

EVALUATING INDICATORS OF JOB PERFORMANCE: DISTRIBUTIONS
AND TYPES OF ANALYSES

by

Richard J. Chambers II, B.S.

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

COLLEGE OF EDUCATION
LOUISIANA TECH UNIVERSITY

November 2016

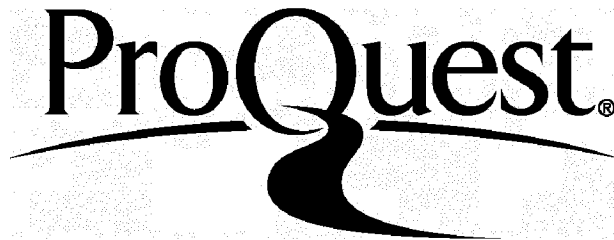
ProQuest Number: 10307765

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

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



ProQuest 10307765

Published by ProQuest LLC(2017). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code.
Microform Edition © ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

LOUISIANA TECH UNIVERSITY

THE GRADUATE SCHOOL


8-1-2016

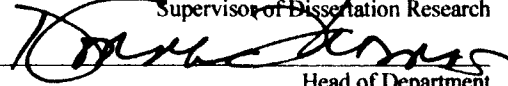
Date

We hereby recommend that the dissertation prepared under our supervision
by Richard Chambers, II

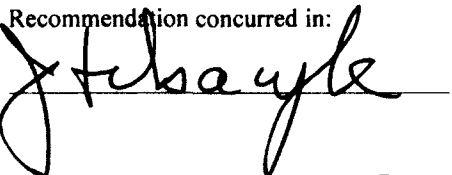
entitled EVALUATING INDICATORS OF JOB PERFORMANCE:
DISTRIBUTIONS AND TYPES OF ANALYSES

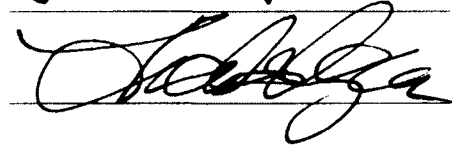
be accepted in partial fulfillment of the requirements for the Degree of
Doctor of Philosophy



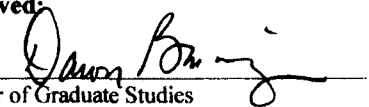
Supervisor of Dissertation Research


Head of Department
Psychology and Behavioral Sciences
Department

Recommendation concurred in:




Advisory Committee

Approved: 

Director of Graduate Studies

Approved: 

Dean of the Graduate School



Dean of the College

ABSTRACT

Distributions of job performance indicators have historically been assumed to be normally distributed (Aguinis & O'Boyle, 2014; Schmidt & Hunter, 1983; Tiffin, 1947). Generally, any evidence to the contrary has been attributed to errors in the measurement of job performance (Murphy, 2008). A few researchers have been skeptical of this assumption (Micceri, 1989; Murphy, 1999; Saal, Downey, & Lahey, 1980); yet, only recently has research demonstrated that in certain specific situations job performance is exponentially distributed (Aguinis, O'Boyle, Gonzalez-Mulé, & Joo, 2016; O'Boyle & Aguinis, 2012). To date there have been few recommendations in the Industrial-Organizational Psychology literature about how to evaluate distributions of job performance to determine whether they fit an exponential curve. There also has not been substantial justification in the literature as to why distributions of job performance would be expected to be normally distributed versus exponentially distributed. Furthermore, recent research about job performance distributions has narrowly focused only on a few specific types of work and on a few specific indicators of performance. Thus, research concerning distributions of job performance indicators is, to date, of limited generalizability.

The current research attempts to close the gaps in the literature by identifying high fidelity methods and applying them to classify distributions of various indicators of job performance on a continuous spectrum from normal to exponential. In this research,

multiple types of indicators of performance (and indices computed from combinations of indicators) were found to produce exponential distributions. More specifically, managerial indicators of job performance were found to best fit a normal distribution whereas objective measures, as well as composite measures of performance consisting of objective and subjective indicators, were found to best fit an exponential distribution. This study provides researchers and practitioners with new suggestions for classifying job performance distributions as well as new techniques for better differentiating between top and bottom performers.

APPROVAL FOR SCHOLARLY DISSEMINATION

The author grants to the Prescott Memorial Library of Louisiana Tech University the right to reproduce, by appropriate methods, upon request, any or all portions of this Dissertation. It was understood that "proper request" consists of the agreement, on the part of the requesting party, that said reproduction was for his personal use and that subsequent reproduction will not occur without written approval of the author of this Dissertation. Further, any portions of the Dissertation used in books, papers, and other works must be appropriately referenced to this Dissertation.

Finally, the author of this Dissertation reserves the right to publish freely, in the literature, at any time, any or all portions of this Dissertation.

Author Richard J. Chambers IV
Date 10-26-2016

DEDICATION

To my wife, Alecia, who has tolerated “I can’t do that now, I am working on my dissertation” for far too long.

TABLE OF CONTENTS

Abstract.....	iii
Dedication.....	vi
List of Tables.....	xi
List of Figures.....	xii
Chapter One Introduction.....	1
The Problem with Performance Distributions.....	5
The Importance of Distributions of Job Performance.....	6
The Proposed Study.....	12
Chapter Two History of Job Performance.....	14
Why Job Performance?.....	15
In the Beginning: Measuring Outcomes.....	19
Conceptual Advancement: Actual Job Performance versus Criteria.....	22
Theoretical Criterion Advancements.....	25
Improving Methods and Tools for Measuring Job Performance.....	32
Most Common Methods for Measuring Job Performance.....	37
Absolute Methods.....	38
Essays.....	38
Critical incidents.....	39
Comparative methods.....	39
Simple rank order.....	39

Alternating rank order	40
Paired comparisons.....	40
Relative percentile	40
Forced distribution.....	41
Comparing Absolute and Comparative Methods.....	42
Error Introduced by Raters.....	44
Most Common Rater Errors	45
Similar-to-me error	45
Contrast error.....	45
Leniency error	46
Severity error.....	47
Central tendency error.....	48
Halo/Horns error	49
Negativity error.....	50
Recency error.....	50
Primacy error.....	51
First impression error.....	51
Stereotype error	52
Attribution error.....	52
Modeling Error in Ratings of Job Performance.....	54
Rater Training.....	56
Behavioral observation training	57
Frame of reference training	58
Calibration meetings.....	59
Judgments Versus Ratings	62

Alternatives to Traditional Ratings of Job Performance	63
Modern Conceptualizations of Job Performance.....	66
Campbell's great eight.....	67
Task performance and contextual performance	68
Organizational citizenship behavior.....	70
Counterproductive work behavior	73
Adaptive performance.....	74
A general factor of job performance.....	75
Evaluating and Classifying Distributions of Job Performance	77
Evaluating the normality of distributions.....	78
Evaluating exponential distributions.....	79
Hypotheses.....	80
Chapter Three Methods.....	84
Participants.....	84
Procedures	85
Managerial ratings of short-term and long-term objectives.....	85
Manager quality	87
360° performance feedback.....	88
Time in role.....	89
Time prior to promotion	89
Measures.....	89
Manager short-term	90
Manager long-term	90
Manager quality	91
360° Performance Tool.....	92

Chapter Four Results	93
Organizing Performance Data for Analysis.....	93
Statistical Tests and Data Examination Procedures	95
Summary of Results by Hypothesis	99
Hypothesis 1	101
Hypotheses 2, 3, and 4.....	108
Hypothesis 2.....	116
Hypothesis 3.....	118
Hypothesis 4.....	120
Summary of Results.....	121
Chapter Five Discussion	124
Theoretical and Practical Implications.....	129
Study Limitations	134
Future Research.....	135
Conclusions.....	136
References	137
Appendix IRB Approval	164

LIST OF TABLES

Table 1. Groups of Performance Measures.....	94
Table 2. Summary of Criteria and Parameters.....	100
Table 3. Descriptives for Groups A, B, and C	102
Table 4. Descriptives for Groups D, E, F, G, and H.....	110

LIST OF FIGURES

<i>Figure 1.</i> Normal Distribution of Job Performance.....	3
<i>Figure 2.</i> Non-normal Exponential Distribution of Job Performance	9
<i>Figure 3.</i> Relationship Between Ultimate and Actual Criterion	30
<i>Figure 4.</i> Frequency Distribution of Group A.....	102
<i>Figure 5.</i> Frequency Distribution of Group B	103
<i>Figure 6.</i> Frequency Distribution of Group C	103
<i>Figure 7.</i> Q-Q plot of Group A.....	105
<i>Figure 8.</i> Q-Q plot of Group B.....	105
<i>Figure 9.</i> Q-Q plot of Group C.....	106
<i>Figure 10.</i> Frequency Distribution of Group D.....	110
<i>Figure 11.</i> Frequency Distribution of Group E	111
<i>Figure 12.</i> Frequency Distribution of Group F.....	111
<i>Figure 13.</i> Frequency Distribution of Group G.....	112
<i>Figure 14.</i> Frequency Distribution of Group H.....	113
<i>Figure 15.</i> Q-Q Plot of Group D	113
<i>Figure 16.</i> Q-Q Plot of Group E.....	114
<i>Figure 17.</i> Q-Q Plot of Group F	114
<i>Figure 18.</i> Q-Q Plot of Group G	115
<i>Figure 19.</i> Q-Q Plot of Group H	116
<i>Figure 20.</i> Graph of All Ratings Provided by SMEs.....	123

CHAPTER ONE

INTRODUCTION

Over time, the nature of work has evolved. Lerman and Schmidt (1999) point out that within the past one hundred years work in many parts of the world has shifted from being predominantly physical labor to being largely service-oriented. The shift has resulted in jobs becoming more vague, nebulous, and difficult to define. By extension, it has become increasingly difficult to articulate to employees what is expected of them and what “good” performance is. Along these same lines, it has become more difficult to evaluate and differentiate between high and low performing employees. However, current buzzwords and phrases, such as *top talent*, *talent wars*, and *Hi-Po* (i.e., high-potential employee), are indicators that organizations have an interest in identifying the best performers within organizations.

For Industrial and Organizational (I-O) Psychologists, the interest in assessing job performance has been a focal point over the last century. A search of PsycInfo, an online research search engine, for ‘job performance’ revealed 14,776 articles where job performance was a major theme. While job performance continues to receive much attention in the research literature, there are many points of contention between researchers. Austin and Villanova (1992) discussed four controversies in the job performance literature:

- What to consider as job performance (e.g., behaviors versus outcomes).
- How to measure job performance (e.g., descriptive versus quantitative methods, absolute versus relative ranking systems).
- What theoretically constitutes job performance and which theoretical models to apply.
- How job performance is distributed.

Austin and Villanova (1992) point out that lack of agreement over these four questions is understandable because performance can take on vastly different meanings, is conceptually abstract, and is extremely difficult to measure. Considering the complexity and lack of agreement concerning job performance, and because job performance is a broad and abstract concept, the following definition is a guide to help understand job performance: “Job performance is conceptualized as those actions and behaviors that are under the control of the individual and contribute to the goals of the organization” (Rotundo & Sackett, 2002, p. 66).

The fourth controversy, regarding how job performance is distributed, is particularly important when trying to evaluate employee job performance and when trying to specifically identify top performers within an organization. Understanding the meaning of a *distribution of job performance* is vital to determining differences in the performance of employees. A distribution of job performance is the result of natural variation that occurs in employees’ proficiency on the job; that is, different employees are bound to be more or less effective in their work when compared to others in the same role. Conceptually this may be straightforward; yet, as may be concluded from the ongoing debate in I-O research, it is difficult to convert conceptual variation between

employees' proficiency into practical solutions that reflect actual variation in job performance that organizations can leverage in a meaningful way. One attempt at a solution is for organizations to graph employees' job performance. Graphing employee job performance generates a visually discernable physical distribution that organizations can use to aid the process of understanding and differentiating between employees' varying levels of proficiency (see Figure 1).

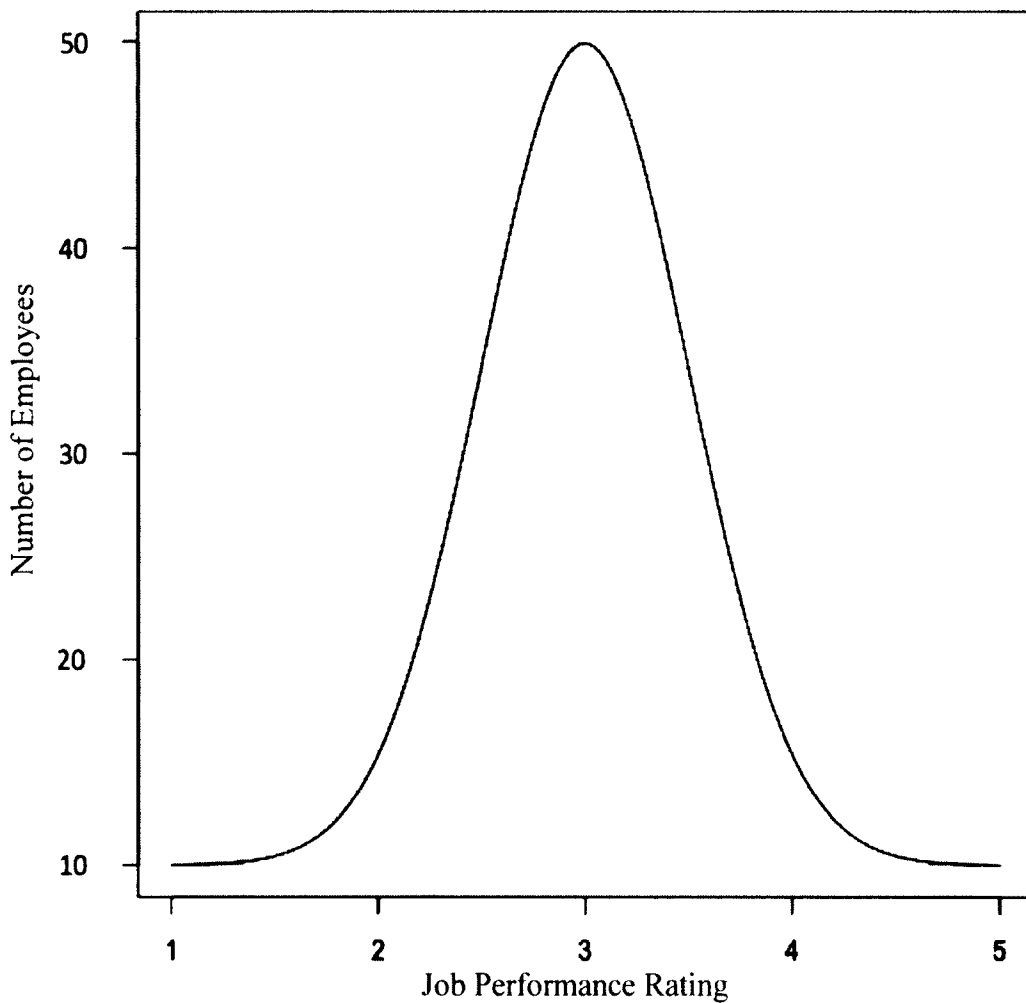


Figure 1. Normal Distribution of Job Performance.

To generate a physical distribution of employees' job performance, organizations can plot job relevant performance "scores" for employees that perform similar roles. However, in order to create a physical distribution of employees' job performance, each employee's proficiency must be quantified. Thus, one underlying problem in the debate about *how job performance is distributed* is muddled by the fact that anytime organizations attempt to quantify something conceptual, such as job performance, at least some degree of accuracy is lost. Job performance is a construct, which means that it is intangible and not *directly* measurable (Ronan & Prien, 1971). On the other hand, when attempting to measure the construct of job performance, indicators of the construct of job performance are being measured, which invariably include error (Ronan & Prien, 1971). As a result, all attempts to measure job performance will include some degree of error.

The present research focuses on improving the accuracy of physical distributions of job performance. Specifically, this study seeks to provide practical guidance on methods that researchers and practitioners alike can use to increase accuracy when attempting to analyze distributions of job performance and which researchers and practitioners can broadly apply to different job types and organizational contexts. This remains a gap in literature as there appears to be no prior studies that have proposed and tested an analytical approach to use across multiple jobs and performance dimensions (Aguinis & O'Boyle, 2014; Aguinis, O'Boyle, Gonzalez-Mulé, & Joo, 2016; Campbell & Wiernik, 2015; Crawford, Aguinis, Lichtenstein, Davidsson, & McKelvey, 2015). This study also seeks to provide evidence that job performance should not *always* be expected to be normally distributed. To underscore the importance of the need to represent job performance distributions accurately, there will also be a discussion on the impact that

different types of distributions of job performance can have on identifying and subsequently evaluating and managing individual job performance. Furthermore, none of the four controversies outlined operates in isolation, which is why it is particularly important to review them in depth. This is especially true for how job performance is distributed. That is, observed distributions of job performance are dependent upon how each of the first three controversies is ultimately approached and resolved.

The Problem with Performance Distributions

For decades, researchers and practitioners alike assumed that job performance would *always* be normally distributed (Aguinis & O'Boyle, 2014; Schmidt & Hunter, 1983; Tiffin, 1947). A normal distribution resembles a “bell shape,” is symmetrical around its mean, and has a mean, median, and mode which are all equal (Heiman, 2013). The problem with assuming job performance will *always* be normally distributed is that this assumption has not been consistently and critically evaluated. Researchers and practitioners continue to accept the assumption that job performance is *always* normally distributed when investigating human performance. This is the case despite the fact that organizations work, in general, and that how organizations select employees has changed. For instance, recall that work in many parts of the world has shifted from being predominantly physical labor to being largely service-oriented (Lerman & Schmidt, 1999). Organizations have also evolved and some organizations now operate on a global scale. Many organizations have also put into practice some of the vast I-O research on employee selection (e.g., Ryan & Ployhart, 2014), which theoretically should impact distributions of job performance because the use of validated selection systems should result in the selection of high opposed to low performing employees. Furthermore, over

the last century, instead of critically evaluating the assumption that job performance is *always* normally distributed, many studies on job performance became a sort of self-fulfilling prophecy whereby researchers set out to confirm this assumption. Thus, for the most part, the notion that job performance is normally distributed became unquestioned. In an effort to more accurately capture job performance, many researchers and practitioners came to believe that any indication that job performance was not normally distributed was the result of measurement error, the influence of irrelevant factors, or worse, the impact of intentional distortions (Murphy, 2008). If departures from normality in the measurement of job performance are actually the result of sampling error, then there would be justification to correct for normality (Anastasi & Urbina, 1997). However, given the changing landscape of work over the last century, a normal distribution of performance may no longer be the only, or most appropriate, distribution to represent measures of job performance. In turn, distributions of job performance that deviate from normality could actually be accurate representations of distributions of job performance. If this is true, then assuming a normal distribution of job performance or correcting distributions of job performance that deviate from normality may introduce additional error rather than correct for measurement error (Anastasi and Urbina, 1997).

The Importance of Distributions of Job Performance

How distributions of job performance are viewed (e.g., normal and non-normal) is important because this premise can influence many areas of I-O Psychology practice and research (Aguinis & O'Boyle, 2014; O'Boyle & Aguinis, 2012). For instance, practitioners may design performance-rating systems that force managers to assign performance ratings that fit a normal distribution of job performance, irrespective of

actual differences in employee job performance. This practice can artificially decrease the variation of job performance ratings, making it difficult to accurately differentiate between performers. This matters because employee performance ratings directly tie to organizational logistics such as compensation and promotions, as well as benefits such as inclusion to selective developmental opportunities.

As an example, organizations may instruct managers to distribute only a pre-specified percentage of each of the possible performance ratings among his or her employees (Motowidlo & Borman, 1977; Reilly & Smither, 1985; Schneier, 1977a, 1977b). However, it may not be appropriate to assume a predefined percentage of employees perform at certain levels under a normal curve. In fact, if an organization requires a realistic minimum level of performance in order to continue employment it might be reasonable to suggest that the lowest performers will exit the organization through attrition. In situations where organizations can retain higher performing employees at a rate greater than it is retaining lower performing employees, the result should be that very few employees would fall within the lowest end of any performance distribution.

In another example, if an organization leverages a valid selection system to select new employees, the result should be selection of a greater number of high performers, as opposed to low performers. If the selection system is valid, then through an iterative process of attrition and selection, over the natural course of time, a performance distribution should become more positively skewed (more high performers), as opposed to normally distributed (e.g., Beck, Beatty, & Sackett, 2014; Meyer, 1980; Schmidt & Hunter, 1983; Tiffin, 1947). Given these examples, *assuming* job performance should

resemble a specific type of distribution may result in a great deal of error. By assuming a pre-specified distribution of job performance, organizations may mistakenly differentiate employees and make poor or inappropriate decisions based on their inaccurate data.

An accurate understanding of the distributions of job performance, free of predetermined assumptions, has many benefits for individual employees as well as organizations. For example, increased realization and differentiation between top performers and bottom performers, opportunities to increase retention of a greater number of top performers, recognition of higher levels of attrition among specific sub groups, easier workforce and succession planning, and a greater ease of demonstrating the value of selection systems to upper management. Many of these potential advantages stem from the fact that an assumed normal distribution of job performance may be providing erroneous results concerning the performance differentiation between employees.

Additionally, there has been a narrow focus thus far in the literature when challenging the assumption of normality (e.g., Micceri, 1989; Murphy, 1999; Saal, Downey, & Lahey, 1980). Current research challenging the assumption of normality has primarily focused on non-normal, exponential distributions of job performance in a very limited number of occupations. A non-normal distribution can be any distribution that does not resemble the distribution in Figure 1, but the present research is interested specifically in the positively skewed and leptokurtic, exponential distribution. Given that the exponential distribution is of specific interest in the current research, it will be referenced directly going forward instead of using the broader terminology of non-normal distribution. An exponential distribution (see Figure 2) is a positively skewed, leptokurtic

distribution. Positively skewed means that there are a few scores that are substantially larger than the rest of the scores, which pulls the mean up. This also makes the mean greater than the median. Leptokurtic means that there is a large 'peak,' or very large mode. For example, a leptokurtic distribution can result when a large number of employees receive similarly high job performance ratings. Examples of the occupations used in previous research include actors, academics and professional athletes, which only make up a very small proportion of occupations (e.g., Aguinis & O'Boyle, 2012).



Figure 2. Non-normal Exponential Distribution of Job Performance.

Some researchers have argued to exclude the most common method of evaluating employee job performance, managerial performance ratings, as being able to produce exponential distributions of job performance, even in situations where exponential distributions of job performance have been found using performance ratings other than

managerial ratings (Aguinis & O'Boyle, 2014; O'Boyle & Aguinis, 2012). Yet, because managerial ratings of performance are the most common method used currently to assess job performance (Aguinis, 2013; Murphy, 2008; O'Boyle & Aguinis, 2012; Viswesvaran, Schmidt, & Ones, 2005), this is a major restriction limiting the availability of more appropriate performance distributions.

The use of managerial ratings of performance, which are a subjective type of rating, may make it difficult for organizations to identify exponential distributions of performance. Objective indicators of performance, as opposed to subjective indicators, tend to be more readily available for only a small set of occupations (e.g., salespeople) where organizations can use objective output (e.g., the number of sales) as an indicator of performance. An objective indicator is a measure of job performance that does not require a judgment or interpretation. Subjective indicators of job performance require a rater to make a judgment about how well an employee has performed a specific behavior. For example, a manager could provide a rating of job performance on how well a salesperson demonstrated the use of specific selling techniques.

Although managerial ratings of job performance are currently the most widely used method for assessing job performance, there has been a growing trend to completely do away with managerial ratings of job performance (Coens & Jenkins, 2002; Pulakos, Hanson, Arad, & Moye, 2015; Pulakos & O'Leary, 2011). This makes it imperative to identify additional alternative methods of measuring job performance that apply to a broad number of occupations. Furthermore, there are also subjective indicators of job performance that have yet to be thoroughly explored with an impartial lens (i.e., not presupposing a normal distribution), that may be readily accessible to organizations

interested in a holistic representation of employee performance and that may be exponentially distributed. Examples of these include managerial ratings of performance, upward ratings of manager quality, and 360° performance evaluations (e.g., multi-rater or multi-source feedback).

Not only is there a growing trend to do away with managerial ratings of job performance, but organizations that use managerial ratings of job performance may assume job performance ratings should be normally distributed and attempt to force a normal distribution of job performance to produce a distribution similar to the one illustrated in Figure 1. This occurs when organizations explicitly suggest to managers what percentage of employees should receive each potential rating. The distribution of performance ratings that results from this practice may not be accurate (Balzer & Sulsky, 1992; Cooper, 1981; Guion, 2011; Landy & Farr, 1980; Murphy & Cleveland, 1995; Murphy, Jako, & Anhalt, 1993; Murphy & Reynolds, 1988; Wallace, 1974). For example, organizations may provide the suggestion that approximately 40 percent of their employees should receive a rating of *three*, 25 percent should receive a rating of *two*, 25 percent should receive a rating of *four*, five percent of employees should receive a rating of *one*, and five percent of employees should receive a rating of *five*. The result of this practice would be a distribution of performance that closely resembles Figure 1. This example demonstrates that assuming and or suggesting a distribution of job performance can have a direct impact on employee performance ratings with no regard for whether or not job performance is *actually* normally distributed.

Some researchers have been skeptical that individual performance is normally distributed (e.g., Micceri, 1989; Murphy, 1999; Saal, Downey, & Lahey, 1980). O'Boyle

and Aguinis (2012) recently challenged the assumption that job performance is always normally distributed by using objective measures to produce exponential distributions of job performance. The researchers used objective measures of job performance to demonstrate exponential distributions of job performance because, unlike subjective measures, objective measures typically have an unrestricted maximum value. The greater the maximum value, the greater the opportunity for a positively skewed distribution. Subsequently, this also creates a greater opportunity for an exponential distribution to exist, because exponential distributions are characterized by a positive skew and leptokurtic peak. Thus, these authors demonstrated that in some instances job performance best resembles an exponential distribution.

The Proposed Study

In summary, the majority of past research has argued that job performance is normally distributed, with few challenges to this assumption. This has resulted in organizations suggesting to managers who provide ratings of job performance that their ratings should essentially resemble a normal distribution (Motowidlo & Borman, 1977; Reilly & Smither, 1985; Schneier, 1977a, 1977b). This typically results in distributions of job performance that are normally distributed, but that may not necessarily represent an accurate distribution of job performance (Balzer & Sulsky, 1992; Cooper, 1981; Guion, 2011; Landy & Farr, 1980; Murphy & Cleveland, 1995; Murphy, Jako, & Anhalt, 1993; Murphy & Reynolds, 1988; Wallace, 1974). Only recently has research begun to successfully challenge the previous assertion that job performance is always normally distributed (Aguinis & O'Boyle, 2014; O'Boyle & Aguinis, 2012).

The distribution of job performance is important not only because it can impact the individual job performance evaluation and management for each employee, but it can also impact other important aspects of work such as pay, promotion, retention, selection utility, team dynamics, team performance, succession planning and overall firm performance and culture (Pulakos & O'Leary, 2011). Therefore, accurately identifying the performance distribution of jobs should be a top priority. As top performers have the most job opportunities, provide the most value to organizations, and the war for talent is at an all-time high (Gravett & Caldwell, 2016), the current study suggests a paradigm shift. The proposed study argues for increased emphasis on differentiating and identifying top performers by embracing exponential distributions of job performance when and where appropriate. This is in contrast to assuming normal distributions of job performance. It is important to challenge the widely held assumption that job performance is normally distributed in order to ensure that the most fitting distribution given to a set of employee performance data in any given context is achieved (Aguinis & O'Boyle, 2014; Crawford et al., 2015). The current study seeks to close this knowledge gap by 1) *identifying* and *applying* statistical methods for determining whether the distribution of performance data is normal or exponential, and 2) investigating the potential of multiple types of indicators and combinations of indicators to produce exponential distributions.

CHAPTER TWO

HISTORY OF JOB PERFORMANCE

To place modern struggles of understanding job performance into context, the following discussion first reviews key historical influences on our understanding of job performance. Second, the discussion focuses on four broad topics related to job performance: 1) conceptualizing performance, 2) measuring performance, 3) addressing performance measurement error/contamination, and 4) the current state of affairs as related to job performance. To appreciate the importance of accurately identifying and appropriately analyzing different performance distributions, which is the focus of the current study, it is crucial to understand these four broader topics. In addition, each of these four topics provides insights into each of the four controversies introduced in Chapter One. Specifically, the first topic, conceptualizing performance, will provide insight on the controversy over what to consider job performance. The next topic, measuring job performance, and the third topic, addressing performance measurement error/contamination, will help guide understanding of the controversy surrounding measuring job performance. The fourth topic, the current state of affairs as related to job performance, will also aid understanding of the controversy around measuring job performance. Holistically, all four topics will provide insight into the controversy over how job performance is distributed.

Why Job Performance?

The drive to perform and produce results is an innate characteristic of all living things (Alchian, 1950). In the animal kingdom, performance includes securing a stable source of food or displaying the knowledge, skills, and abilities to defend one's self and others from predators. Natural selection dictates that the better an organism performs the better chance it has at reproducing and passing on its genes to future generations (Darwin, 1859). Thus, people in general may be motivated to increase their performance in order to provide for their basic needs and wants. For example, prior to growing crops, people had to hunt and gather food to survive. When people discovered how to grow crops, people were able to spend more time working towards achieving other goals. With additional time afforded, people were able to develop new technologies such as tools, which made it easier to grow crops and essentially continue to increase performance and achieve progress (Richerson & Boyd, 2008).

From an organizational perspective, performance is important because, just as an organism's performance generally dictates its success in passing on its genes, the perpetuation of an organization is dependent on the performance of its employees. The more top performers an organization can identify, select, develop, and retain, the more likely the organization is going to be successful (Aguinis & O'Boyle, 2014; Aguinis et al., 2016; Boudreau & Ramstad, 2007; Call, Nyberg, & Thatcher, 2015). Changing an organization's view of employee performance as being normally distributed to being exponentially distributed may allow the organization to more easily and accurately identify employees who are most likely to contribute to the organization's success. This is possible because an exponential distribution provides greater differentiation between

top performers than a normal distribution. An exponential distribution is more likely to differentiate a greater number of employees that are in the upper echelons of performance, while a normal distribution would compress many of these employees toward the mid-point (Pulakos & O'Leary, 2011).

