

Louisiana Tech University

Louisiana Tech Digital Commons

Doctoral Dissertations

Graduate School

Spring 5-2022

Analyzing Extreme Cases: How Quantile Regression can Enhance Our Ability to Identify Productivity Stars

Evan Theys

Follow this and additional works at: <https://digitalcommons.latech.edu/dissertations>

**ANALYZING EXTREME CASES: HOW QUANTILE
REGRESSION CAN ENHANCE OUR ABILITY TO
IDENTIFY PRODUCTIVITY STARS**

by

Evan R. Theys, B.A..

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

COLLEGE OF EDUCATION
LOUISIANA TECH UNIVERSITY

May 2022

LOUISIANA TECH UNIVERSITY

GRADUATE SCHOOL

March 25, 2022

Date of dissertation defense

We hereby recommend that the dissertation prepared by

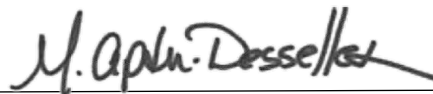
Evan Theys

entitled **Analyzing Extreme Cases:**

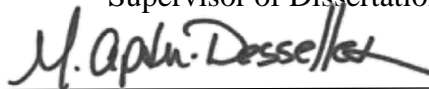
How Quantile Regression can Enhance Our Ability to Identify Productivity Stars

be accepted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Industrial/Organizational Psychology



Dr. Marita Apter-Desselles
Supervisor of Dissertation Research



Dr. Marita Apter-Desselles
Head of Psychology and Behavioral Science

Doctoral Committee Members:

Dr. Tilman Sheets

Dr. Frank Igou

Approved:



Don Schillinger
Dean of Education

Approved:



Ramu Ramachandran
Dean of the Graduate School

ABSTRACT

Recent research suggests that individual productivity may not be normally distributed and is best modeled by a power law, a form of a heavy-tailed distribution where extreme cases on the right side of the distribution affect the mean and skew the probability distribution. These extreme cases, commonly referred to as “star performers” or “productivity stars,” provide a disproportionately positive impact on organizations. Yet, the field of industrial-organizational psychology has failed to uncover effective techniques to identify them during selection accurately. Limiting factors in the identification of star performers are the traditional methods (e.g., Pearson correlation, ordinary least squares regression) used to establish criterion-related validity and inform selection battery design (i.e., determine which assessments should be retained and how those assessments should be weighted). Pearson correlation and ordinary least squares regression do not perform well (i.e., do not provide accurate estimates) when data are highly skewed and contain outliers. Thus, the purpose of this dissertation was to investigate whether an alternative method, specifically the quantile regression model (QRM), outperforms traditional approaches during criterion-related validation and selection battery design. Across three unique samples, results suggest that although the QRM provides a much more detailed understanding of predictor-criterion relationships, the practical usefulness of the QRM in selection assessment battery design is similar to the OLS regression.

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 is 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 _____

Date _____

DEDICATION

I dedicate this dissertation to my wife, Claire, who is my biggest supporter and source of inspiration. This accomplishment would not be possible without your many sacrifices, especially during our time in Ruston. You are my person, and I am forever grateful for you. I also dedicate this work to my parents, Sue Sark and Bob Theys, who instilled the importance of education early in my life and have always done everything possible to support me.

TABLE OF CONTENTS

ABSTRACT.....	iii
APPROVAL FOR SCHOLARLY DISSEMINATION	iv
DEDICATION	v
LIST OF TABLES	ix
LIST OF FIGURES	xi
ACKNOWLEDGMENTS	xii
CHAPTER 1 INTRODUCTION	1
Operationalization of Job Performance	4
The Distribution of Individual Productivity	5
Selection Test Validation	8
Limitations of Traditional Approaches	12
Influence of Outliers	12
Usefulness at the Extremes	14
The Case for Quantile Regression	15
Overview of QRM	15
Leveraging the QRM to Improve Selection Decisions	19
CHAPTER 2 METHOD	25
Sample 1.....	25
Sample 2.....	26

Sample 3.....	27
Measures	28
Personality.....	28
Individual Productivity	29
Data Analytic Approach	29
Descriptive Statistics.....	29
Assessing the Presence of a Power-Law Distribution	29
Hypothesis Testing.....	31
CHAPTER 3 RESULTS	34
Sample 1.....	34
Descriptive Statistics.....	34
Assessing the Presence of a Power-Law Distribution	34
Hypothesis Testing.....	35
Sample 2.....	40
Descriptive Statistics.....	40
Assessing the Presence of a Power-Law Distribution	40
Hypothesis Testing.....	42
Sample 3.....	46
Descriptive Statistics.....	46
Assessing the Presence of a Power-Law Distribution	47
Hypothesis Testing.....	48
CHAPTER 4 DISCUSSION.....	53
Implications for Theory and Practice.....	55

Limitations and Future Directions	57
Conclusions.....	59
REFERENCES	61
APPENDIX A SAMPLE 1 DEMOGRAPHICS, DESCRIPTIVES, AND PEARSON CORRELATIONS.....	76
APPENDIX B SAMPLE 2 DESCRIPTIVES, AND PEARSON CORRELATIONS	80
APPENDIX C SAMPLE 3 DESCRIPTIVES, AND PEARSON CORRELATIONS	83
APPENDIX D HUMAN USE EXEMPTION LETTER	86

LIST OF TABLES

Table 3-1	<i>Sample 1 Power-Law Distribution Fit and the Corresponding P-Value</i>	35
Table 3-2	<i>Sample 1 t Statistics for OLS Regression and LASSO Quantile Regression Results</i>	38
Table 3-3	<i>Selection Model Details for Sample 1</i>	39
Table 3-4	<i>Percentage of Stars Identified by Model Across Selection Ratios for Sample 1</i>	40
Table 3-5	<i>Mean Productivity of Selected Cases by Model Across Selection Ratios for Sample 1</i>	40
Table 3-6	<i>Sample 2 Power-Law Distribution Fit and the Corresponding p-Value</i>	42
Table 3-7	<i>Sample 2 t Statistics for OLS Regression and LASSO Quantile Regression Results</i>	44
Table 3-8	<i>Selection Model Details for Sample 2</i>	45
Table 3-9	<i>Percentage of Stars Identified by Model Across Selection Ratios for Sample 2</i>	46
Table 3-10	<i>Mean Productivity of Selected Cases by Model Across Selection Ratios for Sample 2</i>	46
Table 3-11	<i>Power-Law Distribution Fit and the Corresponding p-Value for Sample 3 Training Set</i>	48
Table 3-12	<i>Sample 3 t Statistics for OLS Regression and LASSO Quantile Regression Results</i>	50
Table 3-13	<i>Selection Model Details for Sample 3</i>	51
Table 3-14	<i>Percentage of Stars Identified by Model Across Selection Ratios for Sample 3</i>	52

Table 3-15 *Mean Productivity of Selected Cases by Model Across Selection Ratios
for Sample 3*52

LIST OF FIGURES

Figure 1-1	<i>Sample OLS Regression and Quantile Regression Plots</i>	21
Figure 3-1	<i>Histogram of Individual Productivity for Sample 1</i>	35
Figure 3-2	<i>Sample 1 OLS Regression and Quantile Regression Plots</i>	37
Figure 3-3	<i>Histogram of Individual Productivity for Sample 2</i>	41
Figure 3-4	<i>Sample 2 OLS Regression and Quantile Regression Plots</i>	43
Figure 3-5	<i>Histogram of Individual Productivity for Sample 3 Training Set</i>	47
Figure 3-6	<i>Sample 3 OLS Regression and Quantile Regression Plots</i>	49

ACKNOWLEDGMENTS

I would like to thank the following people without whom I would not have been able to complete my dissertation:

- My advisor, Dr. Mitzi Desselles, for her belief in me from the moment I stepped foot on Louisiana Tech's campus, the countless hours she has spent teaching, coaching, and developing me, and for her endless patience through my many missed deadlines. She is the advisor that I wish all graduate students had the luxury of working with.
- My committee members, Dr. Tilman Sheets and Dr. Frank Igou, for creating the program that has enabled the career of my dreams, and for their guidance throughout graduate school and the dissertation process.

My fellow cohort members, Dr. Christopher Patton and Dr. Richard Chambers, for always inspiring me to be better, checking-in on my dissertation progress, and simply being outstanding friends.

CHAPTER 1

INTRODUCTION

Arguably, the most crucial role of researchers and practitioners within the field of industrial-organizational (I-O) psychology is to understand, assess, predict, and improve the performance of individual employees. There is a long-standing assumption within the field that suggests job performance is best modeled by a normal distribution (Hull, 1928; Tiffin, 1947). Historically, when performance data do not follow a normal distribution, the sampled data are deemed to be biased, contain error, or are unrepresentative of the underlying population (Murphy, 2008). As such, “problem” cases within the sample (i.e., outliers) are either transformed or removed to satisfy the assumption of normality (Aguinis, Gottfredson, & Joo, 2013). However, recent research (e.g., Aguinis, Ji, & Joo, 2018; Aguinis & O’Boyle, 2014; Call, Nyberg, & Thatcher, 2015; Crawford, Aguinis, Lichtenstein, Davidsson, & McKelvey, 2015; Joo, Aguinis, & Bradley, 2017; O’Boyle & Aguinis, 2012) challenges the assumption of normality of individual performance.

These researchers suggest that outputs from a small group of employees (i.e., star performers) are “inconsistent with what would be expected using a normal distribution” (Aguinis & O’Boyle, 2014, p. 316), and the prevalence of star performers (i.e., those more than three standard deviations above the mean) far exceeds what would be predicted under a normal distribution (O’Boyle & Aguinis, 2012).

As a result, it has been proposed that the distribution of individual employee performance may be best modeled by a power-law distribution (Joo et al., 2017; O'Boyle & Aguinis, 2012). Power-law distributions are a form of heavy-tailed distributions (i.e., highly skewed and leptokurtic), which in this context, suggests that a small number of employees at the positive tail of the distribution (i.e., star performers) account for a much greater proportion of production than the large group of average performers suggested by a normal distribution (Aguinis, O'Boyle, Gonzalez-Mulé, & Joo, 2016; O'Boyle & Aguinis, 2012).

Some scholars have continued to argue that job performance is normally distributed (e.g., Beck, Beatty, & Sackett, 2014); however, a recent aggregation of arguments from economists, sociologists, and management scholars revealed that market (e.g., technological advances, organizational structure shifts, knowledge-based work), social (e.g., advent of mass communication, ease of collaboration), and individual (e.g., motivation, opportunity) forces provide unique opportunities for star performers to emerge in the modern workplace (Aguinis & O'Boyle, 2014; Call et al., 2015). Despite these drastic changes to the nature of work, our theoretical and methodological approaches to understanding extreme levels of employee productivity still lag. For example, over 75 articles within leading management, sociology, and economic journals have investigated star performers, but the literature has yet to uncover effective techniques to accurately identify them (Call et al., 2015; Terviö, 2009). This is problematic as star performers have a considerable impact on organizational outcomes, such as firm sustainability and survival (Bedeian & Armenakis, 1998; Boudreau & Ramstad, 2007), innovative performance (Grigoriou & Rothaermel, 2014; Kehoe &

Tzabbar, 2015; Tzabbar & Kehoe, 2014; Zucker & Darby, 1996), as well as coworker performance (Oetl, 2012) and career advancement (Malhotra & Singh, 2016). A potential solution is to improve the accuracy of decisions made during the selection process. Traditional methods, such as the Pearson correlation and ordinary least squares (OLS) regression, used to estimate future job performance, do not facilitate the identification of star performers as they provide accurate estimates only under a normal distribution (Aguinis et al., 2013; Kim, Kim, & Ergun, 2015; Stevens, 1984). As such, these methods should not be used when data are unlikely to meet these requirements (O'Boyle & Aguinis, 2012). Thus, the purpose of this research study is to examine an alternative statistical method (i.e., quantile regression) that may: 1) handle data with unstable means and infinite variance(s), with the hopes of reducing error in the prediction of job performance (Li, 2015; O'Boyle & Aguinis, 2012); and 2) avoid the removal and/or downward weighting of star performers to maintain a sample that generalizes to the population from which it was drawn (Aguinis et al., 2013; Becker, Robertson, & Vandenberg, 2019). To achieve this objective, I first address the operationalization of job performance (i.e., behavior versus results). Second, I provide an overview of the assumption that job performance is normally distributed and the recent research challenging this assumption. Third, I summarize the current set of practices for selection test validation and the prediction of job performance. Fourth, I provide a critique of current methods for test validation, emphasizing their limitations, and propose a remedy. Finally, I rigorously test the newly proposed validation procedure to provide evidence that quantile regression provides theoretical and practical advantages over traditional statistical techniques (i.e., OLS regression).

Operationalization of Job Performance

There have been many proposed definitions of job performance, one of which defines the construct as “scalable actions, behavior, and outcomes that employees engage in or bring about that are linked with and contribute to organizational goals” (Viswesvaran & Ones, 2000, p. 216). As may be seen within this definition, there are two ways job performance may be operationalized: behavior- or results-based (Austin & Villanova, 1992). Most research efforts involving job performance rely on supervisor ratings of job-relevant behaviors, whereas others have utilized objective measures of the outcomes of employee behavior, such as sales revenue (Aguinis, 2013). According to Aguinis (2013), neither approach should be considered universally applicable. For example, behavior-based approaches are most appropriate when the link between behavior and results is not apparent, outcomes occur in the distant future, or poor results are due to causes beyond the performer’s control. Conversely, a results-based approach is most appropriate when workers are skilled in the needed behaviors, behaviors and results are clearly related, and when the nature of job performance is equifinal. Considering this, it appears that both behavior- and results-based operationalizations have value depending on the context, but organizations and the nature of work have changed immensely since the turn of the century (Burke & Ng, 2006). Twenty-first-century work has become more knowledge- and service-based, complex, and autonomous (Cascio & Aguinis, 2008), and organizations are faced with the conundrum of having too much objective data on their employees, customers, and other entities (Berry & Linoff, 2004). Previous research has leveraged Aguinis’s (2013) recommendations and argued that a results-based approach is

more appropriate for work typical of the 21st century (Aguinis & O'Boyle, 2014; Aguinis et al., 2016; O'Boyle & Aguinis, 2012).

In the present work, I also adopt a results-based approach to ensure the results are applicable and generalizable to the future of work as the labor market continues to become increasingly dominated by highly complex occupations, such as those found in sales, service, technology, research, and white-collar sectors (Aguinis & O'Boyle, 2014; Barley, Bechky, & Milliken, 2017). Additionally, following the lead of Aguinis et al. (2016), I use the terms “productivity” and “productivity stars” instead of “performance” and “star performers” moving forward when referencing a results-based approach as that is not how job performance is commonly defined in the literature (Beck et al., 2014). With this perspective in mind, the next section provides an overview of the normality (or lack thereof) of individual productivity as well as research-based evidence exemplifying why we should expect to see heavy-tailed productivity distributions moving forward across most samples and contexts.

The Distribution of Individual Productivity

The normal distribution was originally developed by de Moivre in 1738 as an approximation for the binomial distribution. Interestingly, many natural phenomena have been found to (approximately) follow a normal distribution, such as intelligence (Galton, 1889), height (Yule, 1912), body temperature (Shoemaker, 1996), and blood pressure (Orme et al., 1999), which has led to a tendency for scientists to assume normal

distributions. This may seem like a precarious assumption, but it is often a sound approximation because of the central limit theorem¹.

In a recent review of the literature, O'Boyle and Aguinis (2012) cite early work within the performance management literature as the catalyst for the belief that job performance also follows a normal distribution. For example, in his development of a performance appraisal tool for Metropolitan Life Insurance Company, Ferguson (1947) suggested that performance ratings should be distributed similarly to that predicted by a normal distribution. Additionally, several researchers and practitioners went beyond assuming and deliberately forced normal distributions regardless of the actual observed performance or productivity of employees (e.g., Canter, 1953; Schmidt & Johnson, 1973; Schultz & Siegel, 1961). Building on these seminal articles, researchers began publishing work on the causes of non-normal performance distributions, suggesting that variations from normality were the result of biases (e.g., leniency and severity bias), untrained raters, or statistical artifacts (e.g., range restriction), and are in need of an alteration to achieve normality (e.g., removal of outliers, data transformations) (Motowidlo & Borman, 1977; Reilly & Smither, 1985; Schneier, 1977a, 1977b). In other words, practitioners and academics designed practices (e.g., forcing normal distributions of performance appraisal ratings) and used statistical analyses (e.g., data transformations) to create a normal distribution of performance data regardless of the characteristics of the underlying population distribution.

