2019

The Effect of Community Evaluators on the Selection of Entry-Level Police Officers

Eric Hutchison
Walden University

Walden University

College of Social and Behavioral Sciences

This is to certify that the doctoral dissertation by

Eric Sean Hutchison

has been found to be complete and satisfactory in all respects,
and that any and all revisions required by
the review committee have been made.

Review Committee
Dr. James Herndon, Committee Chairperson, Psychology Faculty
Dr. John Schmidt, Committee Member, Psychology Faculty
Dr. Marlon Sukal, University Reviewer, Psychology Faculty

Chief Academic Officer
Eric Riedel, Ph.D.

Walden University
2019

Abstract

The Effect of Community Evaluators on the Selection of Entry-Level Police Officers

by

Eric Sean Hutchison

MS, Walden University, 2017

MBA, Otterbein University, 2004

BBA, Kent State University, 1998

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Psychology

Walden University

May 2019

Abstract

As a method of building relationships with the public, some police forces have integrated community members into the candidate assessment and selection process. The purpose of this quantitative correlational study was to determine the effect of integrating community evaluators as a new method in the assessment and selection process for police officers in a city police force. Media richness theory and general mental ability were used as a framework, and archival data from a large Midwest department of public safety were collected by filing two public records requests. Data from 2,510 police candidates were included. Quantitative data analysis was conducted using correlational and regression tests to examine rater agreement, subgroup differences (gender or race/ethnicity) in selection outcomes, and the predictive validity of a testing method as measured by academy performance with and without the integration of community evaluators. There was no evidence to suggest that integrating community evaluators into the assessment and selection process for entry-level police officers affected rater agreement or subgroup differences in selection outcomes. The findings from this study support positive social change by indicating that integrating the community into a structured assessment process did not impact selection outcomes as measured by gender, race/ethnicity, or academy performance, which may encourage public safety departments to build community relationships by inviting local residents to participate in the assessment and selection process for police officers. Other social change may include the effect that the integration of community members could have on applicant and community perceptions of the assessment and selection process for police officers.

The Effect of Community Evaluators on the Selection of Entry-Level Police Officers

by

Eric Sean Hutchison

MS, Walden University, 2017

MBA, Otterbein University, 2004

BBA, Kent State University, 1998

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Psychology

Walden University

May 2019

Dedication

This dissertation is dedicated to the many people who work to make this world a safer place.

Acknowledgments

I am humbled to acknowledge the many people who have inspired and supported me throughout this journey.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to the Study

Introduction

The assessment and selection of police officers has been a topic of industrial and organizational psychology literature for more than 100 years (Ployhart, Schmitt, & Tippins, 2017). The use of intelligence, pedagogic, and physical testing were first used as the only methods of assessment for the selection of police and fire candidates as noted in the first issue of the *Journal of Applied Psychology* (Terman et al., 1917). Then civil service agencies were considered in the selection process as well as psychological tests of intelligence (Gosnell, 1923). Today, the selection of police officers is still a relevant topic that has been addressed by many articles with suggestions for industrial and organizational psychologists and law enforcement agencies (Bergman, 2016; Chatterjee, 2016; Farley & Thompson, 2016; Herndon, 2016; Jacobs, Phillips, & Gully, 2016; Ruggs et al., 2016; Zabel, Zabel, Olson, & Carlson, 2016).

Many city administrators and police forces are looking for ways to build engagement and relationships between the community and the police force (Gould, 2017). A new method in the assessment and selection of entry-level police officer applicants is the integration of community members as raters (community evaluators) in the assessment process (Ferrell, 2017; Gould, 2017; Rouan, 2017; Simmons, 2012). This chapter includes an overview of the background of police officer selection, the introduction of the community evaluator into the selection process, and the framework, assumptions, and limitations for this study.

Background

For more than 100 years, researchers have been studying the recruiting, assessment, and selection of police officers and firefighters (Ployhart et al., 2017). Most large cities use a noncompensatory multiple hurdle selection process for entry-level police and firefighter positions, where applicants must meet minimum qualifications, compete in a series of tests, and undergo several evaluations (DeCicco, 2000; Potter, 2013). In a noncompensatory multiple hurdle selection process that consists of four hurdles, applicants must consecutively pass hurdles one, two, and three before attempting the fourth hurdle. Failure to pass any of the hurdles results in disqualification from the multiple hurdle selection process. Using the example of a multiple hurdle selection process that consists of four hurdles for entry-level police officers, only candidates who pass all four hurdles are eligible to become recruits in a police academy (DeCicco, 2000; Potter, 2013). However, researchers are still seeking guidance (Annell, Lindfors, & Sverke, 2015), or offering advice (Albrecht, 2017) on the most effective methods of selection for entry-level police officers.

One of the contributing factors for this ongoing discussion about police officer selection methods is the environmental climate of American policing (Bergman, 2018; Chatterjee, 2016; Gould, 2017; Herndon, 2016; Ruggs et al., 2016; Todak, 2017). The U.S. government has sponsored studies to explore methods of improving the relationship between police agencies and communities (Omnibus Crime Control and Safe Streets Act, 1968; President’s Task Force on 21st Century Policing, 2015; Simmons, 2010). The published research, opinions, and funded studies have proposed several tactics for community engagement, better recruitment and selection methods, and transparency within the law enforcement system. One of the recommended tactics is engaging community members in the assessment and selection process of entry-level police officers (Gould, 2017; Simmons, 2010).