Being able to differentiate top performers accurately from the rest of employees gives organizations an opportunity to target and invest in employees who are most likely to help the organization succeed. This affords organizations an opportunity to maximize performance by providing more accurate data in strategic planning. Increased performance and progress helps to ensure an organization's survival. To that end, organizations often provide incentives for employees to perform well and to continue improving their performance. However, not all employees possess the same drive to improve their performance (Latham & Pinder, 2005; Maslach, Schaufeli, & Leiter, 2001). Some employees are satisfied with doing the bare minimum of what is required, or even less. Organizations may let these lowest performing employees go or find some way to motivate them to perform with at least a minimum performance requirement to hold onto a position. As a result, the lower end of job performance distributions should become almost non-existent, further supporting the argument to consider the possibility of exponential distributions of job performance.

The goal of maximizing employee performance and retaining top employees appears to be imperative for organizational success (Becker, Huselid, Pickus, & Spratt, 1997; Combs, Liu, Hall, & Ketchen, 2006; Huselid, 1995; Pfeffer, 1998). Furthermore, there appears to be a relationship between organizations' abilities to achieve this goal and the evolution of technology. As technology advances, there may be greater opportunities

for differentiation between employee levels of performance. Over time, more technologies have become available to employees to help them learn and perform their jobs more effectively and efficiently (Devaraj & Kohli, 2003; Goodhue & Thompson, 1995). When desktop computers first entered organizations, their functionality and purpose was limited. Over time, the number of software applications available to organizations has become almost limitless with each organization leveraging a unique mix of applications and platforms. Theoretically, this means that the greater proficiency an employee possesses for applications and platforms required for a given job, the better the employee should be able to perform. When jobs require proficiency in only one application, there should be less differentiation between employees' performance than when a job requires proficiency in multiple applications.

For example, assume proficiency is measured on a one-to-five scale similar to the performance examples provided in Chapter One. When proficiency is only required on one application, employees can only be differentiated on the single one-to-five scale as related to the single application. However, if proficiency is required on two applications then proficiency can be differentiated for each application on separate one-to-five scales. This results in two separate one-to-five ratings that can be summed to provide an overall score from two through ten. The more behaviors evaluated to assess performance, the greater the amount of true variance between employees may exist and may be measured. As work has evolved from primarily physical labor to being more service oriented, a greater number of skills may be required for many jobs, creating an opportunity for exceptional employees to substantially outperform the rest. It may be reasonable to assume that the majority of employees will possess at least the minimum required skills

to perform a job and a decreasing number of employees will possess an increased level of proficiency of additional skills. This could result in the majority of employees performing at a similar basic level where the lowest performers are forced to exit an organization or perform at a basic level of performance and a few employees perform at a much higher level. Within I-O Psychology, this perspective challenges historical views of job performance established around World War I (WWI; Austin & Villanova, 1992).

For I-O Psychologists, the focus on job performance came about during the turn of the nineteenth century as a result of the Industrial Revolution and WWI (e.g., Bingham, 1926; Kornhauser & Kingsbury, 1924; Link, 1918; Scott, 1917). During WWI, the demand for laborers and soldiers was greater than ever before. This demand provided an opportunity for emerging sciences focusing on human work and performance to catalyze (e.g., Hull, 1928; Kornhauser, 1922; Munsterberg, 1913; Parsons, 1909). As the demand for laborers and soldiers increased, demand to select the *best* laborers and soldiers also increased. Thus, early work researchers sought to find indicators to identify people who would produce the highest levels of job performance (e.g., Link, 1918; Strong, 1918; Yoakum & Yerkes, 1920). In order to find these indicators and select applicants with the greatest potential, job performance had to be quantified (Scott, 1917).

From this demand, performance criteria were created and a refinement process between job performance and indicators of job performance (e.g., intelligence and personality) was formed. Researchers and practitioners began by seeking the best predictors of job performance and then seeking out the best measures of job performance to increase the predictive validity of their indicators (Bingham & Davis, 1924; Fryer, 1922; Viteles, 1925). Work researchers would then return to their indicators and attempt

to increase their predictive validity further, thus starting the process over, establishing the refinement process that has been happening for well over the last century (Austin & Villanova, 1992).

In the Beginning: Measuring Outcomes

In their report on the United States Army's development of selection tests, Yoakum and Yerkes (1920) mentioned that as the United States entered World War I, it had to rapidly build a military force much larger than ever before. The authors went on to explain that military selection and classification systems at that time were not designed to handle the large influx of new soldiers required to fight in the war and, as a result, there was a need to change selection and classification systems. One of the authors, Robert Yerkes, was one of a team of psychologists commissioned by the United States to revamp this process. According to the report, Yerkes developed a new system intended to identify the potential of each military candidate and accurately place him or her into a military position that would best fit the person's innate capabilities.

To do this, Yerkes developed two selection measures known as the Army Alpha and Army Beta. The Army Alpha test was a cognitive battery of tests administered to people who could read, whereas the Army Beta was a similar cognitive battery administered to people who were illiterate. Unfortunately, there were serious problems with these two batteries. The batteries did not accurately assess the abilities of enlisted soldiers, but severely discriminated between soldiers based on race (Brigham, 1930). After the war, the use of these batteries in the Army was abandoned in search of better selection tests. However, immediately following WWI, the private sector as well as United States government agencies did see value in the use of intelligence testing and

began using new refined intelligence tests as a means for selecting employees (Fryer, 1922, 1935). Unfortunately, although these newer intelligence tests were more valid they still possessed the limitation of producing adverse racial impact in selection (Gould, 1984).

From about nineteen-twenty until well into the nineteen-forties, and even to present day, much focus has been placed on predictors, such as traits, to select employees (Ryan & Ployhart, 2014). The goal of industry has been to increase job performance and ultimately production through the use of selection methods. Outcomes of performance were measured as a proxy for job performance in order to validate predictors of performance. As a result, outcomes included criteria selected based on *ease of measurement*, not *job performance* (Jenkins, 1946). Measurement of outcomes could be considered among the earliest forms of objective performance *criteria*.

World War I and the large-scale increase in industrial jobs hastened the realization that organizations needed to select the best employees in order to increase performance. This, in part, is the benefit Yerkes' initial selection batteries provided. Yerkes' selection batteries increased enthusiasm for testing and triggered the realization of the need for better selection systems that could increase job performance (Kingsbury, 1923). This realization may also have been attributable in part to the large-scale increase in mass production factory jobs in the United States. A desire to increase job performance and increase production was likely the motivation, as many new jobs were created during this time¹.

¹ As an aside, it is from the creation of these new jobs that I-O psychologists got their name. Many of these jobs were industrial or factory work, which is where the term Industrial Psychologist was initially derived. It was not until much later that *organizational* was added to the title.

In summary, during the turn of the nineteenth century through WWI there was a push by the industrialist mindset to focus on maximizing job performance (Katzell & Austin, 1992). During this time there was an initial focus on developing ways to predict job performance measured as the number of outputs (Katzell & Austin, 1992). For example, job performance measured objectively as the number of bolts tightened per hour or the number of completed cars produced. The research of this time suggests that the mindset was to improve selection systems so that the candidates with the greatest potential to perform well could be selected (Austin & Villanova, 1992). Yet, evidence was still lacking as to the best way to measure job performance and what types of indicators were important to try to predict. Emphasis was on selection systems used to select employees, but not on improving the measurement of job performance. This made it difficult to validate the impact of new selection systems. There was a gap between the advancement of selection systems and the advancement of how to accurately measure job performance. Researchers could improve selection systems, but were unable to accurately demonstrate the impact of new selection systems because there had not been emphasis placed on improving the measurement of the outcome of interest, job performance. As a result, researchers began to realize that before they could demonstrate the actual impact of new selection systems on job performance they had to also invest in measuring and better understanding job performance. This meant an initial shift from focusing primarily on indicators that could be used for selecting employees to also attempting to understand better, conceptually, what constitutes job performance.

Conceptual Advancement: Actual Job Performance versus Criteria

Before discussing theoretical models of job performance in the next section, there is an important conceptual distinction between job performance and measures of job performance. This section will explain the difference between actual job performance and what is measured as job performance. This section will also present historical events that have resulted in the necessity of this distinction.

Ronan and Prien (1971) define job performance as a latent construct, meaning that it is intangible and not *directly* measurable. Criteria, they say, are quantitatively measured manifestations or indications of latent job performance. For example, a manager may have a list of behaviors that an employee should be performing on the job. A manager could then rate the employee on how well the employee is performing each behavior. Hence, the ratings provided by the manager are signals of an employee's job performance. The distinction is that job performance is "pure," whereas criteria are merely imperfect indications of job performance. Furthermore, criteria can be both objective as well as subjective. Objective measures of job performance may focus on outcomes such as bolts tightened per hour. Subjective measures, on the other hand, refer to measurement that requires judgments about things such as behaviors, like manager's ratings used in previous examples.

Unfortunately, both objective and subjective criteria are inheritably plagued with error, which can distort what organizations measure as job performance. Both objective and subjective criteria also tend to lack completeness in their measurement; in other words, there are many factors that are not measurable that could contribute to an employee's job performance. It is also not feasible to control all sources of error. When

using subjective measures, error may be introduced in many forms. For example, similar-to-me bias is when a rater provides higher ratings to employees that are similar to him or herself. As a result, criteria are likely to capture only some of the many factors that comprise job performance as well as various sources of error.

Perhaps more problematic, it is likely that some of the factors measured as job performance are not necessarily related to job performance at all (Landy & Farr, 1980). For instance, what is being measured as performance can be influenced by environmental factors (Murphy, 2008). In a factory, this could be due to many influences, such as differences in the equipment used by employees or differences in the rate at which employees receive materials from others in a factory line preceding them. If two employees are equally able to perform, but one has a newer machine, then one employee may produce more widgets per hour than another employee due to this environmental difference. In this scenario, equipment would be considered an environmental factor.

The issues of dealing with incomplete job performance criteria and their associated error can also be placed into historical context. In the early part of the 1900s, researchers and practitioners were at a crossroads where they had to decide for the first time how to measure job performance. The first solution to this problem was to simply count tangible outcomes such as number of bolts tightened per hour or number of buttons sewn on a shirt per hour. From the turn of the nineteenth century through WWI, this approach appeared to work. However, research and practice of that time focused narrowly on the industrial worker (Katzell & Austin, 1992). More recent research has demonstrated that measuring *outcomes* opposed to *processes* (i.e., behaviors) may be a deficient method for evaluating employee job performance in most jobs and situations where additional factors

outside the immediate control of an employee are at play (Aguinis, 2013; Beck et al., 2014). As research and practice on measuring and predicting performance continued, huge deficits have been identified within the practice of only using objective indicators of performance.

While outcomes as an indicator of job performance has been the accepted method of performance measurement, over time, through arguments over what constitutes job performance, there has been a gradual switch toward evaluating employee behavior as an indicator of job performance (Austin & Villanova, 1992). Research has found that the use of subjective measures in many situations may be better than objective measures (Aguinis, 2013; Campbell & Campbell, 1988). Although the environment can influence both objective and subjective ratings, behaviors, not outcomes, tend to be more under the control of employees. Employees may demonstrate all of the behaviors required to have optimal performance and the associated outcomes may not always be ideal. However, the use of subjective measures to measure job performance is not addressed until before the start of WWII, when researchers like Bingham (1926) and Viteles (1932) began to challenge conceptualizations of job performance. For instance, Bingham (1926) and Viteles (1932) presented evidence that the criteria being used by organizations to measure job performance did not align completely with the standards employees thought should be used to evaluate their performance. In essence, employees did not believe that the criteria being used to measure performance was face valid. Face validity refers to whether something looks like it is measuring what it is supposed to be measuring in the eyes of the person being evaluated (Mosier, 1947). Thus, employees did not feel that the criteria used by organizations to evaluate their performance were accurately measuring

performance. The conceptualization of job performance began to advance beyond the mere use of objective outcomes as criteria because of the critiques provided by researchers like Bingham (1926) and Viteles (1932).

In sum, there is error in the measurement of objective outcomes and subjective measures of job performance. The reason we continue to use criteria, despite the known deficiencies, is that they currently represent the best approach available to measure job performance (Pulakos et al., 2015). Viswesvaran, Ones, and Schmidt (1996) would argue that statistically, error can be accounted for and corrected. However, Murphy (2008) argues that this is an oversimplified and inadequate view of dealing with error in job performance measurement.

Theoretical Criterion Advancements

The advancement of criteria used to measure job performance did not see large strides until the start of World War II, shortly after the critiques of Bingham (1926) and Viteles (1932), when the U.S. government again invested a great deal of resources into developing better measures of job performance such as combat performance (Marquis, 1944). During this time, researchers such as Toops (1944) and Thorndike (1949) also began to provide some of the first theoretical models of job performance.

Toops (1944) helped lay the groundwork that would make later researchers work, such as Thorndike's theoretical contribution, possible. Toops established that although a unidimensional criterion (i.e., a sole criterion that captures all aspects of job performance) is desirable, all criteria are likely influenced by many sub-criteria. Similar to Thorndike's (1949) theory, which argues that a nearly infinite number of behaviors across time comprise an ultimate criterion, Toops identified a myriad of factors that could

affect a unidimensional performance criterion. Examples of these sub-criteria include wages, production, quality of work, rate of acquisition of new skills, supervisory judgments, knowledge, job tenure, supervisory and leadership ability, job satisfaction, and amount of supervision required. It is from these initially identified sub-criteria that later researchers more easily understood the complexity and unattainable nature of an ultimate criterion.

Building upon the work of Toops, Thorndike (1949) proposed an *ultimate* conceptual criterion and contrasted it with an *intermediate* criterion. His emphasis was not on how employees are rated for their performance, but instead with how employees actually perform. Thorndike's ultimate criterion accounts for every possible factor that influences job performance (noise in the work environment, job tenure, motivation, job satisfaction, knowledge, every behavior performed by the employee, etc.). The ultimate criterion would also account for an individual's performance throughout that person's entire tenure in a specific job. According to Thorndike, to identify an ultimate criterion, a group of the most qualified subject matter experts (SMEs) would have to provide every possible objective, related behavior, and weights for each behavior for a particular job and come to unanimous agreement. As Thorndike details, this process alone has the potential to be drawn-out, laborious, and impractical to the needs of organizations.

Thorndike's (1949) intermediate criterion, on the other hand, is not as comprehensive as an ultimate criterion but is more feasible to attain. Intermediate criteria are intended to capture as much of the conceptual space of the ultimate criterion as is reasonable. Although, Thorndike does not elaborate on what he means by "reasonable," a literal interpretation may be identifying the point at which measuring additional

behaviors no longer adds to the incremental validity of the measurement of job performance. In other words, it is that point where there no longer is a difference between job performance ratings that include X number of behaviors versus X+1 number of behaviors. For example, if there is no difference in job performance ratings that comprise five behaviors versus six behaviors, but there is a difference between measuring four behaviors and five behaviors, the intermediate criteria may be best conceptualized as comprising only five behaviors. Furthermore, it may be possible to capture a great deal of ultimate criterion space. However, just because something is possible does not make it practical. In turn, “reasonable” may also refer to a non-scientifically derived judgment that needs to be made by organizations about the point at which the organization thinks it will cost them more money and time than they deem necessary to adequately measure performance.

Thorndike (1949) also specifies that an intermediate criterion captures part of the ultimate criterion and includes measurement of a behavioral component and a time component. This means that a *behavior* must be measured at multiple points *across time*. As a contrast between an ultimate and intermediate criterion, consider the work of a heart surgeon. Comprising an ultimate criterion, over the course of a surgeon’s entire career, there are many possible types of heart-related surgeries that the surgeon must be able to perform, a number of complications that might be encountered during surgery, skills and abilities to work as part of an interdependent team, as well as many additional factors. Using an ultimate criterion that comprises all factors of performance a surgeon could experience would not be possible. Conversely, it would not be wise to evaluate the performance of the heart surgeon based solely on a performance during only one type of

heart surgery at one point in time. Yet, an intermediate criterion provides a sensible compromise between these two alternatives. For instance, one could comprise a list of the most common heart surgeries and most common complications and measure a surgeon's performance during these surgeries within a specific period, such as within a one-year span. Measurement of these factors would meet the requirements of an intermediate criterion.

According to Thorndike (1949), there is a third type of criterion, which he refers to as an immediate criterion. It is easiest to understand this type of criterion as an intermediate criterion lacking measurement across time. An immediate criterion would include the measurement of multiple behaviors related to the theoretical ultimate criterion, but these behaviors would be measured only at one point in time. This would be comparable to measuring a surgeon's performance based on multiple behaviors demonstrated during only one surgery. In terms of quality, the immediate criterion may be considered acceptable in some situations, the intermediate criterion better, and the ultimate criterion the best, albeit unattainable.

Thorndike (1949) also challenged developers of intermediate criterion to think logically and rationally when choosing a set of behaviors and a time interval. Considerations should include the ease and feasibility of measuring the behavior and the expected interrelationships between behaviors. Ideally, related behaviors should correlate strongly and unrelated behaviors should not correlate. This recommendation is also observable during the 1960's when measurement of job performance becomes the emphasis of research. Research focused on measurement of job performance began by not only improving the validity and reliability of measuring job performance, but also

focused on constructing measures that were logical and easy for organizations to implement and leverage. Thorndike's theoretical conceptualization of performance criteria and his related recommendations are among the most influential work in this area. While these practices may seem to be lacking scientific rigor by today's standards, Thorndike's proposed approach to developing performance criteria is consistent with the standards of present-day criteria development (Hoffman et al., 2012). Present-day approaches to criteria development will also be discussed further when reviewing methodological advancements in measuring job performance.

Progressing through time and approaching the methodological revolution of job performance, Brogden and Taylor (1950) advanced understanding and differentiation between Thorndike's ultimate criterion and what is *actually* measured. Brogden and Taylor proposed that job performance could be comprised of two main parts, actual performance and theoretical performance. *Actual performance* is everything measured or operationalized as job performance. This is conceptually similar to Thorndike's intermediate criterion. *Theoretical performance* is conceptually similar to Thorndike's ultimate criterion. The unique contribution that Brogden and Taylor provided is in how they explained the relationship between ultimate and actual criteria.

In Figure 3, the space to the far left that is labeled criterion deficiency represents the portion of the ultimate criterion that is not measured by the actual criterion. Criterion deficiency is everything related to the ultimate criterion that is not being measured. As a brief example, if a heart surgeon's performance were measured based on one procedure, performance in all of the procedures not measured would fill the criterion deficiency space.

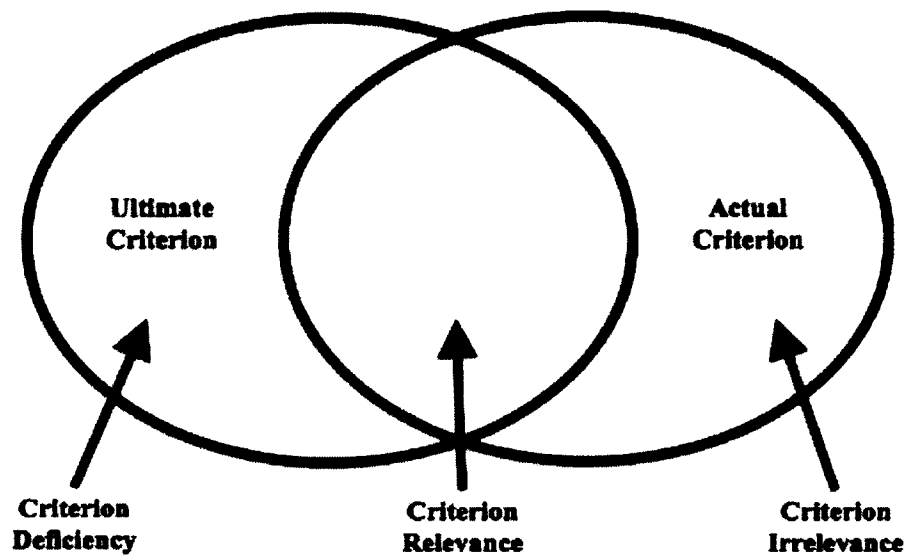


Figure 3. Relationship Between Ultimate and Actual Criterion.

The space on the far right in Figure 3 is criterion irrelevance, sometimes referred to as criterion contamination. It represents the portion of actual criterion measured but not related to the ultimate criterion. Criterion irrelevance represents the unintended measurement of things not related to performance. In measurements of job performance, this represents *error*. The vast amount of job performance research during the 1960's focused on the methodology of measuring job performance as well as specifically reducing error (criterion irrelevance) in job performance measurement (Hoffman et al., 2012; Murphy, 2008). An example of error would be bias introduced by a rater, such as Halo Error. Halo Error occurs when a rater observes an employee performing one behavior well and then provides positive ratings for all behaviors (Balzer & Sulsky, 1992). Observing an employee performing one behavior well can act as a lens to interpret other behaviors. Bias such as Halo Error can act as a contaminant (criterion irrelevance), effectively distorting ratings (actual criteria).

The overlapping space in the middle of Figure 3 is criterion relevance and represents how much of what is measured is actually related to job performance. Criterion relevance is the part of the ultimate criterion that is actually measured. It is easy to conceptually pull apart criterion relevance and criterion irrelevance, but in practice, this is much more difficult. For instance, measuring a surgeon's success based on the execution of certain techniques during surgery may fall into the criterion relevance category, but the performance ratings made by the person observing the surgery and evaluating the surgeon's execution of certain techniques could be impacted by many different types of error. For example, if the rater recently watched a different surgeon misuse techniques that resulted in a fatality, then other surgeons also receiving ratings may appear to perform a great deal better in comparison, even if their performance is only a little bit better. This is why the actual criterion is comprised of both criterion relevance and irrelevance.

Brogden and Taylor (1950) seemed to agree with Thorndike about his ultimate criterion, but as Figure 3 demonstrates, both sets of authors likely differed on their views of actual performance criteria. Thorndike argued that intermediate or an actual criterion captures a portion of an ultimate criterion. Brogden and Taylor seemed to agree that an actual criterion is only measuring a portion of an ultimate criterion, but that an actual criterion is also measuring irrelevant criteria that are not related to the ultimate criterion. Anything measured as part of an actual criterion that is not part of the ultimate criterion is error.

The primary point is that measures of job performance, because of how they are operationalized and measured, will likely introduce some amount of error. This applies to

all forms of measurement (Borman & Motowidlo, 1997; Heneman, Moore, & Wexley, 1987; Landy & Farr, 1980; Murphy, 2008). It is the role of researchers and practitioners to limit the amount of contamination or error and to increase the validity of the measurement. When trying to measure job performance, it is the duty of researchers and practitioners to ensure that job performance is what is actually being measured.

As a result of theoretical contributions made from Toops (1944), Thorndike (1949), and Brogden and Taylor (1950), among others (e.g., Bolanovich, 1946; Creager & Harding, 1958; Ewart, Seashore, & Tiffin, 1941; Grant, 1955; Wherry, 1952), it became clear to researchers and practitioners alike that more emphasis needed to be placed on improving the measurement of job performance. In turn, the coming decades from the 1960's forward are characterized by an emphasis on improving how practitioners develop, implement, and measure job performance in organizations. In decades following methodological advancements of measuring job performance, the emphasis evolves to focusing on further reducing error introduced into the measurement of job performance by individual raters of performance.

Improving Methods and Tools for Measuring Job Performance

Thus far on the historical journey of job performance from the nineteenth century up to the early 1960's, there has been emphasis placed on maximizing outputs as well as on trying to predict job performance by developing psychometrically sound selection tests (Schmidt & Hunter, 1998). To validate and demonstrate the value of job performance predictors, job performance had to be measured. This led to conceptual explorations of job performance and an interest in determining what should be measured and considered job performance. Initial conceptual explorations resulted in vast

conceptual improvements around what job performance is and what it is not. Yet, up to the early 1960's there had not been much advancement related to how job performance is measured and the tools that were available to raters of job performance, especially in comparison to the rigor subjected to predictive employment tests. In turn, predictive validities of employment tests remained fairly low (Schmidt & Hunter, 1998). However, now armed with a better theoretical understanding of job performance, I-O Psychologists were able to start to direct focus on improving the measurement of job performance.

As research up to the early 1960's had worked towards disentangling conceptual issues surrounding job performance, I-O Psychologists in the early 1960's onward began to place more emphasis on improving the measurement of job performance. For instance, Dunnette (1963) began arguing that current measures of that time were unreliable and not valid measures of performance. Measures of job performance up to that point in time provided inconsistent ratings and were very contaminated with error (i.e., considerations unrelated to job performance or irrelevant criteria). Other researchers during this time such as Campbell, Dunnette, Lawler, and Weick (1970) also begin to shift the focus of what measures of job performance should be capturing, away from outcomes as indicators of performance and more toward behaviors as indicators of performance. Campbell and colleagues argued that job relevant behaviors were less contaminated with irrelevant criteria, such as environmental factors, than more common criteria of the time, such as performance outcomes. Rating performance based on behavior, however, presented its own challenges. For example, to provide ratings of job performance based on behavior, raters were required, which introduced human error. Additional researchers also began to search for alternative ways to measure job performance, beyond measuring

mere performance outcomes, which included attending to environmental and other extraneous factors.

Remnants of early research that began the transition period from disentangling conceptual issues to focusing on methodological concerns trace back in time to Wherry (1952). Wherry drew from previous research done in psychometrics and cognitive psychology to develop a systematic procedure for rating job performance. This system involved observing employees, parsing observations, and making quantitative ratings. Where previous research had focused on outcomes as objective indicators of job performance, Wherry attempted to take subjective observations of behavior and methodically make them more objective. Unfortunately, Wherry's ideas may have been ahead of his time. While his ideas were novel, they gained little traction in advancing this line of research during the 1950's.

Research emphasizing the assessment of behavior over outcomes did not gain traction until approximately a decade later, in the 1960's, when P. C. Smith and Kendall (1963) introduced behaviorally anchored rating scales (BARS). BARS are used to make objective ratings of job performance based on *behaviors* performed by employees instead of making ratings based on outcomes. BARS are job performance ratings scales that are comprised of behavioral examples of varying levels of job performance. Each BARS would typically include three to five behavioral examples of performance related to a specific dimension of a job, ranging from poor to excellent performance. Raters select the behavioral example that best resembles the employee's actual behavior. Seminal research by P. C. Smith and Kendall on BARS essentially prompted an entirely new vein of job performance research. Thus, Wherry (1952) laid the groundwork for P. C. Smith and

Kendall development of an easy-to-replicate process for developing a tool that could be utilized by practitioners.

BARS have made a large impact on the science and practice of assessing job performance because BARS are prescriptive and because they drew on other common practices that were already well established at the time (e.g., job analysis; Flanagan, 1954; Levine, Ash, Hall, & Sistrunk, 1983). The method for developing BARS, as prescribed by P. C. Smith and Kendall (1963), built directly on common job analysis practices of that time, interviews with subject matter experts (SMEs). SMEs are people who are knowledgeable about the job of interest. SMEs are typically job incumbents who have experience in a given job or managers of the job of interest.

Given that BARS draw on other common practices, they are considered fairly convenient and intuitive to use in practice. This is demonstrated by how quickly managers were to adopt BARS in practice and by BARS current wide spread use today (Debnath, Lee, & Tandon, 2015). The convenience and the clear relevance of BARS give them many advantages over the use of outcomes for measuring job performance. For instance, one advantage BARS are believed to have over the use of outcomes when measuring job performance is that BARS are believed to possess less criterion irrelevance than outcomes (Landy & Farr, 1983; Murphy & Cleveland, 1995). Secondly, BARS focus on behaviors that employees can demonstrate. Even if an employee demonstrates good behavior, the outcome may not always reflect behavior when measuring outcomes (Aguinis, 2013). As a result, BARS are more likely to be readily accepted by employees that are receiving ratings on BARS because they focus on measuring behavior that is in the control of employees, as opposed to outcomes. The bottom line is that BARS are

perceived as more fair to employees as an indicator of job performance (Dickinson & Zellinger, 1980).

The observation and measurement of behavior is also more applicable and easily implemented across a larger variety of jobs (e.g., industrial and office jobs) than the measurement of outcomes. Until the 1950's, the United States had a fairly industrialized workforce. During the 1950's and early 1960's, as the United States' economy strengthened post-WWII, more office jobs were created. This resulted in a need for different method for measuring performance. Office jobs, for instance, do not necessarily have clear and easily measurable performance outcomes. However, all jobs require employees to demonstrate behaviors that can be classified by SMEs as reflecting various levels of performance.

Following the creation of BARS, similar graphical rating scales were also proposed, such as Behavioral Observation Scales (BOS; Latham & Wexley, 1977). BOS are similar to BARS in the sense that both rely on behavioral descriptions to judge performance. BARS have three or more incremental behavioral statements related to one behavior and performance raters choose one behavioral description that best describes an employees' performance. BOS, on the other hand, may have three similar behavioral descriptions that ask a rater to provide one through five ratings for each. For example, managers would be presented with a behavioral description and would have to rate an employee based on how often they display the behavior, one 'never' through five 'always'. BOS are likely to use multiple items that all measure similar behaviors. This results in multiple numerical ratings for similar behaviors, which allows for increased certainty in the reliability of measurements.

To summarize, the main difference between BARS and BOS is that BOS present one descriptive behavioral anchor and require a numerical rating related to that behavior, whereas BARS present multiple descriptive and incremental behavioral anchors within one item and force raters to choose the best anchor. Research has demonstrated, however, that despite the differences between BARS and BOS, no single rating format results in superior ratings (Jacobs, Kafry, & Zedeck, 1980).