¹ The central limit theorem states the following: given a population with a finite mean and non-zero variance, the sampling distribution of the mean approaches a normal distribution if you have a reasonably large sample (Urda, 2017).

Despite the prevailing assumption that job performance is normally distributed within the I-O literature, researchers have sought to evaluate whether this assumption is rooted in data. Early research on the normality of job performance commonly uncovered normal distributions, seemingly affirming the assumption within the I-O literature (Aguinis, 2013; Chambers, 2016). However, Aguinis and O'Boyle (2014) later argued that these findings were a function of the pervasive use of supervisory ratings as criteria (i.e., behavior-based approach), rater training practices (e.g., supervisors being instructed to place raters on a normal distribution), data preparation techniques (e.g., removal of outliers, data transformations), and/or reliance on studies leveraging samples from low-complexity jobs (e.g., manufacturing; Schmidt & Hunter, 1983). Further, Aguinis and O'Boyle (2012) proposed that in roles with "increased job complexity, reduced situational constraints, and flexible hierarchies, the distribution of individual performance will be better modeled by a power law" (p. 319). When reviewing the extant literature for studies that contain samples aligning with Aguinis & O'Boyle's proposition, there is a preponderance of evidence supporting the presence of heavy-tailed distributions, like a power law (e.g., Aguinis et al., 2016, 2018; Crawford et al., 2015; Grant, 2013; Grant & Sumanth, 2009; Hunter, Schmidt, & Judiesch, 1990; O'Boyle & Aguinis, 2012; Ryazanova, McNamara, & Aguinis, 2017; Toliver & Constable, 1998). More specifically, these studies have produced and replicated the non-normality of individual productivity across dozens of academic disciplines, roles within the movie and TV industries, authors, musicians, professional and amateur athletes (e.g., baseball, basketball, football, tennis), elected officials in several countries (e.g., Australia, Canada, Ireland, Estonia, United States), military positions, entrepreneurs, dentists, physicians, attorneys, sales

professionals, bank tellers, call center employees, and higher-complexity blue-collar jobs (e.g., pelt pullers, electrical fixture assemblers, wirers) (Joo et al., 2017).

In sum, there is considerable support for the presence of heavy-tailed productivity distributions for several professions and organizational settings that reflect the knowledge-based and complex nature of the current labor market and economy. Given these findings, it appears that the presence of non-normal distributions may likely be the new norm when researchers leverage a results-based approach to performance (i.e., productivity) and samples contain occupations with higher productivity ceilings (e.g., monopolistic productivity, job autonomy, job complexity) (Aguinis et al., 2016). When considering the implications of non-normal productivity distributions, researchers have argued that many theories and current organizational practices (e.g., employee selection, performance management, training, compensation) may need to be revisited (Aguinis et al., 2016; O'Boyle & Aguinis, 2012). One of which, the current set of best practices for predicting individual productivity based on assessment results, is unquestionably challenged by a shifting distribution as the statistical techniques commonly used do not provide accurate estimates when distributions deviate from normality (Aguinis et al., 2013). In the subsequent sections, I discuss the current set of best practices recommended to validate and predict individual productivity and highlight their deficiencies when heavy-tailed productivity distributions are present.

Selection Test Validation

Personnel selection may be defined as the “process of collecting and evaluating information about an individual to extend an offer of employment” (Gatewood & Field, 2001, p. 3). Pre-employment testing for personnel selection purposes has become

ubiquitous across organizations in the 21st century. Currently, organizations spend billions of dollars on employee selection processes each year and over 60% of organizations use assessments to help identify candidates that have the appropriate knowledge, skills, and abilities (KSAs) to perform well in the role and distinguish between qualified and unqualified candidates (Guion, 2011). According to the *Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards*), selection assessments ought to be based upon an understanding of the objectives for a test's use, job requirements, and test validity (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014). The last component, validity, is the most important consideration when developing and evaluating selection assessments (Society for Industrial and Organizational Psychology [SIOP], 2018). Validity is “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA et al., 2014, p. 11). In other words, the validity of a selection assessment or method is based upon the accumulation of evidence that supports and defends the various inferences demanded of a test (Lawshe, 1985; SIOP, 2018).

There are three primary sources of evidence that contribute to the understanding of the inferences that may be drawn from a selection assessment: construct validity, criterion-related validity, and content validity (AERA et al., 2014; Binning & Barrett, 1989). The *Standards* argue that the primary inference in personnel selection is that a score on the selection assessment predicts future work outcomes regardless of the strategy employed. To expand, even when the validation strategy does not include an empirical link (e.g., correlation) between the selection assessment and future work

behavior or outcome, such as in a content validation study, there is still an inferred link between the scores from the predictor (i.e., assessment) and the criterion (i.e., work behavior or outcome). While the *Standards* and other guidelines (e.g., Equal Employment Opportunity Commission [EEOC], Civil Service Commission, Department of Labor, & Department of Justice, 1978; SIOP, 2018) argue that validity is a unitary concept where different sources of evidence (e.g., content relevance, construct meaning, criterion relatedness) contribute to understanding, this study solely focuses on criterion-related validity evidence. As such, the remainder of this discussion will focus there.

Criterion-related validity dates to Binet's work on intelligence testing (Binet, 1903; Binet & Simon, 1908) and is the earliest form of validity discussed within the literature (Anastasi, 1986; Sireci, 2009). Evidence for criterion-related validity is commonly established by demonstrating an empirical relationship between scores on a selection assessment (i.e., predictor) and some criteria, such as work-relevant behavior or outcomes (AERA et al., 2014; EEOC et al., 1978; SIOP, 2018). To conduct a criterion-related validation study, one must first determine whether it is feasible. According to the *Principles*, there are three key factors to determining feasibility: availability of appropriate criterion measures, a representative research sample, and adequate statistical power. Most importantly, criteria must be relevant (i.e., accurately reflects an employees' standing on an outcome critical to job success) and the sample must not only be (reasonably) representative of the current workforce and broader candidate pool, but also large enough to meet the desired level of statistical power (SIOP, 2018). If a criterion-related validation study is feasible, then the researcher must determine a design. The *Uniform Guidelines, Principles, and Standards* all reference two potential designs:

predictive or concurrent. Operationally, predictive and concurrent designs differ based on the presence or absence of a time lapse between data collection of the predictor and criterion. More specifically, in a predictive design, there is a time interval present (e.g., approximately six months to ensure performance levels of new hires have stabilized) and data are commonly collected on candidates, whereas in a concurrent design, predictor and criterion data are commonly collected at the same time (or in close proximity) and incumbents make up the sample. Regardless of the design employed, the data analyses required to examine the empirical relationship between the predictor and criterion are identical (assuming the same predictors and criterion are used). The *Uniform Guidelines* state that a “criterion-related validity study should consist of empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance” (Section 1607.5 B). Further, the *Principles* recommend that the analysis should provide three pieces of information: (1) effect size; (2) statistical significance of the predictor-criterion relationship; and (3) confidence intervals or standard errors for the respective relationships. As such, academics and practitioners commonly rely on the Pearson correlation and OLS regression as these models align with legal requirements and are easy to calculate and interpret (i.e., make judgments about the strength of the relationship between the predictor and criterion) (Aguinis, Pierce, Bosco, & Muslin, 2009; Li, 2015; Pulakos, 2005; van Zyl & de Bruin, 2018). To illustrate, the I-O literature has widely adopted “rules of thumb” to denote low, moderate, and high levels of criterion-related validity which are based upon the Pearson correlation coefficient, where low levels of validity are .20 or less and high validities are .40 or more (Pulakos, 2005). While the Pearson correlation and OLS regression, have

attractive properties and are widely used, they are not without limitations (Aguinis et al., 2013; Hao & Naiman, 2007; Koenker & Bassett, 1978; Li, 2015). In the following section, I discuss these limitations with an emphasis on the challenges and shortcomings as they pertain to predicting individual productivity in the presence of heavy-tailed distributions.

Limitations of Traditional Approaches

The statistical methods (i.e., Pearson correlation, OLS regression) commonly employed to conduct criterion-related validation analyses possess two key limitations that inhibit our ability to accurately predict individual productivity and identify productivity stars. Namely, Pearson correlation and OLS regression are sensitive to outliers (Aguinis et al., 2013; Kruschke, Aguinis, & Joo, 2012) and offer limited usefulness at non-central locations on the distribution (i.e., where productivity stars are located) (Hao & Naiman, 2007; Li, 2015; van Zyl & de Bruin, 2018).

Influence of Outliers

As reviewed previously, individual productivity has been shown to follow a power law, a form of heavy-tailed distribution, where extreme cases (i.e., productivity stars) affect the mean and skew the probability distribution (Aguinis et al., 2016, 2018; Aguinis & O'Boyle, 2014; O'Boyle & Aguinis, 2012). Due to their heavier tails, power-law distributions predict that extreme cases are far more common than under a normal curve, so outliers (e.g., values more than three standard deviations from the mean) should be retained and studied (rather than deleted) when the underlying distribution is assumed to be a power law (Aguinis et al., 2013). Unfortunately, the recommended statistical methods used to conduct criterion-related validation analyses (i.e., Pearson correlation

and OLS regression) are sensitive to outliers (Aguinis & Edwards, 2013; Cohen, Cohen, West, & Aiken, 2003; Hunter & Schmidt, 2004), and researchers have concluded that even as few as one to two extreme cases may substantially affect results (e.g., drastically change parameters, increase errors in estimation) (Kim et al., 2015; Stevens, 1984). Simply put, Pearson correlation and OLS regression are not robust enough to tolerate extreme cases that characterize power-law distributions and produce accurate estimates (O'Boyle & Aguinis, 2012).

Given these limitations, researchers frequently use ill-advised data cleaning and preparation practices, such as the removal of outliers and nonlinear data transformations (NLTs), to ensure predictor and criterion data are more normal, and therefore, more suitable for the statistical analyses used in criterion-related validation studies (Aguinis et al., 2013; Becker et al., 2019). Research has shown these approaches may have a dramatic impact on results and the subsequent conclusions drawn from the analyses, such as increasing the likelihood of rejecting null hypotheses, changing parameter estimates, and linearizing relationships (i.e., changing nonlinear relationships into linear ones) (Box & Cox, 1964; Cortina, 2002; Emerson, 1983; Hollenbeck, DeRue, & Mannor, 2006; Tukey, 1957; Yeo & Johnson, 2000). When considering individual productivity, researchers have argued that these practices are especially problematic for the internal and external validity of criterion-related results (Aguinis et al., 2013; Becker et al., 2019; O'Boyle & Aguinis, 2012). For example, Becker and colleagues (2019) state that “findings using NLTs may not generalize to other samples or situations. This potential lack of external validity occurs because NLTs may create distributions that do not exist in the real world” (p. 832). Likewise, O'Boyle and Aguinis (2012) argued that “dropping

influential cases excludes the top performers responsible for the majority of the output and doing so creates a sample distribution that does not mirror the underlying population distribution” (p. 110). In short, these practices only exacerbate the issues with accurately predicting individual productivity and identifying stars during selection by reducing external validity. As such, a more optimal approach for conducting criterion-related validity analyses would be to use an analytic technique that is robust against the influence of outliers or violations of normality, eliminating the need for NLTs and removal of extreme cases, and provides accurate estimates when the criterion’s underlying distribution is a power law.

Usefulness at the Extremes

Traditional approaches used to investigate the relationship between selection assessments and individual productivity, such as Pearson correlation, OLS regression, and other methods of conditional means modeling, produce a single summary statistic (i.e., effect size) to describe the full distributional impact of the predictor(s) on the criterion. More specifically, OLS regression, which is based upon the conditional mean framework, examines the *average* degree to which a predictor relates to the criterion (Petscher, Logan, & Zhou, 2013), and assumes that the relationship between predictor and criterion is uniform across the entire distribution (Aguinis, Petersen, & Pierce, 1999). When OLS regression’s assumptions are met (e.g., normally distributed residuals, homoscedasticity) this notion is sound, and the estimated parameters are unbiased (Greene, 2008). However, when these assumptions are violated, research has shown that the magnitude of the relationship tends to vary across the distribution (Li, 2015).

Individual productivity data are likely to violate an OLS regression's assumptions due to the underlying distribution commonly being heavy tailed (Aguinis & Edwards, 2013; Cohen et al., 2003). Consequently, criterion-related validation results leveraging OLS regression and other conditional mean-based methods are unlikely to generalize to non-central locations of individual productivity. Given the disproportionate value that productivity stars add to organizations (Aoyama, Yoshikawa, Iyetomi, & Fujiwara, 2010; Crain & Tollison, 2002; O'Boyle & Aguinis, 2012), one may argue that when researchers and practitioners examine the relationship between selection assessments and individual productivity, they should certainly try to determine what is happening in the heavy tail. Despite the potential value in being able to understand the nuanced relationship between selection assessments and individual productivity where productivity stars are located, the methods we historically employ simply do not allow for it and constrain our ability to make accurate selection decisions and identify stars (Li, 2015). That said, a more ideal approach for conducting criterion-related validity analyses would be to use an analytic technique that is not based upon the conditional mean framework.

The Case for Quantile Regression

In this section, I provide an overview of the quantile regression model (QRM) and offer justification for its use during criterion-related validation studies. Notably, I discuss why the QRM overcomes the limitations described previously and propose a process that utilizes QRM during criterion-related validation studies to inform selection assessment battery design and improve our ability to predict individual productivity and identify stars.

Overview of QRM

While the QRM is not a new concept (Koenker & Bassett, 1978), it has been rarely used in psychological research, let alone in the mainstream journals associated with I-O psychology (Li, 2015; van Zyl & de Bruin, 2018). As such, it is easiest to explain the QRM by comparing it with the more familiar OLS regression. The QRM may be viewed as a semiparametric extension of an OLS regression to estimate rates of change across the entire distribution of the criterion. Further, the QRM serves as a viable alternative when assumptions for OLS regression are not met (e.g., linearity, homoscedasticity, normality) (Koenker & Bassett, 1978)². OLS regression models the relationship between one or more predictors and the conditional mean of the criterion, meaning it examines the average degree to which the independent variable predicts the dependent variable (Petscher, Logan, & Zhou, 2013). Conversely, the QRM models the relationship between one or more predictors and specific quantiles (i.e., percentiles) of the criterion. In other words, OLS regression uses the conditional mean to define central tendencies, whereas the QRM is based on the conditional quantile, which allows the QRM to find parameters (e.g., intercepts, β coefficients) for each quantile (e.g., 0.5th quantile or the median) and provide a nuanced investigation of the relationship between predictor and criterion (Koenker & Bassett, 1978).

Equation 1-1 displays the formula used to represent the OLS regression model, where p is the number of predictors and n is the total number of data points. However, to determine the best fitting regression line and compute parameters, the OLS regression

² The QRM is semiparametric, so while it does not assume any distribution of the error term, it does share some assumptions with an OLS regression, such as independence of observations and that the dependent variable is continuous (Koenker, 2005).

assumes a normal distribution and minimizes the summed squares of the residuals (i.e., mean squared error, MSE) (Equation 1-2). Given this assumption, OLS regression does not perform well (e.g., parameters are drastically influenced, increased errors in prediction), when data are highly skewed and contain outliers which has been reported extensively in the literature (e.g., Cohen et al., 2003; Kim et al., 2015; Li, 2015; Stevens, 1984).