Though there is considerable research on the effectiveness and fairness of measures like the constructed response multimedia test to measure problem-solving and interpersonal skills in selecting entry-level police officers (Arthur & Villado, 2008), there is a lack of research on the effectiveness of community members as a rating method in the entry-level police officer selection process (Simmons, 2012). Research has included measurements in the selection process that includes the small differences between ethnic subgroups for the constructed response multimedia test when compared to the cognitive ability test, language proficiency test, personality inventory, structured interview, and role play (De Soete, Lievens, Oostrom, & Westerveld, 2013). Additionally, research has suggested that the verbal response mode outperformed the written response mode regarding verbal and written responses for police officer academy cadets using a constructed response multimedia test (Lievens, De Corte, & Westerveld, 2015). Further research has indicated that rater and ratee characteristics, as defined by race and sex, did not have a statistically significant effect on applicant scores for a behavioral-personnel assessment device (B-PAD; Doerner & Nowell, 1999). Because there is little research on community participation in the entry-level police officer selection process, this study was necessary to address a gap in the literature. This topic is explored further in Chapter 2.

Problem Statement

Based on U.S. government guidelines, court cases, and professional standards, the evaluation of personnel assessment and selection methods uses a test of adverse impact, psychometric adequacy, and use of alternative devices (De Soete et al., 2013; Highhouse, Doverspike, & Guion, 2016; Wolgast, Backstrom, & Bjorklund, 2017). Alternative devices are often a replacement for, or complement to, multiple-choice job knowledge testing and can involve the use of work samples, situational judgment tests, oral boards, and constructed response multimedia tests (Cucina, Su, Busciglio, Thomas, & Peyton, 2015; De Soete et al., 2013; Lievens et al., 2015; Riccucci & Riccardelli, 2015). When dealing with high stakes, public sector testing, such as police and fire personnel, the procedures and alternative devices come under scrutiny (De Soete et al., 2017; Guajardo, 2014; Gustafson, 2013; Hoffman, 2018; Kringen, 2016; Riccucci & Riccardelli, 2015; Riccucci & Saldivar, 2014). The scrutiny of selection procedures and alternative devices is one of the reasons for calls to include diverse members of the community in the assessment and selection process of police officers (Gould, 2017; Simmons, 2010). Although community members have participated as evaluators in the police officer assessment and selection processes in the past (Ferrell, 2017; Rouan, 2017; Simmons, 2012), there is a lack of research on the effectiveness of this approach.

Purpose of the Study

The purpose of this quantitative study was to determine the effect of integrating community evaluators as an adjunct to the assessment and selection process for entry-level police officers in Columbus, Ohio (see Appendix A). The study included an exploration of whether rater agreement and selection outcomes were influenced by the introduction of community evaluators into one phase of an entry-level police officer assessment and selection process. The second hurdle of the 10-hurdle selection process, applicant testing, consists of four phases. The community evaluator was integrated into Phase 3 of this hurdle. In Phase 3, the constructed response multimedia test was designed to measure the problem-solving and interpersonal skills of the candidate as a predictor of performance in the Columbus Police Academy. The goal of this study was to determine whether selection outcomes in Phase 3 were influenced by the introduction of community evaluators into the assessment and selection method based on measurements of adverse impact indicators and psychometric adequacy.

Research Questions and Hypotheses

Quantitative methods were used to answer the following research questions to determine the effect of the community evaluator on the assessment and selection of entry-level police officers in Columbus. The three research questions were intended to measure predictors of candidate performance on the Columbus Civil Service Commission (CSC) assessment, subgroup differences (gender and race/ethnicity) in assessment and selection outcomes, and predictors of performance in the Columbus Police Academy.

Research Question 1: Does evaluation method type and/or candidate demographic characteristics predict the score on the Columbus Civil Service Commission constructed response multimedia test for candidates between 2015–2017?

$H_0$1: Evaluation method type (community evaluator presence or absence) and/or candidate demographic characteristics (gender and race/ethnicity) do not significantly predict the score on the Columbus Civil Service Commission constructed response multimedia test for candidates between 2015–2017.

H_a1: Evaluation method type (community evaluator presence or absence) and/or candidate demographic characteristics (gender and race/ethnicity) significantly predict the score on the Columbus Civil Service Commission constructed response multimedia test for candidates between 2015–2017.

Research Question 2: Does evaluation method type, candidate demographic characteristics, and/or score on the Columbus Civil Service Commission constructed response multimedia test predict Academy performance for recruits who were candidates between 2015–2017?

H_02: Evaluation method type (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test do not significantly predict Academy performance for recruits who were candidates between 2015–2017.

H_a2: Evaluation method type (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test significantly predict Academy performance for recruits who were candidates between 2015–2017.

Research Question 3: Does evaluation method type (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test predict Academy graduation for recruits who were candidates between 2015–2017?

H$_0$3: Evaluation method (community evaluator presence or absence of), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test do not significantly predict Academy graduation for recruits who were candidates between 2015–2017.

H$_a$3: Evaluation method (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test significantly predict Academy graduation for recruits who were candidates between 2015–2017.

Theoretical Framework

A theoretical framework of media richness theory (MRT; Daft & Lengel, 1986) and general mental ability (GMA; Schmidt & Hunter, 1998, 2004) was used in this study to interpret the findings of the community evaluator on the assessment and selection of entry-level police officers. This framework aligns the consistent process associated with administering a media-rich assessment (Cucina et al., 2015; De Soete et al., 2013; Lievens et al., 2015), structured method of rating (see Wolgast et al., 2017), and the predictive validity of similar assessments (see Corey, MacAlpine, Rand, Rand, & Wolf, 1996; Doerner & Noell, 1999). Both MRT and GMA are present in the current research on entry-level police officer selection; however, the combination of these two theories as a framework was not found when conducting an extensive literature review. Chapter 2 of this dissertation includes an analysis of MRT and GMA to demonstrate the relevance of these two theories to the police officer selection process and why using this framework is significant to the current study.