In general, BARS, BOS and other similar scales tend to result in similar ratings of job performance irrespective of specific format (Greene, Bernardin, & Abbott, 1985; Jacobs et al., 1980). Since their inception, behavioral scales have continued to grow in popularity among managers. However, many different methods have been developed over approximately the last half-century. The most common methods still in use today will be discussed in the next section. Each of the methods discussed has the potential to provide different challenges related to exponential distributions of job performance. For the current study, it is important to be aware of the advantages and disadvantages of each behavioral method for assessing job performance in order to understand the challenges of identifying exponential job performance distributions.

Most Common Methods for Measuring Job Performance

There are two main groups of behavioral methods that are typically utilized to measure job performance. These are comparative and absolute ratings. BARS are a form of absolute rating, meaning that an employees' performance is only measured against himself or herself (Wagner & Goffin, 1997). Other forms of absolute methods include essays or free form text reviews of employee performance provided by managers and critical incidents, which require managers to provide specific examples of especially

effective and ineffective behaviors. Comparative methods, on the other hand, require raters to make judgments on employee performance relative to other employees (Wagner & Goffin, 1997). An example of a comparative method is the simple rank order method. Bernardin & Wiatrowski (2013) describe this method as one that requires raters to generate a list of all employees, identify the best performer, and rank that employee as number one. Once an employee receives a ranking, he or she is removed from the list, and the rater identifies from the list of remaining employees, the next top performer and so on.

In order to discuss the advantages and disadvantages of comparative and absolute methods and how they may affect the shape of a performance distribution, it is important to have a basic understanding of various types of absolute and comparative methods. Having a basic understanding will allow for critical evaluation of the advantages and disadvantages of each method. In turn, the following sections list different types of absolute and comparative methods and provide basic descriptions and examples of each.

Absolute Methods

Essays. Essays are likely an uncommon form of assessing job performance. They typically consist of a narrative written by an employees' manager about ones' job performance (B. N. Smith, Hornsby, & Shirmeyer, 1996). Essays provide deep insight into an employee's job performance, but require a lot of time on the part of managers (Huber, 1983). As a result, they lack practicality. The other main disadvantage of essays is that they are not standardized, which makes it very difficult to make comparisons between employees (Brutus, 2010). Being able to easily compare employees is important when making employment decisions (e.g., deciding who should receive a promotion or a

raise relative to other employees). Although not an essay, many other types of numerical ratings that are provided by managers may be accompanied by a few additional free form sentences about an employee's job performance.

Critical incidents. Critical incidents are also very time consuming for managers. According to Flanagan (1954), the critical incidents method requires managers to observe the behavior of employees and record across time specific behaviors they see as being especially effective or ineffective. These examples are then used to provide feedback to employees. Critical incidents on their own can be laborious for managers and can also make it difficult to make comparisons between employees, which is something managers typically need to do when identifying employees for promotion or rewards. Instead of leveraging critical incidents as a method for providing ongoing performance ratings, they may be done once by multiple SMEs and then used to create BARS or BOS (Bernardin & Smith, 1981). In terms of performance management, critical incidents and essays are not very practical to perform on a regular basis, but can result in very rich descriptive behavioral examples that can be used to help facilitate performance feedback to employees and subsequently be invaluable to the development of employees.

Comparative methods. In opposition to absolute ratings of job performance, comparative measures, as the name implies, relies on relative comparisons to be made between employees to determine the performance of each individual employee (Wiese & Buckley, 1998). Examples of these systems include simple rank order, alternating rank order, paired comparisons, relative comparisons, and forced distribution.

Simple rank order. According to Aguinis (2013), the simple rank order method requires managers to sit down with a list of their employees and simply rank order them

from best to worst. A manager would begin by identifying the top performing employee and ranking that employee as first. Then the manager would look through the remaining list of employees and identify the next top performer. That employee would receive a two. This process would be repeated until all employees have been rated.

Alternating rank order. The alternating rank order method is very similar to the simple rank order method. According to Miner (1988), to use this method a manager would again begin with a list of employees. Then the manager would identify the top performer in the list. Then, instead of identifying the next best performer, the lowest performing employee is identified and put at the bottom of the list. Once the manager has identified the best performer and lowest performer, the manager identifies the second best performer and the second lowest performer. This process continues until the manager has ranked all employees on the list.

Paired comparisons. The paired comparisons method was presented by Siegel (1982) and takes a unique approach relative to the simple rank order and alternating rank order methods. Based on the method used by Siegel, a manager would need to write out every possible comparison between all employees. The manager then reviews all possible comparisons and choosing the top performer in each comparison. In every comparison, the top performer should receive a score of one while the other employee receives a score of zero. The score of every employee across all comparisons is then summed. The higher the total score of an employee, the higher the ranking.

Relative percentile. In general, the relative percentile method asks managers to consider the job performance of all employees simultaneously (Goffin, Jelley, Powell, & Johnston, 2009). Ideally, the manager would be able to identify the employee that is

directly in the middle of the job performance distribution relative to all other employees (Goffin, Gellatly, Paunonen, Jackson, & Meyer, 1996). Once an employee is identified as a midpoint, the manager is to assume that fifty percent of remaining employees should have better performance and fifty percent of remaining employees should have lower performance (Goffin et al., 1996). The manager then could repeat this process for the top fifty percent of employees and the bottom fifty percent of employees (Goffin et al., 2009). This process could be repeated until all employees have been relatively assigned (Goffin et al., 1996).

Forced distribution. The final comparative method of assigning job performance ratings is the forced distribution method. This method typically leverages the assumptions of normal distributions (Blume, Baldwin, & Rubin, 2009; Stewart, Gruys, & Storm, 2010). However, this method could also assume a distribution of any shape. For example, using a five-point rating scales, managers are instructed that sixty percent of employees must receive ratings of three, eighteen percent of employees must receive ratings of two, eighteen percent must receive ratings of four, two percent must receive ratings of one and two percent must receive ratings of five. This method is not exclusive, meaning that it does not need to be the only method utilized, it can be paired with other methods (Chattopadhyay & Ghosh, 2012). For example, forced distribution instructions could be applied to the instructions provided with a BARS or BOS. In such instances, managers may start by providing an absolute rating using BARS or BOS and then attempt to force each employee into a pre-specified distribution. While placing employees into the pre-specified distribution, managers may have to go back and change their initial ratings for some employees so that the requirements of the distribution can be met.

Comparing Absolute and Comparative Methods

Understanding the advantages and disadvantages of each method makes it easier to understand the impact that each rating method can potentially have on resulting distributions of job performance. Some advantages of comparative methods include that they are fairly straightforward and easy to explain to managers as well as to employees. Comparative methods also make it clear how an employee's performance relates to other employees' performance (Goffin et al., 1996). This makes it easy to identify and justify which employees are more deserving of pay increases, bonuses, promotions, etc., based on performance. Another advantage of comparative methods is that they can help control for various rater errors or biases that influence ratings of job performance, and resulting distributions of job performance. Much of the research on rater error and biases occurred in decades following methodological advancements in how job performance is measured.

A commonly noted disadvantage of using comparative systems is that relative to absolute methods, little research has been conducted on them (Goffin et al., 2009). Second, typically when comparative methods are implemented as the sole method for rating job performance they focus only on overall ratings of job performance (Wagner & Goffin, 1997). From research on criterion theory, as discussed previously, job performance is complex and is comprised of many different factors across time. As a result, it may be difficult for raters to provide accurate overall measures of job performance. This is in opposition to absolute methods such as BARS or other graphical rating scales, which typically result in many ratings on multiple types of behaviors (P. C. Smith & Kendall, 1963). However, absolute ratings can also result in only one overall rating of job performance given at a single point in time.

Another drawback to comparative methods is that they result in employee job performance rankings, but not necessarily actual scores, unless used in tandem with an absolute method. This means that the rankings that result from using comparative methods assume equal distance between rankings, although this may not be true (Aguinis, 2013; Murphy, 2008; Stewart et al., 2010). The top performer may be twice as good as the second best performer and the second best performer may be three times as good as the third best performer.

As a result, this means that the distribution of performance may be greatly distorted. As a result, the benefits that exponential distributions of job performance may offer, such as better differentiation between high and low performers, are potentially lost if comparative methods are utilized to measure job performance.

Absolute rating methods, on the other hand, do offer the unique advantage of not being bound necessarily by the constraints of a normal distribution or some other prescribed distribution. When leveraging absolute ratings of job performance, each job incumbent receives a rating of job performance, which can be plotted. This allows underlying distributions of job performance to be evaluated. Absolute ratings are able to represent any type of distribution. Therefore, if job performance does better represent an exponential distribution rather than a normal distribution, absolute ratings are more likely to provide such evidence beyond what comparative methods would be able to demonstrate.

In addition to the *unique* advantages and disadvantages of comparative and absolute methods already discussed, absolute and comparative methods also share a common disadvantage. Both methods rely on subjective evaluations or ratings by design. While

the limitations of subjective methods were alluded to earlier, a more thorough discussion of the potential errors/biases inherent in subjective evaluations is needed. The key takeaway is that job performance is complex and any method used to make ratings of job performance has its own set of issues. No method for rating job performance is without error. It is important to remember that all job performance ratings are only indicators of actual job performance plus criteria irrelevance.

Error Introduced by Raters

Rater error is error introduced to the measurement of job performance by the person providing ratings. Research on rater error initially became very popular in the 1980s as a direct result of a scathing critique of job performance rating methods provided by Landy and Farr (1980). Landy and Farr argued that no method for measuring job performance was accurate because they fail to account for a multitude of additional factors that may influence ratings of job performance. The factors discussed by the authors included the cognitive processes used by raters while providing ratings of job performance. These cognitive processes often result in error in the measurement of job performance. There are many kinds of error that raters may inadvertently introduce to the rating process, such as similar-to-me errors, contrast errors, leniency errors, central tendency errors, severity errors, halo errors, and many others. Neither comparative nor absolute rating methods are void of rater error. Although some methods, such as comparative methods, are more resistant to certain types of error such as leniency, severity, and central tendency errors (Stewart et al., 2010). To varying degrees, rater error can affect any tool that requires a human to make judgments (Landy & Farr, 1980). Furthermore, it is likely that no tool used to measure job performance that requires raters to make judgments is devoid of all

rater error (Austin & Villanova, 1992). As briefly acknowledged above, some rating methods such as the forced distribution method may be less susceptible to certain types of rater errors. However, methods less susceptible to certain types of rater errors do still present their own unique set of challenges. For example, the forced distribution method may only be successful if the prescribed distribution actually represents a true performance distribution (Boyle, 2001; Murphy, 2008; Scullen, Bergey, & Aiman-Smith, 2005; Stewart et al., 2010).

Most Common Rater Errors

Similar-to-me error. Similar-to-me errors occur when managers give higher job performance ratings to employees that they view to be more like themselves (Latham, Wexley, & Pursell, 1975). This also means that employees that have less in common with their managers may receive lower job performance ratings (Rand & Wexley, 1975). For example, if a manager enjoys fishing as a hobby, the manager may engage in conversations with select employees that also like fishing and as a result provide employees that like fishing, better performance ratings than employees that do not enjoy fishing. If this happens, error is introduced into the rating process as irrelevant criterion, because something other than job performance is being captured in the job performance rating.

Contrast error. Contrast errors occur when a manager unintentionally makes comparisons between employee's levels of performance, which can affect and magnify differences in job performance ratings (Palmer & Feldman, 2005). As an example, imagine a manager with two employees to rate. Employee A is a star performer and deserves a rating of five out of five. Employee B is an above average performer and

deserves a rating of four out of five. While providing ratings, the manager may rate Employee A first and provides a rating of five. While providing performance ratings for Employee B, the manager unintentionally makes comparisons between Employee B's performance and Employee A's performance. This comparison magnifies the difference in performance between Employee A and Employee B. This results in a rating of three for Employee B. Although Employee B is truly an above average employee and deserves a rating of four out of five, the lower rating is due to a comparison to Employee A and not related to Employee B's true performance.

Contrast errors could also result in an employee receiving a rating of job performance that is higher than his or her actual performance (Maurer & Alexander, 1991). For example, Employee A could be an average performer and Employee B could be a below average performer. If a manager compares Employee A and Employee B side-by-side, Employee A could appear to be a much better performer than Employee B. The result would be Employee A receiving a rating of four when his or her performance more accurately deserves a rating of three.

Leniency error. Leniency errors can occur for many reasons and happen when a manager rates employees very favorably, even when employees do not perform favorably (DeCotiis, 1977; Saal & Landy, 1977). If a manager injects leniency error into his or her ratings it may be negatively skewed. Negatively skewed means that the majority of ratings would be very positive (fours and fives) and only a few ratings would be found on the lower end of the scale (ones and twos). There may be virtually no ratings on the left side of the scale – most employees would receive positive ratings, primarily fours and fives. A common reason this occurs is because managers may want to make themselves

look better, where better performance of subordinates results in a better reflection on the manager providing ratings (Klimoski & Inks, 1990). This type of error may also be easier to identify because, if present, the distribution of job performance may be leptokurtic and negatively skewed (Landy, Farr, Saal, & Freytag, 1976). Although, this does not mean that every distribution of job performance that is leptokurtic and negatively skewed is plagued with leniency error. It is possible that a distribution of job performance with these characteristics also resembles a distribution of job performance measured without leniency error. For instance, a rater may inject leniency error for only or two employee ratings, because some employees do not respond well to negative ratings and the manager providing the ratings may want to avoid conflict. In which case, leniency error may not be discernable by the resulting distribution of job performance. Regardless, leniency error may only be present for one or two employees and not always easily discernable by viewing a distribution of performance ratings (Sharon & Bartlett, 1969).

Severity error. Severity errors are the exact opposite of leniency errors (Saal et al., 1980). This type of error can also be flagged in severe cases once ratings of job performance are reviewed holistically as a graphical distribution. This type of error may also be less common, but results when a manager provides low ratings of job performance even when employees deserve higher ratings (Lunenborg, 2012). At first pass, this may sound very similar to the exponential distribution. However, the case of severity error is distinct for two reasons. First, in extreme instances of severity error, the resulting distribution will have virtually no ratings above the midpoint on the rating scale (Bernardin, LaShells, Smith, & Alvares, 1976; Sharon & Bartlett, 1969); whereas the exponential distribution will. In an exponential distribution that accurately reflects job

performance it would be expected that the majority of employees would be around the mid to low points of the rating scale, virtually no employees on the lowest end of the scale, and a fairly large number of employees above the midpoint of the scale.

Additionally, if severity error is present in a distribution there will likely be very limited variance in ratings, this is something which can and should be tested (Borman & Dunnette, 1975; Saal et al., 1980). The second major difference is that a distribution that results from severity error is the *result of rater error* operating at an individual level (i.e., most individuals are down-rated), whereas an exponential distribution should result from accurate ratings of job performance. Additionally, similar to leniency error, severity error may only occur for a few employees. For instance, if a manager wants to “send a message” to a certain employee that he or she needs to improve performance a manager may provide an underserved exceptionally low rating.

Central tendency error. There is also a third related type of rater error that resides in-between leniency and severity error known as central tendency error. Central tendency errors result when a manager provides most employees with a rating that is equal to the midpoint of the rating scale that is being used and virtually no employees receive ratings above or below the middle point on the rating scale (Aguinis, 2013). This is different from a normal distribution because there are an extreme number of employees that receive a mid-point rating. In the case of central tendency error, even employees that deserve higher or lower ratings would still receive the midpoint rating (Lunenburg, 2012). A normal distribution would still consist if a sufficient and equal number of ratings above and below the midpoint. Furthermore, in less extreme instances of central tendency error a normal distribution could result because the rater may be unsure of an

employee's true performance level and defer to giving that employee a rating equal to the midpoint. Therefore, if a distribution is normally distributed, central tendency error could still be present. Similar to most rater errors that may influence performance ratings, central tendency error could result for many different reasons. Central tendency error may commonly result because of organizational norms or suggested distributions of performance that a manager is expected to follow. This may be an additional justification for organizations to not suggest or attempt to force distributions of job performance. When organizations suggest or recommend that managers provide ratings, which resemble a specific type of distribution, they may introduce additional types of rater error into ratings of job performance.

A potentially effective way to check for any extreme instances of these three errors, severity, leniency, or central tendency is through evaluating the variance associated with the ratings assigned by each rater (Borman & Dunnette, 1975). When there is less within rater variance, one of these types of errors may also be present. Evaluating the amount of variance, however, may only work in the most extreme cases. Leveraging the variance technique to check for extreme instances of these errors would entail calculating the standard deviation, which is a function of variance, for each rater and evaluating the standard deviation in relation to the mean (Borman & Dunnette, 1975; Saal et al., 1980). If there are a few number of raters that appear to have less variance than the majority of raters, the raters with smaller variances may be adding leniency, severity, or central tendency error to his or her ratings.

Halo/Horns error. Halo/Horns errors are two types of different, but closely related errors. Halo and horns errors are also similar to, but distinct from, leniency and severity

errors. Halo error occurs when a rater provides positive ratings on all attributes being measured for an employee because the employee performs one attribute very well (Balzer & Sulsky, 1992). The rater makes the assumption, which results in error because the employee is really good at one thing (King, Hunter, & Schmidt, 1980; Lance, LaPointe, & Stewart, 1994). Similarly, Horns error occurs when a rater attributes an employee's negative performance on one task to being generalizable across all tasks (Turnipseed & Rassuli, 2005). As a result, the rater assumes that the employee performs negatively on all tasks because of how the employee performs one specific task (Turnipseed & Rassuli, 2005). Halo/Horns errors may be more likely to occur when raters only have an opportunity to witness first hand some of an employee's behavior. Halo and Horns errors along with the remaining types of errors that will be discussed are even more difficult to identify statistically.

Negativity error. Negativity error is similar to Horns error. Negativity error occurs when a rater places a greater emphasis on negative behaviors than positive behaviors (Ganzach, 1995; Skowronski & Carlston, 1989). For instance, when providing an overall performance rating for an employee a rater may recall an equal number of positive as well as negative examples of behaviors, but instead of weighting both types of examples equally, the rater weights the negative behaviors higher than the positive behaviors, which results in a lower than deserved rating of job performance for the employee.

Recency error. Recency error occurs when a rater bases all ratings for an employee only on the employee's most recent performance (Latham et al., 1975). For instance, based on criterion theories such as Thorndike's intermediate criterion, if an employee receives a performance review once every year then the rater should provide ratings

based on performance across the entire year. Unfortunately, when recency error occurs, a rater only provides ratings based on, for example, the employee's last month worth of performance or worse last week. As a result, not only is it important to train raters on this potential error, but it is also important to develop job performance evaluation methods that require raters to provide ratings at multiple points in time instead of one overall rating after 12 months (Steiner & Rain, 1989). For instance, to subdue the effects of recency error, raters may be asked to provide ratings using graphical rating scales such as BARS, once a month or more realistically quarterly, over a period of 12 months. Scores from each monthly or quarterly rating can then be aggregated to provide an overall average yearly performance score that is less impacted by recency error. It is conceded that each monthly or quarterly rating may still be impacted by recency error to some extent; however, the more measurements that are made over time the more likely an accurate measurement of job performance will be made by raters.

Primacy error. Primacy error can be understood, in some sense, as the opposite of recency error. Instead of only accounting for the most recent performance of an employee the rater only accounts for the employee's performance during the initial phases of that performance review period (Murphy, Balzer, Lockhart, & Eisenman, 1985). For example, the rater may only recount the employee's performance during the first week or two of the performance-rating period. Again, to help reduce the potential impact of primacy error, rater training as well as multiple measurements of job performance across time is important (Steiner & Rain, 1989).

First impression error. First impression error, is also harder to counter with measurement over time, making rater training again, even more important. First

impression error is one reason that first impressions are so important. First impression error occurs when a rater makes all future performance ratings based on his or her first encounter of the employee (Latham et al., 1975). This type of error can be very hard for an employee to overcome, which is why it is always important to strive to make a positive first impression.

Stereotype error. Stereotype error is when a rater applies any type of stereotype when making performance ratings for employees (Bauer & Baltes, 2002). Common types of stereotypes that can occur include race and gender, but could also include other stereotypes of things such as education (Dipboye, 1985; Ferris, Yates, Gilmore, & Rowland, 1985; Rosen & Jerdee, 1976; Sackett, DuBois, & Noe, 1991; Schwab & Heneman, 1978). A rater may have a negative view of women, believing that women are not invested in their careers and that they are poor performers. On the other hand, a rater may believe that employees with advanced degrees such as PhDs are only academic and not business savvy, which results in low performance. As a result, the rater integrates these personal beliefs into the rating process without regard for whether or not they accurately reflect employees' performance.

Attribution error. Attribution error is one of the most common errors that people in general make about others (Feldman, 1981; Green & Mitchell, 1979; Mitchell & Wood, 1980). An attribution error occurs when a rater attributes an employee's performance directly to that employee's behavior without considering the possibility that the employee's performance may more accurately be the result of additional factors not under the control of the employee receiving the rating (Ilgen & Favero, 1985). As an example, consider an employee that is responsible for producing monthly reports and

consistently completes the reports late. The rater providing performance ratings may automatically assume that the employee is lazy or does not manage well. In reality, the reports are late because the employee relies on data that is provided late from another employee. Thus, the employee being rated is not entirely responsible for producing the reports late; instead, the reports are late because the employee does not have all the resources needed in time to meet the deadline. This is an especially dangerous error because the root of the problem may never be addressed and resolved. If the rater made the correct attribution, then the rater could work with the employee to improve the process instead of making the wrong attribution.

In summation, all of these forms of rating error can have a significant negative impact on providing accurate ratings of job performance. Consequently, these errors can also greatly impact the observed distribution of job performance. Additionally, many of these errors are likely to be present in ratings of job performance simultaneously and to varying degrees (Borman, 1977, 1978; Iqbal, Akbar, & Budhwar, 2015; Viswesvaran et al., 2005). For instance, spillover error and halo error could both be active during rating, resulting in raters making overall positive job performance ratings based on an employee's past job performance, when in fact the employee is no longer performing as well as he or she used to perform. Although rating error can have a great impact on the accuracy of job performance ratings and resulting distributions of job performance, there are also approaches available to help improve rater accuracy. In addition, there are statistical methods and models available to help better identify and understand the impact of errors in ratings of job performance (Feldman, 1986; Landy, Vance, Barnes-Farrell, & Steele, 1980; Murphy, 2008; Schmidt, Viswesvaran, & Ones, 2000; Viswesvaran et al.,

1996; Viswesvaran et al., 2005). For example, the amount of measurement error in ratings of job performance can be estimated and used to provide corrections for ratings of job performance.

Modeling Error in Ratings of Job Performance

There are three main categories of models that can explain error in job performance ratings. These three types of models are comprehensively reviewed by Murphy (2008) and include One-Factor Models, Multi-Factor Models, and Mediated Models. These types of models are very similar to classical test theory, where observed test scores are the result of measurement error plus actual ability. The One-Factor Models posit that performance ratings are the direct result of actual performance when accounting for measurement error. Viswesvaran et al. (2005) conducted a meta-analysis on job performance research spanning the previous century and found that, when accounting for measurement error in predictors of job performance, job performance could represent a unidimensional model. In this case, unidimensional means that all of the various complex components of job performance are related and correlate strongly together into one overall construct of job performance. This finding demonstrates that for most performance measures, both objective and subjective, significant direct relationships between predictors and measures of job performance can be made when accounting for measurement error. This evidence provides direct support for One-Factor Models.

The second category of models is Multi-Factor Models, which treat performance ratings as the result of actual performance, system characteristics, and individual characteristics when accounting for measurement error (DeNisi, Cafferty, & Meglino, 1984; Landy & Farr, 1980; Wherry & Bartlett, 1982). This model corrects some of the

issues mentioned in Landy and Farr's (1980) scathing critique of job performance measures. Landy and Farr argued that there were no good measures of job performance because measures of job performance include for a myriad of additional factors that influence ratings of job performance.

The third category of models, Mediated Models, builds on Multi-Factor Models. Similar to Multi-Factor Models, Mediated Models treat ratings of job performance as the result of actual performance, system characteristics, and individual characteristics when accounting for measurement error (Murphy, 2008; Murphy & Cleveland, 1991, 1995). The difference is that Mediated Models also treat the relationship between these factors and job performance as mediated by rater goals and intentions. While Multi-Factor Models treat rater errors introduced into the measurement of job performance as the result of unintentional cognitive process, Mediated-Models include intentional distortions provided by raters as rater error in addition to unintentional errors. Murphy and Cleveland argue that raters possess unique goals that may influence ratings of job performance in addition to unintentional errors. As an example, consider the impact that different organizational political factors may have on influencing a rater. To elaborate, managers rating subordinates may want to create the image that they are good leaders by artificially inflating the ratings of all subordinates. On the other hand, managers may believe that their subordinates have too much work and that there is a need to hire additional employees. As a result, managers may provide deflated ratings as additional justification that their subordinates have too much work and cannot meet expectations with current head count.

These three theoretical models of job performance ratings take different approaches to the three main factors that may distort ratings of job performance. These factors include measurement error, unintentional rater errors, and intentional rater errors. Although not initially evident, Ones, Viswesvaran, and Schmidt (2008) argue that One-Factor Models are sufficient and do account for both unintentional and intentional sources of rater error. Regardless, the primary issue is still the same, there is always error in the measurement of job performance. Despite this gloomy conclusion, there are methods that can be used to help increase the accuracy of job performance ratings.

Rater Training

Although all types of rating errors can have a strong negative impact on job performance ratings, there are techniques that can help counteract their impact. These techniques do not safeguard entirely against rater error, but they can have a strong positive effect, if implemented appropriately (Woehr & Huffcutt, 1994). The most common method used to increase the reliability and validity of job performance ratings is the use of rater training. Starting in the mid-to-late 1970's and going through the mid-1980's, research emphasized the benefit of providing raters with various types of training such as behavioral observation training, frame-of-reference training, and calibration meetings (Bernardin & Buckley, 1981; Borman & Dunnette, 1975; Ivancevich, 1979; Latham et al., 1975; Pulakos, 1984, 1986; Pulakos & O'Leary, 2011; Roch, Woehr, Mishra, & Kieszczynska, 2012). These types of training may result in more accurate measurement of job performance, which could mean more accurate differentiation of employee performance and more accurate distributions of job performance.

Behavioral observation training. Behavioral observation training teaches raters how to evaluate employee behavior (Noonan & Sulsky, 2001). Recall, many of the most common methods for evaluating job performance, such as behaviorally anchored rating scales, require managers to make ratings based on rater's observations of employee's job related behaviors. These types of rating methods are most likely to be impacted by unintentional types of error that are made while observing employees (McIntyre, Smith, & Hassett, 1984). Behavioral observation training aims to decrease the impact of unintentional rating errors by teaching raters how to observe, store, and recall information about employee performance (Latham et al., 1975). This means teaching raters about the type of behaviors employees are most likely to display (Noonan & Sulsky, 2001). One approach to observing employee behavior is the critical incidents technique, used in the development of BARS or a BOS (Pulakos, 1986). This type of training teaches raters how to watch for these behaviors and how to properly take notes about observed behaviors so that notes can be referred to later on during formal rating processes. Behavioral observation training also typically provides raters with guidance on how frequently to take notes on behaviors (Thornton & Zorich, 1980). This is important because formal job performance evaluations or ratings may only be administered once or twice a year, in practice. As a result, raters may need to incorporate information about employee behaviors that are up to a year old. Therefore, it is important that raters make observations frequently and at appropriate intervals *throughout* the year. Behavioral observation training teaches raters how to appropriately prepare so that more accurate ratings can be made during the formal job performance rating process (Hedge & Kavanagh, 1988).

Frame of reference training. The second most common type of rater training is “frame of reference” training. Frame of reference training attempts to make sure that all raters of job performance are looking for similar behaviors and rating the same employee behaviors similarly (Hauenstein & Foti, 1989; McIntyre et al., 1984; Schleicher & Day, 1998; Sulsky & Day, 1992). Within organizations there are typically multiple employees reporting to different managers, performing similar tasks. Thus, it is important that different managers who are providing job performance ratings on similar tasks are making ratings in a similar way. If two raters observe the same behaviors, but one rater provides an employee a rating of two and another provides an employee a rating of three, then the validity of the job performance rating system will be reduced (Pulakos, 1984; Sulsky & Day, 1994). It is important that raters have a common frame of reference so that while providing job performance ratings for employees, they provide similar ratings for similar types of behaviors (Day & Sulsky, 1995; Pulakos, 1986; Stamoulis & Hauenstein, 1993; Uggerslev & Sulsky, 2008).

Frame of reference training can be broken down into five steps. The following five steps are an example process that can be followed and are paraphrased from a similar method used by Pulakos (1986).

1. The trainer explains each performance dimension on which the raters will have to make judgments.
2. The trainer provides examples and discusses with raters the behaviors that illustrate various performance levels. The purpose of this step is to have raters understand and agree on behaviors that reflect various levels of effectiveness for different behaviors.

3. Participants view video that includes behaviors reflective of various performance dimensions. Raters are then required to provide performance ratings for the employees in the video.
4. After all raters have made ratings, the raters share their ratings with the rest of the group. A discussion should also occur, especially when there are discrepancies between ratings so that the raters can further develop a common theory (reference) of what behaviors reflect which ratings.
5. The trainer presents to participants the correct ratings. The trainer should also talk through and provide explanations related to any discrepancies made by raters.

Calibration meetings. It is also important to note that frame of reference training is different from another common practice that organizations may engage in known as calibration meetings. A calibration meeting is a meeting that typically occurs after all raters have already made ratings of job performance, but before sharing the ratings with the employees. During the meeting raters discuss the ratings they provided and attempt to develop a common standard of rating (Sammer, 2008). For example, raters may attempt to reach agreement on what behaviors should signify a rating of three on a one through five scale.