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} \quad i = 1, \dots, n \quad \text{EQ. 1-1}$$

$$MSE = \frac{1}{n} \sum_{i=1}^n \left(y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) \right)^2 \quad \text{EQ. 1-2}$$

The QRM equation (Equation 1-3) takes a similar form and structure to the OLS regression, where $Q_\tau(y_i)$ is the conditional quantile of τ , which may be any point on the distribution of the criterion (e.g., 0.5th quantile or the median). Since the QRM may find parameters for each quantile across the distribution, the β coefficient is a function of the specific quantile instead of being constant like in an OLS regression. For example, if the strength of the relationship between the predictor and criterion is not perfectly uniform across the distribution, Equation 3 would have different β coefficients per quantile, τ . Despite this difference, QRM parameters (e.g., β coefficients) are interpreted exactly like those from OLS regression (Buchinsky, 1998).

$$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \dots + \beta_p(\tau)x_{ip} \quad i = 1, \dots, n \quad \text{EQ. 1-3}$$

$$MAD = \frac{1}{n} \sum_{i=1}^n \rho_\tau \left(y_i - (\beta_0(\tau) + \beta_1(\tau)x_{i1} + \dots + \beta_p(\tau)x_{ip}) \right) \quad \text{EQ. 1-4}$$

To estimate the parameters of each quantile, the QRM minimizes the sum of absolute residuals (i.e., median absolute deviation, MAD) and does not make a distributional assumption to the error term. The formula to calculate the MAD for the QRM is shown in Equation 1-4. Rho (ρ) represents the non-parametric weighting that

QRM applies to residuals that protects the QRM against outliers and skewed data. The formula for calculating ρ at a single data point, u , is depicted in Equation 1-5.

$$\rho_{\tau}(u) = \tau(u, 0) + (1 - \tau) \max(-u, 0) \quad \text{EQ. 1-5}$$

At a specific quantile, τ , the QRM assigns a weight of τ for positive residuals (i.e., data points that fall above the regression line), and a weight of $1 - \tau$ for negative residuals (i.e., data points that fall below the regression line). For instance, if you are calculating parameter estimates at $\tau = .25$ (i.e., 0.25th quantile or the 25th percentile), 75% of the errors should be positive and the remaining 25% should be negative. To find the smallest median absolute residuals and ensure that condition is true, weights must be added. In our example, the negative residuals would be weighted by a factor of .75, whereas the positive residuals would have a weight of .25. This process of weighting data ensures parameters for each quantile are estimated using data from the whole sample (Koenker, 2005).

To recap, the QRM offers distinct advantages over OLS regression and other conditional mean approaches commonly used to estimate criterion-related validity. Notably, the QRM does not make assumptions about the distribution of error terms, and it may provide parameter estimates for each quantile across the entire distribution. In total, these characteristics ensure that the QRM parameter estimates are robust to outliers and highly skewed data (Li, 2015). Further, the QRM supports the internal and external validity of results as heavy-tailed criteria do not need to be altered (e.g., removal of outliers, NLTs) to satisfy assumptions. This also allows extreme cases, such as productivity stars, to be studied explicitly as they may remain present in the dataset and parameters may be calculated for the heavy tail (i.e., where stars are located; 0.9th

quantile and above). As such, researchers have concluded that the QRM is “an extremely powerful tool for understanding the nuanced relationships between dependent variables with heavy-tailed distributions and their predictors” (Li, 2015, p. 72). Given the presence of heavy-tailed productivity distributions reported in the I-O psychology literature, this study posits that similar benefits will be realized when investigating the relationship between selection assessments and individual productivity.

***Hypothesis 1:** The QRM will produce a more detailed conceptualization of the predictive validity between selection assessments and individual productivity.*

Next, I present a previously unreported process that leverages the QRM during criterion-related validation studies to inform selection assessment battery design and improve our ability to make accurate selection decisions and identify stars.

Leveraging the QRM to Improve Selection Decisions

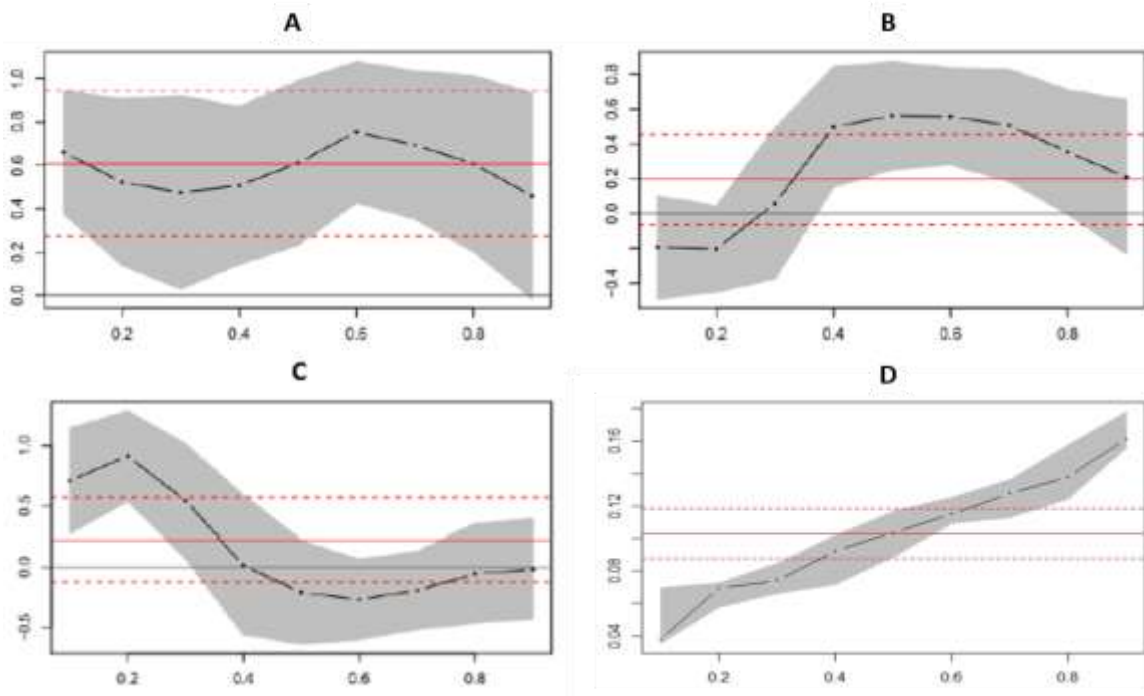
Results from criterion-related validation studies not only provide empirical evidence to support an assessment’s job-relatedness (AERA et al., 2014; EEOC et al., 1978; SIOP, 2018) but also inform selection battery design, such as determining which assessments should be retained (i.e., used to make selection decisions) and how those assessments should be weighted to optimally identify candidates that are most likely to succeed on the job (Cascio & Aguinis, 2004). In operational selection settings, assessments that exhibit significant relationships or add incremental validity are retained and combined to build a final composite (Cascio & Aguinis, 2004; Guion, 2011). Although several methods are used in practice (e.g., unit weighting, regression weighting, multiple cut-offs), regression-weighted composites (i.e., weighting each predictor based on its relationship with individual productivity) are viewed as most effective because they

are more likely to ensure the incremental validity of additional predictors is realized (Sackett, Dahlke, Shewach, & Kuncel, 2017). However, given the inherent limitations with OLS regression described previously (e.g., sensitivity to outliers, lack of usefulness at extremes), researchers and practitioners are likely making critical decisions on selection batteries and developing composites without an accurate or complete understanding of predictor-criterion relationships. Research efforts on regression-weighted composites support this claim as results indicate they are influenced by the presence of outliers which impact robustness under cross-validation (Schmidt, 1971; Wainer, 1976). Therefore, I propose that the QRM may be used during criterion-validation studies to overcome limitations with current approaches and facilitate a more intelligent design of selection batteries, ultimately improving our ability to make accurate selection decisions and identify productivity stars.

To illustrate the use of the QRM during criterion-related validation and selection battery design, consider Figure 1-1 below. Beta estimates are indicated on the y -axis, and the conditional quantiles of the criterion are plotted on the x -axis (i.e., 0.2, 0.4, 0.6, 0.8). The solid and dotted red horizontal lines reflect the OLS regression coefficient and the 95% confidence interval for the predictor. The black dots and broken lines represent the estimated regression coefficients for the quantile indicated on the x -axis. Lastly, the shaded grey background represents the 95% confidence interval for the estimated quantile regression coefficients.

Figure 1-1

Sample OLS Regression and Quantile Regression Plots



Immediately, one observes that the OLS regression estimates do not appear representative of the relationship between predictors B, C, and D and the criterion across the entire distribution of the criterion. Simply put, these relationships are heterogeneous (i.e., the strength of the relationship varies across the distribution), whereas the relationship between predictor A and the criterion is homogenous (i.e., the strength of the relationship is uniform across the distribution). The OLS regression results suggest that predictors B and C are non-significant predictors of the criterion (i.e., 95% confidence interval for the OLS regression includes zero). However, the QRM shows that both B and C are useful predictors for some, but not all, quantiles of the criterion. More specifically,

predictor B has a significant relationship with the criterion between the .4th and .8th quantiles, whereas predictor C is significantly related to the criterion up to the .3th quantile but not beyond. Finally, while the OLS regression uncovered a significant, positive relationship between predictor D and the criterion (i.e., 95% confidence interval does not include zero), this relationship is also heterogeneous as the OLS regression overestimates the strength of the relationship at lower levels of the criterion and underestimates it at the higher levels, which is demonstrated by the quantile-specific estimates falling below the OLS regression confidence interval until the 0.3th quantile and above it after the 0.8th quantile.

Here, the QRM uncovered important differences in the relationships between predictors B, C, and D and the criterion that may be used to make better decisions regarding these predictors' inclusion and weight in the final battery. Considering the sample results presented in Figure 1-1, a traditional approach would lead researchers to include predictors A and D, with each being weighted by their respective β coefficients from the OLS regression to compute the final composite. In contrast, the QRM's granular investigation of the predictor-criterion relationships suggests that each predictor may add value. However, the value of some predictors is limited to a subset of quantiles (e.g., predictor C is predictive up to the .3th quantile).

To establish a predictive model and forecast future outcomes (e.g., individual productivity) using the QRM results, one must approximate a global estimate (i.e., a single summary statistic of the predictor-criterion relationship), like an OLS regression, by combining the quantile-based information (i.e., local estimates) (Judge, Hill, Griffiths, Lutkepohl, & Lee, 1988). Although this concept is new to the I-O psychology literature,

other academic disciplines that commonly encounter heavy-tailed criteria (e.g., economics, finance) have developed procedures to predict future outcomes based on the QRM for decades (Gastwirth, 1966; Judge et al., 1988; Tukey, 1977). Results from studies examining the forecasting capabilities of the QRM have been shown to outperform OLS regression consistently (i.e., predict future outcomes more accurately) (Furno, 2011; Lima & Meng, 2017; Meligkotsidou, Panopoulou, Vrontos, & Vrontos, 2021; Sayegh, Munir, & Habeebullah, 2014). One such procedure, developed by Lima and Meng (2017), is particularly attractive for determining which assessments should be retained and how those assessments should be weighted.

Lima and Meng's approach, the post-*LASSO* quantile combination (PLQC), identifies weak and partially weak predictors by applying the L_1 -penalized (*LASSO*) quantile regression (Belloni & Chernozhukov, 2011). First, predictors that are significant at the various quantile functions are selected. Next, quantile regressions with only the selected predictors are estimated, resulting in the post-penalized quantiles, which are then combined to obtain the final composite. To expand, if an assessment helps predict all quantiles, such as predictors A and D from Figure 1-1, then it is deemed to be strong. If a given assessment predicts some, but not all, quantiles, such as predictors B and C from Figure 1-1, it is labeled as partially weak, whereas an assessment that does not predict any quantiles is fully weak. The *LASSO* quantile regression identifies and sorts predictors, or in our example, assessments, according to this classification scheme. After the predictors are sorted, a prediction equation is created (i.e., composite) by averaging the quantile results. Fully weak predictors are removed from the model (i.e., weights are set to zero) and the coefficients of partially weak predictors are adjusted to reflect their

relative contribution compared to strong predictors. By accounting for partially weak predictors, the PLQC has improved prediction accuracy over OLS regression and other models (Lima & Meng, 2017). Further, *LASSO* quantile regression has been shown to outperform other penalized regression approaches when sample sizes are smaller, such as those commonly used during criterion-related validation studies (Tibshirani, 1996). In fact, *LASSO* quantile regressions have been found to produce robust estimates in samples as small as 10 – 20, and only require that the number of predictor variables that may be selected by the method is smaller or equal to the total sample size (Ismail, 2015; Kirpich et al., 2018). Given the likelihood that many selection assessments are partially weak predictors of individual productivity (i.e., due to the presence of heavy-tailed criterion and small-to-moderate linear relationships with individual productivity) and criterion-related validation studies often leverage smaller sample sizes for complex computational techniques, the PLQC procedure provides accurate global estimates for the QRM needed to improve the prediction of individual productivity. Thus, I hypothesize that selection batteries designed using the QRM and PLQC procedure will result in greater practical benefits over and above OLS regression.

***Hypothesis 2:** QRM-informed selection batteries will result in a higher proportion of productivity stars being identified.*

***Hypotheses 3:** QRM-informed selection batteries will result in more accurate selection decisions.*

CHAPTER 2

METHOD

This study used archival data³ from criterion-related validation studies conducted at three organizations (i.e., a global professional services company, a multinational business directory and advertising firm, and a national quick-service restaurant). The roles included in this study (i.e., account executive, account manager, franchisee owner) are somewhat representative of the 21st-century workplace as each is highly complex, autonomous, as well as knowledge- and service-based. Moreover, given previous findings reported in the literature (e.g., Aguinis et al., 2016; O’Boyle & Aguinis, 2012), these samples should allow productivity stars to emerge, thus providing an appropriate arena to test the hypotheses proposed in this study.

Sample 1

Sample 1 contains data from a concurrent validation study of account executives at a global professional services company. Incumbents were identified and invited to participate based upon a stratified random sample that accounted for location, tenure, and other demographics (i.e., age, race/ethnicity, and gender) to ensure a robust, representative sample.

³ For each sample all personally identifying information was removed (i.e., deidentified), and participants were assigned a random participant code prior to the primary researcher obtaining the data.

Altogether, 209 account executives were included in the study and have both assessment and productivity data. Individual productivity was operationalized as the average percentage of sales goal attainment across the incumbent's tenure in the account executive role. This measure of individual productivity was chosen as it controls for short-term variability in sales and other extraneous factors (e.g., market, location) that influence an account executive's sales expectations (i.e., goals). Demographic data (i.e., age, race/ethnicity, gender) were made available from the host organization for this study.

Sample 2

The second sample contains data from a concurrent validation study for account managers at a multinational business directory and advertising firm. Incumbents were randomly sampled across four locations in the United States (i.e., two locations in the Northeast, one in the Midwest, and one in the South), totaling 90 account managers with paired data (i.e., personality assessment and individual productivity data). Account managers maintain, cultivate, and expand current customer relationships, so renewal and upsell totals were combined to measure individual productivity. Since the study participants operate in different regions, sales metrics were adjusted to control for location to mitigate the impact of extraneous market factors (e.g., market share, population density) on individual productivity. To do this, Z-scores were computed for each incumbent using the mean and standard deviation of individual productivity at their location. Z-scores were then converted to T-scores for easier interpretation. Note, demographic data (e.g., age, race/ethnicity, gender) were not available from the host organization for this study.

Sample 3

The final sample contains data from a concurrent validation study for franchisee owners at a national quick-service restaurant chain. A stratified random sample was drawn to ensure incumbents included in the study were representative of the broader population on several key variables, such as tenure, location, and restaurant type (e.g., free-standing, mall, airport). In total, 281 incumbents completed the personality assessment and have paired productivity data. Due to the sufficient sample size, Sample 3 was randomly split into separate training and testing datasets. The training set was used to develop the models, and the testing set was used to see how well the models perform when applied to new data. When using a hold-out method for cross-validation, 80% of data is commonly used for training and the remaining 20% for testing (Lever, Krzywinski, & Altman, 2016). As such, the training set contained 225 cases, leaving 56 cases for the testing set. Individual productivity was measured using total sales for the franchise location over the past calendar year (i.e., past 12 months) prior to conducting the validation study. Given restaurants operate in different markets, total sales figures were adjusted to control for menu pricing, location, and other extraneous variables that influence sales figures (e.g., restaurant type). To do this, total sales were first divided by the deviation from average menu pricing for that franchise location. For example, if a location's average menu pricing was 30% higher than the average menu pricing across all locations, then total sales were divided by 1.30. Next, Z-scores were computed using the mean and standard deviation of total sales for the franchise's location and restaurant type. Z-scores were then converted to T-scores for easier interpretation. Note, demographic

data (e.g., age, race/ethnicity, gender) were not available from the host organization for this study.