Nature of the Study

I used a quantitative research design in this nonexperimental study. Quantitative methods enable measurement of the effect of a rater on selection outcomes and validity by using the demographic characteristics, assessment scores, and performance in a police academy (Corey et al., 1996; Cucina et al., 2015; Doerner & Nowell, 1999; Lievens, 2015; Park, 2013). In this study, an applicant becomes a candidate once they have passed the first hurdle in a 10-hurdle selection process (see Appendix A; Columbus, 2019d). The candidate becomes a recruit once they have passed all 10 hurdles and are hired to participate in the Columbus Police Academy. In the second hurdle, the Columbus CSC uses a noncompensatory examination process consisting of four exams:

1. a multiple-choice test;
2. a written work sample;
3. a constructed response multimedia test;
4. a physical fitness test.

The results of the second hurdle are used to determine which candidates are eligible to participate in the remaining eight hurdles that precede a notification of appointment for the Columbus Police Academy (Columbus, 2019c).

An evaluation of the results from Phase 3, the constructed response multimedia test, was conducted from 2015–2017 to examine rater reliability, indicators of adverse impact, and the predictive validity of the assessment as measured by performance in and graduation from the police academy (Field, 2013; Frankfort-Nachmias & Leon-Guerrero, 2015; Warner, 2013). Data were collected by submitting a public records request (Columbus, 2019e) to the CSC Public Safety Divsion of Columbus, Ohio and the Columbus Police Academy. Conducting this analysis enabled me to determine whether the introduction of the community evaluator into the testing process in 2017 made a statistically significant difference in rater reliability, selection outcomes, and on the validity of the assessment when compared to 2015 and 2016.

The testing method that is the focus of this study was administered by a CSC public safety team that is responsible for creating, implementing, administering, and scoring several steps of a multiple hurdle selection process when screening police officer applicants to determine who will move forward to the academy. This noncompensatory multiple hurdle process is a common theme in the literature on the assessment and selection of police officers (Annell et al., 2015; Columbus, 2019d; Cucina et al., 2015; DeCicco, 2000; Hoffman, 2018; Kringen, 2016; Potter, 2013; Riccucci & Riccardelli, 2015; Ryan, Sacco, McFarland, & Kriska, 2000). The phase evaluated in this study is a constructed response multimedia test, designed after the principles of a B-PAD, which is also a common method of testing for police officer applicants (Corey et al., 1996; Cucina et al., 2015; De Soete et al., 2017; Doerner & Nowell, 1999; Lievens et al., 2015).

For a constructed response multimedia test, applicants are presented one of three versions of eight prerecorded scenarios (City of Columbus CSC, 2012). Applicant responses to each scenario are videotaped and evaluated by raters using behaviorally anchored rating scales (BARS; Pulakos, 2007). Columbus CSC employees and

Columbus police officers worked together on three-person panels to assess applicants in 2015 and 2016 (Ferrell, 2017; Rouan, 2017). In 2017, the structure of the panel was changed to include two Columbus police officers (uniform evaluators) and one community evaluator, with a Columbus CSC employee serving as a moderator for the three-person panel (Columbus, 2019d; Ferrell, 2017; Rouan, 2017). Adverse impact, reliability, and validity were examined using data from the preemployment process from 2015–2017 and the Academy from 2015–2018.

Three groups were examined in this study. The first group (candidates) were participants in the entry-level police officer multiple hurdle testing process who have met the minimum requirements (see Appendix B) and participated in the Columbus Oral Police Exam (COPE), which is the third phase of the testing hurdle, from 2015–2017. The second group (recruits) were a subset of candidates who passed the fourth test and subsequent hurdles of the selection process (Columbus, 2019c) and were recruits who participated in, or graduated from, the Columbus Police Academy from 2016–2018. The third group (evaluators) rated candidate responses to COPE (Columbus, 2012).

Definition of Terms

Community evaluator: Based on information from the public safety test team manager, a community evaluator is a citizen of the local community who passed an interview and background screening before being selected, trained, and engaged as a rater for the constructed response multimedia test in the Columbus, Ohio entry-level police officer assessment and selection process.

Columbus Oral Police Exam (COPE): The COPE is a constructed response multimedia test designed to evaluate candidate problem-solving and interpersonal skills (Columbus, 2019d). COPE is administered in Phase 3 of the second multiple hurdle selection step that precedes the remaining eight steps of the entry-level police officer selection process in Columbus, Ohio (see Appendix A; Columbus, 2019c).

Critical incident: A critical incident is a scenario where the behaviors and interpersonal skills of the employee can influence the effectiveness of the outcome (Harvey, Anderson, Baranowski, & Morath, 2007).

Moderator: Based on information from the public safety test team manager, the term *moderator* refers to the position of a CSC employee during Phase 3 of the testing process in 2017. A moderator's responsibilities included playing applicant video responses, ensuring rating forms were thoroughly completed by all three raters, and reassigning applicants to other panels if a rater disclosed a conflict of interest.

Realistic job preview: A realistic job preview is when applicants are given an opportunity to learn specific details about the environment, procedures, policies, and traits for a job (Breaugh & Billings, 1988).

Restriction of range: The term to explain a scenario where only a specific selection of the data for the entire assessment and selection process is under evaluation (Markus & Lin, 2010).

Situational judgment test: A method of evaluating an applicant's problem-solving techniques or responses to one or more critical incidents (Anastasi & Urbina, 1997; Christian, Edwards, & Bradley, 2010; Ployhart & MacKenzie, 2011; Tuzinski, 2013; U.S.

Office of Personnel Assessment, 2007). Situational judgment tests can be a multidimensional method of evaluating a candidate's problem-solving and interpersonal competencies (Ployhart & MacKenzie, 2011).

Structured interview: A selection method where each applicant receives a similar set of questions or scenarios in the assessment process (Huffcutt & Youngcourt, 2007).

Uniform evaluator: Based on information from the public safety test team manager, a uniform evaluator is a sworn police officer or sergeant with the Columbus Division of Police who is selected, trained, and engaged as a rater for COPE.