Calibration meetings are used to ensure that raters are providing similar ratings to employees for similar types of behaviors (Park & Kim, 2013). Although calibration meetings have the potential to provide similar results as frame of reference training, calibration meetings may provide an additional opportunity for another type of error to be introduced into job performance ratings. For instance, because calibration meetings take place after initial ratings, there is a greater risk of raters reaching agreement on an

incorrect common frame of reference. The second potential error may result from group pressure. Raters may adjust ratings to reach group consensus based on group norms, group dynamics, and or group pressure (Obidinnu, Ejiofor, & Ekechukwu, 2014). This could result in adjusting accurate ratings to fit inaccurate group perceptions. There may be group consensus, but the consensus may be inaccurate compared to initial ratings. This means that raters may make accurate ratings of job performance, but then may introduce new error into their ratings based on the views of other raters. It is also possible that during calibration meetings raters end up better aligning and producing more accurate ratings (Sammer, 2008). However, there is still the risk that group dynamics could introduce new error to the measurement of job performance ratings. Thus, it may not always be appropriate to take the added risk calibration meetings present, making it even more important that raters receive frame of reference training regularly before evaluating employee performance.

More research is needed on the impact and potential errors that can occur when using calibration meetings. Despite the fact that calibration may introduce additional sources of error in the actual measurement of job performance calibration is also likely to have many positive effects for organizations (Obidinnu et al., 2014). Calibration can help assure that monetary compensation is more uniformly distributed (Pulakos & O'Leary, 2011). Calibration can also help diminish perceptions of unfairness that employees may have about performance rating systems (Pulakos & O'Leary, 2011). This is important because if ratings are perceived to be unfair, employees are less likely to respect performance rating systems and are more likely to feel dissatisfied and be potentially less

engaged (Sammer, 2008). As a result, calibration may be advantageous for organizations, but may ultimately introduce error into the measurement process.

By leveraging both frame of reference training as well as behavioral observation training, the impact of various types of error can be reduced. As a result, the reliability and validity of job performance ratings by multiple raters can be increased. However, research has demonstrated the effects of these types of trainings can decay over time (Ivancevich, 1979). Thus, it is important to provide periodic refresher training for raters.

Although, training can help reduce the impact error has on ratings of job performance, it is important to reiterate that training does not guarantee valid and reliable measurement of job performance. Following the surge of research on the benefits of rater training during the mid-1970's, Landy and Farr (1980) argued that research needed to begin to shift focus from methodology (e.g., rating scales) onto raters and the potential error that raters introduce to the rating process. Landy and Farr (1980) acknowledged that rater training was positive, but that another issue still influenced ratings of job performance during the decision-making process. Rater training helps to ensure that raters have the skills necessary to make accurate ratings, but rater training does not ensure that raters will make the decision to use the skills they have learned. In addition to the common rater errors already discussed, Landy and Farr (1980) also argued that research needed to focus on the decision-making processes used by raters while making ratings. This resulted in a stream of research during the 1980's and beyond focusing on the cognitive processes underlying job performance ratings.

Judgments Versus Ratings

By distinguishing between judgments and ratings, authors Murphy and Cleveland (1991, 1995) explain the issues surrounding the decision-making process used by raters. A judgment is a private type of evaluation of job performance that is made by raters. Raters can make observations and conclusions about employee job performance. However, judgments may be different from actual ratings of job performance that raters share with employees. For example, a manager may believe that an employee is a top performer, but the manager may only give the employee a three-out-of-five performance rating. Unlike judgments, which are private evaluations of employee job performance, ratings are public evaluations of employee job performance. Ratings, unlike judgments, are more likely influenced by additional factors such as rater motivation (Murphy & Cleveland, 1995). A judgment is closer to being a pure evaluation of job performance than an actual rating of job performance (Murphy & Cleveland, 1991, 1995). When a raters provides ratings of job performance they may not be motivated to provide the most accurate rating of job performance for an employee (Wong & Kwong, 2007). For instance, consider a manager who has five employees that all work on the same team. Four of the employees perform at a high level of performance while one employee performs at a mediocre level of performance. The manager, who provides ratings of job performance, may be motivated to provide all five employees with higher ratings of job performance because the rater is afraid of the discord that may result if only one employee receives lower ratings. Managers who provide performance ratings may also be motivated to inflate job performance ratings for employees depending on the performance rating structure used. For example, if a manager's job performance ratings

are based, in part, on the performance of the ratings provided to his or her employees, the manager providing ratings may also be motivated to inflate the job performance ratings provided to his or her subordinates. The point is that the goals of raters can motivate raters to introduce additional forms of error into their ratings independent of judgments (Spence & Keeping, 2013; St-Onge, Morin, Bellehumeur, & Dupuis, 2009; Wong & Kwong, 2007). This means that rater training may help raters to make better private ratings of job performance, but that actual ratings of job performance may still be distorted by motivations. As a result, a few authors have begun debating the merit of providing any job performance ratings at all (e.g., Hantula, 2011; Hauenstein, 2011; Jones & Culbertson, 2011; Mone, Price, & Eisinger, 2011; O'Leary & Pulakos, 2011; Pulakos et al., 2015; Pulakos & O'Leary, 2011).

Alternatives to Traditional Ratings of Job Performance

Previous research has demonstrated the positive impact that rater training can have on increasing the accuracy of job performance ratings; yet, some authors such as Pulakos and O'Leary (2011) have argued that performance rating systems have strayed too far from their original mark. Specifically, Pulakos and O'Leary (2011) argue that the original goal of performance rating was to provide accurate ratings of job performance to facilitate employee development and despite research attempting to reduce error in performance measurement performance ratings systems have strayed too far from their original mark. They further suggest that organizations may be better off moving away from the use of formal job performance rating systems.

Given that ratings of job performance are notoriously inaccurate, time consuming, and often occur only once or twice a year, it is difficult to use ratings of job performance

to shape employee behavior. Yet, being able to provide employees with specific examples of their performance can help inform employees about specific types of behaviors they need to change to improve performance (Cannon & Witherspoon, 2005; Kim & Hamner, 1976). This has led to arguments of doing away with performance ratings in favor of other methods to improve employee performance for organizations. It also leads to arguments in favor of not necessarily moving away from formal methods of providing job performance ratings, but at least shifting the focus of performance management from providing ratings to alternatives such as: improving communications between employees and managers, building trust relationships, and delivering regular, timely, and candid feedback (O'Leary & Pulakos, 2011). The main point of these arguments is to shift the focus from *evaluation*, rating performance, to performance management and improvement *processes* (Pulakos et al., 2015).

By assuming normal distributions of job performance and attempting to force or “correct” ratings to reflect a normal distribution, additional error may be added to distributions of job performance. Furthermore, as discussed earlier, if an exponential distribution of job performance is observed, there may be many positive implications for organizations, such as easier identification and better differentiation of higher performing employees.

The current study will examine observed distributions of job performance, accepting that job performance ratings possess a degree of error (Landy & Farr, 1980; Murphy, 2008; Pulakos & O'Leary, 2011; Scullen, Mount, & Goff, 2000; Viswesvaran et al., 1996; Viswesvaran et al., 2005), and apply statistical tests to address whether observed distributions are better classified as normal or exponential. Being able to

identify whether a distribution of job performance is normal or exponential will allow organizations to interpret ratings differently. One inherent advantage of identifying exponential distributions is the greater differentiation between top and bottom performers within organizations (Aguinis & O'Boyle, 2014; Aguinis et al., 2016; O'Boyle & Aguinis, 2012). It is also important to note that the error found in normal versus exponential distributions may also be meaningful. For instance, consider the following example of missing data. When data are missing from statistical analyses, data may be missing at random or missing in a meaningful way that may provide additional information. When data are missing at random, additional information may not be derived, but when data are missing in a meaningful way, it means that there may be a third unmeasured variable that is reflected in the missing data (Roth, 1994). The same may be true of distributions of job performance. The error in distributions of job performance that are not forced into a specific distribution may be meaningful error, potentially reflecting a third variable that was not directly measured. This error could be useful for organizations and researchers alike when attempting to understand error in the measurement of job performance.

Distributions of job performance may vary in shape because of error. However, ratings of job performance can become more accurate as a result of methods such as rater training and calibration. Thus, if exponential distributions are identified, they could be the result of a third factor such as the complexity of job performance. Job performance is a complex construct made up of various lower order constructs such as task performance, organizational citizenship behaviors, and counterproductive work behaviors (Borman & Motowidlo, 1993; Kidwell & Bennett, 1993; Rotundo & Sackett, 2002; C. Smith, Organ, & Near, 1983). Given that job performance is complex, it may be possible that varying

distributions of job performance are due to the many different factors that make up job performance. Job performance may be unidimensional, all lower order constructs of job performance are correlated and reflect overall job performance; yet, job performance is still complex because it is comprised of multiple lower order factors. As a result, varying distributions of job performance may result because of the complexity of job performance. For example, not all the factors that comprise job performance are captured in each rating. Understanding the many different factors of job performance that are detailed in the next section should help to better understand how the complexity of job performance may impact ratings of job performance and ultimately observed distributions of job performance.

Modern Conceptualizations of Job Performance

The history of job performance literature can be parsed into three main categories that loosely fit chronologically: a) initial research defining and measuring job performance, b) improving methods and the measurement of performance, and c) contemporary conceptualization. During the early research on job performance, there was a clear need to measure job performance, but no best way to measure it and no clear understanding of it. This period (e.g., early 1900's to the late 1950's) also marked the beginning of interest in studying various areas related to job performance, such as methods to measure job performance as outcomes and conceptualizations such as Thorndike's Ultimate Criterion. The second period of job performance research, the method and measurement era, is when large strides in tools used to measure job performance such as graphic rating scales and a better understanding of the myriad of errors that influence raters occurred. This era (e.g., the 1960's throughout the 1980's)

established many of the best practices (e.g., rater training) that are still used today. The late 1980's mark the beginning of the third period in job performance research history, contemporary conceptualizations of job performance which tends to emphasize the question of *what* to measure.

With the exception of a few early researchers such as Seashore, Indik, and Georgopoulos (1960) and James (1973), the many researchers historically ascribed to a one-dimensional conceptualization of job performance that did not consist of multiple lower order factors. The one-dimensional view of job performance primarily focused on what researchers now refer to as task performance. Much of the groundwork that defined the ideological shift of the contemporary conceptualization era began in the late 1980's and early 1990's.

Campbell's great eight. Campbell, Mcloy, Oppler, and Sager (1993) were at the forefront of the shift in conceptualizing job performance to mean more than only task performance. They distinguished between outcomes and behaviors of job performance by relating outcomes to results or effectiveness. Behaviors, on the other hand, were considered by Campbell and colleagues to produce performance. Better behaviors can increase performance and can help produce better results and outcomes, which can result in increased effectiveness (Borman, Klimoski, & Ilgen, 2003). Campbell et al. (1993) further developed this idea by defining eight behavioral dimensions of job performance that they believed encompassed all potential lower-order or more specific behavioral components of job performance. These eight lower-order dimensions include the following:

- *Job-specific task proficiency* is how well an employee performs tasks that are specific to that employee's job. These specific tasks differentiate one employee's job from other employees' jobs.
- *Non-job-specific task proficiency* is how well an employee performs on specific tasks that are unique to that employee's organization, but not unique to that employee's specific job. These types of tasks would apply broadly to many employees within an organization.
- *Written and oral communication* is how well an employee is able to write and speak with others.
- *Demonstrating effort* is how much commitment and persistence an employee demonstrates on the job.
- *Maintaining personal discipline* is how well an employee refrains from engaging in behavior that negatively impacts him or her as well as others and the organization.
- *Facilitating team and peer performance* is how much support an employee provides others within the organization to ensure others as well as the organization as a whole are successful.
- *Supervision* is the amount of positive influence an employee exerts on subordinates.
- *Management and administration* is how well an employee performs administrative and oversight tasks that are beneficial to the organization.

Task performance and contextual performance. Borman and Motowidlo (1993) took a different approach to conceptualizing job performance. Similar to Campbell et al.

(1993), Borman and Motowidlo agreed that performance is comprised of multiple related lower order factors, but disagreed on the appropriate number of factors that comprise job performance. Borman and Motowidlo argued for a two-factor structure of job performance comprised of task and contextual performance. The authors defined task performance as activities performed by employees that contributed to the success of an organization by providing the organization with necessary materials and services. Task performance is essentially how well an employee performs his or her job duties, the day-to-day tasks that are assigned (Rotundo & Sackett, 2002). Contextual performance, on the other hand, is how well an employee engages in behaviors that help facilitate the success of individual task performance as well as the task performance of other employees (Borman & Motowidlo, 1997; Borman, Penner, Allen, & Motowidlo, 2001; Coleman & Borman, 2000). Borman and Motowidlo (1993) described five specific types of contextual performance in which an employee may engage. These behaviors include a) voluntarily agreeing to go out of one's way to perform tasks that are not formally part of one's role, b) putting in extra time or effort without complaint to ensure tasks are completed successfully, c) supporting other employees, d) adhering to the rules of the organization even when rules are inconvenient to oneself, and e) putting the goals of the organization above the goals of oneself (Borman et al., 2001).

Borman and Motowidlo (1997) also argued that task and contextual performance differ in three distinct ways. First, although *different* jobs require different tasks to be performed, different jobs can still require *similar* contextual behaviors. Second, tasks are usually defined at the role level, whereas contextual behaviors are defined at the organizational level. Third, task performance is typically believed to be the result of

cognitive ability, whereas contextual behaviors are most likely associated with an employee's personality. For instance, employees' altruism, honesty, and/or integrity may influence whether they put the goals of the organization in front of personal goals.

Organizational citizenship behavior. Prior to Borman and Motowidlo (1993, 1997) proposing a two factor structure of task and contextual performance, C. Smith et al. (1983) suggested a construct known as Organizational Citizenship Behavior (OCB), which is similar but arguably unique from contextual performance (Borman & Motowidlo, 1997; Werner, 2000). In 1983, when Smith et al. first offered a definition and conceptualization of OCB, C. Smith et al. (1983) suggested OCB was comprised of two factors, altruism and generalized compliance. Building on the research of C. Smith et al. (1983), Organ (1988) provided a definition of OCB as behavior that is discretionary, not directly or explicitly recognized by formal reward systems, and that in the aggregate promotes the effective function of the organization. At the time, Organ (1988) suggested OCB included five dimensions: altruism, conscientiousness, sportsmanship, courtesy, and civic virtue. However, following evidence provided by Borman and Motowidlo (1993, 1997) on contextual performance, Organ (1997) updated the definition of OCB to more closely resemble Borman and Motowidlo (1993, 1997) definition and conceptualization of contextual performance. Organ (1997) updated and refined the definition of OCB to include contributions made to the maintenance and enhancement of social and psychological context that support task performance. This definition is still widely used (e.g., Bolino, Hsiung, Harvey, & LePine, 2015; Frazier & Bowler, 2015; Gajendran, Harrison, & Delaney-Klinger, 2015; Lemoine, Parsons, & Kansara, 2015; Shah, Cross, & Levin, 2015; Takeuchi, Bolino, & Lin, 2015; Trougakos, Beal, Cheng, Hideg, & Zweig,

2015). Organ (1997) also refined the construct of OCB to fit a two-factor structure consisting of interpersonal OCBs, contributions that are targeted toward an individual, and other OCBs, which are behaviors that demonstrate no immediate aid to any specific person, but that, demonstrate high standards for attendance, punctuality, conservation of organizational resources, and use of time while at work.

Coleman and Borman (2000) eventually refined the two-factor taxonomy of job performance proposed by Borman and Motowidlo (1997) to include three lower-order factors of contextual performance. This refinement came after Organ (1988, 1997) and other authors (e.g., Conway, 1999; Hodson, 1999; Konovsky & Pugh, 1994; P. M. Podsakoff, Ahearne, & MacKenzie, 1997; P. M. Podsakoff & MacKenzie, 1997; Van Dyne, Graham, & Dienesch, 1994) expounded on OCB. The three factors of contextual performance described by Coleman and Borman (2000) were empirically derived and include interpersonal support, organizational support, and job-task conscientiousness.

In 2001, Borman et al. decided to refine contextual performance factors further using a larger sample of job performance. This study provided additional support for the three-factor structure of contextual performance, which included interpersonal support, organizational support, and job-task conscientiousness found by Coleman and Borman (2000). However, the results of Borman et al. (2001) did result in a slight relabeling of two of the categories. The new categories were labeled personal support, organizational support, and conscientious initiative. The main point is that through research, multiple studies have found support for similar factor structures of job performance that are comprised of a task performance component and a contextual or citizenship component

(Johnson, 2001; Motowidlo, Borman, & Schmit, 1997; Motowidlo & Van Scotter, 1994; Rotundo & Sackett, 2002).

Personal support as contextual job performance includes behaviors such as helping other employees by offering suggestions, teaching, sharing knowledge, and even performing some of their tasks (LePine, Erez, & Johnson, 2002). Examples of organizational support include behaviors such as representing, defending, positively promoting the organization, and sticking it out with the organization through difficult times (N. P. Podsakoff, Whiting, Podsakoff, & Blume, 2009). Finally, conscientious initiative includes the display of additional effort even during difficult times and doing whatever is needed to complete objectives even if it requires doing things outside of one's role (Carpenter, Berry, & Houston, 2014).

Although there have been many studies advancing and refining the conceptualization of OCB and contextual performance, the distinction between OCB and contextual performance is still not clearly defined. This may be because at face value OCB and contextual performance appear similar. As a result, there is controversy surrounding the relationship and distinctiveness of these two constructs of job performance, contextual and OCB (Motowidlo, 2000). In part, this may be because when the construct of OCB was initially proposed by C. Smith et al. (1983) and Organ (1988) it appeared distinct from contextual performance, but following the updated definition of OCB by Organ (1997) both constructs begin to blend. The definitions of both constructs may have semantic differences, but the behaviors associated between both constructs share a great deal of overlap. The primary distinction between these constructs may lie within the full spectrum of behaviors associated with each construct. For instance, OCBs

are thought to only account for positive behaviors whereas contextual performance is believed to be the aggregated total of both positive and negative behaviors (Kell & Motowidlo, 2012). Negative contextual behaviors may be considered counterproductive work behaviors (CWB; Robinson & Bennett, 1995). Contextual performance is then the net performance between OCBs and CWBs. In essence, contextual performance may be considered a higher order factor that encompasses both OCBs and CWBs.

Counterproductive work behavior. Kidwell and Bennett (1993) proposed the idea of a withholding effort or CWB, which is the antithesis of OCB. Examples of CWB proposed by Kidwell and Bennett (1993) include shirking, social loafing and free riding. Sackett (2002) offered a definition of CWB as any intentional behavior on the part of the employee viewed by the organization as contrary to its legitimate interests. Robinson and Bennett (1995) elaborated on CWBs, suggesting a taxonomy consisting of four categories: production deviance, property deviance, political deviance, and personal aggression. Production deviance is a minor form of organizational deviance characterized by behavior such as leaving early. Property deviance is a more severe form of organizational deviance, such as stealing from the organization. Political deviance is a minor form of interpersonal deviance, for example, showing favoritism. Personal aggression is a severe form of interpersonal deviance, for example, sexual harassment.

Thus far three factors of job performance have been introduced, task performance, contextual/OCB, and CWB. However, up until the early 2000's, *how* the three factors fit together to comprise overall job performance remained a question in the literature. To help answer this question, Rotundo and Sackett (2002) conducted a study to offer additional clarity to the conceptualization of job performance by attempting to explain

how these three factors fit together. These authors started by reviewing twenty years of research and concluded that the same three factors, task performance, OCB, and CWB, are the three main factors comprising job performance. These authors then designed a study to understand how these three main factors of job performance are related. The results of Rotundo and Sackett's (2002) study found that about sixty percent of the variance in job performance ratings could be explained by task performance and CWB; where about thirty percent of the variance was explained by task performance and about thirty percent of the variance was explained by CWB. OCB, on the other hand, only explained between four and twenty percent of the variance. When totaled, sixty percent and four to twenty percent do not total to one hundred percent. This is because there is error in the measurement of job performance and because an additional factor or factors are not measured by the three main factors. The results of this study suggest that raters consider task, CWB, and OCB as part of job performance when providing ratings of job performance, , but do not put equal weight on all three factors when providing overall ratings. These findings also suggest that the weight raters place on OCB may fluctuate widely, from being sparsely considered to being weighted heavily. This provides support for previous research that job performance may be better conceptualized as two factors, task and contextual performance where contextual performance is comprised of both CWB and OCB.

Adaptive performance. Although task, OCB, and CWB are considered to be the three primary categories of job performance, other more recent research has also identified adaptive performance as a potentially prominent type of job performance (Pulakos, Arad, Donovan, & Plamondon, 2000). Adaptive performance was originally

proposed by Pulakos et al. (2000, p. 615) as “altering behavior to meet the demands of the environment, an event or new situation.” Their research identified eight adaptive performance dimensions. Campbell (2012) offered additional support for adaptive performance as a unique type of job performance by demonstrating that traditional task and contextual performance dimensions (i.e., OCB and CWB) do not subsume it.

However, additional research is still needed to help explain how adaptive performance fits within the broader taxonomy of job performance relative to task, CWB, and OCB.

A general factor of job performance. Thus far, three main constructs of job performance have been introduced, adaptive performance, task performance, and contextual performance (comprised of OCB and CWB). Although each main construct of job performance may appear to be differentiated substantively, evidence has demonstrated that each of these constructs converge onto one general factor reflecting job performance to some degree (Viswesvaran et al., 2005). Viswesvaran et al. (2005) conducted a meta-analysis on over ninety-years of research and were demonstrated that even when controlling for various types of rater error, there remained a general factor of job performance able to account for sixty percent of the total variance in job performance ratings. This suggests that even when evaluating different types of performance there should be shared variance between different measures and types of job performance (e.g., task performance, adaptive performance, and contextual performance), if all measures are truly measuring job performance. This also means that when conducting research on job performance that it may be acceptable to combine and evaluate multiple measures of job performance as an indicator of overall job performance (Rotundo & Sackett, 2002; Viswesvaran, 1993; Viswesvaran et al., 2005).

Being able to combine multiple measures of job performance into an overall indicator of job performance is paramount for the present study. Recall that managerial performance ratings are the most commonly used method for assessing employee performance (Aguinis, 2013; O'Boyle & Aguinis, 2012; Viswesvaran et al., 2005) and that organizations typically suggest that job performance should be normally distributed (Motowidlo & Borman, 1977; Reilly & Smither, 1985; Schneier, 1977a, 1977b). As a result, it would be unlikely to find exponential distributions of job performance using managerial ratings as the only measure of performance (O'Boyle & Aguinis, 2012). However, objective indicators of performance may exist within organizations that are inherently free from being forced into any specific distribution (Aguinis & O'Boyle, 2014; Aguinis et al., 2016; O'Boyle & Aguinis, 2012). Although, evidence has suggested that some objective indicators tend to be exponentially distributed (Aguinis & O'Boyle, 2014; Aguinis et al., 2016; Campbell & Wiernik, 2015; Crawford et al., 2015; O'Boyle & Aguinis, 2012).

There may not be specific guidance in the literature on the most appropriate ways to combine subjective and objective measures; yet, authors have suggested it would be appropriate (Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995). In a meta-analysis by Bommer et al. (1995), it was determined that both objective and subjective measures of job performance contain error, but integrating both types of measures may lead to better assessment of performance. Findings by Viswesvaran (1993) and Viswesvaran et al. (2005), provide evidence it would be appropriate to combine various measures such as managerial ratings and objective indicators of performance into one overall measure of job performance that may better reflect the actual distribution of

performance. However, given that in most cases managerial ratings are expected to reflect a normal distribution and objective measures are expected to be exponentially distributed, integrating the two distributions would result in a distribution distinct from the distributions of which it is comprised (Stephens, 2000). In this specific instance, integrating a normal and exponential distribution would be expected to produce an exponentially modified Gaussian distribution (Pauls & Rogers, 1977)². An exponentially modified Gaussian distribution is essentially a weighted average of the normal and exponential distributions that have been integrated (Pauls & Rogers, 1977). As a result, how well the resulting integrated distribution resembles a normal or exponential distribution depends on the degree that managerial ratings resemble a normal distribution and the degree that objective measures resemble an exponential distribution.

Evaluating and Classifying Distributions of Job Performance

Although it may be appropriate to combine objective and subjective measures of job performance, guidelines about how to evaluate and classify the resulting distribution of integrated measures is lacking in the I-O Psychology literature. There are recommendations for evaluating the normality of job performance distributions (Shapiro, Wilk, & Chen, 1968; Thode, 2002). However, because job performance has generally been assumed to be normally distributed, further recommendations for evaluating job performance distributions beyond normality are less common in I-O Psychology. In light of recent research demonstrating that some objective measures of job performance are exponentially distributed (e.g., Aguinis et al., 2016; O'Boyle & Aguinis, 2012), further recommendations for identifying and evaluating whether a distribution is exponential is

² A Gaussian distribution is another name for a normal distribution. The normal distribution was given this name in honor of the German mathematician Carl Friedrich Gauss (Dunnington, Gray, & Dohse, 2004)

needed. Other disciplines provide recommendations that have potential in I-O Psychology research. In turn, this section reviews accepted methods for evaluating the normality of distributions. This section also reviews potential methods for evaluating whether distributions are exponential.

Evaluating the normality of distributions. There are many commonly used statistical tests (e.g., the *t*-test, analysis of variance, and regression) which all require that the data being “tested” is normally distributed (Tabachnick & Fidell, 2013). As a result, many statistical tests can be applied to assess the normality of data. Thode (2002), while acknowledging that his list was not comprehensive, identified over forty statistical tests for testing normality. However, not all statistical tests for evaluating normality are equally robust at identifying departures from normality. According to Razali and Wah (2011), one test of normality, the Shapiro-Wilk test, has been identified to have the most power, the ability to identify departures from normality if they truly exist, relative to other tests of normality. The Shapiro-Wilk test, tests the null hypothesis that a set of data came from a normally distributed population (Shapiro & Wilk, 1965). However, because the Shapiro-Wilk test relies on null hypothesis testing, it can be biased if the size of the sample is too large (e.g., greater than 2000 data points; Royston, 1982). This means that if a sample size is too large, there is increased chance of rejecting the null hypothesis when it should be accepted (i.e., making a Type 1 error).

Given that statistical tests of normality, including the Shapiro-Wilk test, have the potential to be biased, it is important to leverage more than one method to assess the normality of a distribution of data. According to Tukey (1977, p. 43), “there is no excuse for failing to plot and look.” Graphical methods are a powerful tool for verifying the

accuracy of statistical tests (Chambers, Cleveland, Kleiner, & Tukey, 1983), and probability plots are a specific type of graphical tool that can be used to assess normality (Thode, 2002). The Q-Q plot (quantile-quantile plot) is a specific type of probability plot that is recommended for assessing normality (Razali & Wah, 2011). The use of the Shapiro-Wilk test and a Q-Q plot are an accepted standard for adequately assessing normality (Tabachnick & Fidell, 2013).

Evaluating exponential distributions. Similar to tests of normality, many different statistical tests can be used to test whether data are exponentially distributed. An exponential distribution is defined by its infinite variance and a greater proportion of extreme events (O'Boyle & Aguinis, 2012). While many distributions can meet these criteria, the current research is interested in a specific type of exponential distribution, the pareto distribution, also known as the power-law distribution, which is characterized by a leptokurtic bulge and positive skew (Choulakian & Stephens, 2001). When evaluating whether data fit a specific type of exponential distribution, such as the power-law distribution, a modified version of the Shapiro-Wilk test has been shown to have the most power (Uthoff, 1970). This modified Shapiro-Wilk test, similar to the Shapiro-Wilk test for testing normality, is also a null hypothesis test. The modified Shapiro-Wilk test tests the null hypothesis that data came from a specified a priori exponential distribution.

A modified Shapiro-Wilk test, however, should not have the final say on whether data are exponentially distributed. Similar to testing for normality, it is important to leverage multiple methods to verify the shape of a distribution. Again, graphical methods such as the Q-Q plot should be used to verify the accuracy of the statistical test

(Chambers et al., 1983). A modified Shapiro-Wilk test and graphical methods should provide sufficient evidence for data being exponentially distributed.

Hypotheses

Hypothesis 1: Managerial ratings of job performance are normally distributed.

In summary, there have been three major milestones in the history of job performance research. The first was the birth of research relating to job performance, which began in the early 1900s. Large industry and world wars drove research to focus on developing selection systems. In order to develop better selection systems processes for measuring job performance also had to advance. This led to the second major historical job performance milestone, conceptualizing job performance. This milestone was marked by the distinction between actual job performance and what is measured as job performance. Actual job performance is an intangible ideal construct that researchers and practitioners alike strive to measure. Indicators of job performance are what are actually measured. The third milestone in the history of job performance research involves theoretical advancements in the understanding of job performance. One of the most important contributions of these theoretical advancements is the acknowledgement that there is error in all measurement of job performance. Finally, the current research attempts to further the understanding of job performance by building on current research that posits that job performance may be exponentially distributed instead of normally distributed.

As detailed by O'Boyle and Aguinis (2012), managerial ratings of job performance historically have restricted amounts of variance and therefore would be an unlikely place to identify exponential distributions of performance. Managerial rating scales that are

provided by organizations for managers to provide ratings typically only include a small set of discrete anchors (e.g., 1, 2, 3, 4, or 5), which will likely not include enough anchors to allow for adequate differentiation of employees. The expected result would be a normal distribution of managerial performance ratings. Organizations could use rating scales that have more anchors and may result in more variance, but as research has demonstrated, this is ineffective because raters cannot accurately distinguish performance when many anchors are present (Cox, 1980). For example, on a scale with anchors from one to one hundred, raters are not able to distinguish accurately an employee that deserves a rating of 78 versus 79. Raters are not able to accurately and meaningfully distinguish between employees on scales with more than about five to seven anchors (Cox, 1980). As a result, managerial ratings of job performance are more likely to have constricted variance.

Hypothesis 2: A composite multi-rater multi-method measure of subjective job performance indicators will demonstrate an exponential distribution of job performance.

To circumvent the issue that managerial ratings are likely to be normally distributed and test whether exponential distributions exist, more variance may be needed. To capture more variance organizations could combine subjective ratings from multiple raters using different methods to increase the amount of variance between employees.