Measures

Personality

Across each sample, study participants completed the same personality assessment, which scholars have recognized as an outstanding example of I-O psychology in the workplace (e.g., International Personnel Assessment Council's 2015 Innovations in Assessment Award, SIOP's 2016 M. Scott Myers Award for Applied Research in the Workplace). The assessment measures 15 unique aspects of personality, 10 of which are directly related to the Big Five model and five additional aspects reflecting traits relevant for the workplace, leadership, and high-potential performance not directly related to the Big Five. The personality measure consists of 100-items and employs a pairwise multidimensional forced-choice (MFC) format, which requires participants to choose between two statements representing different personality aspects that are matched based on social desirability (e.g., "I am always on time for appointments" or "I make friends easily"). This response format has been shown to reduce participant response distortion and impression management (i.e., faking), while also being less cognitively loaded than MFC items using a greater number of statements (Christiansen, Burns, & Montgomery, 2005; Vasilopoulos, et al., 2006). To avoid issues with ipsativity and to return normative trait scores, the assessment uses the General Graded Unfolding Model (GGUM) to estimate personality statement parameters and multi-unidimensional pairwise preference (MUPP) IRT model to adaptively administer and score pairwise MFC items (Roberts, Donoghue, & Laughlin, 2000; Stark, 2002).

Research has shown this approach generates reliable and valid personality scores and allows for secure administration of MFC items in high-volume, unproctored settings (Chernyshenko et al., 2009; Drasgow, Stark, Chernyshenko, Nye, & Hulin, 2012; Martin & Theys, 2019).

Individual Productivity

Individual productivity is operationalized using a results-based approach for each sample included in this study. Given the inherent differences in roles and organizational context, individual productivity is defined and operationalized using different metrics for each sample; however, all relate to sales outcomes. Steps were taken to remove the influence of extraneous factors and confounding variables (e.g., location, market, pricing) that may contaminate sales outcomes and ensure each objective metric was representative of individual productivity.

Data Analytic Approach

Descriptive Statistics

Descriptive statistics, the mean, standard deviation, skewness, and kurtosis were calculated and examined for all variables across the three samples. Additionally, Pearson correlations were calculated to examine the relationship between variables in each sample.

Assessing the Presence of a Power-Law Distribution

To understand the distribution of individual productivity across the samples included in this study, I assessed for the presence of a power-law distribution. In the examination of power-law data, several methods have been developed (e.g., Adler, Feldman, & Taqqu, 1998; Arnold, 1983; Clauset, Shalizi, & Newman, 2009; Resnick,

2006). However, many methods for analyzing power-law data, such as least-squares fitting, may produce substantially inaccurate estimates of parameters for power-law distributions. Even in cases where such methods return accurate answers, they are still unsatisfactory because they give no indication of whether the data obey a power law at all (Clauset et al., 2009). As a result, Clauset and colleagues developed a framework to assess the presence of power-law distributions that marries maximum-likelihood estimation (MLE) with the Kolmogorov-Smirnov (K-S) goodness-of-fit statistic.

The first step in the procedure is to estimate the scaling exponent (α), which provides information about how quickly the distribution's right tail "falls" (i.e., rate of decay). Lower values (i.e., those closer to 1) indicate the distribution's right tail is heavier. For example, a distribution with $\alpha = 2$ has a heavier tail than a distribution with $\alpha = 4$ (Aguinis et al., 2018). To estimate the scaling exponent, the Hill estimator is used (Hill, 1975). This process uses MLE based on running a semiparametric Monte Carlo bootstrap calculation 1,000 times. This process estimates α and provides information regarding the rate of decay, ultimately indicating the weight or "heaviness" of the distribution's tail.

After the size of the scaling exponent, α , is calculated, the next step in Clauset and colleagues' procedure is to assess the likelihood each distribution follows a power law, which is done via the K-S statistic. The K-S statistic is a nonparametric goodness-of-fit index that may be used in accordance with its p -value to assess the probability that the sampled simulated distributions follow a power law. Lower K-S statistic values and higher p -values suggest a better fit to a power-law distribution because the null

hypothesis suggests there are no differences between the observed and underlying power-law distribution (Clauset et al., 2009).

Hypothesis Testing

Hypothesis 1 was evaluated using output from two separate analyses. First, by directly comparing OLS regression and QRM results using plots like Figure 1-1. Specifically, when the 95% confidence interval from the OLS regression (i.e., dotted red lines) and the QRM estimates (i.e., black dots and broken lines) do not overlap, this suggests that the OLS regression is over- or underestimating the predictor-criterion relationship at that quantile on the criterion distribution (van Zyl & de Bruin, 2018). Second, by observing differences in quantile-level results from the LASSO quantile regressions used in the PLQC procedure. In other words, when results vary across quantiles, this suggests the predictor-criterion relationships are heterogeneous or non-uniform. Support for Hypothesis 1 will be shown if predictor-criterion relationships do not exhibit uniform relationships across the distribution and discrepancies between the OLS regression and QRM estimates are observed.

To evaluate the practical benefits of the QRM, I assessed the increase in selection decision accuracy following a procedure established by Bing and colleagues (2007)⁴. The general procedure is as follows: 1) Composite scores on the assessment battery, \hat{Y} , were calculated separately based on results from the OLS regression and QRM approaches; 2) cases were rank-ordered and selected based on \hat{Y} for both models; and 3) average

⁴ This procedure has also been used by other researchers aimed at evaluating utility and practical benefits associated with alternative approaches in the selection space (e.g., Carter et al., 2014).

observed individual productivity, Y , of the selected cases was compared across the models. This procedure is outlined in greater detail below.

Two models were developed per sample to calculate the composite scores on the assessment battery, \hat{Y} . The first model reflects a traditional, local validation effort leveraging an OLS regression. For the OLS regression model, I used regression weighting to generate a composite (i.e., overall selection battery score). The second model used the newly proposed procedure for the QRM described previously. \hat{Y} was calculated for each case in the sample and saved for later use. Next, cases were rank-ordered and selected based on \hat{Y} for both models using a top-down selection procedure. For each study, I used selection ratios of .1, .3, .5, and .7. These conditions result in several realistic selection scenarios that are commonly observed in practice (Roth et al., 2014; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997).

Hypothesis 2 was tested by calculating the percentage of stars identified within each cohort. For this study, a productivity star was defined as any case that is at least 1.5 standard deviations above the mean on individual productivity. Support for Hypothesis 2 will be shown if the QRM-based model identifies a higher percentage of stars per cohort. Lastly, Hypothesis 3 was tested by calculating the mean of Y for each cohort. To the extent that a model results in more accurate selection decisions, the mean of Y (i.e., the observed level of individual productivity) should be higher. As such, support for Hypothesis 3 will be shown if the QRM-based model results in a larger mean than the OLS regression-based model.

Even though it may appear logical to conduct a statistical test of the mean differences between cohorts (i.e., independent samples t -test), it is not warranted. For

example, an independent samples t -test and two-sample z -test assumes that each sample is an independent random sample (Boneau, 1960). This assumption would be violated in this study because the models will be derived and used to select a pool of cases from the same sample. It is highly likely that at least some of the same cases will be identified as having the highest expected productivity (\hat{Y}) across the two models. In this instance, the same case(s) would appear in both groups, violating the assumption of independence. Given this, Bing and colleagues' (2007) procedure serves as a viable and established alternative to assess the increase in selection decision accuracy.

CHAPTER 3

RESULTS

Sample 1

Descriptive Statistics

Demographics of sample participants, descriptive statistics (e.g., mean, standard deviation, skewness, kurtosis) and Pearson correlations between study variables may be found in the Appendix in Tables A-1, A-2, and A-3 for Sample 1.

Assessing the Presence of a Power-Law Distribution

Figure 3-1 contains a histogram of individual productivity and Table 3-1 summarizes the power-law distribution fit based on Clauset and colleagues' (2009) procedure for Sample 1. While individual productivity (i.e., average percentage of sales goal attainment across the incumbent's tenure) was found to be non-normally distributed (i.e., skewness = 1.49, kurtosis = 3.18), the power-law distribution was found to be a poor fit for the data as α , the scaling parameter, was quite large ($\alpha = 11.91$), suggesting a lighter tail, and the associated p -value from the goodness-of-fit test was less than .10 (i.e., recommended cutoff) (Aguinis et al., 2018; Clauset et al., 2009). Specifically, $p = .00$ indicates there is a "near-zero" probability that the data really follow a power law (Clauset et al., 2009). Despite this, individual productivity still appears to be skewed and leptokurtic.

Figure 3-1

Histogram of Individual Productivity for Sample 1

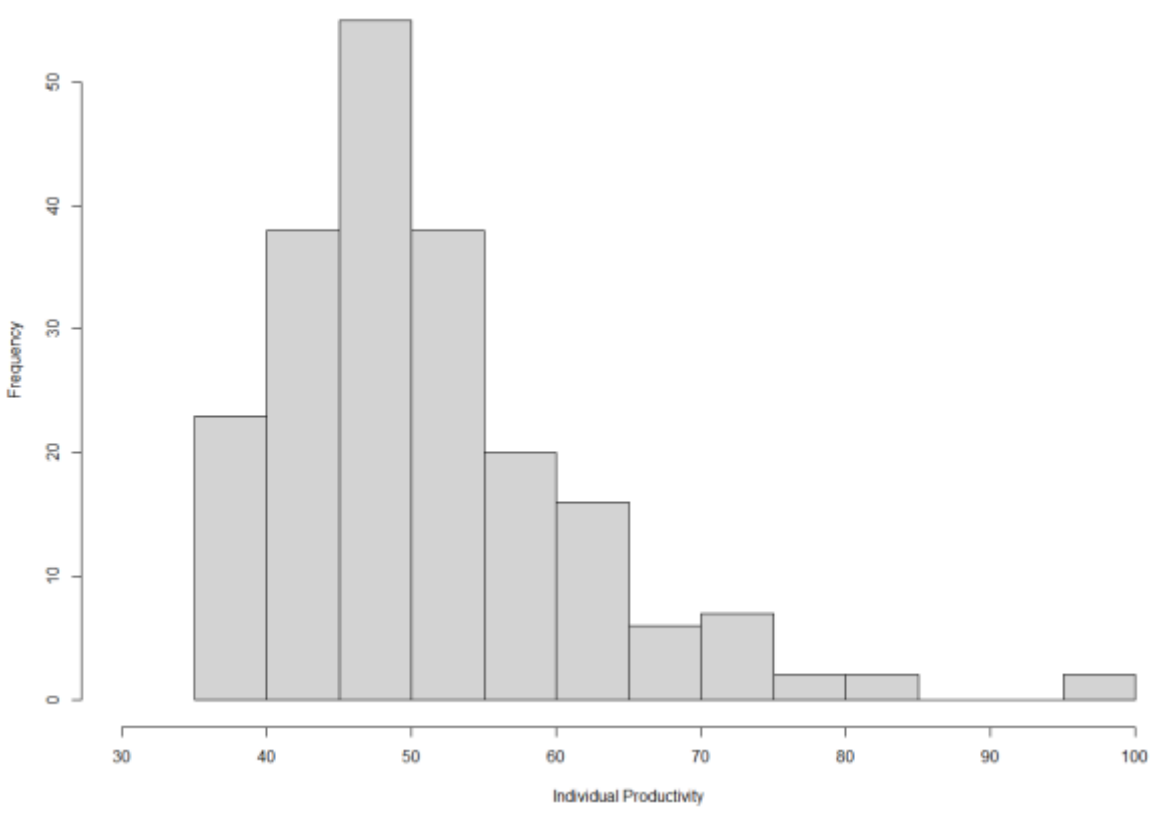


Table 3-1

Sample 1 Power-Law Distribution Fit and the Corresponding P-Value

	<u>K-S</u>	<u>Xmin</u>	<u>α</u>	<u>p</u>
Productivity	.09	54.64	11.91	.00

Note: N = 209

Hypothesis Testing

To test Hypothesis 1, I compared OLS regression and QRM results using the plots found in Figure 3-2. Results suggest that the OLS regression estimate was not representative of the relationship between Conceptual (INT), Mastery (MST), Drive

(IND), and Awareness (AWR) and individual productivity. For example, the OLS regression underestimates the relationship between Conceptual and individual productivity starting at the .8th quantile as the 95% confidence interval from the OLS regression (i.e., dotted red lines) and the QRM estimates (i.e., black dots and broken lines) do not overlap. Moreover, the results from the LASSO quantile regressions presented in Table 3-2 highlight additional discrepancies regarding the predictive validity of the predictors (i.e., personality traits) with individual productivity. Specifically, Conceptual (INT), Mastery (MST), Ambition (ACH), Composure (CMP), Awareness (AWR), Liveliness (ENT), and Assertiveness (ASR) significantly predict individual productivity at some quantiles but were not found to be significant predictors by the OLS regression. For example, Composure was not found to significantly predict individual productivity based on the OLS regression results ($t = .51, p = .61$); however, the LASSO quantile regression results suggest that Composure has a significant, negative relationship at extreme levels (i.e., .9th quantile) of individual productivity ($t = -2.34, p = .02$). Given that 8 out of 15 predictors (53.33%) have heterogeneous relationships with individual productivity, this provides support for Hypothesis 1 in Sample 1.

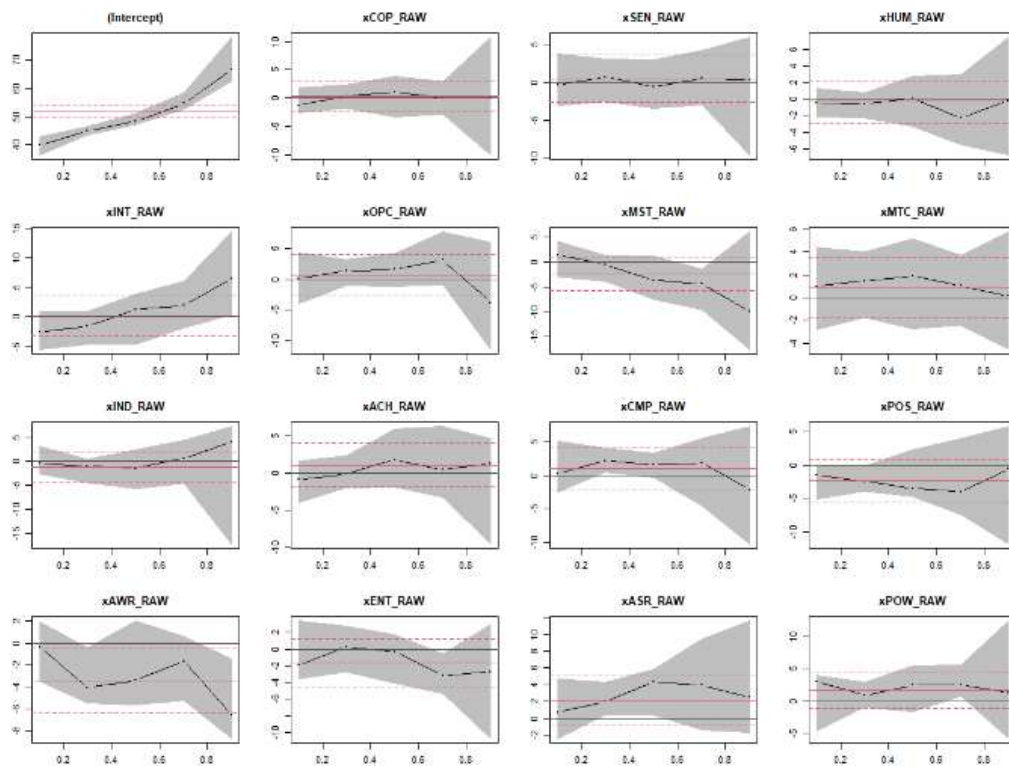
Figure 3-2*Sample 1 OLS Regression and Quantile Regression Plots*

Table 3-2*Sample 1 t Statistics for OLS Regression and LASSO Quantile Regression Results*

<u>Predictor</u>	<u>LASSO Quantile Regression</u>					
	<u>OLS</u>	<u>.1th</u>	<u>.3th</u>	<u>.5th</u>	<u>.7th</u>	<u>.9th</u>
COP	.17	.00	.25	.96	.26	.00
SEN	.24	.00	.10	-1.34	.00	.00
HUM	-.20	-.22	-.37	-.53	-1.31	.00
INT	.13	-.97	-1.07	.34	1.46	4.85**
OPC	.35	-.16	1.20	1.10	1.57	.00
MST	-1.15	.00	-.40	-1.72	-2.74**	-5.30**
MTC	.53	.00	.74	.36	1.07	.45
IND	-.59	-.76	-1.02	-.64	.00	.71
ACH	.60	-.23	.46	.42	.00	2.07*
CMP	.51	.00	1.65	.18	.84	-2.34*
POS	-1.21	-1.47	-1.70	-1.61	-1.58	-1.45
AWR	-1.94	.00	-3.20**	-2.09*	-.97	-4.00**
ENT	-.93	-.98	.00	.00	-1.60	-2.67**
ASR	1.17	-.85	1.25	2.75**	1.97	.45
POW	1.01	1.04	.87	.32	2.77**	.00

*Note: N = 209. *p < .05. **p < .01*

To test Hypotheses 2 and 3, two models were developed: (1) a regression-weighted model based on OLS regression results; and (2) a QRM-based model using the PLQC procedure. Table 3-3 contains the selected predictors from this process and their assigned weights. Note, since none of the predictors significantly predicted individual productivity at $p < .05$ using the OLS regression, those with p -values below .10 were considered for the OLS regression model. As such, the OLS regression model only contains Awareness (AWR) ($t = -1.94, p = .05$). In contrast, the PLQC procedure

identified eight predictors to include in the model; however, each were determined to be partially weak (i.e., predict some, but not all quantiles).

