all necessary information into CAD</td>
</tr>
<tr>
<td>Follow court orders in regard to evidence</td>
</tr>
</tbody>
</table>
Table 1 (cont’d)
Repeat task items

<table>
<thead>
<tr>
<th>Item</th>
</tr>
</thead>
</table>

Repeat Items Within Miscellaneous Category, cont’d

Maintain communication with Communication Center about location, destinations, and outcomes using appropriate 10 codes

Maintain evidence locker

Maintain or replace equipment

Meet with assigned Troopers to brief them on their duties

Notify and update communication center to advise them of pursuit, give vehicle description, directions, etc.

Obtain DMV inquiries

Record witnesses’ statements

Respond to public notification of road problem or violation

Take defendant’s personal property

Transport arrestee to Magistrate’s or hospital
Table 2
Distractor task items

<table>
<thead>
<tr>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authorize out-of-state travel</td>
</tr>
<tr>
<td>Discuss disciplinary actions with complainant</td>
</tr>
<tr>
<td>Fire warning shots</td>
</tr>
<tr>
<td>Pursue violator across state lines at officer’s discretion</td>
</tr>
<tr>
<td>Unarrest defendant</td>
</tr>
</tbody>
</table>
Table 3
Descriptive statistics of reliability and accuracy indices

<table>
<thead>
<tr>
<th>Index</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>1181</td>
<td>24.146</td>
<td>25.000</td>
<td>26.000</td>
<td>3.34</td>
</tr>
<tr>
<td>Accuracy</td>
<td>1181</td>
<td>4.659</td>
<td>5.000</td>
<td>5.000</td>
<td>.740</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Index</th>
<th>100 % Max</th>
<th>75% Q3</th>
<th>25% Q1</th>
<th>0% Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>27</td>
<td>26</td>
<td>23</td>
<td>3</td>
</tr>
<tr>
<td>Accuracy</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>
# Table 4
Correlations among accuracy and reliability index scores and individual difference variables

<table>
<thead>
<tr>
<th>n</th>
<th>Mean r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1. Reliability</td>
<td>1181</td>
</tr>
<tr>
<td>2. Accuracy</td>
<td>1181</td>
</tr>
<tr>
<td>3. Position Tenure</td>
<td>1158</td>
</tr>
<tr>
<td>4. Organization Tenure</td>
<td>1157</td>
</tr>
<tr>
<td>5. Prior Education</td>
<td>1153</td>
</tr>
<tr>
<td>9. Post Education</td>
<td>1131</td>
</tr>
</tbody>
</table>

* $p < .0001$
Table 5
Mean reliabilities for task level and significance ratings of selected incumbent groups

<table>
<thead>
<tr>
<th>All Incumbents (N = 1181)</th>
<th>Reliable Incumbents (N = 519)</th>
<th>Accurate Incumbents (N = 916)</th>
<th>Reliable &amp; Accurate Incumbents (N = 463)</th>
</tr>
</thead>
<tbody>
<tr>
<td>n  r</td>
<td>n  r</td>
<td>n  r</td>
<td>n  r</td>
</tr>
</tbody>
</table>

Task level ratings

| 27 .70 | 27 .86 | 27 .69 | 27 .87 |

Task significance ratings

| 27 .72 | 27 .85 | 27 .73 | 27 .84 |

* In computing the average reliability, mean correlations were converted to Fisher-z scores, averaged, and then converted back to correlations.
Table 6
Descriptive statistics of reliability profile analysis using task level ratings

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted Z</td>
<td>1167</td>
<td>1.447</td>
<td>1.284</td>
<td>2.647</td>
<td>.776</td>
</tr>
<tr>
<td>$r^*$</td>
<td>1167</td>
<td>.90</td>
<td>.86</td>
<td>.99</td>
<td>.226</td>
</tr>
</tbody>
</table>

100% Max      | 75% Q3 | 25% Q1 | 0% Min |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted z</td>
<td>2.647</td>
<td>1.648</td>
<td>.854</td>
</tr>
<tr>
<td>$r^*$</td>
<td>.99</td>
<td>.93</td>
<td>.69</td>
</tr>
</tbody>
</table>

$^a$ In computing the average reliability, mean profile correlations were converted to Fisher-z scores, averaged, and then converted back to correlations.
Table 7  
Descriptive statistics of reliability profile analysis using task significance ratings