Hypothesis 3: Objective measures of job performance will result in exponential distributions of job performance.

Alternatively, in order to demonstrate that job performance may be exponentially distributed, organizations could use objective indicators of job performance. A few types of objective indicators of performance that result in exponential distributions of

performance have already been identified (Aguinis et al., 2016; O'Boyle & Aguinis, 2012). The problem is that these objective indicators only apply to a small number of jobs. There are, however, additional objective indicators of performance that have yet to be thoroughly explored. As an example of an objective indicator of job performance that has yet to be tested and that could be applied broadly to many different types of jobs, organizations could look at the amount of time an employee has spent in each role prior to receiving a new role or promotion. Less time spent by an employee in a role prior to receiving a promotion may indicate better performance.

Hypothesis 4: A composite of subjective and objective job performance measures will demonstrate an exponential distribution of job performance.

As an alternative to using only objective indicators of job performance to demonstrate that job performance may be exponentially distributed, organizations could use a composite of objective indicators and subjective indicators to capture job performance holistically. Previous research has demonstrated that irrespective of the method used for measuring job performance, a single higher-order factor of job performance should exist (Viswesvaran, 1993; Viswesvaran et al., 2005). Therefore, multiple methods of measuring job performance should still be measuring the same thing. Furthermore, using multiple methods to converge on performance should reduce the amount of error and idiosyncratic rater effects, providing an acceptable method for collecting data to accurately determine the shape that the distribution of job performance best fits. By combining ratings from multiple methods, actual job performance may essentially be triangulated and the amount of distinction should be increased between employees. However, because previous research has yet to demonstrate the existence of

exponential distributions for many jobs, combining multiple ratings of performance captured using multiple methods and objective measures should circumvent this critique and allow the opportunity to adequately assess the distribution of job performance.

CHAPTER THREE

METHODS

This study used archival job performance data from a large multi-national organization. The archival job performance data included multiple types of performance ratings provided by managers (e.g., assessing performance on short-term and long-term goals), subordinate ratings of performance (e.g., assessing manager quality), peer performance ratings and objective performance indicators. In total, six different indicators of job performance were used. These six indicators of job performance are frequently used in most large organizations (Aguinis, 2013). Four of the indicators were subjective measures and were evaluated individually as well as combined into a single composite measure. All indicators of performance were examined using multiple statistical tests of normality and exponentiality. Importantly, all measures of job performance were chosen because they fit Rotundo and Sackett's (2002) definition of performance as actions and behaviors under the control of an individual that contribute to the goals of an organization.

Participants

The archival data set included only mid-level managers of people (62 women, 141 men, $M_{age} = 45.76$ years) who had at least three direct reports. The primary organizational function of all the managers was information technology (IT). The sample included

managers that resided in many different countries and managed both employees within their respective country as well as employees located in different parts of the world.

Fifty-six percent of the sample was comprised of participants from four countries: the United States ($n = 90$), Spain ($n = 16$), Mexico ($n = 24$), and Russia ($n = 14$).

In part, this sample was selected to meet criteria put forth by Beck et al. (2014), which if ignored could unintentionally result in non-normal distributions of job performance. The one criterion for identifying an appropriate sample dictates that the sample consists of employees with comparable jobs. This criterion is believed to be met because all employees in the sample shared the same job function (i.e., Information Technology), were from the same organization, and had a similar overarching responsibility of managing people.

Procedures

The procedures in this section detail how all six archival measures were collected. The six archival measures of job performance include: managerial ratings of performance on short-term objectives (Manager Short-Term); managerial ratings of performance on long-term objectives (Manager Long-Term); subordinate ratings of manager quality (Manager Quality); 360-ratings of manager performance comprised of equally weighted ratings from subordinates, peers, and managers (360), average time in role prior to receiving a new role (Time in Role, TIR) and average time prior to receiving a promotion (Time Prior To Promotion, TPTP).

Managerial ratings of short-term and long-term objectives. Managerial ratings of short-term and long-term objectives are two separate ratings, but were collected using the same process.

Short-term and long-term objective ratings are both a part of the organization's performance management system that is administered on an annual basis. It includes four phases that start in January of each year and conclude in March of the next year. As a result, the cycle that begins in, for example 2015 has some overlap with the cycle that begins in 2016. Starting in January of each year employees meet with their supervisors and discuss their performance and objectives from the previous year. During this time, managers and employees collaborate to create both long-term and short-term objectives for the year. Some objectives may cascade from managers ensuring alignment with the organization's overarching goals, but this is not required.

In April, the second phase begins. During the second phase, employees meet again with their managers and conduct a career conversation. A career conversation is an opportunity for the employee to review developmental feedback they have received up to that point in the year, to discuss career preferences, potential career paths, mobility, strengths, opportunity areas, and key areas of development.

The third phase begins in July. During this phase employees and managers conduct a mid-year conversation. This conversation is essentially a mid-year review of the employee's performance. Employees review their short-term and long-term objectives and provide feedback on how they believe their performance has been to date on each objective. The employee and manager create a plan to redirect and make adjustments going forward to help aid performance towards reaching each short-term and long-term objective.

The fourth and final phase of the performance management process begins in October. This phase includes a development reconnect, self-input on both short-term and

long-term objectives, as well as manager input and calibration. Self-input is where the employees provide input on how well they believe each short-term and long-term objective was executed. This is done in conjunction with managers also providing input. Calibration, the fourth part of phase four refers to calibration meetings conducted by managers that have similar employees. Recall calibration meetings are similar to frame of reference training. Calibration involves multiple managers and human resources associates meeting and aligning to provide job performance ratings for employees on short-term and long-term objectives. Managers are expected to provide ratings of job performance on short-term and long-term objectives independently and then meet collectively with other managers to align and adjust ratings. This process ensures that managers are providing similar ratings for similar performance on objectives for employees in similar roles.

Manager quality. The measure of Manager Quality is administered annually. It is typically administered during the middle of the year, but the exact date varies each year based on business needs. All managers with at least three direct reports participate in the process and receive feedback ratings. If a manager has input on an employee's short-term and long-term performance objectives, the manager may invite the employee to provide feedback ratings. The manager may also invite direct reports, matrixed employees, and other employees that have reported to the manager within the last three months. Additionally, all raters are required to have reported to the manager for a minimum of six months.

Raters are allowed approximately three weeks to provide feedback. All ratings are provided anonymously. Feedback reports are released to human resources, the

participants, and their managers. Managers discuss the feedback they received with their raters and identify managerial behaviors that they can improve. Based on the Manager Quality feedback report, participants work with their managers to identify long-term development objectives. As a result, Manager Quality ratings have some direct impact on the managerial performance ratings for long-term objectives. This alignment is positive and important. It helps to ensure that similar types of performance are measured between the different indicators of performance used in the current study.

360° performance feedback. The 360° performance feedback tool is administered annually, but individual employees typically do not participate every year. The expectation is that employees participate once every few years. However, if an employee experiences a significant change in role, the recommendation is that the employee participates sooner instead of waiting for a year or two to pass. The 360 process typically begins with the identification of employees that have not recently participated in the 360 process and that meet various criteria. Example criteria include, having been in their current role for at least six months and having at least three direct reports. Once a list of participants is completed, an invitation email is sent out to managers inviting them to participate in the 360 process. Once invited, managers log into an online tool and select raters. Raters should include all subordinates, multiple peers, direct manager and matrixed manager if they have one and others such as external clients. The recommendation is that the participant has worked closely with all raters for at least six months. Participants are advised to select at least three subordinates and three peers. After the participant's raters have been selected, raters are sent an email inviting them to provide feedback. Raters have approximately two weeks to respond and provide ratings.

After all ratings have been made, a feedback report is generated that is shared with the participant and his or her manager. In line with best practices and to protect confidentiality, average scores are not produced at the rater category level unless at least three raters have responded. However, rater categories with less than three respondents will be included in the overall average 360 score. Results are typically also used to help develop at least one PDR objective for the participants.

Time in role. Every time an employee begins a new role, the start date for the new role is recorded in a database. Employees' previous role start date was subtracted from their current role start date and divided by three-hundred sixty-five. This procedure provided an employee's time in role in number of years. This procedure was applied to every role an employee has had within the organization. If an employee only has had one role within the organization then the employee's first role within the organization, the employee's hire date, was subtracted from the current date and divided by three-hundred sixty-five. An average time in role was calculated for each employee.

Time prior to promotion. A similar procedure was used to calculate time prior to promotion as was used to calculate time in role. The distinction between time in role versus time prior to promotion is that an employee may have multiple roles prior to receiving a promotion. As a result, the average time prior to promotion for each employee may be greater than the average time in role for each employee.

Measures

In this section, an overview of all four subjective archival measures is provided: Manager Short-Term, Manager Long-Term, Manager Quality, and 360° ratings of manager performance. The two objective archival methods, time in role and time prior to

promotion, are based on the calculation process previously described and do not include a formal measure administered to employees. As a result, they will not be reviewed in this section, but the subjective measures of performance will be.

Manager short-term. These ratings are provided based on two types of objectives related to delivering a business plan and creating efficiency: employees may create their own objectives or leverage objectives provided by their manager. Objectives related to delivering on the business plan refer to the impactful objectives each person can take to achieve annual operating plan metrics and typically reflect things the employee can influence. An example objective may include: “Collaborate with cross-functional teams and follow new processes to implement 3 innovations as defined in the annual operating plan that will generate net revenue = \$500,000 by year-end.” All manager ratings provided based on employee objectives are based on how successful the employee is at achieving the objective.

The second type of short-term objective is related to creating efficiency. This refers to the realization of initiatives that, when executed, ensure sustainable performance. The focus is on progress that the employee makes during a given year, even though the impact may not come to fruition immediately. An example ‘create efficiency’ objective may include: “Streamline efforts by creating, aligning, and communicating by quarter 2 a training to improve production by 8%.” Manager short-term ratings are provided on a one through five scale where one represents the lowest level of performance and five represents the highest level of performance.

Manager long-term. These objectives fall into four separate categories: drive future business success, drive organizational health, develop others, and develop self.

Employees should have at least two short-term objectives and four long-term objectives. The first long-term objective, drive future business success, is about identifying and achieving progress against strategic business plans to achieve long-term growth and strengthen innovation. An example of a *drive future business success* objective may include: “Assemble project team and begin product development in collaboration with Region A as measured by 100% approval for launch by June 3rd.” The second category, drive organizational health, is the ability of an organization to align, execute, and renew itself to sustain exceptional performance over time. An example of this type of objective may include: “Build organization roadmap to unlock synergies and streamline structure measured by leadership alignment; implement key milestones for end of year.” The third category, develop others, is based on career level and degree of managerial responsibility. These objectives should require significant effort over an extended period of time, not one-time activities. An example would be: “Establish partnership with teams to build capability to implement changes across the organization; measured by successful adoption of change.” The final category, develop self, should focus on the development of initiatives that require significant effort over an extended period of time, not one-time activities to increase capabilities. An example would be: “Develop verbal communication skills by applying insights from a local college course; measured by regular feedback from my manager and quarterly feedback from my peers.” Manager long-term ratings are provided on a one through five scale where one represents the lowest level of performance and five represents the highest level of performance.

Manager quality. The manager quality performance tool was developed in-house by the organization but is similar to other upward manager feedback tools (e.g., tools that

provide feedback from employees to their managers) used across many organizations. The tool consists of twelve behavioral items that are intended to measure three broad leadership dimensions: moving toward people, moving against people, and moving away from people. An example item is: "Leaves big decisions up to others." Each item is rated by subordinates in reference to their manager. Ratings are provided on a five-point Likert style scale that ranges from "no extent" to "a very great extent." An average overall score was calculated and used.

360° Performance Tool. The 360° performance tool consists of 58 items and nine dimensions that are equally weighted and averaged into a total score. Only the total score was used. The nine dimensions include: decision making, innovating, driving for results, creating an inclusive culture, building trust, motivating and inspiring others, collaborating and influencing, acting with integrity, and inspiring trust. An example item is, "Takes the initiative to find ways to get better results." Each item is measured on a five-point Likert-style scale that ranges from "small extent" to "great extent."

CHAPTER FOUR

RESULTS

The current study tested four hypotheses to determine the distribution of job performance scores and generates practical methods that can be employed by organizations. To test the four hypotheses, two main questions had to be addressed. First, were the indicators of job performance normally or exponentially distributed? Second, did the indicators of job performance specifically fit an exponential distribution? To answer these two questions multiple statistical and visual tests were applied.

Organizing Performance Data for Analysis

The first step required combining various measures of job performance into performance measure groups that could be used to test each hypothesis. This first step resulted in eight performance measure groups (see Table 1). To test Hypothesis 1 there was one group for each of the managerial measures of performance and one composite group consisting of both managerial performance ratings. To test Hypothesis 2 there was one group that consisted of all subjective performance measures. Hypothesis 3 used three groups one for each objective measure and one composite group consisting of both objective measures. Finally, Hypothesis 4 used one overall composite measure consisting of all measures of job performance.

Table 1

Groups of Performance Measures

Group	Hypothesis	Performance Measures	Type of Measure
A	1	Manager Short-Term Ratings	Subjective
B	1	Manager Long-Term Ratings	Subjective
C	1	Manager Short-Term Ratings Manager Long-Term Ratings	Subjective
D	2	Manager Short-Term Ratings Manager Long-Term Ratings Manager Quality Ratings 360 Ratings	Subjective
E	3	Time in Role	Objective
F	3	Time Prior to Promotion	Objective
G	3	Time in Role Time Prior to Promotion	Objective
H	4	Manager Short-Term Ratings Manager Long-Term Ratings Manager Quality Ratings 360 Ratings Time in Role Time Prior to Promotion	Subjective Objective

Group E and Group F were different from the other groups because for these groups, a lower score was associated with more positive job performance. As a result, both Group E and Group F scores were reversed. To reverse score for these two groups, each score was subtracted from the maximum value for each group. Group E and Group F also used a different scale than the other measures. As a result, Group E and Group F additionally required a linear transformation.

Group C consisted of a composite measure of managerial ratings of job performance. Following the recommendation of Viswesvaran (1993) a weighted composite approach was used to generate this group. This approach was similar to weighting test score items based on their relationship with an overall test score. Furthermore, this approach was used for all groups that utilized a composite of multiple indicators of performance. The first step to calculating the weighted composite score for this group was to sum both indicators of job performance into an overall score. The next step was to correlate both indicators of performance with the overall score. Then each indicator of performance was multiplied by its correlation coefficient with the overall score. The resulting values were then summed and averaged to create an overall weighted score for each case.

Group G consisted of a composite measure of objective indicators of job performance. This group used a weighted composite approach similar to the method used to calculate Group C. The values from Group E and Group F were combined to create an overall score that was correlated with both Group E and Group F to generate weights for each group. Once the values of Group E and Group F were weighted, they were summed and averaged to create a weighted composite score for each case. Group D was a composite of all subjective indicators of performance and Group H was a composite of all measures of job performance, both subjective and objective indicators. To create Groups D and H, a weighted composite approach was again used.

Statistical Tests and Data Examination Procedures

Four tests were used to test each of Hypotheses 1 through 4, these included histograms, Q-Q plots, the Shapiro-Wilk test of normality, and the modified

Shapiro-Wilk test for exponentiality. Histograms and Q-Q plots are visual approaches of evaluating the shape of a distribution whereas both Shapiro-Wilk tests rely on a test of statistical significance to determine whether a set of data fit a prespecified distribution.

To elaborate, a histogram is a bar graph of frequencies based on an empirical distribution of data. For example, there are a finite number of short-term manager ratings. An employee can only receive a value between one and five. The histogram would consist of five bars, one for each potential value. The height of each bar is dependent on how many employees receive a one, two, three, four, or five. Depending on the height of each bar, holistically the histogram takes on different shapes, which represent the underlying distribution of the data. If the data were normally distributed, the histogram resembles a normal distribution and if the data were exponentially distributed the histogram resembles an exponential distribution. To rigorously evaluate how well each histogram resembles a normal distribution versus an exponential distribution seven subject matter experts (SMEs) were asked to provide two separate ratings for each histogram. The ratings from all seven SMEs were then averaged into two overall scores for each histogram. The first rating was based on agreement with the statement, "This histogram is normally distributed." The second rating was based on agreement with the statement, "This histogram is exponentially distributed." The agreement scale used by all seven SMEs had five anchors that ranged from one "strongly agree" to five "strongly disagree." Average ratings of less than three were considered support for each statement. A separate histogram was generated for all eight groups and the same seven SMEs provided both ratings for all eight groups. SMEs were I-O Psychology doctoral students who all had training in advanced statistical analysis.

The second visual approach to assessing the shape of a distribution was the Q-Q plot. A Q-Q plot is a special type of probability plot. Probability plots are used to graphically compare the similarity of two distributions. Q-Q plots are a non-parametric approach generally used for assessing goodness of fit (Thode, 2002). To create the Q-Q plot the first step is to calculate quantiles. Quantiles are the cut-points in a data set that separate the data into four equal groups based on the three quantiles. The second step is to sort the data in increasing order, lowest scores to highest scores. A second set of artificially generated data that is normally distributed is also utilized. The artificially generated data is also sorted in increasing order. The observed data from each group are all paired with the normally distributed data and plotted. If the data from each group closely resemble the data from the normally distributed data then the plotted data on the Q-Q plot follow a 45-degree angle where $X=Y$. For example, if there were thirty ratings in hypothetical Group Z, the artificially generated normally distributed data would also include thirty ratings. Each rating in Group Z would be ordered in increasing values. Each rating in the normally distributed data would also be ordered in increasing values. X and Y coordinates to be plotted would be generated by pairing the two lowest ratings in Group Z and the artificially generated normally distributed data, and by pairing the second two lowest ratings in Group Z and the normally distributed data and so on until all values have been paired. The pairs are used as X and Y coordinates to be plotted. If the two sets of data have the same distribution, then the plotted data would resemble a straight line at a 45-degree angle. Another way to think about a Q-Q plot is as a correlation. If the two sets of data correlate strongly, the plotted data would resemble a straight line. Visual analysis of the Q-Q plot comprised of ratings from SMEs was used to

evaluate the normality and exponentiality of the plots. Each Q-Q plot was evaluated by using the same procedure previously used to evaluate how well each histogram resembles a normal distribution versus an exponential distribution. The same seven SMEs provided two separate ratings for all eight Q-Q plots on the same five-point rating scale used previously. The SMEs rated: How well each plot followed a 45-degree line and how much each plot curved away from a 45-degree line.

An intraclass correlation coefficient (ICC) was calculated based on all of the ratings that were used to test all four hypotheses provided by all SMEs. The ratings included how well each histogram for all eight groups fit a normal distribution, as well as how well each distribution fits an exponential distribution. The ratings also apply to Q-Q plots that were used as the second criteria to test all four hypotheses. Ratings related to Q-Q plots included how well each Q-Q plot fit a 45-degree line, as well as how much each Q-Q plot curves away from a 45-degree line. ICC is an indicator of agreement between ratings provided by all SMEs. A high degree of agreement was found between all seven SMEs. The average ICC was .92 with a 95% confidence interval ranging from .87 to .96 ($F(31, 186) = 12.95, p < .001$). This magnitude of ICC suggests that there was strong agreement among all seven SMEs for all ratings.

The third set of criteria used to evaluate each hypothesis were Shapiro-Wilk tests. The Shapiro-Wilk test for normality is a statistical test that evaluates whether a set of data came from a normally distributed population. It tests the null hypothesis that the sample did come from a normally distributed population. Therefore, if a p -value of less than .05 were obtained, the null hypothesis was rejected. This would provide support for a distribution being non-normally distributed. The test statistic is calculated by “dividing

the square of an appropriate linear combination of the sample order statistics by the usual symmetric estimate of variance” (Shapiro & Wilk, 1965, p. 591). This procedure is very similar to the procedure that is used by Q-Q plots. The difference between the Shapiro-Wilk test and a Q-Q plot is that the Shapiro-Wilk test goes one step further and summarizes the Q-Q plot into a test statistic. The Shapiro-Wilk test statistic is essentially a summary statistic of how different the sample data are from the artificially generated normally distributed comparison data.

The modified Shapiro-Wilk test for exponentiality is very similar to the Shapiro-Wilk test for normality. The only difference between the two tests is that the Shapiro-Wilk test for normality compares an observed set of data to a normally distributed set of data, while the modified Shapiro-Wilk test for exponentiality compares an observed set of data to an exponential distribution. The null hypothesis for the modified Shapiro-Wilk test for exponentiality is that the observed sample did come from an exponential distribution. Therefore, when using the modified Shapiro-Wilk test for exponentiality, a p -value of .05 or greater would provide support for the assumption that the observed sample was exponentially distributed.

Summary of Results by Hypothesis

There were various criteria used to evaluate each of the four hypotheses. Table 2 provides a summary of the criteria used to evaluate each hypothesis. The following paragraphs provide a detailed review of each criterion as it relates to each hypothesis as well as a summary of how well each hypothesis was supported.

Table 2

Summary of Criteria and Parameters

Hypothesis	Completely Supported	Partially Supported	Not Supported
(1) Managerial ratings of job performance will be normally distributed.	All groups - A, B, and C demonstrate all of the following: <ul style="list-style-type: none"> • Normally Distributed Histogram • Q-Q plot resembles straight line at forty-five-degree angle • Non-significant Shapiro-Wilk test • Significant-Modified Shapiro-Wilk test 	At least one group - A, B, or C demonstrates at least one of the following: <ul style="list-style-type: none"> • Normally Distributed Histogram • Q-Q plot resembles straight line at 45-degree angle • Non-significant Shapiro-Wilk test 	None of the groups - A, B, or C demonstrate any of the following: <ul style="list-style-type: none"> • Normally Distributed Histogram • Q-Q plot resembles straight line at 45-degree angle • Non-significant Shapiro-Wilk test
(2) A composite multi-rater multi-method measure of subjective job performance indicators will demonstrate an exponential distribution of job performance.	Group D demonstrates all of the following: <ul style="list-style-type: none"> • Exponentially Distributed Histogram • Q-Q plot displays an upward curve • Non-significant Modified Shapiro-Wilk test • Significant Shapiro-Wilk test 	Group D demonstrates at least one of the following: <ul style="list-style-type: none"> • Exponentially Distributed Histogram • Q-Q plot displays an upward curve • Non-significant Modified Shapiro-Wilk test 	Group D does not demonstrate any of the following: <ul style="list-style-type: none"> • Exponentially Distributed Histogram • Q-Q plot displays an upward curve • Non-significant Modified Shapiro-Wilk test
(3) Objective measures of job performance will demonstrate exponential distributions of job performance.	All groups – E, F, and G demonstrate all of the following: <ul style="list-style-type: none"> • Exponentially Distributed Histogram • Q-Q plot displays an upward curve • Non-significant Modified Shapiro-Wilk test • Significant Shapiro-Wilk test 	At least one group – E, F, or G demonstrates at least one of the following: <ul style="list-style-type: none"> • Exponentially Distributed Histogram • Q-Q plot displays an upward curve • Non-significant Modified Shapiro-Wilk test 	None of the groups – E, F, or G demonstrates any of the following: <ul style="list-style-type: none"> • Exponentially Distributed Histogram • Q-Q plot displays an upward curve • Non-significant Modified Shapiro-Wilk test
(4) A composite of subjective and objective job performance measures will demonstrate an exponential distribution of job performance.	Group H demonstrates all of the following: <ul style="list-style-type: none"> • Exponentially Distributed Histogram • Q-Q plot displays an upward curve • Non-significant Modified Shapiro-Wilk test • Significant Shapiro-Wilk test 	Group H demonstrates at least one of the following: <ul style="list-style-type: none"> • Exponentially Distributed Histogram • Q-Q plot displays an upward curve • Non-significant Modified Shapiro-Wilk test 	Group H does not demonstrate any of the following: <ul style="list-style-type: none"> • Exponentially Distributed Histogram • Q-Q plot displays an upward curve • Non-significant Modified Shapiro-Wilk test

Hypothesis 1

To test Hypothesis 1 (i.e., managerial ratings of job performance will be normally distributed), four methods were used. Managerial measures of job performance and the weighted average composite measure of managerial performance for groups A, B, and C were used to test Hypothesis 1. Means, standard deviations, skewness, and kurtosis are shown in Table 3. First, Groups A, B, and C were plotted as separate histograms and can be reviewed here respectively as Figures 4, 5, and 6. To evaluate these histograms SMEs provided ratings to indicate how well each histogram resembled a normal distribution as well as how well each histogram resembled an exponential distribution.

For Hypothesis 1, the average ratings provided by SMEs for how well each histogram resembled a normal distribution are as follows: Group A ($M = 2.57$), Group B ($M = 3.29$), and Group C ($M = 1.43$). The scale ranges from one through five where a lower rating indicates SME judgments of a better fit of the data to a normal distribution. Ratings were also provided by SMEs for how well each histogram resembles an exponential distribution and the average ratings are as follows: Group A ($M = 4.14$), Group B ($M = 3.00$), and Group C ($M = 4.86$). The scale ranges from one through five where lower ratings indicate SME judgments of a better fit of the data to an exponential distribution.

Based on the ratings from SMEs, none of the resulting histograms for Groups A, B, and C perfectly resembles a normal distribution and they possess limited characteristics of normality. Group B appears to deviate the most from normality, followed by Group A, while Group C appears to best resemble a normal distribution compared to groups A and

B individually. Furthermore, based on SME ratings, evidence also suggests that Groups A, B, and C are not exponential. This provides limited support for Hypothesis 1.

Table 3

Descriptives for Groups A, B, and C

Group	M	SD	Skewness	Kurtosis
Group A	3.30	0.51	0.32	-0.68
Group B	3.59	0.53	-0.57	-0.74
Group C	2.71	0.32	-0.08	-0.65

Note. $N = 203$. Group A = Manager Short-Term Ratings, Group B = Manager Long-Term Ratings, Group C = Weighted Average Composite of Manager Short-Term Ratings and Manager Long-Term Ratings.

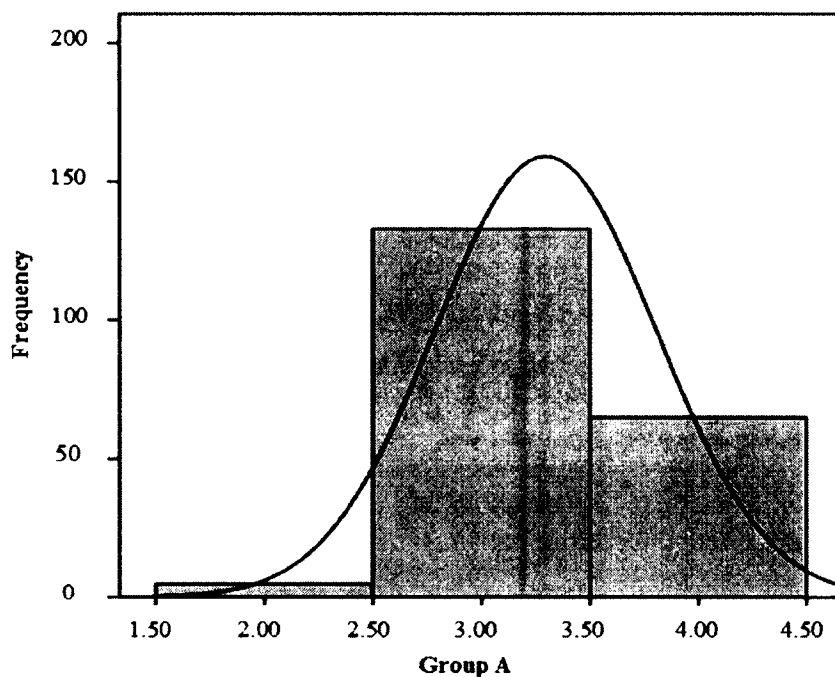


Figure 4. Frequency Distribution of Group A (Manager Short-Term Ratings of Job Performance). A higher score indicates better performance. The overlaying line indicates how closely the distribution resembles a normal distribution.

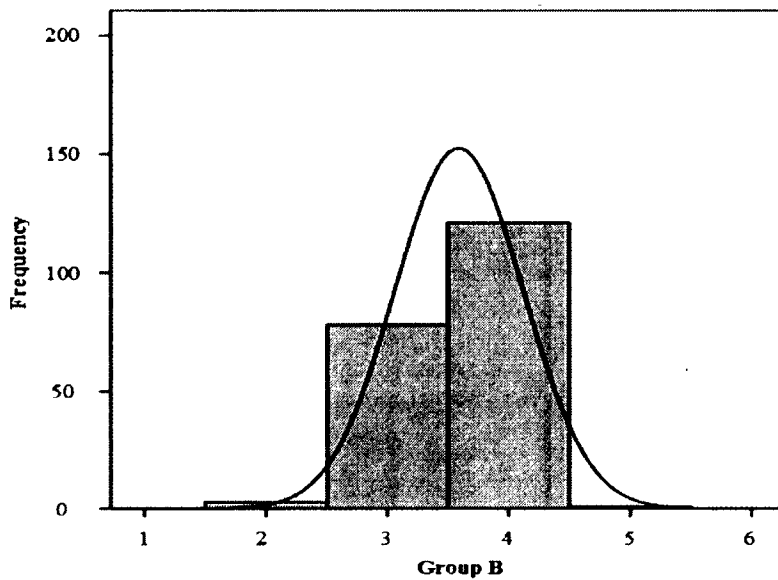


Figure 5. Frequency Distribution of Group B (Manager Long-Term Ratings of Job Performance). A higher score indicates better performance. The overlaying line indicates how closely the distribution resembles a normal distribution.