Table 3-3

Selection Model Details for Sample 1

<u>Model</u>	<u>AWR</u>	<u>MST</u>	<u>ASR</u>	<u>ENT</u>	<u>POW</u>	<u>INT</u>	<u>ACH</u>	<u>CMP</u>
OLS	-1.00	--	--	--	--	--	--	--
QRM	-.24	-.24	.14	-.11	.11	.09	.06	.01

Note: The weights presented in this table are a ratio of the β coefficients. The absolute value of these ratios sum to 1 and accurately reflects the relationship with individual productivity for each predictor included in the model.

Next, I followed the procedure established by Bing and colleagues (2007) and calculated the percentage of stars identified as well as the mean productivity of the selected cohort for each selection ratio (i.e., .1, .3, .5, .7). Table 3-4 and Table 3-5 summarize the results for Sample 1. Results show the QRM model identified a higher percentage of productivity stars in two out of four selection ratio scenarios (i.e., .1 and .3) and higher mean productivity of the selected cohort in three out of four (i.e., .1, .3, and .7). While the results appear to provide partial support for Hypotheses 2 and 3, it must be noted that the differences in mean productivity of the selected cases are quite small and there is considerable overlap between the 95% confidence intervals of the sampled means for each selection ratio, indicating marginal or negligible practical gains of the QRM in Sample 1.

Table 3-4*Percentage of Stars Identified by Model Across Selection Ratios for Sample 1*

<u>Model</u>	<u>Selection Ratio</u>			
	<u>.1</u>	<u>.3</u>	<u>.5</u>	<u>.7</u>
OLS	15.79%	36.84%	73.68%	84.21%
QRM	21.05%	52.63%	73.68%	84.21%

Note: The number of selected cases per selection ratio varied, with 10% = 21, 30% = 63, 50% = 105, and 70% = 146. $N = 19$ productivity stars were identified (i.e., any case that is at least 1.5 standard deviations above the mean on individual productivity).

Table 3-5*Mean Productivity of Selected Cases by Model Across Selection Ratios for Sample 1*

<u>Model</u>	<u>Selection Ratio</u>			
	<u>.1</u>	<u>.3</u>	<u>.5</u>	<u>.7</u>
OLS	54.93 [47.50 – 62.35]	52.82 [49.72 – 55.92]	52.77 [50.46 – 54.94]	51.60 [49.85 – 53.35]
QRM	55.38 [49.27 – 61.49]	53.34 [50.32 – 56.36]	52.48 [50.24 – 54.72]	52.15 [50.40 – 53.90]

Note: The number of selected cases per selection ratio varied, with 10% = 21, 30% = 63, 50% = 105, and 70% = 146. 95% confidence interval of selected cohort's mean productivity is reported in the brackets.

Sample 2

Descriptive Statistics

Descriptive statistics (e.g., mean, standard deviation, skewness, kurtosis) and Pearson correlations between study variables may be found in the Appendix in Table B-1 and Table B-2 for Sample 2.

Assessing the Presence of a Power-Law Distribution

Figure 3-3 contains the histogram of individual productivity and Table 3-6 summarizes the power-law distribution fit based on Clauset and colleagues' (2009)

procedure for Sample 2. Unlike Sample 1, the power-law distribution appeared to be a good fit for the data as the associated p -value from the goodness-of-fit test was greater than .10 (i.e., $p = .43$); however, high p -values should be interpreted with caution when the sample size is small (i.e., $N < 100$) as it is difficult to rule out the power-law distribution in these scenarios (Clauset et al., 2009). Given the larger scaling exponent (i.e., $\alpha = 8.67$) and moderate skewness (i.e., .91), it is unlikely individual productivity (i.e., renewal and upsell totals) follows a power-law distribution in Sample 2.

Figure 3-3

Histogram of Individual Productivity for Sample 2

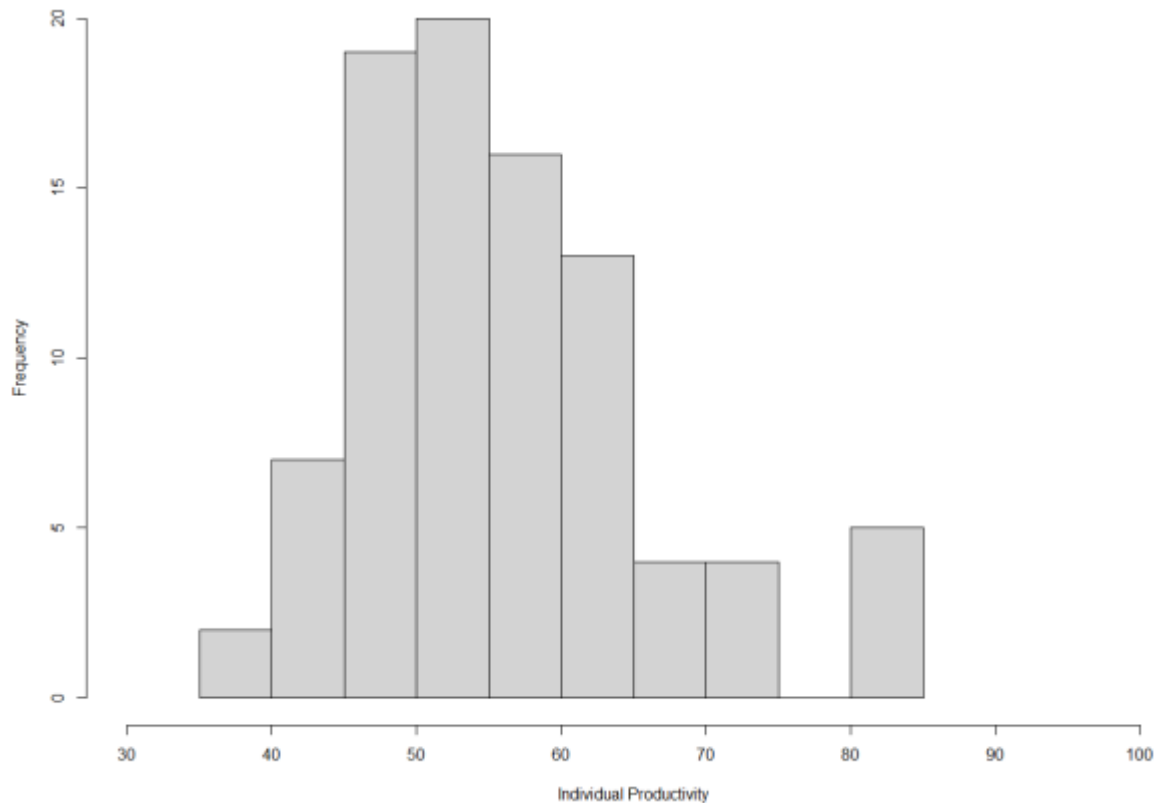


Table 3-6*Sample 2 Power-Law Distribution Fit and the Corresponding p-Value*

	<u>K-S</u>	<u>Xmin</u>	<u>α</u>	<u>p</u>
Productivity	.08	57.12	8.67	.43*

*Note: N = 90. *p > .10*

Hypothesis Testing

Like Sample 1, the plots in Figure 3-4 suggest that the OLS regression estimates were not representative of several predictor-criterion relationships. Specifically, Sensitivity (SEN), Ambition (ACH), Positivity (POS), and Assertiveness (ASR) had QRM estimates at various quantiles (i.e., black dots and broken lines) fall outside the 95% confidence intervals from the OLS regressions (i.e., dotted red lines). For example, the OLS regression underestimates the relationship between Sensitivity and individual productivity at lower levels of individual productivity (i.e., .4th quantile and below). Additionally, results in Table 3-7 from the LASSO quantile regression show that most predictors have heterogeneous or non-uniform relationships with individual productivity (e.g., estimates differed across quantiles). For example, Cooperativeness (COP), Sensitivity (SEN), Humility (HUM), Conceptual (INT), Mastery (MST), Structure (MTC), Drive (IND), Ambition (ACH), Positivity (POS), Liveliness (ENT), Assertiveness (ASR), and Power (POW) predict individual productivity at some quantiles but were not found to be statistically significant predictors by the OLS regression. Overall, 12 out of 15 predictors (i.e., 80.00%) have relationships with individual productivity that are heterogeneous or are unrepresentative of the OLS regression estimate. Given this, Hypothesis 1 is supported for Sample 2.

Figure 3-4

Sample 2 OLS Regression and Quantile Regression Plots

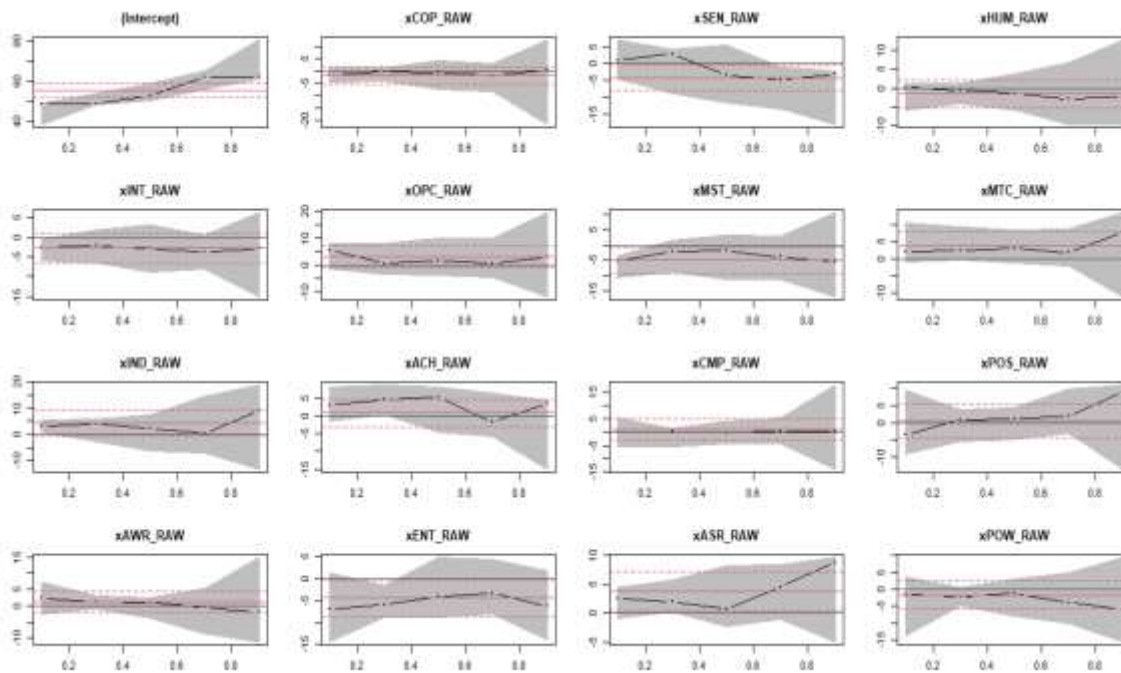


Table 3-7*Sample 2 t Statistics for OLS Regression and LASSO Quantile Regression Results*

Predictor	OLS	<u>LASSO Quantile Regression</u>				
		.1th	.3th	.5th	.7th	.9th
COP	-.77	-1.99*	-.59	-1.03	-.47	.00
SEN	-1.82	.00	2.35*	-1.10	-3.88**	-9.17**
HUM	-.69	2.22*	-1.22	-.33	-1.40	-4.14**
INT	-1.15	-3.63**	-.53	-1.14	-3.12**	.00
OPC	-1.21	1.78	.00	.85	.00	.66
MST	-1.91	-5.26**	-1.18	.00	-1.82	-6.30**
MTC	1.66	.56	2.03*	2.76**	1.60	.00
IND	1.39	2.13*	3.62**	2.90**	.00	8.85**
ACH	.31	3.64**	2.62*	1.66	-.11	-.40
CMP	.43	-.35	.00	.00	.00	1.91
POS	.22	-1.37	.48	.00	2.30*	2.82**
AWR	.62	.00	1.17	1.07	.00	.00
ENT	-1.77	-5.46**	-5.37**	-3.26**	-3.90**	-7.96**
ASR	1.91	.00	1.96*	.24	3.59**	4.43**
POW	-.65	.00	-1.54	-.63	-3.69**	-1.17

*Note: N = 90. *p < .05. **p < .01*

Next, an OLS regression-weighted model and a QRM-based model using the PLQC procedure were developed to test Hypotheses 2 and 3. Table 3-8 contains the selected predictors from this process and their assigned weights. As with Sample 1, none of the personality traits were found to significantly predict (i.e., $p < .05$) individual productivity based on the OLS regression results. As such, those with p -values less than .10 were considered, leading to a model that contains Mastery (MST), Liveliness (ENT), Sensitivity (SEN) and Assertiveness (ASR). In contrast, the PLQC procedure selected 12

out of the 15 possible predictors, with Liveliness (ENT) being the sole strong predictor (i.e., predicts all quantiles).

Table 3-8

Selection Model Details for Sample 2

<u>Model</u>	<u>MST</u>	<u>ENT</u>	<u>SEN</u>	<u>ASR</u>	<u>IND</u>	<u>INT</u>	<u>MTC</u>	<u>ACH</u>	<u>POW</u>	<u>HUM</u>	<u>POS</u>	<u>COP</u>
OLS	-.29	-.25	-.24	.22	--	--	--	--	--	--	--	--
QRM	-.11	-.21	-.10	.09	.13	-.07	.07	.06	.06	-.04	.03	-.03

Note: The weights presented in this table are a ratio of the standardized β coefficients. The absolute value of these ratios sum to 1 and accurately reflects the relationship with individual productivity for each predictor included in the model.

After following the procedure developed by Bing and colleagues (2007), the percentage of stars identified and mean productivity of the selected cohort for each selection ratio (i.e., .1, .3, .5, .7) were calculated. Results presented in Table 3-9 and Table 3-10 provide partial support for Hypotheses 2 and 3 as the QRM-based model generally resulted in more desirable selection outcomes than the OLS regression model. Specifically, the QRM model identified a higher percentage of productivity stars in smaller selection ratios (i.e., .1 and .3) and higher mean productivity in all four. Like Sample 1, the differences in mean productivity of the selected cases are small (e.g., ~2% improvement on mean productivity across all selection scenarios), and there is considerable overlap between the 95% confidence intervals of the sampled means for each selection ratio. As such, the practical gains associated with the QRM appear to be minor in Sample 2.

Table 3-9*Percentage of Stars Identified by Model Across Selection Ratios for Sample 2*

<u>Model</u>	<u>Selection Ratio</u>			
	<u>.1</u>	<u>.3</u>	<u>.5</u>	<u>.7</u>
OLS	38.46%	69.23%	100.00%	100.00%
PLQC	46.15%	76.92%	100.00%	100.00%

Note: The number of selected cases per selection ratio varied, with 10% = 9, 30% = 27, 50% = 45, and 70% = 63. $N = 13$ productivity stars were identified (i.e., any case that is at least 1.5 standard deviations above the mean on individual productivity).

Table 3-10*Mean Productivity of Selected Cases by Model Across Selection Ratios for Sample 2*

<u>Model</u>	<u>Selection Ratio</u>			
	<u>.1</u>	<u>.3</u>	<u>.5</u>	<u>.7</u>
OLS	66.97 [56.09 – 77.85]	61.82 [57.32 – 66.32]	60.58 [57.30 – 63.87]	57.96 [55.30 – 60.62]
PLQC	71.06 [63.40 – 78.72]	62.77 [58.02 – 67.52]	60.64 [57.50 – 63.78]	58.07 [55.39 – 60.75]

Note: The number of selected cases per selection ratio varied, with 10% = 9, 30% = 27, 50% = 45, and 70% = 63. 95% confidence interval of selected cohort's mean productivity is reported in the brackets.