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted Z</td>
<td>1176</td>
<td>1.428</td>
<td>1.344</td>
<td>2.647</td>
<td>.723</td>
</tr>
<tr>
<td>r*</td>
<td>1167</td>
<td>.89</td>
<td>.87</td>
<td>.99</td>
<td>.213</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>100% Max</th>
<th>75% Q3</th>
<th>25% Q1</th>
<th>0% Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted z</td>
<td>3.273</td>
<td>1.828</td>
<td>.928</td>
<td>-.549</td>
</tr>
<tr>
<td>r*</td>
<td>1.00</td>
<td>.95</td>
<td>.73</td>
<td>-.50</td>
</tr>
</tbody>
</table>

*a In computing the average reliability, mean profile correlations were converted to Fisher-Z scores, averaged, and then converted back to correlations.
### Table 8
Mean reliabilities for task ratings by inventory position grouping

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean r</th>
<th>Mean Z</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Task level ratings</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task items 1-50</td>
<td>50</td>
<td>.62</td>
<td>.729</td>
</tr>
<tr>
<td>Task items 51-100</td>
<td>50</td>
<td>.57</td>
<td>.650</td>
</tr>
<tr>
<td>Task items 101-150</td>
<td>50</td>
<td>.50</td>
<td>.542</td>
</tr>
<tr>
<td>Task items 151-200</td>
<td>50</td>
<td>.56</td>
<td>.637</td>
</tr>
<tr>
<td>Task items 201-250</td>
<td>50</td>
<td>.60</td>
<td>.692</td>
</tr>
<tr>
<td>Task items 251-300</td>
<td>50</td>
<td>.62</td>
<td>.724</td>
</tr>
<tr>
<td>Task items 301-345</td>
<td>50</td>
<td>.58</td>
<td>.664</td>
</tr>
<tr>
<td><strong>Task significance ratings</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task items 1-50</td>
<td>50</td>
<td>.64</td>
<td>.755</td>
</tr>
<tr>
<td>Task items 51-100</td>
<td>50</td>
<td>.55</td>
<td>.615</td>
</tr>
<tr>
<td>Task items 101-150</td>
<td>50</td>
<td>.46</td>
<td>.501</td>
</tr>
<tr>
<td>Task items 151-200</td>
<td>50</td>
<td>.51</td>
<td>.564</td>
</tr>
<tr>
<td>Task items 201-250</td>
<td>50</td>
<td>.58</td>
<td>.667</td>
</tr>
<tr>
<td>Task items 251-300</td>
<td>45</td>
<td>.55</td>
<td>.620</td>
</tr>
<tr>
<td>Task items 301-345</td>
<td>45</td>
<td>.56</td>
<td>.634</td>
</tr>
</tbody>
</table>

In computing the average reliability, mean correlations were converted to Fisher-$z$ scores, averaged, and then converted back to correlations.
Table 9
Mean reliabilities of task ratings by in-role and extra-role categorizations

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean $r$</th>
<th>Mean $Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Task level ratings</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-role mean correlation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>random sample 1</td>
<td>30</td>
<td>.60</td>
<td>.696</td>
</tr>
<tr>
<td>random sample 2</td>
<td>30</td>
<td>.56</td>
<td>.630</td>
</tr>
<tr>
<td>random sample 3</td>
<td>30</td>
<td>.57</td>
<td>.643</td>
</tr>
<tr>
<td>Extra-role mean correlation</td>
<td>30</td>
<td>.46</td>
<td>.505</td>
</tr>
<tr>
<td><strong>Task significance ratings</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-role mean correlation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>random sample 1</td>
<td>30</td>
<td>.54</td>
<td>.610</td>
</tr>
<tr>
<td>random sample 2</td>
<td>30</td>
<td>.52</td>
<td>.573</td>
</tr>
<tr>
<td>random sample 3</td>
<td>30</td>
<td>.54</td>
<td>.610</td>
</tr>
<tr>
<td>Extra-role mean correlation</td>
<td>30</td>
<td>.45</td>
<td>.479</td>
</tr>
</tbody>
</table>

a In computing the average reliability, mean correlations were converted to Fisher-$z$ scores, averaged, and then converted back to correlations.
Figure 1. A model of the potential factors impacting the validity of job analysis ratings
Figure 2. A graph depicting changes in mean reliability of job analysis ratings throughout job analysis inventory.