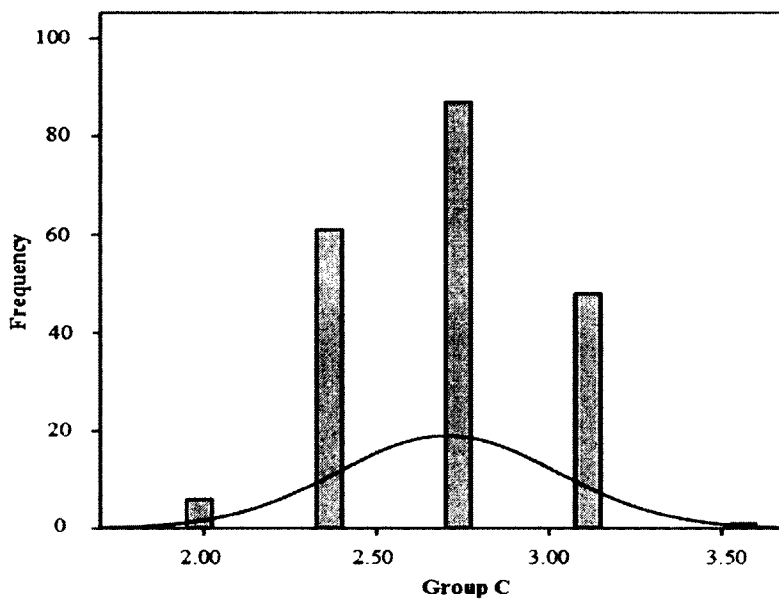


Figure 6. Frequency Distribution of Group C (Weighted Average Composite of Manager Short-Term Ratings and Manager Long-Term Ratings of Job Performance). A higher score indicates better performance. The overlaying line indicates how closely the distribution resembles a normal distribution.

The second criterion used to evaluate Hypothesis 1, how well each group of data fits a normal distribution, is a Q-Q plot. Q-Q plots were generated for Groups A, B, and C and are shown as Figures 7, 8, and 9 respectively. Each Q-Q plot was evaluated for how well it fit a straight line at a 45-degree angle. The better the data fit a straight line at a 45-degree angle, the stronger the evidence that the data are normally distributed. The average ratings provided by SMEs for how well each Q-Q plot fit a straight line at a 45-degree angle are as follows: Group A ($M = 3.00$), Group B ($M = 3.29$), and Group C ($M = 1.86$). Again, the scale ranges from one through five; the lower the rating the better the fit to a straight line at a 45-degree angle indicating better fit to a normal distribution. Ratings were also provided by SMEs for how much each Q-Q plot curves away from the 45-degree line, which indicates an exponential distribution. The average ratings are as follows: Group A ($M = 4.14$), Group B ($M = 3.00$), and Group C ($M = 4.86$). The scale ranges from one through five; the lower the rating the more curved the Q-Q plot, indicating the data are exponentially distributed.

All three plots for Groups A, B, and C tend to resemble a straight line at a 45-degree angle. However, the ratings are less polar for groups A and B. This is likely because only a few data points are produced for these Q-Q plots. The number of data points is limited because the number of potential scale points is limited (i.e., one, two, three, four, & five). Recall that only one data point is produced for each scale point that is within the sample. In turn, a fewer number of data points may make it more difficult to visually discern a pattern and as a result more difficult to interpret the Q-Q plots. However, consistent with evidence from the histograms for Groups A, B, and C, Group A and Group B have greater departures from normality than Group C. Evidence from the

Q-Q plots also suggests that the data are not exponential for any of the three groups.

These findings support Hypothesis 1. Yet, given the limited number of potential scale options within the sample, it may have been difficult to identify an exponential distribution if it did exist.

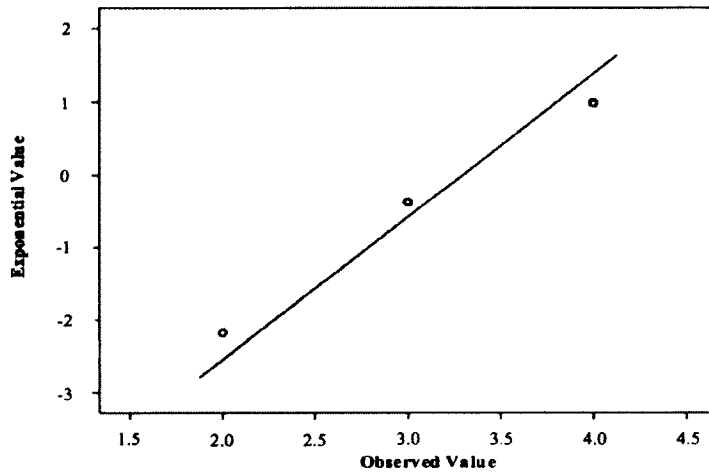


Figure 7. Q-Q plot of Group A (Manager Short-Term Ratings of Job Performance). The line through the center of the plot is at a 45-degree angle and indicates how closely the distribution resembles a normal distribution based on how closely the data points follow the angle of the line.

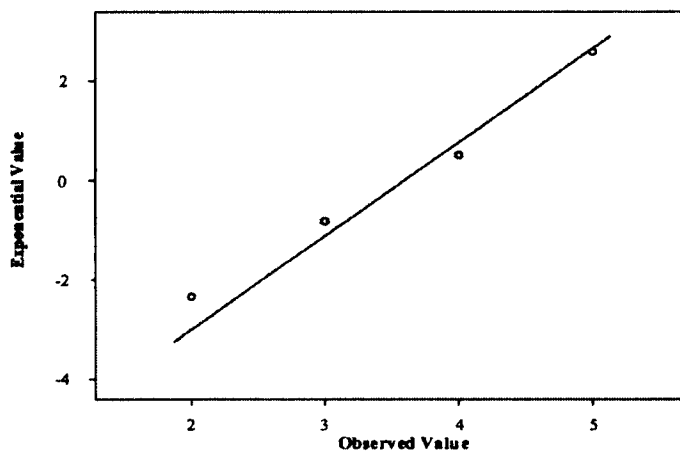


Figure 8. Q-Q plot of Group B (Manager Long-Term Ratings of Job Performance). The line through the center of the plot is at a 45-degree angle and indicates how closely the

distribution resembles a normal distribution based on how closely the data points follow the angle of the line.

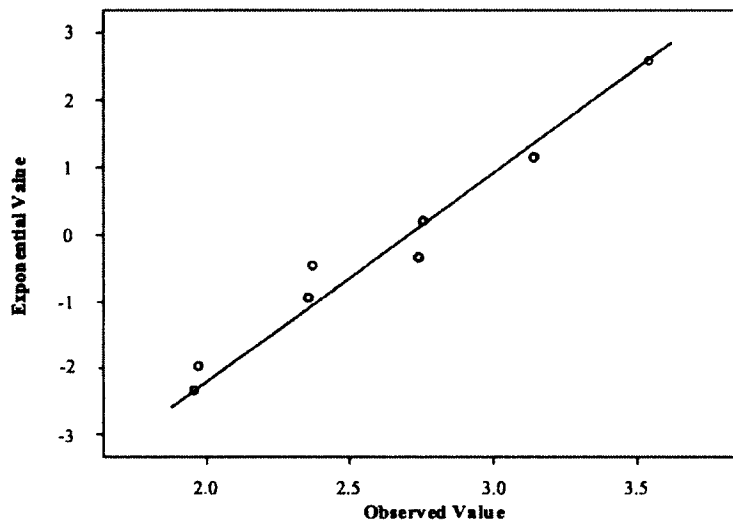


Figure 9. Q-Q plot of Group C (Weighted Average Composite of Manager Short-Term Ratings and Manager Long-Term Ratings of Job Performance). The line through the center of the plot is at a 45-degree angle and indicates how closely the distribution resembles a normal distribution based on how closely the data points follow the angle of the line.

The third criterion used to evaluate Hypothesis 1 was the Shapiro-Wilk test for normality. The Shapiro-Wilk test for normality tests the null hypothesis that the sample came from a normally distributed population. Therefore, a non-significant test statistic would provide further evidence that the data from Groups A, B, and C are normally distributed. However, all three Groups A ($W(203) = 0.67$), B ($W(203) = 0.68$), and C ($W(203) = 0.87$) are statistically significant ($p < .001$). These findings do not support Hypothesis 1.

A modified Shapiro-Wilk test for exponentiality was also used to test Groups A, B, and C. The modified Shapiro-Wilk test for exponentiality tests the null hypothesis that

the sample came from an exponential distribution. Therefore, a significant test statistic would provide additional support for Hypothesis 1. However, consistent with the evidence from the Shapiro-Wilk test for normality, the modified Shapiro-Wilk test for exponentiality was non-significant ($p = 1.00$) for all three Groups A ($W(203) = 0.03$), B ($W(203) = 0.04$), and C ($W(203) = 0.03$). These non-significant results provide support that the data from all three Groups A, B, and C may meet the criteria of an exponential distribution.

Results from the Shapiro-Wilk test for normality provide evidence that none of the three groups were normally distributed and findings from the modified Shapiro-Wilk test for exponentiality suggest that all three groups may be exponentially distributed. For groups A and B, the histograms and Q-Q plots corroborate evidence provided by the Shapiro-Wilk test for normality that groups A and B are not normally distributed. However, the histograms and Q-Q plots suggest that exponential may not be the most accurate classification for groups A and B, despite their non-significant modified Shapiro-Wilk test for exponentiality.

For Group C evidence provided by the histogram and Q-Q plot tends to conflict with evidence from the Shapiro-Wilk test for normality. Evidence from the histogram and Q-Q plot for Group C suggests that the data from this group may be more normally distributed than exponentially distributed. However, similar to groups A and B, it may be that Group C is also not best characterized by either a normal distribution or an exponential distribution. Interpreting all criteria holistically suggests that Groups A, B, and C are not best characterized by a normal distribution nor an exponential distribution,

but do present partial characteristics of both distributions. As a result, only limited support for Hypothesis 1 was found.

Hypotheses 2, 3, and 4

A similar procedure used to evaluate Hypothesis 1 was used to evaluate Hypotheses 2, 3, and 4, but different outcomes were predicted based on each hypothesis. Hypothesis 1 hypothesized normal distributions to result, whereas Hypotheses 2, 3, and 4 hypothesized exponential distributions. Hypothesis 2 used Group D; Hypothesis 3 used Groups E, F, and G; and Hypothesis 4 used Group H.

The first step to evaluate Hypothesis 2, Hypothesis 3, and Hypothesis 4 was to repeat the procedures used to generate histograms and Q-Q plots. However, a different result was hypothesized for both the histogram and Q-Q plot criteria when evaluating Hypotheses 2, 3, and 4. When plotting the data for Groups D, E, F, G, and H the hypothesized distribution was expected to resemble an exponential distribution (Figure 2). If the resulting distributions resembled an exponential distribution, the null hypotheses would be rejected, providing support for Hypotheses 2, 3, and 4. Similarly, if a Q-Q plot was rated as curved, this was evidence that an exponential distribution was present and was further evidence to support Hypotheses 2, 3, and 4.

Finally, in addition to the visual inspections used to evaluate Hypotheses 2, 3, and 4, the Shapiro-Wilk test for normality was used to verify that each group was not normally distributed and a modified Shapiro-Wilk test for exponentiality was used to determine whether the distribution of each group was exponential. A significant (p value of less than .05) Shapiro-Wilk test for normality and a non-significant (p value of .05 or

greater) modified Shapiro-Wilk test for exponentiality would provide additional support for Hypotheses 2, 3, and 4.

Means, standard deviations, skewness, and kurtosis for groups D, E, F, G, and H are reported in Table 4. Figures 10, 11, 12, 13, and 14 are histograms that correspond to Groups D, E, F, G, and H respectively and represent the first criterion used to evaluate Hypotheses 2, 3, and 4. Each histogram found in Figures 10, 11, 12, 13, and 14 was evaluated by SMEs based on how well it resembled an exponential distribution versus a normal distribution. Figures 15, 16, 17, 18, and 19 are Q-Q plots that correspond to Groups D, E, F, G, and H respectively and represent the second criterion used to evaluate Hypotheses 2, 3, and 4. Each Q-Q plot found in Figures 15, 16, 17, 18, and 19 was evaluated by SMEs according to how much it curved away from a 45-degree line versus followed a 45-degree line. Curving away from a 45-degree line indicates the presence of an exponential distribution.

Table 4

Descriptives for Groups D, E, F, G, and H

Group	M	SD	Skewness	Kurtosis
Group D	2.22	0.16	-0.11	-0.42
Group E	3.87	0.68	-2.13	6.71
Group F	4.12	0.69	-2.54	8.70
Group G	3.26	0.46	-1.68	4.02
Group H	1.92	0.15	-1.12	2.89

Note. $N = 203$. Group D = Weighted Average Composite of all four subjective ratings, Group E = Average Time in Role, Group F = Average Time Prior to Promotion, Group G = Weighted Average Composite of Average Time in Role and Average Time Prior to Promotion, Group H = Weighted Average Composite of all indicators of Performance.

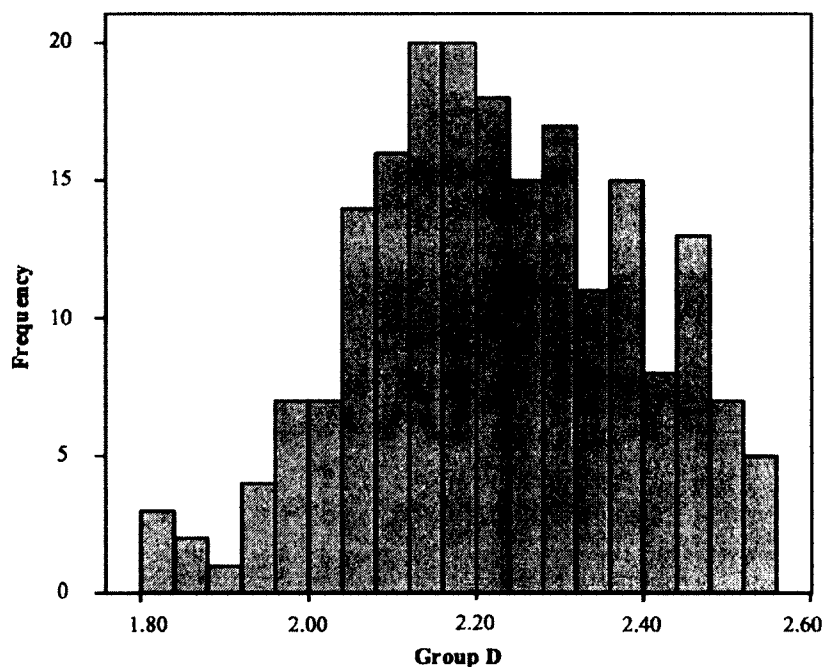


Figure 10. Frequency Distribution of Group D (Weighted Average Composite of all four subjective ratings). A higher score indicates better performance.

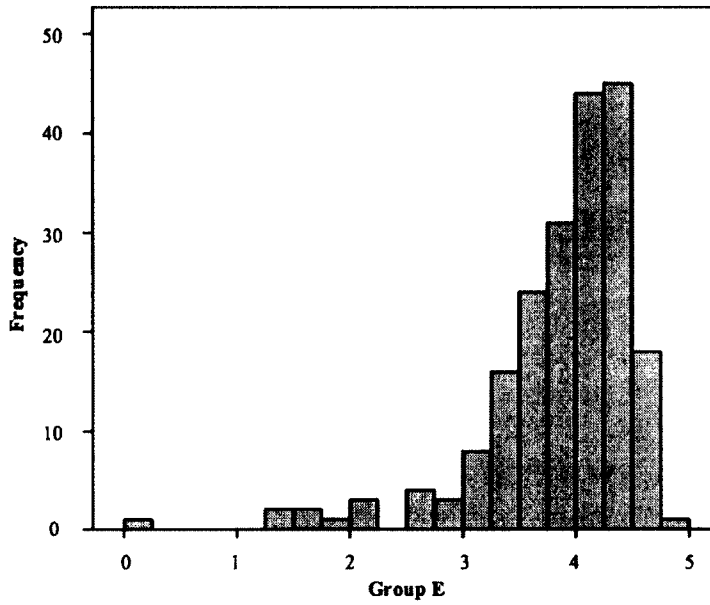


Figure 11. Frequency Distribution of Group E (Average Time in Role). To create this group average time in role was reverse scored and scaled down to use the same measurement scale as the subjective measures. A higher score indicates a lower average time spent in role and better performance.

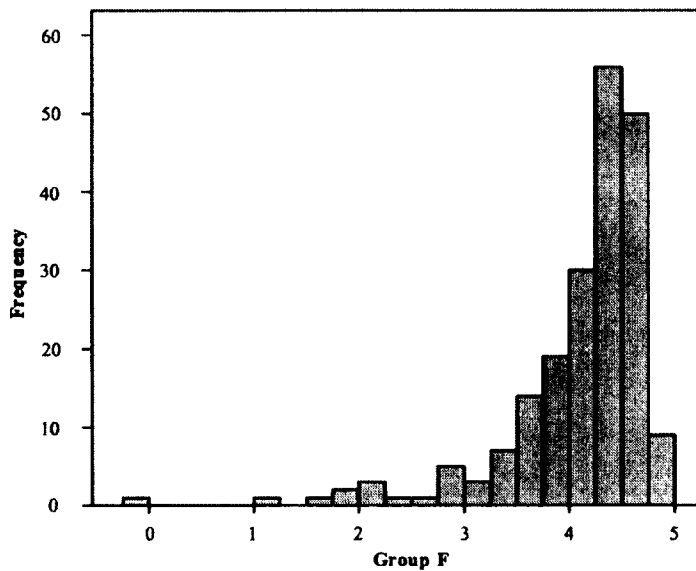


Figure 12. Frequency Distribution of Group F (Average Time Prior to Promotion). To create this group average time prior to promotion was reverse scored and scaled down to

use the same measurement scale as the subjective measures. A higher score indicates a lower average time spent in a role prior to promotion and better performance.

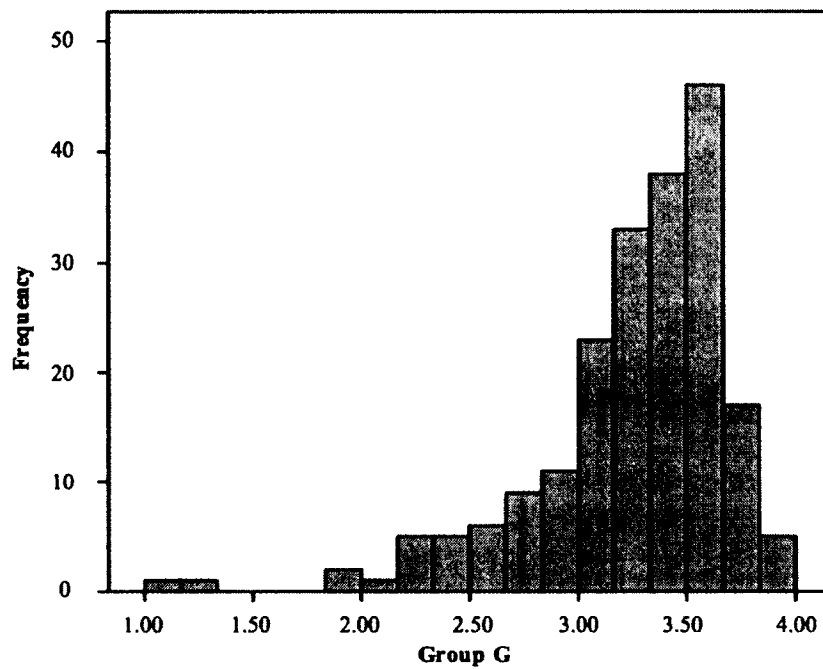


Figure 13. Frequency Distribution of Group G (Weighted Average Composite of Average Time in Role and Average Time Prior to Promotion). A higher score indicates better performance.

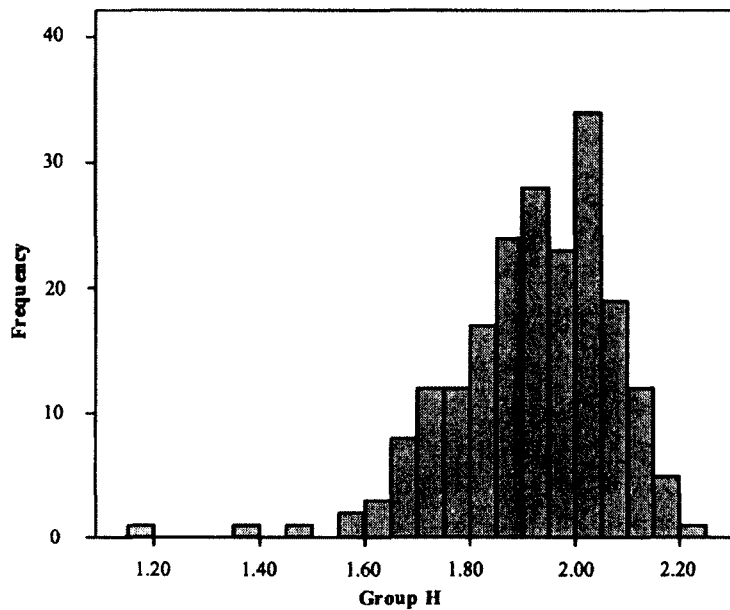


Figure 14. Frequency Distribution of Group H (Weighted Average Composite of all indicators of Performance). A higher score indicates better performance.

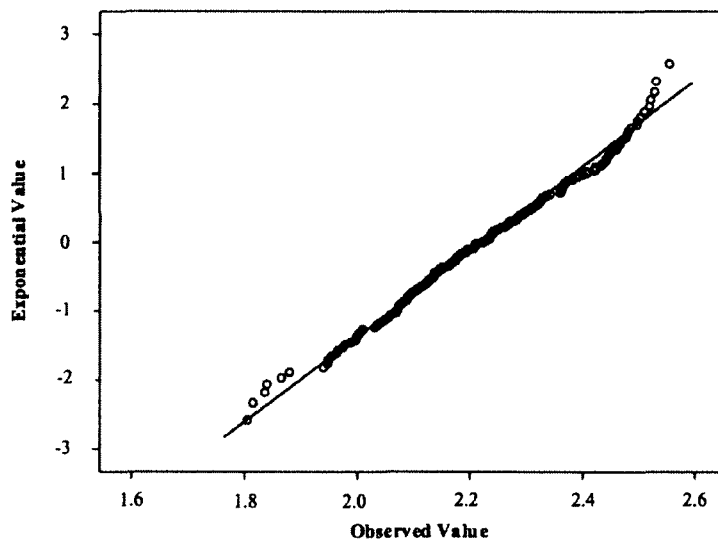


Figure 15. Q-Q plot of Group D (Weighted Average Composite of all four subjective ratings). The line through the center of the plot is at a 45-degree angle. The closer the data follow this line the more likely the data are normally distributed. However, the more the data points depart from this line, curving upwards or curving downwards on the tails, provides evidence that the data are exponentially distributed.

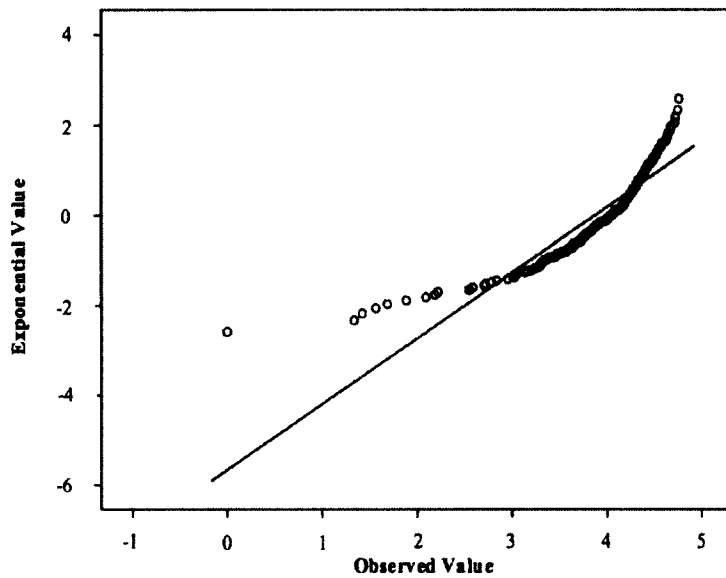


Figure 16. Q-Q plot of Group E (Average Time in Role). The line through the center of the plot is at a 45-degree angle. The closer the data follow this line the more likely the data are normally distributed. However, the more the data points depart from this line, curving upwards or curving downwards on the tails, provides evidence that the data are exponentially distributed.

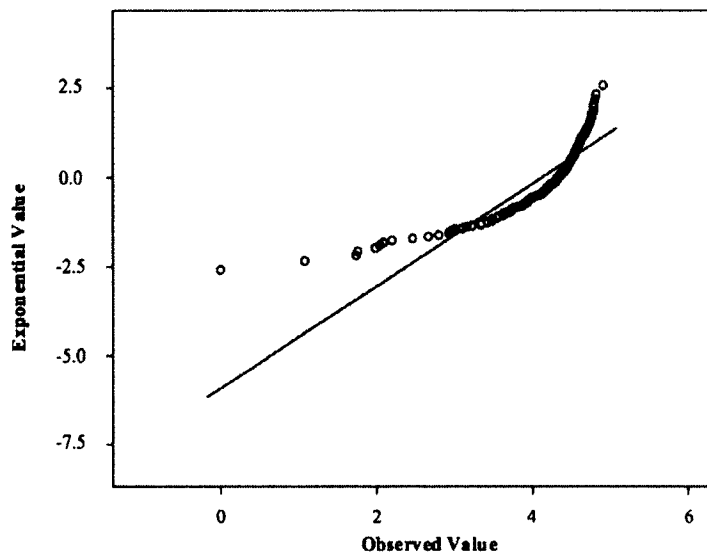


Figure 17. Q-Q plot of Group F (Average Time Prior to Promotion). The line through the center of the plot is at a 45-degree angle. The closer the data follow this line the more likely the data are normally distributed. However, the more the data points depart from

this line, curving upwards or curving downwards on the tails, provides evidence that the data are exponentially distributed.

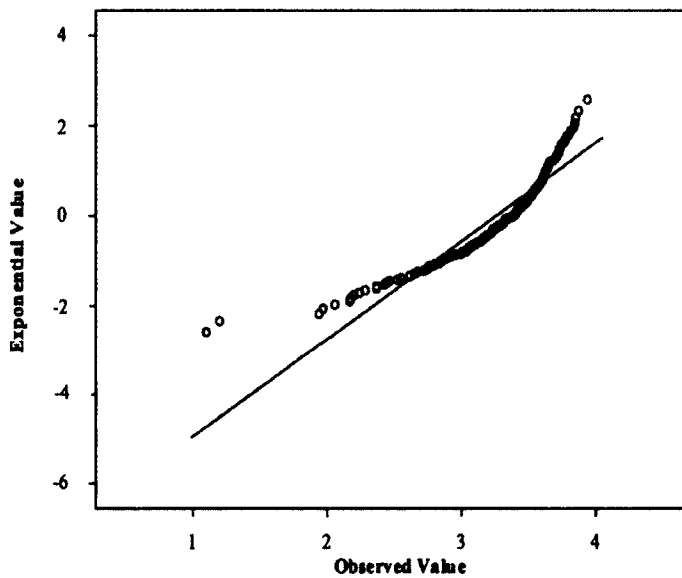


Figure 18. Q-Q plot of Group G (Weighted Average Composite of Average Time in Role and Average Time Prior to Promotion). The line through the center of the plot is at a 45-degree angle. The closer the data follow this line the more likely the data are normally distributed. However, the more the data points depart from this line, curving upwards or curving downwards on the tails, provides evidence that the data are exponentially distributed.

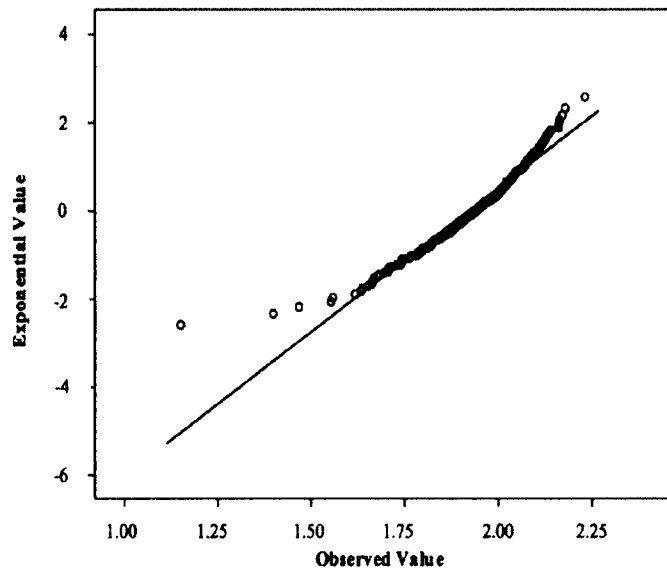


Figure 19. Q-Q plot of Group H (Weighted Average Composite of all indicators of Performance). The line through the center of the plot is at a 45-degree angle. The closer the data follow this line the more likely the data are normally distributed. However, the more the data points depart from this line, curving upwards or curving downwards on the tails, provides evidence that the data are exponentially distributed.

Hypothesis 2

Group D was used to test Hypothesis 2; a composite multi-rater multi-method measure of subjective job performance indicators will demonstrate an exponential distribution of job performance. A mean rating of 3.43 was provided by SMEs for how well the histogram from Group D resembles a normal distribution. The average rating suggests that the data from Group D do not strongly resemble a normal distribution. A mean rating of 4.00 was also provided by SMEs for how well the histogram resembles an exponential distribution. Together these ratings suggest that the Group D may not be normally distributed, but also that Group D is not exponentially distributed. This does not provide support for Hypothesis 2.

The Q-Q plot for Group D found in Figure 15 provides limited evidence that the data are not normally distributed and may be exponential (Group D – data points perfectly follow a 45-degree line, $M = 3.14$; data points significantly curve away from a 45-degree line, $M = 2.71$). The Q-Q plot appears to have a slight upward curve indicating a potential departure from normality, but it does not appear to be conclusive evidence of exponentiality. Thus, only limited support for Hypothesis 2 is provided by the Q-Q plot.