Sample 3

Descriptive Statistics

Sample 3 was split into training and testing sets to understand robustness under cross-validation and how the models perform on new data. As such, descriptive statistics (e.g., mean, standard deviation, skewness, kurtosis) and Pearson correlations between study variables were conducted on the training set and may be found in Table C-1 and Table C-2 in the Appendix.

Assessing the Presence of a Power-Law Distribution

Figure 3-5 contains the histogram of individual productivity and Table 3-11 summarizes the power-law distribution fit for the training set in Sample 3. Following Clauset and colleagues' (2009) procedure, the power-law distribution was determined to be a poor fit for the data. The scaling exponent, α , indicated a lighter tail (i.e., 10.06), and the p -value associated with the goodness-of-fit test suggests a “near-zero” probability that the data follow a power law, like Sample 1 (Aguinis et al., 2018; Clauset et al., 2009). Despite this, individual productivity (i.e., total sales for the franchise location) was still found to be non-normally distributed as it is positively skewed (i.e., skewness = 1.51) and leptokurtic (i.e., kurtosis = 1.58).

Figure 3-5

Histogram of Individual Productivity for Sample 3 Training Set

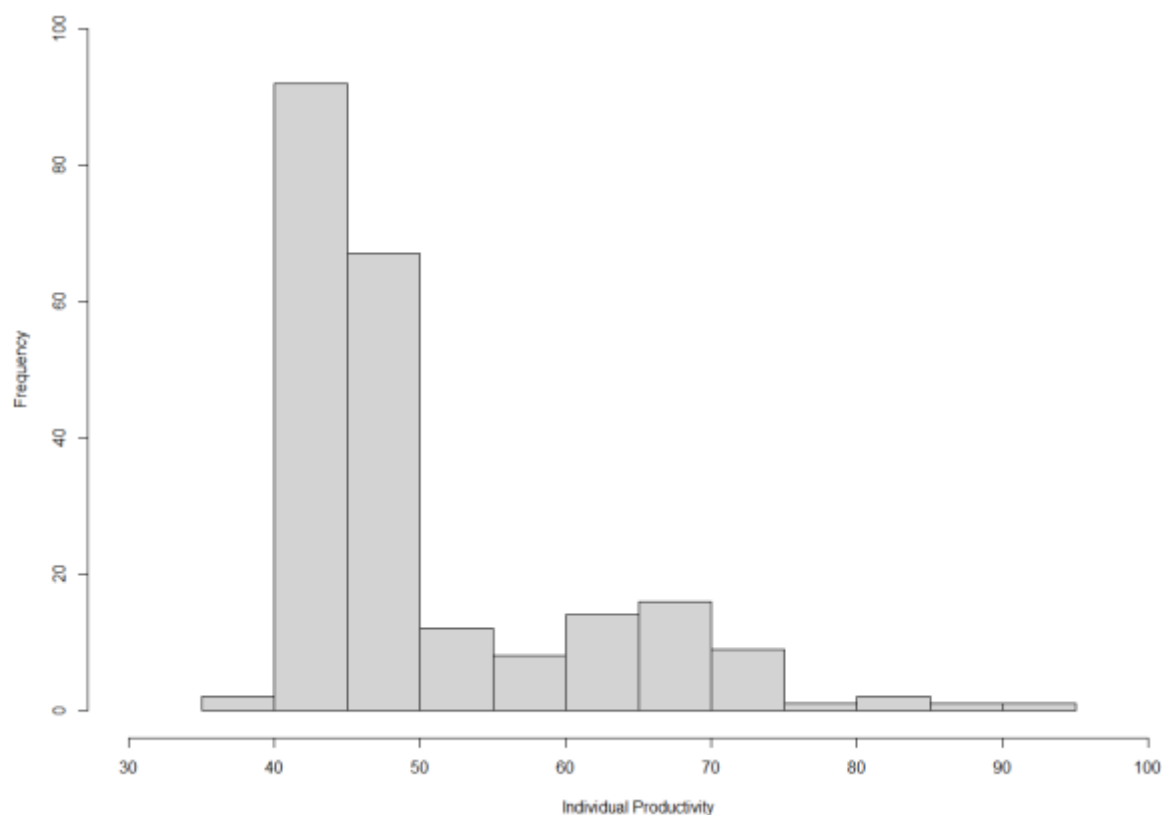


Table 3-11*Power-Law Distribution Fit and the Corresponding p-Value for Sample 3 Training Set*

	<u>K-S</u>	<u>Xmin</u>	<u>α</u>	<u>p</u>
Sales Goal Attainment	.07	46.99	10.06	.01

*Note: N = 225***Hypothesis Testing**

Similar to Sample 1 and 2, I compared OLS regression and QRM results using the plots found in Figure 3-6 to test Hypothesis 1 for the training set in Sample 3. Seven predictors, Sensitivity (SEN), Conceptual (INT), Mastery (MST), Drive (IND), Ambition (ACH), Positivity (POS), and Awareness (AWR), exhibited relationships with individual productivity that were not accurately represented by the OLS regression estimate as the 95% confidence interval from the OLS regression (i.e., dotted red lines) and the QRM estimates (i.e., black dots and broken lines) do not overlap. For example, the OLS regression overestimates the relationship between Conceptual and individual productivity at the .4th quantile and below. Further, when looking at the results comparing the OLS regression to the LASSO quantile regression in Table 3-12, Humility (HUM), Flexibility (OPC), Mastery (MST), Drive (IND), Ambition (ACH), Awareness (AWR), and Power (POW) significantly predict individual productivity at various quantiles but were not found to be significant predictors by the OLS regression. Additionally, Sensitivity (SEN) and Conceptual (INT) were found to be significant predictors of individual productivity by the OLS regression, but results from the LASSO quantile regression suggest these relationships were heterogeneous (i.e., Sensitivity and Conceptual are not significant predictors at lower levels of individual productivity). In total, 11 out of 15 predictors (i.e.,

73.33%) have relationships with individual productivity that are heterogeneous or not representative of the OLS regression estimate, providing support for Hypothesis 1.

Figure 3-6

Sample 3 OLS Regression and Quantile Regression Plots

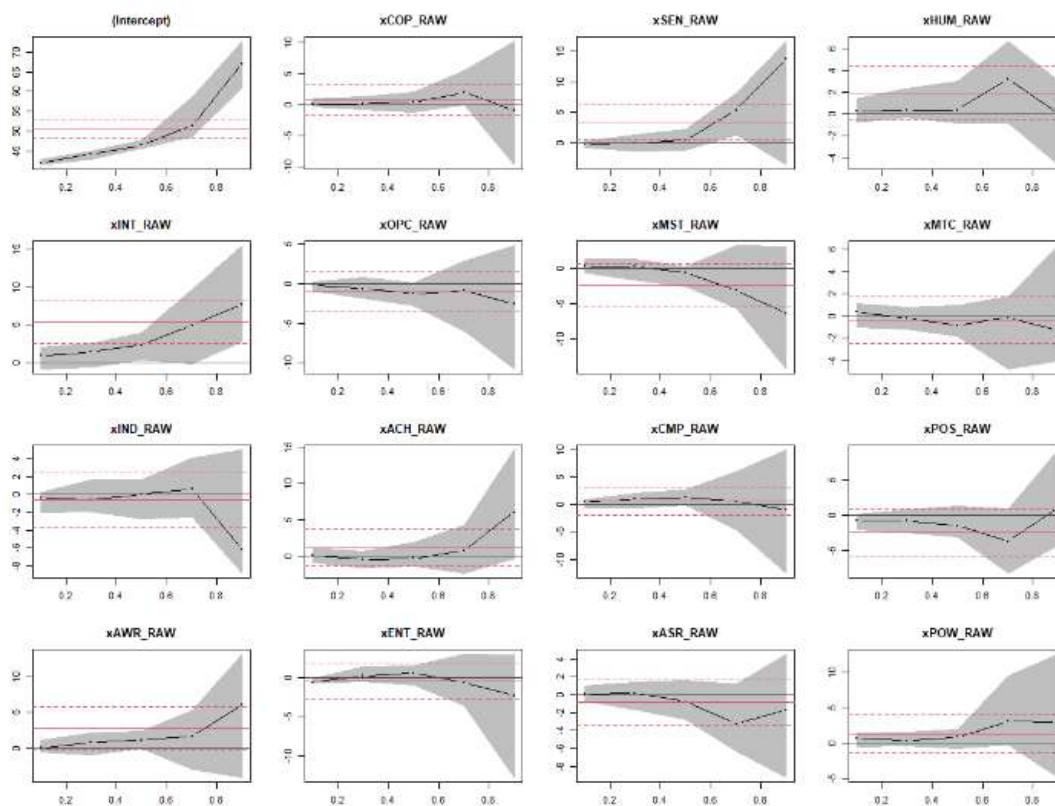


Table 3-12*Sample 3 t Statistics for OLS Regression and LASSO Quantile Regression Results*

<u>Predictor</u>	<u>OLS</u>	<u>LASSO Quantile Regression</u>				
		<u>.1th</u>	<u>.3th</u>	<u>.5th</u>	<u>.7th</u>	<u>.9th</u>
COP	.52	.00	.02	.06	1.64	.00
SEN	1.98*	-.08	-.09	.20	4.56**	9.03**
HUM	1.30	.00	.44	.29	1.98*	-1.33
INT	3.07**	.82	1.18	1.77	3.80**	5.36**
OPC	-.68	.00	-.44	-1.13	.00	-2.38*
MST	-1.32	.84	.20	-.51	-1.68	-5.84**
MTC	-.24	.34	-.12	-.47	-.42	-.07
IND	-.28	-.31	-.19	.00	.00	-1.98*
ACH	.76	.00	-.46	-.59	.23	2.11*
CMP	.44	.00	1.00	1.27	.00	.00
POS	-1.20	-.72	-.73	-.90	-1.71	.00
AWR	1.55	-.16	.71	.97	1.71	5.67**
ENT	-.33	-.66	.10	.98	-.73	-1.02
ASR	-.48	.11	-.02	-.82	-1.85	.00
POW	.81	.82	.18	.65	1.84	2.57*

Note: $N = 225$. * $p < .05$. ** $p < .01$

To test Hypotheses 2 and 3, I followed the procedure established by Bing and colleagues (2007) to demonstrate the practical superiority of the QRM. Unlike Sample 1 and 2, I employed a hold-out sample given the sufficient sample size. As such, both models (i.e., OLS regression and QRM) were built using the training set ($N = 225$) and evaluated on a separate testing set ($N = 56$). Table 3-13 reports the selected predictors and their assigned weights for the OLS regression and QRM-based model. The OLS regression model contains two predictors: Conceptual (INT) and Sensitivity (SEN),

whereas the QRM-based model contains nine. Note, all nine predictors selected using the PLQC procedure were determined to be partially weak as they predicted some, but not all quantiles.

Table 3-13

Selection Model Details for Sample 3

<u>Model</u>	<u>INT</u>	<u>SEN</u>	<u>AWR</u>	<u>MST</u>	<u>POW</u>	<u>OPC</u>	<u>IND</u>	<u>ACH</u>	<u>HUM</u>
OLS	.61	.39	--	--	--	--	--	--	--
QRM	.21	.24	.15	-.13	.10	-.07	-.04	.03	.02

Note: The weights presented in this table are a ratio of the standardized β coefficients. The absolute value of these ratios sum to 1 and accurately reflects the relationship with individual productivity for each predictor included in the model.

After generating the models on the training set, the percentage of stars identified and mean productivity of the selected cohort for each selection ratio (i.e., .1, .3, .5, .7) were calculated on the testing set following Bing and colleagues' (2007) procedure. Table 3-14 and Table 3-15 summarize the results for Sample 3. In contrast to Samples 1 and 2, the OLS regression model resulted in more desirable selection outcomes than the QRM-based model. Specifically, the OLS regression identified a higher percentage of productivity stars across each selection ratio and resulted in higher mean productivity for three out of four selected cohorts (i.e., .1, .3, and .5). As such, Hypotheses 2 and 3 were not supported in Sample 3; however, it is worth noting that the differences between the OLS regression and PLQC were again marginal as the OLS regression model resulted in an ~3% improvement on mean productivity across all selection scenarios when compared to the QRM-based model.

Table 3-14*Percentage of Stars Identified by Model Across Selection Ratios for Sample 3*

	<u>.1</u>	<u>.3</u>	<u>.5</u>	<u>.7</u>
OLS	16.67%	50.00%	66.67%	83.33%
QRM	0.00%	16.67%	50.00%	66.67%

Note: The number of selected cases per selection ratio varied, with 10% = 6, 30% = 17, 50% = 28, and 70% = 39. $N = 6$ productivity stars were identified (i.e., any case that is at least 1.5 standard deviations above the mean on individual productivity) in the testing set.

Table 3-15*Mean Productivity of Selected Cases by Model Across Selection Ratios for Sample 3*

	<u>.1</u>	<u>.3</u>	<u>.5</u>	<u>.7</u>
OLS	50.87 [39.07 – 62.67]	51.22 [46.09 – 56.35]	50.23 [46.66 – 53.80]	49.76 [46.81 – 52.72]
QRM	46.35 [41.49 – 51.21]	49.94 [45.70 – 54.18]	50.09 [46.73 – 53.45]	49.76 [46.91 – 52.61]

Note: The number of selected cases per selection ratio varied, with 10% = 6, 30% = 17, 50% = 28, and 70% = 39. 95% confidence interval of selected cohort's mean productivity is reported in the brackets.

CHAPTER 4

DISCUSSION

Given the importance of productivity stars to team performance (Oettl, 2012; Volmer & Sonnentag, 2011) and organizational success (Boudreau & Ramstad, 2007; Grigoriou & Rothaermel, 2014; Kehoe & Tzabbar, 2015; Tzabbar & Kehoe, 2014), it is paramount that organizations identify and retain them. Unfortunately, the literature has yet to uncover effective techniques to accurately identify productivity stars, meaning they may be overlooked during the selection process (Call et al., 2015; Terviö, 2009). A potential solution to this challenge is to improve the methods used to estimate future job performance during criterion-related validation studies given the limitations associated with current approaches (e.g., estimates are greatly influenced by outliers) (Aguinis & Edwards, 2013; Cohen et al., 2003; Hunter & Schmidt, 2004). As such, the purpose of this study was to investigate whether an alternative statistical method, the QRM, could improve selection-decision accuracy and star identification due to its robustness against outliers and ability to provide a more meaningful understanding of predictor-criterion relationships.

Regarding Hypothesis 1, in Sample 1 (account executives at a global professional services company), Sample 2 (account managers at a multinational business directory and

advertising firm), and Sample 3 (franchisee owners at a quick-service restaurant chain), the QRM consistently produced a more thorough conceptualization of the predictive validity between study variables (i.e., personality assessment scores and individual productivity). For example, across all three samples, 31 out of 45 (68.89%) predictors exhibited heterogeneous relationships with individual productivity, meaning the strength of these relationships varied across the distribution of individual productivity.

Regarding Hypotheses 2 and 3, the QRM-based model and OLS regression model resulted in similar selection outcomes across all three samples. Specifically, the PLQC procedure resulted in very small improvements for within-sample-results (e.g., star identification, cohort productivity) when compared to the OLS regression in Samples 1 and 2. However, the OLS regression model outperformed the QRM when tested out-of-sample (i.e., applied to new data) via the testing set in Sample 3. While these results provide partial support for Hypotheses 2 and 3, the use of hold-out sampling in Sample 3 provided a more robust model validation than Samples 1 and 2, suggesting that the QRM-based model may not add practical value above and beyond the OLS regression.

Overall, the results from Sample 3 are surprising as the QRM has consistently been shown to outperform OLS regression in forecasting efforts in the external literature (i.e., provide more accurate estimates on new data) (Furno, 2011; Lima & Meng, 2017; Meligkotsidou et al., 2021; Sayegh et al., 2014). There is, however, a potential explanation for the lack of support for Hypotheses 2 and 3 in Sample 3. While the overall sample size was reasonably large for a criterion-related validation study ($N = 281$) in Sample 3, the hold-out sample (i.e., testing set) used to validate the model was quite

small (i.e., $N = 56$), potentially leading to underlying differences between the training and testing sets and the broader population.