Assumptions

Archival data were used for the statistical analysis in this study. Therefore, several assumptions about these data were made and relied upon throughout this study. First, applicants completed a preemployment questionnaire that included their demographic information, which was assumed to be correct because these data are vital to measuring subgroup differences. Second, the assumption was made that the development of the constructed response multimedia test and BARS complied with the *Principles for the Validation and Use of Personnel Selection Procedures 5th Edition* (Society of Industrial and Organizational Psychology, 2018). Third, it was assumed that adequate methods of rater training were conducted to ensure comprehension of the assessment process and use of BARS to mitigate rater bias and misinterpretation of the scales (Dessler, 2011; Pulakos, 2007). Finally, the integrity and accuracy of the data were also assumptions because the CSC and Academy are credible agencies that have demonstrated consistency and fairness in the assessment, selection, and development of police officers based upon accreditation by the Commission on Accreditation for Law Enforcement Agencies (2010).

Scope and Delimitations

The specific aspect of the research problem addressed in this study is the use of a community evaluator as an adjunct to an existing method in the assessment and selection of entry-level police officers. This focus was selected because the use of community evaluators as stakeholders in the selection process has occurred in other cities (Simmons, 2012) before Columbus, Ohio (see Ferrell, 2017; Rouan, 2017) and the effect of this method is unknown. Therefore, research on the change to this testing method is necessary to determine whether integrating community members results in a change to selection outcomes based on gender, race/ethnicity, and performance in a police academy.

The samples included in this study were limited to the raters who participated in one phase of the assessment and selection process, entry-level police officer candidates in Columbus, Ohio from 2015–2017, and Academy recruits from 2015–2018. This study did not include a measurement of candidate or rater perceptions. This study did not include assumptions about subgroup differences relating to performance on the assessment, or in the Academy, as part of a determination of adverse or disparate impact. Instead, if indicators of adverse impact were identified in the calculations of the study, the results would have been reported. However, an adverse impact determination would have required further investigation beyond the scope of this dissertation.

Because the B-PAD and other constructed response multimedia tests are standard in entry-level police officer testing (Corey et al., 1996; Cucina et al., 2015; De Soete et al., 2013; Lievens et al., 2015; Riccucci & Riccardelli, 2015), the findings of this study can be used when evaluating the effect of the rater on the assessment and selection outcomes. This research is not intended to be generalizable to the population of entry-level police officer applicants, CSC public safety testing divisions, or police academy participants.

Limitations

There were several limitations associated with this study. First, it was unknown whether the predictive validity of the constructed response multimedia test used in this study has been demonstrated to be a statistically significant predictor of performance in the Academy. Second, there was limited research on the reliability and agreement of three or more raters using BARS to assess entry-level police officer candidates. Third, restriction of range limits the sample of data available for analysis because only the candidates who passed the first two phases of the second hurdle (see Appendix A) were scored on COPE. Therefore, restriction of range was considered a weakness because it was unknown how well the applicants who did not pass the first two phases would have performed on COPE, which could influence the predictive validity component of this study.

Another limitation is that there could be confounding variables that influence attrition throughout the multiple hurdle selection process that were not evaluated in this study (Ryan et al., 2000). The four phases of testing determine who is eligible to participate in the subsequent eight steps of the multiple hurdle selection process that precedes the Academy. The raw scores are adjusted to *z* scores for the purpose of banding candidate scores into three categories, and candidates invited to move onto the third hurdle are selected from the highest band first (see Appendix A).

The sample size of the study was another limitation. The population of test applicants, candidates, and Academy recruits was a fixed size and recruiting additional participants was not an option for this study. Differential validity and differential prediction analysis studies often face challenges relating to statistical power because of the difficulty associated with recruiting and collecting a large, diverse sample of participants (Berry, Sackett, & Sund, 2013).

In addition to the limitations, a potential for researcher bias is also necessary to disclose. I work as a personnel analyst at the CSC that is the focus of this study. However, I was not involved in the development of the assessment or the selection of raters. I worked as a panel moderator for one out of eight rating panels on two out of the five evaluation days in 2017. I have not received, nor intend to receive, any compensation or guarantee of employment from the City of Columbus based upon the work, or results, related to this dissertation.

Significance

This study addresses a gap in the literature through evaluation of two different rating methods used by the Columbus CSC for the selection of entry-level police officer candidates. This study is unique for several reasons. First, several researchers have identified the need for an analysis of CSC practices (Guajardo, 2014; Gustafson, 2013;

Hoffman, 2018; Kringen, 2016; Riccucci & Riccardelli, 2015). Second, there is a lack of scientific evidence on the reliability of using community evaluators as stakeholders in the entry-level police officer assessment and selection process (Simmons, 2012). Third, this study builds on a need for the evaluation of a constructed response multimedia test that includes an examination of diversity and validity (Cucina et al., 2015; De Soete et al., 2013). In addition, research on predictor variables in law enforcement selection has declined since Aamodt's (2004) meta-analysis (Bullock, Latham, & Aamodt, 2018).

The results of this study can provide insights into the effect of evaluation methods on selection outcomes and effectiveness of an entry-level police officer assessment. Insights from this study could aid Columbus CSCs and other entry-level police officer selection committees when identifying the best assessment and panel structure for mitigating the risk of adverse impact while predicting performance in a police academy. Implications for positive social change include selecting the most qualified recruits who will attend, demonstrate high levels of performance in, and graduate from a police academy. Selecting the most qualified recruits, while mitigating the risk of adverse or disparate impact, provides equal access to all applicants in the selection process (Columbus, 2019b) and can reduce the costs associated with poor performance, or attrition, in a police academy.

Summary and Transition

This chapter has introduced the study. A brief background on police officer selection was provided as an overview to present the problem and purpose of the study and are more fully explored in Chapter 2. The research questions demonstrate how the variables are measured and align with the framework and nature of the study. The definitions are limited to terms that are referenced multiple times and have more than one meaning outside of this study. The assumptions, scope and delimitations, and limitations of this study are necessary for the transparency of the research. The significance of this study emphasizes the importance of this work as a contribution to the body of knowledge on entry-level police officer selection and positive social change.