Finally, the Shapiro-Wilk test for normality was non-significant ($W(203) = 0.99$; $p = .096$) and the modified Shapiro-Wilk test for exponentiality was also not significant ($W(203) = 0.03$; $p = 1.00$). These findings directly conflict because the Shapiro-Wilk test for normality suggests that the data are normally distributed while the modified Shapiro-Wilk test for exponentiality suggests that the data are exponentially distributed. Group D does appear to have a few extreme cases in its tails that can be characteristic of an exponential distribution and could conceivably influence the test statistic produced by the modified Shapiro-Wilk test for exponentiality. This is an example of why it is important to graph and visually examine data and not rely solely on statistical tests. Statistical tests, in rare cases, can be influenced by special cases of data. Based on the findings for Group D, one of the two Shapiro-Wilks tests is producing results leading to Type II error. As a result, no support for Hypothesis 2 was found based on this evidence.

Overall, the criteria used to test Hypothesis 2 produced mixed results. The ratings for Group D's histogram and Q-Q plot suggest that the data are not well characterized as normal and also not characterized well as exponential. Group D's histogram and Q-Q plot both suggest that Group D possesses characteristics of both normal and exponential distributions. This finding complements and may explain the findings from both of the

Shapiro-Wilk tests that suggest that the data are both normal and exponential. As a result, no clear support for Hypothesis 2 was found.

Hypothesis 3

Groups E, F, and G were used to test Hypothesis 3, objective measures of job performance will demonstrate exponential distributions of job performance scores. The average ratings provided by SMEs for how well each histogram resembles a normal distribution are as follows: Group E ($M = 4.43$), Group F ($M = 4.86$), and Group G ($M = 4.86$). The scale ranges from one through five where the lower the rating the more normal the histogram. Ratings were also provided by SMEs for how well each histogram resembles an exponential distribution and the average ratings are as follows: Group E ($M = 1.14$), Group F ($M = 1.14$), and Group G ($M = 1.14$). Again, the scale ranges from one through five where the lower the rating the more exponential the histogram. The ratings provided by the SMEs suggest that all three histograms for Groups E, F, and G (depicted in Figures 11, 12, and 13) are all exponentially distributed and not normally distributed. However, although it is evident that all three histograms are exponentially distributed, they also appear to be negatively skewed exponential distributions. Although directionality was not explicitly hypothesized, the justification provided for an exponential distribution argued in Chapter Two assumed a positively skewed exponential distribution, not a negatively skewed distribution. Based on this evidence, it is concluded that Hypothesis 3 is only partially supported.

The Q-Q plots for Groups E, F, and G (depicted in Figures 16, 17, and 18) are all curved and have substantial departures from a 45-degree line. Each Q-Q plot was evaluated for how much it curves away from the 45-degree line, which would indicate an

exponential distribution. The average ratings are as follows: Group E ($M = 1.14$), Group F ($M = 1.00$), and Group G ($M = 1.29$). The scale ranges from one through five where the lower the rating the more curved the Q-Q plot, indicating the data are exponentially distributed. Each Q-Q plot was also evaluated on how well it fit a straight line at a 45-degree angle. The average ratings provided by SMEs for how well each Q-Q plot fit a straight line at a 45-degree angle are as follows: Group E ($M = 5.00$), Group F ($M = 4.86$), and Group G ($M = 4.86$). Again, the scale ranges from one through five where the lower the rating the better the fit to a straight line at a 45-degree angle indicating better fit to a normal distribution.

Based on the ratings provided by the SMEs it is clear that all three Q-Q plots curve away from a 45-degree line indicating exponential distributions. However, the Q-Q plots curve on both ends with greater curves on the lower tails. Although this does suggest the data in all three groups are exponential, the data are more negatively skewed than positively skewed. Based on this evidence, it is concluded that Hypothesis 3 is only partially supported.

The modified Shapiro-Wilk tests for exponentiality for Group E ($W(203) = 0.16$, $p = 1.00$), Group F ($W(203) = 0.17$, $p = 1.00$), and Group G ($W(203) = 0.11$, $p = 1.00$) were not significant. This provides evidence that Groups E, F, and G are exponentially distributed and provides support for Hypothesis 3. Groups E, F, and G were also tested using the Shapiro-Wilk test for normality. Group E ($W(203) = 0.82$), Group F ($W(203) = 0.75$), and Group G ($W(203) = 0.87$) were all significant ($p < .001$). This corroborates evidence from the modified Shapiro-Wilk test for exponentiality and provides further support that these distributions are not normally distributed.

In summary, all three criteria used to test Hypothesis 3 provide consistent evidence that Groups E, F, and G are exponentially distributed. However, after reviewing the histograms for all three groups it is clear that these groups are negatively skewed and not positively skewed. This finding conflicts with the argument outlined in Chapter Two that explains the rationale for anticipating positively skewed exponential distributions in performance data. Although, directionality was not specified in Hypothesis 3, it is appropriate to conclude that Hypothesis 3 was only partially supported.

Hypothesis 4

Group H was used to test Hypothesis 4, a composite of subjective and objective job performance measures will demonstrate an exponential distribution of job performance. An average rating of 3.71 was provided by SMEs for how well the histogram from Group H resembles a normal distribution. This average rating suggests that the data from Group H do not resemble a normal distribution. An average rating of 2.86 was also provided by SMEs for how well the histogram resembles an exponential distribution. These ratings indicated that the histogram for Group H (Figure 14) was slightly exponential. However, this histogram was again negatively skewed. Although directionality was not explicitly hypothesized, the justification provided for an exponential distribution argued in Chapter Two would have assumed a positively skewed exponential distribution, not a negatively skewed distribution. Based on the evidence from this criteria, it is concluded that Hypothesis 4 is only partially supported.

The Q-Q plot for Group H (Figure 19) is curved and does depart from a 45-degree line (Group H – data points perfectly follow a 45-degree line, $M = 4.29$; data points significantly curve away from a 45-degree line, $M = 1.57$). The curved plot does indicate

an exponential distribution. However, the Q-Q plot curves on both ends with a greater curve on the lower tail. Although this does suggest the data for Group H is exponential, it appears that the data are more negatively skewed than positively skewed. Based on this evidence, it is concluded that Hypothesis 4 is only partially supported.

In addition, the modified Shapiro-Wilk test for exponentiality ($W(203) = 0.13$) was not significant ($p = 1.00$)³ and the Shapiro-Wilk test of normality ($W(203) = 0.95$) was significant ($p < .001$). This provides support for Hypothesis 4, that Group H is exponentially distributed.

All three criteria used to test Hypothesis 4 indicate that Group H was exponentially distributed. However, given the directionality of the distribution for Group H, it is concluded that Hypothesis 4 is only partially supported. Similar to the other hypotheses, Hypothesis 4 did not specify directionality, but based on the arguments for exponentiality in Chapter Two a positively skewed exponential distribution would have been assumed.

Summary of Results

Some evidence supporting all four hypotheses was found. Figure 20 provides a summary of ratings provided by SMEs for each group of performance indicators. Hypotheses 1 was only partially supported. Graphical evidence was found that managerial ratings of performance may better resemble a normal opposed to exponential distribution of performance. However, evidence from test statistics also suggest that the managerial ratings used in the current study did depart, to some extent, from normality and may have characteristics of an exponential distribution.

³ A p -value of 1.00 is a convention of the software program used to calculate the Shapiro-Wilk Test. A 1.00 p -value means that the p -value was so close 1.00 that the program rounded it to 1.00 similar to how a p -value that is approaching 0.00 is typically rounded to 0.00.

Hypothesis 2 had the least support of all four hypotheses. Graphical, as well as statistical evidence, produced mixed results. The lack of evidence found for this hypothesis may suggest that for the present data, multi-trait multi-rater indicators of performance produce a distribution that is best characterized by neither a normal or an exponential distribution. Instead, it may be concluded that this type of data results in an exponentially modified Gaussian distribution that presents characteristics of both exponential and normal distributions.

The strongest support was found for Hypothesis 3 and Hypothesis 4, suggesting that objective indicators of performance, as well as composite indicators of performance that include both objective as well as subjective indicators of performance, are most likely to be exponentially distributed. All criteria used to evaluate Hypothesis 3 and Hypothesis 4 consistently indicated the presence of exponential distributions.

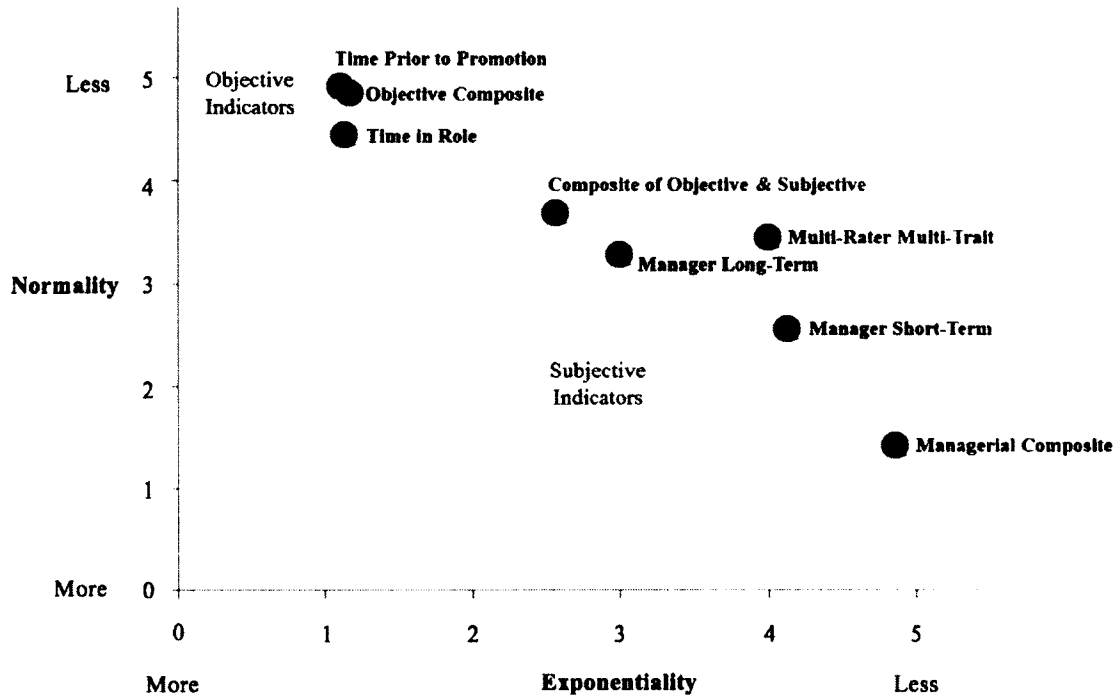


Figure 20. Graph of all ratings provided by SMEs for each indicator of performance. The x-axis represents ratings provided for how well each histogram represents an exponential distribution. The y-axis represents ratings provided for how well each histogram represents a normal distribution. This graph demonstrates a continuous trend between exponentiality and normality as indicators of performance move from objective to subjective.

CHAPTER FIVE

DISCUSSION

Partial support was found for all four hypotheses. Evidence suggests that ratings of performance provided by managers may be characterized as normally distributed.

Evidence also suggests that objective and composite indicators of performance possess characteristics of exponential distributions. Although composite and exponential distributions were negatively skewed, there was still evidence that these indicators were exponentially distributed and provided greater differentiation between top and bottom performers. It is noted that previous research (e.g., O'Boyle & Aguinis, 2012; Aguinis & O'Boyle, 2014), demonstrated positively skewed exponential distributions and the current study argued for the existence of exponential distributions of performance indicators based on this previous research. Thus, the current findings directionally (i.e., negatively skewed distributions) conflict with the directionality of previous research findings (i.e., positively skewed distributions) which may impact the interpretation of results and future theory, but should not negate the value of the current findings (i.e., the exponentiality rather than normality of the distributions).

There are various situational, as well as theory driven explanations, that help explain why current findings obtained negatively skewed opposed to positively skewed distributions of performance. A review of two theoretical explanations followed by

potential situational explanations may help understand these results. Theory-based explanations of negatively skewed distributions will be reviewed in terms of Attraction, Selection, Attrition Theory (ASA; Schneider, 2001), which has close ties to person-environment fit (Kristof-Brown, Zimmerman, & Johnson, 2005) and goal-setting theory (Locke & Latham, 1990).

ASA Theory argues that job applicants that have the best fit to an organization are more likely to apply to work for the organization, are more likely to be selected by the organization, are more likely to have longer tenure with the organization, and are more likely to demonstrate better performance (Schneider, 1987; 2001; Ployhart, Weekley, & Baughman, 2006). The concept of fit has also been generally recognized as a foundation for employee behavior (Saks & Ashforth, 1997).

Based on a meta-analysis by Kristof-Brown, et. al., (2005), there can be many different kinds of fit, such as fit to an organization, fit to a specific job, fit to a manager, and fit to team, to name a few. More specifically, fit can mean the congruence between a person's personality, values, interests, knowledge, skills, and abilities with an organization. This meta-analysis also demonstrated a direct link between fit and job performance, as well as many additional outcomes that are related to better performance such as increases in organizational commitment, job satisfaction, and lower turnover rates.

Through the lens of ASA Theory and fit, it would be assumed that an organization is more likely to possess many employees with a high degree of fit and a fewer number of employees with less fit. Employees with less fit would be more likely to self-select themselves out of an organization, thus leaving a larger number of good fitting

employees. In terms of a negatively skewed performance distribution, this would mean a larger number of employees with good fit and better performance and a decreasing number of employees with less fit and potentially lower performance. This is one potential explanation as to why a negatively skewed performance distribution may be more likely to occur.

The second theory that may explain why a negatively skewed distribution of performance was found instead of a positively skewed distribution of performance is Goal-Setting Theory (Locke & Latham 1990; 2002). In its most basic form, Goal-Setting Theory argues that when goals are specific and difficult, they can lead to higher levels of performance (Locke & Latham, 2006). In the current study, all employees annually created difficult specific goals. Recall that employees created both short-term and long-term goals each year that were evaluated here as managerial ratings of short-term and long-term performance. Although exponential distributions of performance were not found for managerial ratings based on the sole performance of these goals, when performance was evaluated holistically using a composite measure comprised of both objective and subjective indicators of performance, a negatively skewed performance distribution was found.

It seems reasonable to conclude that composite measures of performance, which are likely measuring more of the performance construct space (Viswesvaran, 1992), are also more likely to reflect the impact that the use of difficult specific goals have on performance. Thus, the use of difficult specific goals by everyone in the sample may explain why there were a large number of high performing employees and a smaller number of lower performing employees, resulting in a negatively skewed exponential

distribution. If an organization did not require employees to set difficult specific goals each year, then it is possible that there would be fewer top performers and greater attenuation leading to a positively skewed performance distribution.

Considering the potential impact fit and goal-setting may simultaneously have on distributions of performance may help explain why negatively skewed distributions were found in contrast to previous research (e.g., O'Boyle & Aguinis, 2012). The use of difficult specific goals and the lens of ASA Theory both predict higher levels of performance within organizations. Given that a minority group of employees within an organization are likely to exist that have decreasing degrees of fit and that difficult specific goals do not perfectly predict performance, it may be reasonable to expect the majority of employees will demonstrate high levels of performance. Simultaneously it would be expected that the minority group of employees would demonstrate decreasing levels of performance, resulting in a negatively skewed exponential distribution opposed to positively skewed exponential distribution.

With few exceptions (e.g., Micceri, 1989; Murphy, 1999; Saal, Downey, & Lahey, 1980), I-O psychological research has historically assumed job performance to be normally distributed (Murphy, 2008). Previous literature lacked methods for critically evaluating and classifying distributions of job performance. This study successfully evaluated and classified eight separate groups of performance indicators and examined several methods that can be used by organizations and researchers to determine the distributions of performance data. This study has also demonstrated the impact that different types of indicators of performance can have on observed distributions of performance (see Figure 20). Although all four hypotheses were only partially supported,

current findings demonstrate that combining multiple indicators of performance and leveraging objective indicators of performance can produce greater differentiation between top and bottom performing employees. By combining multiple indicators of performance and by incorporating objective indicators of performance, greater variance was achieved in each distribution. Greater variance inherently allows for greater differentiation, but also altered each distribution such that fewer people were clustered around the center of each distribution and more people were in the tail of the distribution. As a result, this created greater differentiation between top and bottom performers.

Perhaps more importantly, this study revealed a continuum from normal distributions to exponential distributions between multiple types of indicators of job performance, which might represent a conceptual framework for a new classification scheme of job performance measures. This study demonstrated that managerial ratings were most likely to resemble normal distributions, composite subjective indicators were most likely to possess characteristics of normal as well as exponential distributions, and objective indicators were most likely to resemble exponential distributions. Furthermore, composite scores comprised of both objective and subjective measures, which arguably may provide the most accurate measurement of performance (e.g., Viswesvaran, 1993), presented more characteristics of an exponential distribution than a normal distribution. As a result, this study provides insight as to why distributions of job performance have been the subject of debate. Instead of job performance possessing an innate normal distribution, it may be more appropriate to conclude that the observed distribution of job performance is influenced by the types of indicators being used to measure job performance.

Theoretical and Practical Implications

If, as the current study suggests, the type of indicator used to measure job performance does influence the observed distribution of performance, this finding could have various implications. For example, this finding could impact how to choose the best type of analysis when validating selection systems or how to choose the best performance indicators when planning for succession. Present findings may also have implications for future theory related to distributions of performance. For instance, when attempting to validate a selection system, if only subjective indicators are used, it may be most appropriate to use traditional methods and statistics such as linear regression. However, if objective measures of performance are used alone as criteria or in combination with subjective measures, alternative methods and statistics may be more appropriate for validation (e.g., non-parametric tests or non-linear regression). By matching statistical tests to the type of performance measure used as criteria in the validation study or (even more directly) to the observed performance distribution, it may be possible to enhance the validity for selection systems.

In terms of succession planning or being able to better differentiate employee performance, these findings suggest that organizations may be better off leveraging objective indicators of performance or composite measures that account for objective as well as subjective measures. Leveraging objective indicators of performance or composite indicators comprised of objective and subjective performance indicators may achieve greater differentiation between employees. In turn, the use of these types of measures may make it easier for organizations to identify gaps in their talent.

If organizations can better understand their distributions of performance and achieve greater differentiation between top and bottom performers, then they may be better equipped to strategically utilize resources to improve performance. For instance, if the same methods utilized in the current study were applied, organizations could evaluate performance distributions across various functions and roles within their organizations. Through these analyses, organizations would be able to differentiate more easily between top and bottom performers across as well as within functions and roles. Organizations should be better equipped to identify some functions or roles for which they may have many top performers whereas in other functions or roles they may only possess a few top performers. Conversely, some organizations may find that in certain functions they possess a larger degree of low performing employees. Armed with the capability to better differentiate top and bottom performers, organizations will be better prepared to develop new human capital workforce strategies that can focus on retaining top performers and provide resources to improve the performance of the lowest performers. Organizations should also be able to plan for the future and enable activities such as succession planning to better identify where there are gaps in their talent and where they may need to devote resources.

The value of being able to differentiate between top and bottom performers may further be realized by comparing normally distributed indicators of performance such as the managerial ratings to the exponentially distributed indicators, such as the objective indicators and composites of objective and subjective indicators. When comparing these indicators, it becomes evident that when using only managerial ratings of performance not enough variance in performance is captured which results in top and bottom

performers being compressed into the middle part of the distribution (i.e., less differentiation). Objective and composite measures (comprised of objective and subjective indicators), on the other hand, demonstrate greater variance in performance, which helps to generate greater differentiation between employee performance and more easily identify top and bottom performers.

According to Gravett and Caldwell (2016), at a time when the war for talent is at an all-time high, being able to successfully differentiate top and bottom performers may have a large impact on the success of organizations. Instead of identifying a few star performers, these findings identified a large proportion of employees demonstrating a high degree of performance and provided greater differentiation among lower performing employees. One interpretation of these findings is that indicators of performance used in the current study do not differentiate well between employees that demonstrate a high degree of performance. It is also possible that there may be other factors, which were not measured in the current study, such as organizational culture, that may have attracted and retained a large number of top performers. These findings suggest that the organization from which the sample was derived may already be successfully retaining top performers. It is also important to note that in the population job performance could be normally distributed and that samples of employee performance are potentially bound and influenced by the organizations from which they are derived. As a result, the sample from which data are derived and the type of performance indicators used may both play a larger role in the shape of performance distributions than previously assumed. As a result, it may be important for researchers to identify and report potential organizational factors that could influence findings as well as the type of performance indicators leveraged

opposed to attempting to emphasize findings as general phenomenon that can be applied broadly.

Furthermore, given that the present findings were negatively exponentially skewed, the question is raised of whether or not it is reasonable to consistently expect any specific distribution of performance. Recall that Viswesvaran (1993) demonstrated that when combining multiple indicators of performance, more of the theoretical job performance construct space could be measured. Thus, it may be appropriate to assume that the most representative distribution of job performance is derived from multiple indicators of performance. In the current study, Group H represents the combination of multiple indicators of objective as well as subjective performance indicators. This distribution was found to be negatively exponentially skewed. This is in contradiction to previous research that has found positively skewed exponential distributions of performance (e.g., O'Boyle & Aguinis, 2012) and other research which argues that job performance should be normally distributed (e.g., Beck, Beatty, & Sackett, 2014). It is possible that present findings are exponential not because they accurately reflect job performance, but rather because they are severely influenced by error or criterion irrelevance. However, this explanation is improbable because multiple multi-rater, multi-trait and objective indicators were used. In turn, it may be possible that Group H, which was exponential, may best represent an innate distribution of performance.

If Group H is cautiously assumed to be an accurate representation of job performance then it becomes necessary to explore alternatives as to why current findings were not entirely consistent with previous research (e.g., O'Boyle & Aguinis, 2012). For instance, distributions of job performance may not have an innate shape or classification.

Instead, they may be a function of the quality and type of performance indicators, the age of an organization, an organization's ability to attract and retain different types of employees, and the organizations system for selecting employees. To elaborate, if an organization is desirable and offers employees many resources, it may be able to retain a larger proportion of top performers. In this instance, an organization may expect to find a negatively skewed exponential distribution of performance. If an organization has a valid selection system that has been in use for many years, the organization may also expect to find a larger proportion of top performing employees. However, if an organization has only recently started using a valid selection system, the organization may expect a distribution of performance that includes a lower proportion of top performers. Thus, there are many factors that may influence a distribution of job performance and it may not be appropriate to assume that job performance possess an innate distribution that is always normal or exponential regardless of whether or not the data is subjective, objective, or a combination of the two.

Additionally, previous research (e.g., O'Boyle & Aguinis, 2012; Aguinis, et. al., 2016) only relied on objective indicators of performance to identify exponential distributions which resulted in positive skewness. Although the objective indicators in the present study were negatively skewed, a negatively skewed distribution was also found when accounting for subjective indicators of performance. This means that directionality of skew in present findings could be in opposition to previous research because the present findings are accounting for more of the performance construct space. Therefore, it is possible that the present findings were negatively skewed opposed to positively skewed because more of the performance construct space is being measured.

Regardless of why current findings were negatively skewed, differentiation between top and bottom performers was achieved. This differentiation could be used by organizations to more easily identify a group of top performers and with greater ease justify allocation of resources to top performers opposed to low performers. In comparison to a normal distribution, the exponential distributions more clearly differentiate performance between top and bottom performers. Therefore, the results from this study can help organizations more easily identify their lowest performing employees and subsequently remove them or provide them with guidance to improve their performance.

In summary, this study demonstrates that the type of job performance indicator used to measure job performance can affect the observed distribution of job performance and, ultimately, the amount of differentiation between employees. This study also suggests that in terms of understanding distributions of performance, it may also be important for organizations to take into account not only the type of indicator used to measure job performance, but also organizational factors such as selection techniques, age of the organization, and prestige of the organization. Finally, this study reveals the possibility that there is a continuum of distributional forms that underlies job performance indicators that could lead to a new classification scheme for job performance measures.

Study Limitations

Although the results of the current study may have valuable practical and theoretical implications, certain limitations should be acknowledged. First, it is important to comment on the sample used in the current study. The sample was global, but it primarily consisted of employees from a large organization that worked within

information technology and held management positions. As a result, current findings may not generalize well to small organizations, other organizational functions, or non-managerial roles. Future research should attempt to replicate current findings in different organizations of varying size, different organizational functions, and a variety of non-managerial roles. Second, this study is the first of its kind. Only through additional research can the strength and generalizability of current findings be realized. Third, there was likely error in the measurement of job performance, an inherent issue in the measurement of job performance. By ensuring all employees included in the sample had been assessed on the same indicators of performance, the error was likely held constant across all measures allowing differences in distributions of performance to be more likely attributed to the indicator of performance rather than to error. Each indicator of performance may have possessed its own unique type of error, but this error would have then likely been constant across all employees.

Future Research

Based on limitations of the current study, future research should attempt to replicate current findings using samples derived from organizations of varying size, employees of varying organizational functions, as well as leverage non-managerial employees. Future research should also attempt to identify and test additional factors that could influence the distribution of job performance beyond the impact of measurement tools and error while accounting for the type of indicator used to measure job performance. Potential factors that future research could explore that may impact the shape of performance distributions include types of selection systems organizations use, if any, and for what length of time they have been implemented. Other factors may include prestige of the organization,

organizational climate and culture, and external factors such as the demand and opportunity for specific skills external to the organization. Once future research has tested the potential impact of additional factors on job performance, a typology of job performance distributions could be generated that could set a baseline for organizations regarding the type of distributions of performance expected. Current findings could be used as a starting point for a typology of expected performance distributions since the current research has demonstrated that different types of indicators of performance are more likely to generate different types of distributions. Finally, an open system perspective may be one theory used by future research to aid in the interpretation of results.

Conclusions

The current study demonstrates that job performance should not always be assumed to be normally distributed. The findings suggest that, at least in part, distributions of job performance may be influenced by the type of job performance indicator used to measure performance. Initial groundwork has been laid to help organizations better anticipate the type of distribution they can expect to find when leveraging different indicators of performance. This study also identifies high fidelity evaluation tools that can be used to evaluate future indicators of performance and the resulting distributions. Finally, this study provides direction to organizations to enable them to better differentiate between top and bottom performers.

REFERENCES

- Aguinis, H. (2013). *Performance management*. Upper Saddle River, NJ: Pearson/Prentice Hall.
- Aguinis, H., & O'Boyle, E. (2014). Star performers in twenty - first - century organizations. *Personnel Psychology*, 67(2), 313-350.
- Aguinis, H., O'Boyle, E., Gonzalez-Mulé, E., & Joo, H. (2016). Cumulative advantage: Conductors and insulators of heavy-tailed productivity distributions and productivity stars. *Personnel Psychology*, 69(1), 3-66.
- Alchian, A. A. (1950). Uncertainty, evolution, and economic theory. *The Journal of Political Economy*, 58(3), 211-221.
- Anastasi, A., & Urbina, S. (1997). *Psychological Testing*. New York, NY: Pearson.
- Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology*, 77(6), 836-874.
- Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology*, 77(6), 975-985.
- Bauer, C. C., & Baltes, B. B. (2002). Reducing the effects of gender stereotypes on performance evaluations. *Sex Roles*, 47(9), 465-476.
- Beck, J. W., Beatty, A. S., & Sackett, P. R. (2014). On the distribution of job performance: The role of measurement characteristics in observed departures from normality. *Personnel Psychology*, 67(3), 531-566.

- Becker, B. E., Huselid, M. A., Pickus, P. S., & Spratt, M. F. (1997). HR as a source of shareholder value: Research and recommendations. *Human Resource Management*, 36(1), 39-47.
- Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, 6(2), 205-212.
- Bernardin, H. J., LaShells, M. B., Smith, P. C., & Alvares, K. M. (1976). Behavioral expectation scales: Effects of developmental procedures and formats. *Journal of Applied Psychology*, 61(1), 75-79.
- Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored ratings scales (BARS). *Journal of Applied Psychology*, 66(4), 458-463.
- Bernardin, H. J., & Wiatrowski, M. (2013). Performance appraisal. In N. Brewer & C. Wilson (Eds.), *Psychology and Policing*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
- Bingham, W. V. (1926). Measures of occupational success. *Harvard Business Review*, 5(1), 1-10.
- Bingham, W. V., & Davis, W. T. (1924). Intelligence test scores and business success. *Journal of Applied Psychology*, 8(1), 1-22.
- Blume, B. D., Baldwin, T. T., & Rubin, R. S. (2009). Reactions to different types of forced distribution performance evaluation systems. *Journal of Business and Psychology*, 24(1), 77-91.
- Bolanovich, D. J. (1946). Statistical analysis of an industrial rating chart. *Journal of Applied Psychology*, 30(1), 23-31.

- Bolino, M. C., Hsiung, H.-H., Harvey, J., & LePine, J. A. (2015). "Well, I'm tired of tryin'!" Organizational citizenship behavior and citizenship fatigue. *Journal of Applied Psychology, 100*(1), 56.
- Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta - analysis. *Personnel Psychology, 48*(3), 587-605.
- Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20*(2), 238-252.
- Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63*(2), 135-144.
- Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64*(4), 410-421.
- Borman, W. C., & Dunnette, M. D. (1975). Behavior-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology, 60*(5), 561-565.
- Borman, W. C., Klimoski, R. J., & Ilgen, D. R. (2003). Stability and change in industrial and organizational psychology. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology* (Vol. 12, pp. 1-17). Canada: John Wiley & Sons, Inc.
- Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel Selection in Organizations* (pp. 71-98). San Francisco, CA: Jossey-Bass.

- Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance*, *10*(2), 99-109.
- Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001). Personality predictors of citizenship performance. *International Journal of Selection and Assessment*, *9*(1 - 2), 52-69.
- Boudreau, J. W., & Ramstad, P. M. (2007). *Beyond HR: The new science of human capital*. Boston, MA: Harvard Business Press.
- Boyle, M. (2001, May 28). Performance curves ahead. *Fortune*, 187.
- Brigham, C. C. (1930). Intelligence tests of immigrant groups. *Psychological Review*, *37*(2), 158-165.
- Brogden, H. E., & Taylor, E. K. (1950). The theory and classification of criterion bias. *Educational and Psychological Measurement*, *10*(2), 159-186.
- Brutus, S. (2010). Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. *Human Resource Management Review*, *20*(2), 144-157.
- Call, M. L., Nyberg, A. J., & Thatcher, S. (2015). Stargazing: An integrative conceptual review, theoretical reconciliation, and extension for star employee research. *Journal of Applied Psychology*, *100*(3), 623-640.
- Campbell, J. P. (2012). Behavior, performance, and effectiveness in the 21st century. In S. W. Kozlowski (Ed.), *The Oxford handbook of organizational psychology* (pp. 159-194). New York, NY: Oxford University Press.

- Campbell, J. P., & Campbell, R. J. (1988). *Productivity in organizations: New perspectives from industrial and organizational psychology*. San Francisco, CA: Jossey-Bass.
- Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. (1970). *Managerial behavior, performance, and effectiveness*. New York, NY: McGraw-Hill.
- Campbell, J. P., Mcloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel Selection in Organizations* (pp. 35-70). San Francisco, CA: Jossey-Bass.
- Campbell, J. P., & Wiernik, B. M. (2015). The modeling and assessment of work performance. *Annual Review of Organizational Psychology and Organizational Behavior*, 2, 47-74.
- Cannon, M. D., & Witherspoon, R. (2005). Actionable feedback: Unlocking the power of learning and performance improvement. *The Academy of Management Executive*, 19(2), 120-134.
- Carpenter, N. C., Berry, C. M., & Houston, L. (2014). A meta - analytic comparison of self - reported and other - reported organizational citizenship behavior. *Journal of Organizational Behavior*, 35(4), 547-574.
- Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. Boston, MA: Duxbury Press.
- Chattopadhyay, R., & Ghosh, A. K. (2012). Performance appraisal based on a forced distribution system: Its drawbacks and remedies. *International Journal of Productivity and Performance Management*, 61(8), 881-896.

- Choulakian, V., & Stephens, M. (2001). Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*, 43(4), 478-484.
- Coens, T., & Jenkins, M. (2002). *Abolishing performance appraisals: Why they backfire and what to do instead*. San Francisco, CA: Berrett-Koehler Publishers.
- Coleman, V. I., & Borman, W. C. (2000). Investigating the underlying structure of the citizenship performance domain. *Human Resource Management Review*, 10(1), 25-44.
- Combs, J., Liu, Y., Hall, A., & Ketchen, D. (2006). How much do high - performance work practices matter: A meta - analysis of their effects on organizational performance. *Personnel Psychology*, 59(3), 501-528.
- Conway, J. M. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology*, 84(1), 3-13.
- Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90(2), 218-244.
- Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17(4), 407-422.
- Crawford, G. C., Aguinis, H., Lichtenstein, B., Davidsson, P., & McKelvey, B. (2015). Power law distributions in entrepreneurship: Implications for theory and research. *Journal of Business Venturing*, 30(5), 696-713.
- Creager, J. A., & Harding Jr, F. D. (1958). A hierarchical factor analysis of foreman behavior. *Journal of Applied Psychology*, 42(3), 197-203.
- Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. New York, NY: Appleton and Company.

- Day, D. V., & Sulsky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80*(1), 158-167.
- Debnath, S. C., Lee, B. B., & Tandon, S. (2015). Fifty years and going strong: what makes behaviorally anchored rating scales so perennial as an appraisal method. *International Journal of Business and Social Science, 6*(2), 16-25.
- DeCotiis, T. A. (1977). An analysis of the external validity and applied relevance of three rating formats. *Organizational Behavior and Human Performance, 19*(2), 247-266.
- DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33*(3), 360-396.
- Devaraj, S., & Kohli, R. (2003). Performance impacts of information technology: Is actual usage the missing link. *Management Science, 49*(3), 273-289.
- Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. *Journal of Applied Psychology, 65*(2), 147-154.
- Dipboye, R. L. (1985). Some neglected variables in research on discrimination in appraisals. *Academy of Management Review, 10*(1), 116-127.
- Dunnette, M. D. (1963). A note on the criterion. *Journal of Applied Psychology, 47*(4), 251-254.
- Dunnington, G. W., Gray, J., & Dohse, F.-E. (2004). *Carl Friedrich Gauss: Titan of Science*. New York, NY: The Mathematical Association of America.

- Ewart, E., Seashore, S., & Tiffin, J. (1941). A factor analysis of an industrial merit rating scale. *Journal of Applied Psychology*, 25(5), 481-486.
- Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied psychology*, 66(2), 127-148.
- Feldman, J. M. (1986). A note on the statistical correction of halo error. *Journal of Applied Psychology*, 71(1), 173-176.
- Ferris, G. R., Yates, V. L., Gilmore, D. C., & Rowland, K. M. (1985). The influence of subordinate age on performance ratings and causal attributions. *Personnel Psychology*, 38(3), 545-557.
- Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, 51(4), 327-358.
- Frazier, M. L., & Bowler, W. M. (2015). Voice climate, supervisor undermining, and work outcomes: A group-level examination. *Journal of Management*, 41(3), 841-863.
- Fryer, D. (1922). Occupational-intelligence standards. *School & Society*, 16, 273-277.
- Fryer, D. (1935). Intelligence tests in industry. *Personnel Journal*, 13, 321-323.
- Gajendran, R. S., Harrison, D. A., & Delaney - Klinger, K. (2015). Are telecommuters remotely good citizens: Unpacking telecommuting's effects on performance via i-deals and job resources. *Personnel Psychology*, 68(2), 353-393.
- Ganzach, Y. (1995). Negativity (and positivity) in performance evaluation: Three field studies. *Journal of Applied Psychology*, 80(4), 491-499.

- Goffin, R. D., Gellatly, I. R., Paunonen, S. V., Jackson, D. N., & Meyer, J. P. (1996). Criterion validation of two approaches to performance appraisal: The behavioral observation scale and the relative percentile method. *Journal of Business and Psychology, 11*(1), 23-33.
- Goffin, R. D., Jelley, R. B., Powell, D. M., & Johnston, N. G. (2009). Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management, 48*(2), 251-268.
- Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly, 19*(2), 213-236.
- Gould, S. J. (1984). *The mismeasure of man*. New York, NY: Norton.
- Grant, D. L. (1955). A factor analysis of managers' ratings. *Journal of Applied Psychology, 39*(4), 283-286.
- Gravett, S. L., & Caldwell, S. A. (2016). *Learning agility: The impact on recruitment and retention*. New York, NY: Nature America Inc.
- Green, S. G., & Mitchell, T. R. (1979). Attributional processes of leaders in leader—member interactions. *Organizational Behavior and Human Performance, 23*(3), 429-458.
- Greene, L., Bernardin, H. J., & Abbott, J. (1985). A comparison of rating formats after corrections for attenuation. *Educational and Psychological Measurement, 45*(3), 503-515.
- Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions*. New York, NY: Taylor & Francis.

- Hantula, D. A. (2011). What performance management needs is a good theory: A behavioral perspective. *Industrial and Organizational Psychology, 4*(2), 194-197.
- Hauenstein, N. M. (2011). Improving performance management: Take my golf game, please. *Industrial and Organizational Psychology, 4*(02), 176-178.
- Hauenstein, N. M., & Foti, R. J. (1989). From laboratory to practice: Neglected issues in implementing frame-of-reference rater training. *Personnel Psychology, 42*(2), 359-378.
- Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology, 73*(1), 68-73.
- Heiman, G. (2013). *Basic statistics for the behavioral sciences*. Belmont, CA: Cengage Learning.
- Heneman, R. L., Moore, M. L., & Wexley, K. N. (1987). Performance-rating accuracy: A critical review. *Journal of Business Research, 15*(5), 431-448.
- Hodson, R. (1999). Management citizenship behavior: A new concept and an empirical test. *Social Problems, 46*(3), 460-478.
- Hoffman, B. J., Gorman, C. A., Blair, C. A., Meriac, J. P., Overstreet, B., & Atchley, E. (2012). Evidence for the effectiveness of an alternative multisource performance rating methodology. *Personnel Psychology, 65*(3), 531-563.
- Huber, V. L. (1983). An analysis of performance appraisal practices in the public sector: A review and recommendations. *Public Personnel Management, 12*(3), 258-267.
- Hull, C. L. (1928). *Aptitude testing*. Yonkers, NY: World Book.

- Huselid, M. A. (1995). The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of Management Journal*, 38(3), 635-672.
- Ilgen, D. R., & Favero, J. L. (1985). Limits in generalization from psychological research to performance appraisal processes. *Academy of Management Review*, 10(2), 311-321.
- Iqbal, M. Z., Akbar, S., & Budhwar, P. (2015). Effectiveness of performance appraisal: An integrated framework. *International Journal of Management Reviews*, 17(4), 510-533.
- Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology*, 64(5), 502-508.
- Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel Psychology*, 33(3), 595-640.
- James, L. R. (1973). Criterion models and construct validity for criteria. *Psychological Bulletin*, 80(1), 75-83.
- Jenkins, J. G. (1946). Validity for what. *Journal of Consulting Psychology*, 10(2), 93-98.
- Johnson, J. W. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. *Journal of Applied Psychology*, 86(5), 984-996.
- Jones, R. G., & Culbertson, S. S. (2011). Why performance management will remain broken: Authoritarian communication. *Industrial and Organizational Psychology*, 4(02), 179-181.

- Katzell, R. A., & Austin, J. T. (1992). From then to now: The development of industrial-organizational psychology in the United States. *Journal of Applied Psychology, 77*(6), 803-835.
- Kell, H. J., & Motowidlo, S. J. (2012). Deconstructing organizational commitment: Associations among its affective and cognitive components, personality antecedents, and behavioral outcomes. *Journal of Applied Social Psychology, 42*(1), 213-251.
- Kidwell, R. E., & Bennett, N. (1993). Employee propensity to withhold effort: A conceptual model to intersect three avenues of research. *Academy of Management Review, 18*(3), 429-456.
- Kim, J. S., & Hamner, W. C. (1976). Effect of performance feedback and goal setting on productivity and satisfaction in an organizational setting. *Journal of Applied Psychology, 61*(1), 48-57.
- King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology, 65*(5), 507-516.
- Kingsbury, F. A. (1923). Applying psychology to business. *The Annals of the American Academy of Political and Social Science, 110*, 2-12.
- Klimoski, R. J., & Inks, L. (1990). Accountability forces in performance appraisal. *Organizational Behavior and Human Decision Processes, 45*(2), 194-208.
- Konovsky, M. A., & Pugh, S. D. (1994). Citizenship behavior and social exchange. *Academy of Management Journal, 37*(3), 656-669.
- Kornhauser, A. W. (1922). The psychology of vocational selection. *Psychological Bulletin, 19*(4), 192.

- Kornhauser, A. W., & Kingsbury, F. A. (1924). *Psychological tests in business: Materials for the study of business*. Chicago, IL: University of Chicago Press.
- Kristof-Brown, A., Zimmerman, R. D., & Johnson, E. C. (2005). Consequences of individual's fit at work: A meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. *Personnel Psychology, 58*, 281-342.
- Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology, 79*(3), 332-340.
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.
- Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. New York, NY: Academic Press.
- Landy, F. J., Farr, J. L., Saal, F. E., & Freytag, W. R. (1976). Behaviorally anchored scales for rating the performance of police officers. *Journal of Applied Psychology, 61*(6), 750-758.
- Landy, F. J., Vance, R. J., Barnes-Farrell, J. L., & Steele, J. W. (1980). Statistical control of halo error in performance ratings. *Journal of Applied Psychology, 65*(5), 501-506.
- Latham, G. P., & Pinder, C. C. (2005). Work motivation theory and research at the dawn of the twenty-first century. *Annual Review of Psychology, 56*, 485-516.
- Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30*(2), 255-268.

- Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology, 60*(5), 550-555.
- Lemoine, G. J., Parsons, C. K., & Kansara, S. (2015). Above and beyond, again and again: Self-regulation in the aftermath of organizational citizenship behaviors. *Journal of Applied Psychology, 100*(1), 40-55.
- LePine, J. A., Erez, A., & Johnson, D. E. (2002). The nature and dimensionality of organizational citizenship behavior: A critical review and meta-analysis. *Journal of Applied Psychology, 87*(1), 52-65.
- Lerman, R. I., & Schmidt, S. R. (1999). *An overview of economic, social, and demographic trends affecting the US labor market*. (J-9-M-0048). Washington, DC: The Urban Institute Retrieved from http://www.dol.gov/dol/aboutdol/history/herman/reports/futurework/conference/trends/NewTrends_.htm - IVtechnology.
- Levine, E. L., Ash, R. A., Hall, H., & Sistrunk, F. (1983). Evaluation of job analysis methods by experienced job analysts. *Academy of Management Journal, 26*(2), 339-348.
- Link, H. C. (1918). An experiment in employment psychology. *Psychological Review, 25*(2), 116-127.
- Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice-Hall.
- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*, 705-717.

- Locke, E. A., & Latham, G. P. (2006). New directions in goal-setting theory. *Current Directions in Psychological Science, 15*, 265-268.
- Lunenburg, F. C. (2012). Performance appraisal: Methods and rating errors. *International Journal of Scholarly Academic Intellectual Diversity, 14*(1), 1-9.
- Marquis, D. G. (1944). The mobilization of psychologists for war service. *Psychological Bulletin, 41*(7), 469-473.
- Maslach, C., Schaufeli, W. B., & Leiter, M. P. (2001). Job burnout. *Annual Review of Psychology, 52*(1), 397-422.
- Maurer, T. J., & Alexander, R. A. (1991). Contrast effects in behavioral measurement: An investigation of alternative process explanations. *Journal of Applied Psychology, 76*(1), 3-10.
- McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69*(1), 147-156.
- Meyer, H. H. (1980). Self - appraisal of job performance. *Personnel Psychology, 33*(2), 291-295.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.
- Miner, J. B. (1988). Development and application of the rated ranking technique in performance appraisal. *Journal of Occupational Psychology, 61*(4), 291-305.
- Mitchell, T. R., & Wood, R. E. (1980). Supervisor's responses to subordinate poor performance: A test of an attributional model. *Organizational Behavior and Human Performance, 25*(1), 123-138.

- Mone, E. M., Price, B., & Eisinger, C. (2011). Performance management: Process perfection or process utility. *Industrial and Organizational Psychology, 4*(2), 184-187.
- Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*(2), 191-205.
- Motowidlo, S. J. (2000). Some basic issues related to contextual performance and organizational citizenship behavior in human resource management. *Human Resource Management Review, 10*(1), 115-126.
- Motowidlo, S. J., & Borman, W. C. (1977). Behaviorally anchored scales for measuring morale in military units. *Journal of Applied Psychology, 62*(2), 177-183.
- Motowidlo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10*(2), 71-83.
- Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology, 79*(4), 475-480.
- Munsterberg, H. (1913). *Psychology and industrial efficiency*. Boston, MA: Houghton-Mifflin.
- Murphy, K. R. (1999). The challenge of staffing a postindustrial workplace. In D. R. Ilgen & E. D. Pulakos (Eds.), *The Changing Nature of Performance* (pp. 295-324). San Francisco, CA: Jossey-Bass.
- Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology, 1*(2), 148-160.

- Murphy, K. R., Balzer, W. K., Lockhart, M. C., & Eisenman, E. J. (1985). Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology, 70*(1), 72-84.
- Murphy, K. R., & Cleveland, J. N. (1991). *Performance appraisal: An organizational perspective*. Needham Heights, MA: Allyn and Bacon.
- Murphy, K. R., & Cleveland, J. N. (1995). Understanding performance appraisal: Social, organizational, and goal-based perspectives. Thousand Oaks, CA: Sage.
- Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*(2), 218-225.
- Murphy, K. R., & Reynolds, D. H. (1988). Does true halo affect observed halo. *Journal of Applied Psychology, 73*(2), 235-238.
- Noonan, L. E., & Sulsky, L. M. (2001). Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance, 14*(1), 3-26.
- Obidinnu, J., Ejiofor, V., & Ekechukwu, B. (2014). Distributional errors normalisation model for improving the variability of supervisors' appraisals ratings. *African Journal of Computing & ICT, 7*(1), 43-48.
- O'Boyle, E., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology, 65*(1), 79-119.
- O'Leary, R. S., & Pulakos, E. D. (2011). Managing performance through the manager-employee relationship. *Industrial and Organizational Psychology, 4*(2), 208-214.

- Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2008). No new terrain: Reliability and construct validity of job performance ratings. *Industrial and Organizational Psychology, 1*(02), 174-179.
- Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington, MA: Lexington.
- Organ, D. W. (1997). Organizational citizenship behavior: It's construct clean-up time. *Human Performance, 10*(2), 85-97.
- Palmer, J. K., & Feldman, J. M. (2005). Accountability and need for cognition effects on contrast, halo, and accuracy in performance ratings. *The Journal of Psychology, 139*(2), 119-138.
- Park, D. D., & Kim, J. (2013). What are current best approaches companies are using for performance management for wage employees. Student Works ILR.
- Parsons, F. (1909). *Choosing a vocation*. Boston, MA: Houghton Mifflin.
- Pauls, R., & Rogers, L. (1977). Band broadening studies using parameters for an exponentially modified gaussian. *Analytical Chemistry, 49*(4), 625-628.
- Pfeffer, J. (1998). Seven practices of successful organizations. *California Management Review, 40*(2), 96-124.
- Ployhart, R. E., Weekly, J. A., & Baughman, K (2006). The structure and function of human capital emergence: A multilevel examination of the attraction-selection-attrition model. *Academy of Management Journal, 49*, 661-677.
- Podsakoff, N. P., Whiting, S. W., Podsakoff, P. M., & Blume, B. D. (2009). Individual-and organizational-level consequences of organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology, 94*(1), 122-141.

- Podsakoff, P. M., Ahearne, M., & MacKenzie, S. B. (1997). Organizational citizenship behavior and the quantity and quality of work group performance. *Journal of Applied Psychology, 82*(2), 262.
- Podsakoff, P. M., & MacKenzie, S. B. (1997). Impact of organizational citizenship behavior on organizational performance: A review and suggestion for future research. *Human Performance, 10*(2), 133-151.
- Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69*(4), 581-588.
- Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior and Human Decision Processes, 38*(1), 76-91.
- Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85*(4), 612.
- Pulakos, E. D., Hanson, R. M., Arad, S., & Moye, N. (2015). Performance management can be fixed: An on-the-job experiential learning approach for complex behavior change. *Industrial and Organizational Psychology, 8*(1), 51-76.
- Pulakos, E. D., & O'Leary, R. S. (2011). Why is performance management broken. *Industrial and Organizational Psychology, 4*(2), 146-164.
- Rand, T. M., & Wexley, K. N. (1975). Demonstration of the effect, "similar to me," in simulated employment interviews. *Psychological Reports, 36*(2), 535-544.

- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1), 21-33.
- Reilly, R. R., & Smither, J. W. (1985). An examination of two alternative techniques to estimate the standard deviation of job performance in dollars. *Journal of Applied Psychology*, 70(4), 651-661.
- Richerson, P. J., & Boyd, R. (2008). *Not by genes alone: How culture transformed human evolution*. Chicago, IL: University of Chicago Press.
- Robinson, S. L., & Bennett, R. J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal*, 38(2), 555-572.
- Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczyńska, U. (2012). Rater training revisited: An updated meta - analytic review of frame - of - reference training. *Journal of Occupational and Organizational Psychology*, 85(2), 370-395.
- Ronan, W. W., & Prien, E. P. (1971). *Perspectives on the measurement of human performance*. New York, NY: Appleton-Century Crofts.
- Rosen, B., & Jerdee, T. H. (1976). The nature of job-related age stereotypes. *Journal of Applied Psychology*, 61(2), 180-183.
- Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47(3), 537-560.
- Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: a policy-capturing approach. *Journal of Applied Psychology*, 87(1), 66-80.

- Royston, J. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics, 31*(2), 115-124.
- Ryan, A. M., & Ployhart, R. E. (2014). A century of selection. *Annual Review of Psychology, 65*, 693-717.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.
- Saal, F. E., & Landy, F. J. (1977). The mixed standard rating scale: An evaluation. *Organizational Behavior and Human Performance, 18*(1), 19-35.
- Sackett, P. R. (2002). The structure of counterproductive work behaviors: Dimensionality and relationships with facets of job performance. *International Journal of Selection and Assessment, 10*(1), 5-11.
- Sackett, P. R., DuBois, C. L., & Noe, A. W. (1991). Tokenism in performance evaluation: The effects of work group representation on male-female and white-black differences in performance ratings. *Journal of Applied Psychology, 76*(2), 263-267.
- Saks, A. M., & Ashforth, B. E. (1997). A longitudinal investigation of the relationships between job information sources, applicant perceptions of fit, and work outcomes. *Personnel Psychology, 50*, 395-426.
- Sammer, J. (2008). Calibrating consistency. *HR Magazine, 53*(1), 73-75.
- Schleicher, D. J., & Day, D. V. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73*(1), 76-101.

- Schmidt, F. L., & Hunter, J. E. (1983). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. *Journal of Applied Psychology, 68*(3), 407-414.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274.
- Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*(4), 901-912.
- Schneider, B. (1987). The people make the place. *Personnel Psychology, 40*, 437-453.
- Schneider, B. (2001). Fits about fit. *Applied Psychology: An International Review, 50*, 141-152.
- Schneier, C. E. (1977a). Multiple rater groups and performance appraisal. *Public Personnel Management, 6*(1), 13-20.
- Schneier, C. E. (1977b). Operational utility and psychometric characteristics of behavioral expectation scales: A cognitive reinterpretation. *Journal of Applied Psychology, 62*(5), 541-548.
- Schwab, D. P., & Heneman, H. G. (1978). Age stereotyping in performance appraisal. *Journal of Applied Psychology, 63*(5), 573-578.
- Scott, W. D. (1917). A fourth method of checking results in vocational selection. *Journal of Applied Psychology, 1*(1), 61.
- Scullen, S. E., Bergey, P. K., & Aiman-Smith, L. (2005). Forced distribution rating systems and the improvement of workforce potential: A baseline simulation. *Personnel Psychology, 58*(1), 1-32.

- Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*(6), 956-970.
- Seashore, S. E., Indik, B. P., & Georgopoulos, B. S. (1960). Relationships among criteria of job performance. *Journal of Applied Psychology, 44*(3), 195-202.
- Shah, N. P., Cross, R., & Levin, D. Z. (2015). Performance benefits from providing assistance in networks relationships that generate learning. *Journal of Management, 25*(1), 1-33.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3), 591-611.
- Shapiro, S. S., Wilk, M. B., & Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association, 63*(324), 1343-1372.
- Sharon, A. T., & Bartlett, C. (1969). Effect of instructional conditions in producing leniency on two types of rating scales. *Personnel Psychology, 22*(3), 251-263.
- Siegel, L. (1982). Paired comparison evaluations of managerial effectiveness by peers and supervisors. *Personnel Psychology, 35*(4), 843-852.
- Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*(1), 131-142.
- Smith, B. N., Hornsby, J. S., & Shirmeyer, R. (1996). Current trends in performance appraisal: An examination of managerial practice. *SAM Advanced Management Journal, 61*(3), 10-15.
- Smith, C., Organ, D. W., & Near, J. P. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology, 68*(4), 653.

- Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*(2), 149-155.
- Spence, J. R., & Keeping, L. M. (2013). The road to performance ratings is paved with intentions: A framework for understanding managers' intentions when rating employee performance. *Organizational Psychology Review, 3*(4), 360-383.
- St-Onge, S., Morin, D., Bellehumeur, M., & Dupuis, F. (2009). Managers' motivation to evaluate subordinate performance. *Qualitative Research in Organizations and Management: An International Journal, 4*(3), 273-293.
- Stamoulis, D. T., & Hauenstein, N. M. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology, 78*(6), 994-1003.
- Steiner, D. D., & Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. *Journal of Applied Psychology, 74*(1), 136-142.
- Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Annals of Statistics, 28*(1), 40-74.
- Stewart, S. M., Gruys, M. L., & Storm, M. (2010). Forced distribution performance evaluation systems: Advantages, disadvantages and keys to implementation. *Journal of Management & Organization, 16*(1), 168-179.
- Strong, E. K. (1918). Work of the committee on classification of personnel in the Army. *Journal of Applied Psychology, 2*(2), 130-139.

- Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology, 77*(4), 501-510.
- Sulsky, L. M., & Day, D. V. (1994). Effects of frame-of-reference training on rater accuracy under alternative time delays. *Journal of Applied Psychology, 79*(4), 535-543.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.
- Takeuchi, R., Bolino, M. C., & Lin, C.-C. (2015). Too many motives: The interactive effects of multiple motives on organizational citizenship behavior. *Journal of Applied Psychology, 100*(4), 1239-1248.
- Thode, H. C. (2002). *Testing for normality* (Vol. 164). New York, NY: Marcel Dekker.
- Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York, NY: Wiley.
- Thornton, G. C., & Zorich, S. (1980). Training to improve observer accuracy. *Journal of Applied Psychology, 65*(3), 351-354.
- Tiffin, J. (1947). *Industrial psychology*. Oxford, England: Prentice-Hall Industrial psychology.
- Toops, H. A. (1944). The criterion. *Educational and Psychological Measurement, 4*(4), 271-297.
- Trougakos, J. P., Beal, D. J., Cheng, B. H., Hideg, I., & Zweig, D. (2015). Too drained to help: A resource depletion perspective on daily interpersonal citizenship behaviors. *Journal of Applied Psychology, 100*(1), 227.

- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Wesley.
- Turnipseed, D. L., & Rassuli, A. (2005). Performance perceptions of organizational citizenship behaviours at work: A bi - level study among managers and employees. *British Journal of Management, 16*(3), 231-244.
- Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93*(3), 711-719.
- Uthoff, V. A. (1970). An optimum test property of two well-known statistics. *Journal of the American Statistical Association, 65*(332), 1597-1600.
- Van Dyne, L., Graham, J. W., & Dienesch, R. M. (1994). Organizational citizenship behavior: Construct redefinition, measurement, and validation. *Academy of Management Journal, 37*(4), 765-802.
- Viswesvaran, C. (1993). *Modeling job performance: Is there a general factor*. (N00014-92-J-4040). Monterey, CA: Defense Personnel Security Research Center.
- Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*(5), 557-574.
- Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*(1), 108-131.
- Viteles, M. S. (1925). The clinical viewpoint in vocational selection. *Journal of Applied Psychology, 9*(2), 131-138.
- Viteles, M. S. (1932). *Industrial psychology*. New York, NY: W.W. Norton.

- Wagner, S. H., & Goffin, R. D. (1997). Differences in accuracy of absolute and comparative performance appraisal methods. *Organizational Behavior and Human Decision Processes*, 70(2), 95-103.
- Wallace, S. R. (1974). How high the validity. *Personnel Psychology*, 27(3), 397-407.
- Werner, J. M. (2000). Implications of OCB and contextual performance for human resource management. *Human Resource Management Review*, 10(1), 3-24.
- Wherry, R. J. (1952). *The control of bias in ratings: A theory of rating*. (922). Washington, DC: Department of the Army, Adjutant General's Office, Personnel Research Section.
- Wherry, R. J., & Bartlett, C. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology*, 35(3), 521-551.
- Wiese, D. S., & Buckley, M. R. (1998). The evolution of the performance appraisal process. *Journal of Management History*, 4(3), 233-249.
- Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, 67(3), 189-205.
- Wong, K. F. E., & Kwong, J. Y. (2007). Effects of rater goals on rating patterns: Evidence from an experimental field study. *Journal of Applied Psychology*, 92(2), 577-585.
- Yoakum, C. S., & Yerkes, R. M. (1920). *Army mental tests*. New York, NY: Holt.

APPENDIX

IRB APPROVAL



LOUISIANA TECH
UNIVERSITY

MEMORANDUM

OFFICE OF UNIVERSITY RESEARCH

TO: Mr. Richard Chambers and Dr. Tilman Sheets *[Signature]*
 FROM: Dr. Stan Napper, Vice President Research & Development
 SUBJECT: HUMAN USE COMMITTEE REVIEW
 DATE: July 14, 2016

In order to facilitate your project, an EXPEDITED REVIEW has been done for your proposed study entitled:

**"Evaluating the Different Distributions and Types of
Analysis of Job Performance Measures"**

RUC 1446

The proposed study's revised procedures were found to provide reasonable and adequate safeguards against possible risks involving human subjects. The information to be collected may be personal in nature or implication. Therefore, diligent care needs to be taken to protect the privacy of the participants and to assure that the data are kept confidential. Informed consent is a critical part of the research process. The subjects must be informed that their participation is voluntary. It is important that consent materials be presented in a language understandable to every participant. If you have participants in your study whose first language is not English, be sure that informed consent materials are adequately explained or translated. Since your reviewed project appears to do no damage to the participants, the Human Use Committee grants approval of the involvement of human subjects as outlined.

Projects should be renewed annually. *This approval was finalized on July 14, 2016 and this project will need to receive a continuation review by the IRB if the project, including data analysis, continues beyond July 14, 2017.* Any discrepancies in procedure or changes that have been made including approved changes should be noted in the review application. Projects involving NIH funds require annual education training to be documented. For more information regarding this, contact the Office of University Research.

You are requested to maintain written records of your procedures, data collected, and subjects involved. These records will need to be available upon request during the conduct of the study and retained by the university for three years after the conclusion of the study. If changes occur in recruiting of subjects, informed consent process or in your research protocol, or if unanticipated problems should arise it is the Researchers responsibility to notify the Office of Research or IRB in writing. The project should be discontinued until modifications can be reviewed and approved.

If you have any questions, please contact Dr. Mary Livingston at 257-2292 or 257-5066.

A MEMBER OF THE UNIVERSITY OF LOUISIANA SYSTEM

P.O. BOX 3092 • RUSTON, LA 71272 • TEL: (318) 257-5075 • FAX: (318) 257-5079

AN EQUAL OPPORTUNITY UNIVERSITY