Implications for Theory and Practice

First, this study presents new evidence about the underlying distribution of individual productivity in highly complex, autonomous, as well as knowledge- and service-based roles (i.e., Account Executive, Account Manager, Franchisee Owner). Across each sample, results from Clauset and colleagues' (2009) procedure suggested that the power-law distribution was a poor fit for the data, despite Samples 1 and 3 being highly skewed and leptokurtic, and Sample 2 being moderately skewed. As such, results from this study provide some evidence that individual productivity may not be as "heavy-tailed" as the extant literature suggests (e.g., Aguinis et al., 2016, 2018; Crawford et al., 2015; Joo et al., 2017; O'Boyle & Aguinis, 2012; Ryazanova et al., 2017) when precautions are taken to control for extraneous factors and confounding variables (e.g., location, market, pricing for sales outcomes). Given the amount of interest and scholarly attention on the normality of individual productivity over the past decade (e.g., Aguinis et al., 2018; Beck et al., 2014; Joo et al., 2017; Vancouver, Xiaofei, Weindhardt, Steel, & Purl, 2016), these results provide additional insight into the potential causes of heavy-tailed productivity distributions.

Second, this study adds to our understanding of the relationship between personality traits and individual productivity. Results show the QRM provided a more thorough understanding about the relationships between personality and individual productivity than what could be gleaned from conditional means modeling (i.e., OLS regression). For example, 68.89% of the personality-individual productivity relationships

investigated in this study were found to be non-uniform or heterogeneous. Additionally, like findings from van Zyl and de Bruin (2018), who investigated the predictive relationship between personality traits and counterproductive work behaviors (CWBs), the results from this study also show that OLS regression has the tendency to underestimate the magnitude of these relationships at the high end of the distribution. As such, the field of I-O psychology, which has long debated the usefulness of personality assessments for selection (e.g., Morgeson et al., 2007; Ones, Dilchert, Viswesvaran, & Judge, 2007; Tett & Christiansen, 2007), may currently underestimate the value of personality assessments due to the prevalence of conditional means modeling. In other words, when predicting individual productivity, researchers should undoubtedly try to determine what is happening at extreme levels as this is where productivity stars are located. Unfortunately, results across all three samples in this study show this is exactly where the OLS regression fails as it typically underestimates the relationship strength at higher quantiles (e.g., .8th and above).

Third, this study investigated and provided initial evidence regarding the usefulness of a new approach for criterion-related validation and selection assessment battery design. Though the QRM-based models led to marginal improvements during in-sample validation and performed less effectively than the OLS regression model during out-of-sample testing, the results still offer meaningful insights for research and practice. Specifically, the QRM appears to be a viable alternative for criterion-related validation and assessment battery design when an OLS regression's assumptions are violated as the results for star identification and selection-decision accuracy were similar across all three samples.

Limitations and Future Directions

As with all research, this study is not without limitations. First, small sample sizes prevented more robust evaluation and comparison of current approaches used for criterion-related validation and selection assessment battery design and the newly proposed QRM-based process. In fact, larger sample sizes would have enabled an effective use of hold-out samples and provided more valuable insight into the QRM's ability to improve robustness under cross-validation, which has been shown extensively in other fields (e.g., Economics) (Furno, 2011; Lima & Meng, 2017; Meligkotsidou et al., 2021; Sayegh et al., 2014). Second, although individual productivity was moderately to heavily skewed in all three samples included in this study, each had lighter tails than what would be expected under a power-law distribution and what has been reported in the recent literature (e.g., Aguinis et al., 2018; Joo et al., 2017). Due to this, the samples used in this study may have unintentionally suppressed the potential benefits of the QRM over the OLS regression.

Though this study provided a robust evaluation of the proposed QRM-based process for criterion-related validation and selection assessment battery design, there is still a need for continued research to fully understand its effectiveness. First, future research leveraging larger sample sizes and more robust cross-validation techniques are needed. Second, identifying and using samples with more influential cases (i.e., productivity stars) and distributions with heavier tails may provide better insight into the usefulness of the QRM. Third, predictor-criterion relationships were generally quite small as depicted by conditional mean-based methods (i.e., Pearson correlations and OLS regression). Due to these weak, linear relationships, the OLS regression-based models in

Samples 1 and 2 consisted of predictors with p -values between .05 and .10, and most of the predictors were deemed to be partially weak by the LASSO quantile regression across all three samples. As such, future research should seek to understand the practical benefits of the QRM-based process using samples containing stronger predictor-criterion relationships. Fourth, future research should investigate the effect of quantile selection and weighting schemes as previous research suggests these factors affect model performance (e.g., Lima & Meng, 2017). In this study, five quantiles of interest (i.e., .1th, .3th, .5th, .7th, and .9th) were included in model development and weighted equally to obtain a single, global estimate. Due to the importance and impact of productivity stars, one might consider disproportionately weighting quantile results at the high end of the distribution (e.g., .9th) where stars are located. Lastly, future research should investigate the impact of the tuning parameter, lambda (λ), which controls the strength of the penalty term in LASSO regression, on model development and performance. In this study, a default value was selected for λ according to the proposal of Belloni and Chernozhukov (2011). While this approach eliminated some predictors from the model by reducing their coefficients to zero (i.e., shrinkage), larger values for λ may identify a sparser model with a smaller subset of predictors (i.e., remove additional partially weak predictors), potentially leading to improved star identification and selection-decision accuracy.

Beyond the QRM, additional research is needed to vet the viability of other statistical methods that provide robust estimates when faced with heavy-tailed distributions and influential cases like productivity stars. For example, O'Boyle and Aguinis (2012) suggest Bayesian techniques are likely applicable as researchers may test hypotheses without assuming normality as one may specify the distribution of the

criterion a priori (Kruschke et al., 2012). Moreover, given the likelihood that influential cases introduce nonlinearity (O'Boyle & Aguinis, 2012), additional research should examine the use of hierarchical polynomial regressions in selection assessment battery design as well as lower- and upper-bound cut-scores to mitigate the “too-much of a good thing” phenomenon (Carter et al., 2014; Castille, Theys, & Khan, 2016; Le et al., 2011).

Finally, additional research is still needed to better understand the causes of heavy-tailed productivity distributions. To date, most of the literature has focused on work-related factors (e.g., autonomy, job complexity, star's proximity to the organizations strategic core) and the operationalization of job performance (e.g., Beck et al., 2014) that enable productivity stars to emerge (e.g., Aguinis et al., 2016; Aguinis & O'Boyle, 2014; Vancouver et al., 2016). However, more research is needed to understand the effect that extraneous factors and confounding variables (e.g., location differences, tenure) have on the underlying distribution of individual productivity and productivity star emergence.

Conclusions

The purpose of this study was to test the newly proposed QRM-based criterion-related validation procedure and provide evidence that the QRM provides distinct theoretical and practical advantages over traditional approaches when data are non-normally distributed and influenced by the presence of productivity stars. Specifically, I hypothesized that the QRM would produce a more detailed conceptualization of the predictive validity between selection assessments and individual productivity, and that selection assessment batteries designed using the QRM and PLQC procedure would result in greater practical usefulness (i.e., star identification and selection-decision

accuracy) over and above OLS regression. Results showed that the QRM provided a much more comprehensive understanding of the predictor-criterion relationships across all three samples, and that the proposed QRM-based criterion-related validation procedure had similar outcomes to the OLS regression with respect to star identification and selection-decision accuracy. Given the limitations and future directions outlined previously, more research is needed to fully understand the utility of the QRM for assessment and selection purposes.

REFERENCES

- Adler, R. J., Feldman, R. E., & Taqqu, M. S. (1998). *A practical guide to heavy tails: Statistical techniques and applications*. Birkhauser.
- Aguinis, H. (2013). *Performance Management* (3rd ed.). Pearson/Prentice Hall.
- Aguinis, H., & Edwards, J. (2013). Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies*, *51*(1), 143-174.
- Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270-301.
- Aguinis, H., Ji, Y. H., & Joo, H. (2018). Gender productivity gap among star performers in STEM and other scientific fields. *Journal of Applied Psychology*, *103*(12), 1283-1306.
- 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.

- Aguinis, H., Petersen, S. A., & Pierce, C. A. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimated moderating effects of categorical variables. *Organizational Research Methods*, 2(4), 315-339.
- Aguinis, H., Pierce, C. A., Bosco, F., & Muslin, I. S. (2009). First Decade of Organizational Research Methods: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods*, 12(1), 69-112.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37(1), 1-15.
- Aoyama, H., Yoshikawa, H., Iyetomi, H., & Fujiwara, Y. (2010). Productivity dispersion: Facts, theory, and implications. *Journal of Economic Interaction and Coordination*, 5(1), 27-54.
- Arnold, B. C. (1983). *Pareto Distributions*. International Cooperative Publishing House.
- Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917 - 1992. *Journal of Applied Psychology*, 77(6), 836-874.
- Barley, S. R., Bechky, B. A., & Milliken, F. J. (2017). The changing nature of work: Careers, identities, and work lives in the 21st century. *Academy of Management Discoveries*, 32(2), 111-115.

- 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, T. E., Robertson, M. M., & Vandenberg, R. J. (2019). Nonlinear transformations in organizational research: Possible problems and potential solutions. *Organizational Research Methods, 22*(4), 831-866.
- Bedeian, A. G., & Armenakis, A. A. (1998). The cesspool syndrome: How dreck floats to the top of declining organizations. *The Academy of Management Executive, 12*(1), 58-67.
- Belloni, A., & Chernozhukov, V. (2011). L1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics, 39*(1), 82-130.
- Berry, M. J. A., & Linoff, G. S. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Wiley Publishing Inc.
- Binet, A. (1903). *L'étude expérimentale de l'intelligence* [The experimental study of intelligence.] Schleicher.
- Binet, A., & Simon, T. (1908). Le développement de l'intelligence chez les enfants. *Année Psychol, 14*, 1-94.
- Bing, M. N., Stewart, S. M., Davison, H. K., Green, P. D., McIntyre, M. D., & James, L. R. (2007). An integrative typology of personality assessment for aggression: Implications for predicting counterproductive workplace behavior. *Journal of Applied Psychology, 92*(3), 722-744.

- Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, 74(3), 478-494.
- Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, 57(1), 49-64.
- Boudreau, J. W., & Ramstad, P. M. (2007). *Beyond HR: The New Science of Human Capital*. Harvard Business School Publishing.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2), 211-252.
- Boyce, A. S., Conway, J. S., & Caputo, P. M. (2014). *Development and validation of Aon Hewitt's personality model and Adaptive Employee Personality Test (ADEPT-15)*. Unpublished technical report.
- Buchinsky, M. (1998). Recent advances in quantile regression models: A practical guideline for empirical research. *Journal of Human Resources*, 33(1), 88-126.
- Burke, R. J., & Ng, E. (2006). The changing nature of work and organizations: Implications for human resource management. *Human Resource Management Review*, 16(2), 86-94.
- Call, M. L., Nyberg, A. J., & Thatcher, M. B. (2015). Stargazing: An integrative conceptual review, theoretical reconciliation, and extension for star employee research. *Journal of Applied Psychology*, 100(3), 623-640.
- Canter, R. R. (1953). A rating-scoring method for free-response data. *Journal of Applied Psychology*, 37(6), 455-457.

- Carter, N. T., Dalal, D. K., Boyce, A. S., O'Connell, M. S., Kung, M.-C., & Delgado, K. M. (2014). Uncovering curvilinear relationships between conscientiousness and job performance: How theoretically appropriate measurement makes and empirical difference. *Journal of Applied Psychology, 99*(4), 564-586.
- Cascio, W. F., & Aguinis, H. (2004). *Applied Psychology in Human Resource Management*. Prentice Hall.
- Cascio, W. F., & Aguinis, H. (2008). Research in industrial and organizational psychology from 1963 to 2007: Changes, choices, and trends. *Journal of Applied Psychology, 93*(5), 1062-1081.
- Castille, C., Theys, E. R., & Khan, S. (2016). *Too much of a good thing? Nonlinear personality-performance relations*. Poster presented at the annual meeting of the Society for Industrial and Organizational Psychology, Anaheim, CA.
- Chambers, R. C. (2016). *Evaluating indicators of job performance: Distributions and types of analyses* (UMI No. 10307765). [Doctoral dissertation, Louisiana Tech University]. ProQuest Dissertations & Theses Global database.
- Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance, 22*(2), 105-127.
- Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267-307.

- Clauset, A., Shalizi, C. A., & Newman, M. E. T. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661-703.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for behavioral sciences* (3rd ed.). Lawrence Erlbaum.
- Cortina, J. M. (2002). Big things have small beginnings: An assortment of “minor” methodological misunderstandings. *Journal of Management*, *28*(3), 339-362.
- Crain, W. M., & Tollison, R. D. (2002). Consumer choice and the popular music industry: A test of the Superstar Theory. *Empirica*, *29*(1), 1-9.
- 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.
- Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., & Hulin, C. L. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to Support Army Selection and Classification Decisions* (Technical Report 1311). For Belvoir, Virginia: Army Research Institute for the Behavioral and Social Sciences.
- Emerson, J. D. (1983). Mathematical aspects of transformation. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Statistical and Methodological Myths and Urban Legends* (pp. 143-164). Routledge.
- Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register*, *43*, 38294-38309.

- Ferguson, L. W. (1947). The development of a method of appraisal for assistant managers. *Journal of Applied Psychology*, 31(3), 306-311.
- Furno, M. (2011). Goodness of fit and misspecification in quantile regression. *Journal of Educational and Behavioral Statistics*, 36(1), 105-131.
- Galton, F. (1889). *Natural inheritance*. Macmillan and Co.
- Gastwirth, J. L. (1966). On robust procedures. *Journal of the American Statistical Association*, 61(316), 929-948.
- Gatewood, R. D., & Field, S. (2001). *Human Resource Selection*. Harcourt Brace & Company.
- Grant, A. M. (2013). Rocking the boat but keeping it steady: The role of emotion regulation in employee voice. *Academy of Management Journal*, 56(6), 1703-1723.
- Grant, A. M., & Sumanth, J. J. (2009). Mission possible? The performance of prosocially motivated employees depends on manager trustworthiness. *Journal of Applied Psychology*, 94(4), 927-944.
- Greene, H. W. (2008). *Econometric Analysis* (6th ed.). Prentice Hall.
- Grigoriou, K., & Rothaermel, F. T. (2014). Structural microfoundations of innovation the role of relational stars. *Journal of Management*, 40(2), 586-615.
- Guion, R. M. (2011). *Assessment, Measurement, and Prediction for Personnel Decisions*. Lawrence Erlbaum.
- Hao, I., & Naiman, D. Q. (2007). *Quantile regression*. Sage.
- Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5), 1163-1174.

- Hollenbeck, J. R., DeRue, D. S., & Mannor, M. (2006). Statistical power and parameter stability when subjects are few and tests are many: Comment on Peterson, Smith, Martorana, and Owens (2003). *Journal of Applied Psychology, 91*(1), 1-5.
- Hull, C. L. (1928). *Aptitude Testing*. Chicago, IL: World Book Company.
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Sage.
- Hunter, J. E., Schmidt, F. L., & Judiesch, M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75*(1), 28-42.
- Ismail, E. A. R. (2015). Behavior of lasso quantile regression with small sample sizes. *Journal of Multidisciplinary Engineering Science and Technology, 2*(3), 388-394.
- Joo, H., Aguinis, H., & Bradley, K. J. (2017). Not all nonnormal distributions are created equal: Improved theoretical and measurement precision. *Journal of Applied Psychology, 102*(7), 1022-1053.
- Judge, G. C., Hill, R. C., Griffiths, W. E., Lutkepohl, H., & Lee, T. C. (1988). *Introduction to the theory and practice of econometrics*. Wiley.
- Kehoe, R. R., & Tzabbar, D. (2015). Lighting the way or stealing the shine? An examination of the duality in star scientists' effects on firm innovative performance. *Strategic Management Journal, 36*(5), 709-727.
- Kim, Y., Kim, T. H., & Ergun, T. (2015). The instability of Pearson correlation coefficient in the presence of coincidental outliers. *Finance Research Letters, 13*, 243-257.

- Kirpich, A., Ainsworth, E. A., Wedow, J. M., Newman, J. R., Michailidis, G., & McIntyre, L. M. (2018). Variable selection in omics data: A practical evaluation of small sample sizes. *PloS One*, *13*(6).
- Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.
- Koenker, R., & Bassett, G. (1978). Regression quantiles, *Econometrica*, *46*(1), 33-50.
- Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*(4), 722-752.
- Lawshe, C. H. (1985). Inferences from personnel tests and their validity. *Journal of Applied Psychology*, *70*(1), 237-238.
- Le, H., Oh, I.-S., Robbins, S. B., Ilies, R., Holland, E. & Westrick, P. (2011). Too much of a good thing: Curvilinear relationships between personality traits and job performance. *Journal of Applied Psychology*, *96*(1), 113-133.
- Lever, J., Krzywinski, M., & Altman, N. (2016). Points of significance: Model selection and overfitting. *Nature methods*, *13*(9), 703-705.
- Li, M. (2015). Moving beyond the linear regression model: Advantages of the quantile regression model. *Journal of Management*, *41*(1), 71-98.
- Lima, L. R., & Meng, F. (2017). Out-of-sample return predictability: A quantile combination approach. *Journal of Applied Econometrics*, *32*(4), 877-895.
- Malhotra, P., & Singh, M. (2016). Indirect impact of high performers on the career advancement of their subordinates. *Human Resource Management Review*, *26*(3), 209-226.