Chapter 2 includes an exploration of the problem and purpose of this study in relation to the existing body of knowledge on this topic. The literature review includes the synthesis and analysis of peer-reviewed work, dissertations, trade journals, government studies, and newspaper publications. Explanations of themes, gaps, and discrepancies in the literature are also provided in Chapter 2. The research design for this dissertation is addressed in Chapter 3 and includes a description of the sample and statistical methodology that was used to analyze indicators of adverse impact in the selection process, rater agreement, and predictors of performance on a constructed response multimedia test and in a police academy. Chapter 4 includes an analysis of the data and results of the statistical methodology. Chapter 5 consists of an interpretation of the results, limitations, recommendations for future research, social change implications of this study, and conclusions.

Chapter 2: Literature Review

Introduction

The purpose of this quantitative study was to determine the effect of integrating community evaluators into a selection device administered by the Columbus CSC as part of the assessment and selection process for police officers. Although evidence exists that community members have been engaged as evaluators in the police officer assessment and selection process (Ferrell, 2017; Rouan, 2017; Simmons, 2012), there is a lack of evaluations on the effectiveness of this approach. In this chapter, I provide a review of the literature that includes examination of (a) the theoretical framework for this study; (b) evaluation of personnel assessment and selection methods with an emphasis on police officers; (c) the video-based constructed response multimedia test; (d) community involvement in the selection of police personnel; and (e) the importance of understanding the effect of the rater on selection outcomes. This literature review elaborates on the research problem and includes an analysis of studies on police officer selection methods while identifying gaps and discrepancies in the current research on this topic.

Literature Search Strategy

The first search for literature was through the Walden Library using EBSCOhost Thoreau Multi-Database Search (Thoreau) with the following Boolean search: *pre-employment screening* OR *hiring* AND *police** OR *law enforcement*. The search returned more than 22,000 peer-reviewed articles published between 2014–2018. However, refining the search using *police officer* AND *selection* AND *validity* returned 81 peer-reviewed articles published within the past 5 years. Additional databases accessed through Thoreau included ProQuest Dissertations & Theses Global, which provides access to more than 4 million documents. Two additional databases within Thoreau also included PsycINFO and PsycARTICLES, which align with the American Psychological Associations' publications. Other sorting options within Thoreau were *methods* and *instruments* that were used to evaluate approaches to data analysis and "PlumX metrics" and "related information" available through the EBSCOhost search engine.

The Criminal Justice Database is not accessible through Thoreau, so the keyword searches were also repeated for publications specific to the field of criminal justice. The Criminal Justice Database provides access to multiple sources including trade journals, conference papers and proceedings, dissertations and theses, and scholarly journals. Although some of the articles in trade journals are not peer reviewed, they are still useful for understanding the current perceptions and climate in the field of law enforcement. The Encyclopedia of Industrial and Organizational Psychology was also used to research theories, themes, terms, and strategies for the assessment, selection, and validation of methods. Textbooks with the topics of applied psychology, personnel selection and assessment, applied measurement, forensic psychology, and research design and methods were also reviewed as part of the literature search. Boolean and related article searches were conducted within Google Scholar. Google Scholar provides tools to review article citations, number of times an article has been cited, and the ability to review the stream of literature and previous works by researchers.

Theoretical Framework

Media Richness Theory

Researchers use MRT to explain how different types of organizational communications can influence levels of uncertainty and equivocality (Daft & Lengel, 1986). A communication continuum is used to provide examples of media richness. Indirect methods that include preprinted materials and e-mail are considered low in media richness, whereas direct contact methods like video conferencing and in-person meetings would be high in media richness. Communication methods high in media richness can reduce uncertainty and equivocality by providing clarity without the need for additional data (Daft & Lengel, 1986). In personnel selection, pencil and paper tests would be considered low in media richness, whereas structured interviews or video-based methods would be considered high in media richness.

Fidelity is a term often used in the literature to describe media richness and complexity in a video-based assessment, also referred to as a constructed response multimedia test (Cucina et al., 2015; Kroll & Zeigler, 2016; Lievens et al., 2015). A constructed response multimedia test is a method used to present applicants with scenarios that provide opportunities to demonstrate skills in more than one construct. Research has indicated four benefits of using high-fidelity constructed response multimedia tests when compared to low- and moderate-fidelity methods (paper and pencil, verbal, or computer-based tests; Christian et al., 2010). The primary benefit of the high-fidelity method is the ability to portray environmental conditions, visual and verbal clues, and the emotion of a situation to the applicant, which means applicants do not have to read and envision working conditions (Christian et al., 2010; Cucina et al., 2015; Tuzinski, 2013). Work-related scenarios have also been shown to improve the face validity of an assessment and contribute to a realistic job preview (Breaugh & Billings, 1988; Cucina et al., 2015; De Soete et al., 2013; Tuzinkski, 2013). A video-based scenario is also suitable for measuring multiple constructs (Arthur & Villado, 2008; Ployhart & MacKenzie, 2011).

In addition to tests in the assessment and selection process, MRT is also a component in exploring the effectiveness of communication methods within an organization (Dennis & Kinney, 1998). Although MRT is a substitute for providing additional support materials, high-fidelity communications do not equate to better organizational performance. The findings for organizational differences are different from the selection process, where high-fidelity assessments have shown to contribute to smaller subgroup differences and better job performance than low-fidelity methods (Cucina et al., 2015; Kroll & Zeigler, 2016; Lievens et al., 2015).

MRT has been shown to improve the assessment and selection process when applied to the structured simulation of a constructed response multimedia test (Cucina et al., 2015; Lievens et al., 2015). MRT has also been tested to explain the alignment between levels of ambiguity and four distinguishing factors in a selection process. The four factors require the applicants to (a) participate in two-way communication; (b) convey verbal, nonverbal, and paralinguistic cues; (c) demonstrate personal focus; and (d) use their natural language. These four factors are also relevant to assessing social and interpersonal skills (Cucina et al., 2015; Lievens et al., 2015; Tuzinkski, 2013) and are present in the video-based constructed response multimedia test used in Columbus, Ohio (Columbus, 2019d).

In a structured assessment, applicants have similar opportunities to demonstrate their skills to one or more evaluators through direct methods (Tuzinski, 2013). In the selection approach used in Columbus, Ohio, candidates participate in one of three versions of the constructed response multimedia test, each with similar issues and scenarios that relate to the job of a police officer (Columbus, 2019d). Columbus's use of this method in the overall approach to police officer selection is not unusual (see Corey et al., 1996; Cucina et al., 2015; De Soete et al., 2013; Doerner & Noell, 1999; Wolgast et al., 2017). The approach has been shown to be effective because the alignment among scenarios demonstrates a structured approach to situational and behavioral interviewing while providing a realistic job preview (Breaugh & Billings, 1988; Cucina et al., 2015; De Soete et al., 2013; Tuzinkski, 2013). Extensive research demonstrates support for the structured, media-rich approach as a predictor of performance (Corey et al., 1996; Cucina et al., 2015; Doerner & Noell, 1999; Wolgast et al., 2017). For example, Lievens et al. (2015) used MRT to compare the predictive validity of verbal and written responses for police officer academy cadets using a constructed response multimedia test. Though some of their results lacked statistical significance for predictive validity, the study helped identify that high-fidelity test methods were more effective than methods with low-fidelity.

General Mental Ability

In 1904, Spearman introduced the concept of GMA, which is also referred to as intelligence or cognitive ability (Schmidt & Hunter, 1998, 2004). When specific selection measures are combined with a measurement of GMA, the percentage of validity can increase (Schmidt & Hunter, 1998). Additionally, GMA combined with a work-sample, integrity test, or structured interview would yield the highest predictive validity (Schmidt & Hunter, 1998).

Previous research demonstrates that problem-solving serves as a proxy for cognitive ability (Arthur, Doverspike, Barrett, & Miguel, 2013). The attributes of GMA, personality, and experience are also shown to be strong predictors of situational interview performance (Huffcutt, Van Iddekinge, & Roth, 2011). The constructed response multimedia test is expected to demonstrate statistically significant validity as measured by performance in a police academy when the combination of effective problem-solving and interpersonal skills are the constructs being measured (Corey et al., 1996; Cucina et al., 2015; Doerner & Nowell, 1999; Wolgast et al., 2017). COPE was designed to include job-related scenarios for a police officer and the requirement for candidates to demonstrate problem-solving and interpersonal skills (Columbus, 2019d).

Alternative Theories

Researchers have employed other theories as a framework in the evaluation of assessment and selection methods. For example, empowerment theory (Perkins & Zimmerman, 1995), signaling theory (Spence, 1973), and Wherry's theory of rating (Wherry & Bartlett, 1982) were among the many theories reviewed. These theories could apply to other studies on the effect of a community member as a participant in an assessment and selection process. Empowerment theory, a value-based orientation, is appropriate when evaluating the organizational or sociological effect on the community or the perceptions of the evaluators. Signaling theory, how a candidate demonstrates qualifications and the rater's ability to receive these messages, is appropriate for a qualitative study that includes evaluation of rater perceptions (Hilal, Densley, & Jones, 2017). Finally, Wherry's theory of rating suggests that rating is a function of three components: performance of the ratee, observation of performance, and recall of observations by the rater (Wherry & Bartlett, 1982).

Evaluation of Personnel Assessment and Selection Methods

The assessment and selection process for police and firefighter personnel (first responders) has been a subject of personnel psychology research for more than 100 years (Ployhart et al., 2017). Current literature continues to seek guidance (Annell et al., 2015), or offer advice (Albrecht, 2017), on the best methods and constructs of selection for police officers. The climate of American policing contributes to many of the research studies and recommendations for the selection of police officer applicants (Bergman, 2018; Chatterjee, 2016; Ruggs et al., 2016; Todak, 2017). Several research questions about the assessment process range from the validity, reliability, and utility of methods (Lievens et al., 2015; Sackett et al., 2017) to the use of technology (Cucina et al., 2015). However, the most frequently researched topics pertain to whether assessment and selection methods are fair (McLarty & Whitman, 2016) and whether the methods contribute to adverse or disparate impact in the field of law enforcement (De Soete et al., 2013; Guarjado, 2014; Hilal et al., 2017; Kringen, 2016; Riccucci & Riccardelli, 2015; Riccucci & Sadivar, 2018).

Adverse Impact

Adverse, or disparate impact, is the illegal act of discrimination against a group resulting in a disadvantage to their selection for a job or promotion (Civil Rights Act, 1964, 1991). Before the Civil Rights Act of 1964, discrimination in the selection and promotion processes for employees based on race/ethnicity, religion, sex, or national origin was not illegal. The *Uniform Guidelines on Employee Selection Procedures* (1978) state that an indicator of adverse impact is when a low scoring group is less than four-fifths of the higher scoring group. This indicator is often used when making an adverse impact claim or as part of discrimination lawsuits in police and fire departments (Riccucci & Saldivar, 2018).

Another indicator of adverse or disparate impact used in litigation is the identification of subgroup differences (De Soete et al., 2013; Highhouse et al., 2016; Wolgast et al., 2017). Arthur et al. (2013) define subgroup differences as “psychological, scientific phenomena that are represented or conceptualized as standardized mean differences between groups on measures of psychological constructs” (p. 475), whereas adverse impact is the effect of a decision or rule. Subgroup differences are not synonymous with adverse impact (Arthur et al., 2013; Lindsey, King, McCausland, Jones, & Dunleavy, 2013) and can vary by cognitive ability (Wee, Newman, & Joseph, 2014) and situational specificity (McDaniel, Kepes, & Banks, 2011). Factors contributing to subgroup differences in a selection process can include the number of applications, applicant psychological and physical differences, multiple demographics in the pool of applicants, situational variables, and rater performance (Arthur et al., 2013). Two selection strategies available to address the differentiation between subgroup differences and adverse impact are assessment design and scoring (Arthur et al., 2013).