- Martin, N., & Theys, E. R. (2019). *ADEPT-15 Technical Report*. Aon's Assessment Solutions.
- Meligkotsidou, L., Panopoulou, E., Vrontos, I. D., & Vrontos, S. D. (2021). Out-of-sample equity premium prediction: A complete subset quantile regression approach. *The European Journal of Finance*, 27(1-2), 110-135.
- Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology*, 60(4), 1029-1049.
- Motowidlo, S. J., & Borman, W. C. (1977). Behaviorally anchored scales for measuring morale in military units. *Journal of Applied Psychology*, 62(2), 177-183.
- Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology*, 1(2), 148-160.
- 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.
- Oettl, A. (2012). Reconceptualizing stars: Scientist helpfulness and peer performance. *Management Science*, 58(6), 1122-1140.
- Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology*, 60(4), 995-1027.

- Orme, S., Ralph, S. G., Birchall, A., Lawson-Matthew, P., McLean, K., & Channer, K. S. (1999). The normal range for inter-arm differences in blood pressure. *Age and aging, 28*(6), 537-542.
- Petscher, Y., Logan, J. A. R., & Zhou, C. (2013). Extending conditional means modeling: An introduction to quantile regression. In Y. Petscher, C. Schatsneider, & D. L. Compton (Eds.), *Applied quantitative analysis in education and the social sciences*. Routledge.
- Pulakos, E. D. (2005). *Selection assessment methods*. Society for Human Resource Management (SHRM) Foundation.
- 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.
- Resnick, S. I. (2006). *Heavy-tail phenomena: Probabilistic and statistical modeling*. Springer-Verlag.
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3-32.
- Roth, P. L., Le, H., Oh, I. S., Van Iddekinge, C. H., Buster, M. A., Robbins, S. B., & Campion, M. A. (2014). Differential validity for cognitive ability tests in employment and educational settings: Not much more than range restriction? *Journal of Applied Psychology, 99*(1), 1-20.

- Ryazanova, O., McNamara, P., & Aguinis, H. (2017). Research performance as a quality signal in international labor markets: Visibility of business schools worldwide through a global research performance system. *Journal of World Business, 52*(6), 831-841.
- Sackett, P. R., Dahlke, J. A., Shewach, O. R., & Kuncel, N. R. (2017). Effects of predictor weighting methods on incremental validity. *Journal of Applied Psychology, 102*(10), 1421-1434.
- Sayegh, A. S., Munir, S., & Habeebullah, T. M. (2014). Comparing the performance of statistical models for predicting PM₁₀ concentrations. *Aerosol and Air Quality Research, 14*(3), 653-665.
- Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights. *Educational and Psychological Measurement, 31*(3), 699-714.
- 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., & Johnson, R. H. (1973). Effect of race on peer ratings in an industrial situation. *Journal of Applied Psychology, 57*(3), 237-241.
- Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82*(5), 719-730.
- 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.
- Schultz, D. G., & Siegel, A. I. (1961). Generalized Thurstone and Guttman scales for measuring technical skills in job performance. *Journal of applied psychology, 45*(3), 137-142.
- Shoemaker, A. L. (1996). What's normal? Temperature, gender, and heart rate. *Journal of Statistics Education, 4*(2), 4.
- Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Eds.), *The concept of validity: Revisions, new directions, and applications* (pp. 3-22). Information Age Publishing.
- Society for Industrial and Organizational Psychology (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.). SIOP.
- Stark, S. (2002). A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment: The generalized graded unfolding model for multiunidimensional paired comparison responses. *Dissertation Abstracts International, 63*, 1084.
- Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95*(2), 334-344.
- Terviö, M. (2009). Superstars and mediocrities: Market failure in the discovery of talent. *The Review of Economic Studies, 76*(2), 829-850.

- Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007). *Personnel Psychology*, 60(4), 967-993.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1), 267-288.
- Tiffin, J. (1947). *Industrial psychology*. Prentice-Hall.
- Toliver, R. F., & Constable, T. J. (1998). *Die Deutschen jagdflieger-asse* [The German fighter aces] 1939-1945. Motorbuch Verlag.
- Tukey, J. W. (1957). On the comparative anatomy of transformations. *Annals of Mathematical Statistics*, 28(3), 602-632.
- Tukey, J. W. (1977). *Explanatory data analysis*. Addison-Wesley.
- Tzabbar, D., & Kehoe, R. R. (2014). Can opportunity emerge from disarray? An examination of exploration and exploitation following star scientist turnover. *Journal of Management*, 40(2), 449-482.
- Urduan, T. C. (2017). *Statistics in plain English (4th ed.)*. Routledge Publishing.
- van Zyl, C. J. J., & de Bruin, G. P. (2018). Predicting counterproductive work behavior with narrow personality traits: A nuanced examination using quantile regression. *Personality and Individual Differences*, 131(1), 45-50.
- Vancouver, J. B., Xiaofei, L., Weinhardt, J. M., Steel, P., & Purl, J. D. (2016). Using a computational model to understand possible sources of skews in distributions of job performance. *Personnel Psychology*, 69(4), 931-974.

- Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability. *Human Performance, 19*(3), 175-189.
- Viswesvaran, C., & Ones, D. S. (2000). Perspectives on models of job performance. *International Journal of Selection and Assessment, 8*(4), 216-226.
- Volmer, J., & Sonnentag, S. (2011). The role of star performers in software design teams. *Strategic Management Journal, 39*(5), 1239-1267.
- Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*(2), 213-217.
- Yeo, I., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika, 87*(4), 954-959.
- Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75*(6), 579-652.
- Zucker, L. G., & Darby, M. R. (1996). Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry. *Proceedings of the National Academy of Sciences, 93*(23), 12709-12716.

APPENDIX A

**SAMPLE 1 DEMOGRAPHICS, DESCRIPTIVES, AND
PEARSON CORRELATIONS**

Table A-1*Participant Demographics for Sample 1*

<u>Sample Characteristics</u>	<u>N</u>	<u>%</u>	<u>M</u>	<u>SD</u>
Gender				
Female	92	44.02%	--	--
Male	116	55.50%	--	--
N/A	1	.48%		
Race/Ethnicity				
American Indian or Alaska Native	1	.48%	--	--
Asian	6	2.87%	--	--
Black or African American	2	.96%	--	--
Hispanic or Latino	3	1.44%	--	--
Two or More Races	1	.48%	--	--
White	183	87.56%	--	--
N/A	13	6.22%	--	--
Age	--	--	50.00	11.00

Note: N = 209.

Table A-2*Descriptive Statistics for Sample 1*

<u>Variable</u>	<u>Minimum</u>	<u>Maximum</u>	<u>M</u>	<u>SD</u>	<u>Skew</u>	<u>Kurt</u>
COP	-1.23	1.61	.33	.48	.32	.18
SEN	-1.58	1.76	.06	.43	-.37	1.95
HUM	-2.21	1.48	-.02	.50	-.23	2.06
INT	-1.55	1.09	-.20	.43	-.28	.89
OPC	-.80	1.26	.14	.44	.17	-.21
MST	-1.32	1.30	.13	.44	-.01	-.04
MTC	-1.35	1.42	.01	.51	-.01	-.12
IND	-1.03	1.92	.31	.42	.26	1.75
ACH	-1.04	1.57	.26	.48	.26	.13
CMP	-1.02	1.87	.19	.44	-.05	.72
POS	-.83	1.72	.25	.42	.05	.28
AWR	-.98	1.52	.09	.43	.47	.72
ENT	-1.42	2.12	.32	.47	.12	1.96
ASR	-.97	1.52	.15	.46	.23	-.25
POW	-.82	2.34	.40	.49	.80	1.89
Productivity	38.19	97.31	51.20	10.38	1.49	3.18

Note: N = 209.

Table A-3*Pearson Correlations Between Variables in Sample 1*

Variable	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>	<u>13</u>	<u>14</u>	<u>15</u>	<u>16</u>
1. COP	--															
2. SEN	.27	--														
3. HUM	.05	.10	--													
4. INT	-.07	.18	.09	--												
5. OPC	.07	.15	.08	.37	--											
6. MST	-.05	.15	.14	.45	.34	--										
7. MTC	-.09	-.16	-.01	-.16	-.30	-.13	--									
8. IND	-.09	.01	.10	.04	.07	.20	.14	--								
9. ACH	.04	.15	-.17	.10	.17	.26	.10	.33	--							
10. CMP	.11	.01	.04	.10	.21	.16	-.17	.22	.14	--						
11. POS	.09	.22	.00	.08	.08	.18	-.08	.08	.05	.26	--					
12. AWR	.05	.11	.05	.12	.22	.01	-.07	.03	.07	.15	.03	--				
13. ENT	.25	.24	-.02	.12	.25	.12	-.27	-.08	.09	.09	.23	.06	--			
14. ASR	-.01	.18	-.07	.29	.28	.24	-.20	.03	.17	-.10	.25	.06	.22	--		
15. POW	.15	.13	-.09	.14	.20	.27	-.09	.13	.28	.17	.23	.04	.17	.27	--	
16. Productivity	.01	-.01	-.05	-.02	-.01	-.06	.04	-.03	.05	-.02	-.08	-.13	-.07	.06	.06	--

Note: $N = 209$. All bolded correlations are significant at $p < .05$.

APPENDIX B

SAMPLE 2 DESCRIPTIVES AND PEARSON CORRELATIONS

Table B-1*Descriptive Statistics for Sample 2*

<u>Variable</u>	<u>Minimum</u>	<u>Maximum</u>	<u>M</u>	<u>SD</u>	<u>Skew</u>	<u>Kurt</u>
COP	-.98	1.34	.22	.52	-.06	-.38
SEN	-1.24	1.15	.15	.46	-.48	.19
HUM	-1.07	1.14	-.03	.49	.17	-.20
INT	-1.35	1.48	-.20	.49	.49	.99
OPC	-1.32	1.20	.12	.45	-.13	.95
MST	-.89	1.40	.16	.45	-.07	-.42
MTC	-1.00	0.69	-.03	.45	-.30	-.92
IND	-.67	1.49	.35	.36	-.02	.24
ACH	-1.10	1.31	.36	.49	-.56	.52
CMP	-1.00	1.33	.30	.49	-.34	-.17
POS	-.74	1.17	.38	.38	-.64	.47
AWR	-2.12	1.84	.09	.51	-.70	4.50
ENT	-.82	2.20	.39	.51	.94	2.47
ASR	-.88	1.93	.26	.51	.64	.82
POW	-.53	1.28	.28	.42	.46	-.52
Productivity	39.02	84.03	55.99	10.04	.91	.69

Note: N = 90.

Table B-2*Pearson Correlations Between Variables in Sample 2*

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. COP	--															
2. SEN	.28	--														
3. HUM	.24	.13	--													
4. INT	.09	-.08	-.01	--												
5. OPC	.09	-.01	-.13	.25	--											
6. MST	.01	.00	.12	.15	.45	--										
7. MTC	.12	-.04	.20	-.14	-.08	.09	--									
8. IND	-.08	-.16	-.11	-.21	-.03	.11	.01	--								
9. ACH	-.17	-.20	-.26	.29	.07	.12	-.03	.25	--							
10. CMP	.03	.05	-.09	.11	.08	.17	.17	.16	.01	--						
11. POS	.07	.02	.01	.11	.14	.18	.18	.10	-.09	.38	--					
12. AWR	.01	-.03	-.03	-.05	-.05	-.05	-.05	-.07	.04	.17	-.04	--				
13. ENT	.46	.32	.08	.09	.00	.03	.03	.20	.04	.10	.24	.01	--			
14. ASR	-.01	-.06	-.15	.05	-.11	-.02	-.02	.11	.22	-.04	.00	-.08	.10	--		
15. POW	-.07	-.02	-.05	.16	-.05	.24	.24	.13	.21	.00	-.01	-.16	.16	.27	--	
16. Productivity	-.26	-.34	-.21	-.19	-.02	-.17	-.17	.18	.11	.05	-.06	.08	-.33	.17	-.12	--

Note: $N = 90$. All bolded correlations are significant at $p < .05$.

APPENDIX C

SAMPLE 3 DESCRIPTIVES AND PEARSON CORRELATIONS

Table C-1*Descriptive Statistics for Sample 3 Training Set*

<u>Variable</u>	<u>Minimum</u>	<u>Maximum</u>	<u>M</u>	<u>SD</u>	<u>Skew</u>	<u>Kurt</u>
COP	-1.33	1.78	.25	.52	-.10	.22
SEN	-1.48	1.37	.07	.44	-.22	.49
HUM	-.94	1.68	.26	.49	.19	.03
INT	-1.62	1.31	-.18	.46	-.17	.34
OPC	-1.63	1.75	.22	.51	-.09	.49
MST	-1.36	1.10	.12	.43	-.27	-.06
MTC	-1.51	1.56	-.11	.57	-.09	-.07
IND	-1.31	1.48	.14	.42	-.41	.48
ACH	-1.06	1.72	.32	.49	.10	.25
CMP	-1.16	1.44	.16	.48	-.05	.19
POS	-1.20	1.20	.31	.38	-.41	.46
AWR	-1.54	1.02	.04	.40	-.41	.62
ENT	-1.59	1.98	.19	.59	.04	.76
ASR	-1.18	1.71	.26	.47	.04	.64
POW	-.72	2.36	.55	.46	1.10	1.67
Productivity	39.78	90.70	50.22	10.28	1.51	1.58

Note: N = 225.

Table C-2*Pearson Correlations Between Variables in the Training Set for Sample 3*

Variable	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>	<u>13</u>	<u>14</u>	<u>15</u>	<u>16</u>
1. COP	--															
2. SEN	.33	--														
3. HUM	.14	.11	--													
4. INT	.02	.05	-.14	--												
5. OPC	.17	.16	-.05	.36	--											
6. MST	.14	.13	.17	.28	.25	--										
7. MTC	-.09	-.06	.06	-.13	-.15	.08	--									
8. IND	-.06	.04	.14	-.08	-.12	.28	.30	--								
9. ACH	-.05	-.04	-.04	.09	.04	.29	.21	.29	--							
10. CMP	.21	.09	.13	.03	.07	.08	.17	.10	.01	--						
11. POS	.18	.21	.13	.08	.17	.12	.09	.17	.18	.21	--					
12. AWR	.10	.17	.12	-.02	.03	.08	.03	.21	.14	.19	.14	--				
13. ENT	.30	.38	-.04	.32	.32	.19	-.14	-.04	-.03	.04	.29	.03	--			
14. ASR	-.07	-.04	-.05	.25	.09	.14	-.04	.16	.17	-.14	-.05	.05	.10	--		
15. POW	.11	.01	-.02	.17	.22	.19	.03	.15	.18	.19	.19	.20	.05	.21	--	
16. Productivity	.08	.16	.07	.18	.03	.00	-.04	-.02	.03	.07	-.01	.14	.06	.01	.08	--

Note: N = 225. All bolded correlations are significant at $p < .05$.

APPENDIX D

HUMAN USE EXEMPTION LETTER

MEMORANDUM

TO: Mr. Evan Theys and Dr. Mitzi Desselles

FROM: Dr. Richard Kordal, Director of Intellectual Property & Commercialization
(OIPC)
rkordal@latech.edu

SUBJECT: HUMAN USE COMMITTEE REVIEW

DATE: May 19, 2020

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

HUC 20-119

**“How Quantile Regression Can Enhance Our
Ability to Identify Productivity Stars”**

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 May 18, 2020 and this project will need to receive a continuation review by the IRB if the project continues beyond May 18, 2021.*** ANY CHANGES to your protocol procedures, including minor changes, should be reported immediately to the IRB for approval before implementation. Projects involving NIH funds require annual education training to be documented. For more information regarding this, contact the Office of Sponsored Projects.

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 Sponsored Projects or IRB in writing. The project should be discontinued until modifications can be reviewed and approved.