When identifying the knowledge, skills, and abilities necessary for a job, conducting a thorough analysis before designing test instruments is necessary for establishing the construct validity of the assessment (Highhouse et al., 2016; Hoffman, 2018), which is emphasized in the *Uniform Guidelines* (1978). Measures to ensure acceptability of the analysis procedures should include surveying a diverse sample of subject matter experts, ensuring the situations in the selection process resembles the work, and a fair assessment of an individual's competencies (Sinden et al., 2013; Society of Industrial and Organizational Psychology, 2018). Adhering to the analysis and design process can contribute to a legally defensible selection assessment (Highhouse et al., 2016; Riccucci & Riccardelli, 2015; Riccucci & Sadivar, 2018). However, the administration and outcome of a selection process must also demonstrate compliance with the *Uniform Guidelines* (1978) and Civil Rights Act (1964, 1991).

One approach to scoring assessments to mitigate the risk of adverse impact is banding (Murphy & Myors, 1995; Schmidt & Hunter, 1995). Banding is an approach to determine the statistical significance between the highest score and lower scores, thus treating all scores in a range the same (Murphy & Myors, 1995; Schmidt & Hunter, 1995). One criticism of banding is a flaw in the process because bands could potentially overlap, resulting in inconsistency (Schmidt & Hunter, 1995). Benefits of banding include considering lower scores that may have otherwise resulted in rejecting a qualified candidate (Murphy & Myors, 1995) and reducing disparate impact when compared to other selection approaches (Sacket & Roth, 1991).

Adverse or disparate impact in entry-level police officer testing has been the focus of several recent studies (De Soete et al., 2017; Guajardo, 2014; Highhouse et al., 2016; Hilal, Densley, & Jones, 2017; Kringen, 2016; Riccucci & Riccardelli, 2015; Riccucci & Sadivar, 2018). An evaluation of multiple assessment methods for entry-level police officers has resulted in small differences between ethnic subgroups for the constructed response multimedia test when compared to the cognitive ability test, language proficiency test, personality inventory, structured interview, and role play (De Soete et al., 2013). For example, the ethnic differences studied by De Soete et al. (2013) were for Dutch applicants, resulting in a recommendation from the researchers to replicate the methods of their study in a more diverse population. This recommendation by De Soete et al. is essential to this dissertation because the subgroup differences for a large and diverse group of applicants were evaluated based on their performance on a constructed response multimedia test and the alternative approach of community evaluator presence or absence.

Designing and evaluating assessment and selection methods with a focus on validity is essential to selecting the most qualified applicants and mitigating the risk of adverse impact litigation (Arthur et al., 2013; De Corte et al., 2007). An analysis of multiple assessment and selection system strategies address the trade-off between adverse impact and predicting performance (Cucina et al., 2015; De Soete et al., 2013; Finch, Edwards, & Wallace, 2009). Quota hiring, which involves selecting individuals based on their race, ethnicity, or gender to meet requirements set forth in the *Uniform Guidelines* (1978) and the Civil Rights Act (1964, 1991) is not considered an approach that aligns with predictive validity (De Corte et al., 2007; Pynes, 2001). The validity of assessment and selection methods, which includes psychometric adequacy and the use of alternative devices, is the best defense against claims of adverse or disparate impact (Arthur, Edwards, & Barrett, 2002; De Corte et al., 2007).

An analysis of the application of the *Uniform Guidelines* to entry-level police officer selection identified controversies relating to the appropriate statistical methods for scoring applicants, measuring adverse impact, requirements to reduce or eliminate adverse impact, and the importance of moving beyond basic intelligence tests (Pynes, 1991). Since then, multiple studies support the use of entry-level police officer assessments that measure the desired problem-solving and interpersonal skills required for the profession (Aamodt, 2004; Corey et al., 1996; Cucina et al., 2015; Doerner & Nowell, 1999). Studies on the use of a constructed response multimedia test with these measurements are also shown to mitigate the risk of adverse impact (see De Soete et al., 2013) while predicting performance in a police academy (Cucina et al., 2015; Corey et al., 1996).

Alternative Devices and Methods

The use of alternative devices, both methods and constructs, should be evaluated and considered in personnel assessment and selection practices (Arthur & Villado, 2008; Arthur & Woehr, 2013). In police officer testing, the devices and methods are usually an alternative to the paper-and-pencil multiple choice test (Arthur et al., 2002; De Soete et al., 2013). The *Uniform Guidelines* (1978) includes a directive that states:

> Where two or more selection procedures are available which serve the user's legitimate interest in efficient and trustworthy workmanship, and which are substantially equally valid for a given purpose, the user should use the procedure which has been demonstrated to have lesser adverse impact (Section 3B).

The evaluation and consideration of assessment methods should include validity, reliability, and adverse impact (Highhouse et al., 2016; Wolgast et al., 2017), and analyze response modes (Lievens et al., 2015). Although there is a need to evaluate whether adverse impact results from entry-level police officer selection procedures, assessing for ethnic and gender differences alone can have an adverse effect on criterion validity, reliability, utility, and public safety (De Soete et al., 2013). As previously discussed, there are instances where the adverse impact can be explained, such as in a strength assessment. Therefore, researchers and practitioners must not guarantee that adverse impact can be prevented by using alternative devices and methods (Arthur et al., 2013; Arthur & Woehr, 2013; Barrett, Miguel, & Doverspike, 2011).

Methods of Entry-Level Police Officer Testing

Ployhart et al. (2017) explained that the first publication of the *Journal of Applied Psychology* included three articles on personnel selection, one of which focused on psychological assessments of first responder candidates. Ployhart et al. identified personnel selection themes that influenced military operations, business and societal changes, the advancement of technology, diversity and inclusion, and validity.

Researchers and practitioners appear to be more aligned with the theories and practices of selection rather than recruitment (Ployhart et al., 2017). Ployhart et al. also suggested that a challenge with recruitment is that practitioner theories can be forced or are outdated by the time they are published.

Ployhart et al. (2017) identified three recurring questions in the *Journal of Applied Psychology* literature: (a) “How do I determine who has the best knowledge, skills, and abilities to perform a particular job?”; (b) “Where do I find them?”; and (c) “How do I identify people of diverse backgrounds?” (p. 299). Legal, societal, and ethical guidelines include direction for supporting diversity; however, the advancement of global change also requires a commitment to identifying the most qualified applicant regardless of demographic criteria. As a result, researchers and practitioners must be aware of ongoing legal and societal changes that influence the evolution of selection and recruiting practices (Society of Industrial and Organizational Psychology, 2018).

The implications of the review by Ployhart et al. (2017) can be considered relevant for several reasons. The first, identification of appropriate methods of assessment and selection for police officers dates back more than 100 years, which means there is a substantial amount of evidence and recommendations to influence this process. Second, assessing the effect of selection methods on diversity can mitigate risks to adverse impact (Arthur et al., 2002; Arthur & Villado, 2013; Highhouse et al., 2016; Ployhart et al., 2017; Wolgast et al., 2017). Third, the awareness of laws and procedures ensure that legal, ethical, and scientific methods should be incorporated into the evaluation of assessments being used or considered for a selection process (Arthur et al., 2002; Arthur & Villado, 2013; Highhouse et al., 2016; Ployhart et al., 2017; Wolgast et al., 2017). Fourth, identification of knowledge, skills, and abilities from the job design and an application of current research can contribute to evaluations of the validity and reliability of assessment and selection methods (Arthur et al., 2002; Arthur & Villado, 2013; Highhouse et al., 2016; Ployhart et al., 2017; Tuzinski, 2013; Wolgast et al., 2017).

In addition to the historical representation already provided, discussing the extensive development in the standards and processes for police officers can be considered relevant. In the early days of American policing, officers were recruited and funded by political parties (Potter, 2013). The political appointment process for police officers was informal and contributed to inequality (Hilal et al., 2017; Kringen, 2016; Potter, 2013). The inequality contributed to corruption and was an influential factor in the development of the *Omnibus Crime Control and Safe Streets Act* (1968). This act provided federal grant money for the development of plans, programs, and priorities to improve law enforcement.

Science, government regulations, industry guidelines, technology, and changes in police officer responsibilities are instrumental to the standards and methods that are most prevalent in the assessment and selection process today (Potter, 2013). Civilians conduct a component of most of the selection procedures for police officers through personnel departments, CSCs, and as city officials (Kringen, 2016; Potter, 2013). Most police agencies are required to follow city- and state-specific CSC guidelines (DeCicco, 2000). Therefore, it is necessary to provide an overview of the CSC directives.

Civil Service Commission

A CSC employs people who are responsible for establishing, administering, or managing partners associated with the assessment and selection procedures for public safety personnel in a municipality (City of Columbus, 2019a; Hoffman, 2018; Kringen, 2016). In many cities, the CSC is accountable for overseeing the noncompensatory multiple hurdle selection process consisting of a variety of procedures, tests, and interviews that applicants proceed through on a pass/fail basis (Annell et al., 2015; City of Columbus, 2019a; Hoffman, 2018; Kringen, 2016). The goals of a merit process are: (a) protect civil service employees from the political process that contributed to corruption and inequality in the early days of policing, (b) establish rules for hiring, and (c) require that applicants participate in a competitive examination process (Hilal et al., 2017; Kringen, 2016; Potter, 2013).

Researchers calling for an investigation of CSC selection processes and procedures cite multiple court cases on adverse impact as evidence of the need for these studies (Guajardo, 2014; Gustafson, 2013; Hoffman, 2018; Kringen, 2016; Riccucci & Riccardelli, 2015). As of 2015, four states in the United States had eliminated civil service systems at the state level, and four others were in the process of abolishing their systems (Riccucci & Riccardelli, 2015). However, most U.S. states utilize a decentralized civil service approach to assessment and selection for police officers at the city level.

Multiple Hurdle Selection Process

Two types of selection scoring methods are compensatory and noncompensatory (Kehoe, 2007). The noncompensatory method is when there are only two outcomes (pass/fail) for each step of a multiple hurdle selection process, where candidates are screened into the next phase or screened out from the process. The compensatory method is when scores from previous steps are combined and reviewed at each step in the process. For example, a multiple-choice test may be the first assessment, a writing sample in the second phase, and a constructed response multimedia test in the third phase. The compensatory method would be used to select candidates based on a combination of the three scores, whereas the noncompensatory method would be used to select candidates at each phase of testing (Kehoe, 2007).

The goal of a multiple hurdle selection process is to identify the most suitable applicants while screening out those who are unqualified (Annell et al., 2015; Hoffman, 2018; Kehoe, 2007; Kringen, 2016). The multiple hurdle selection process for police officers consists of a variety of assessments and tests that applicants proceed through on a pass/fail or scoring basis. An advantage of the multiple hurdle selection process is the cost savings associated with administering the less-expensive tests at the beginning of the process (Kehoe, 2007). However, the disadvantages to the noncompensatory multiple hurdle selection process can be eliminating candidates too early in the process without evaluating all the eligibility requirements (Kehoe, 2007) and the measurement of the reliability of an individual test (Mendoza, Bard, Mumford, & Ang, 2004).

During the process of screening out candidates, administration of the more expensive tests generally occurs later in the process (Kehoe, 2007). As in the previous example of the three types of consecutive tests, costs to administer and grade multiple-choice tests are lower than reading and evaluating writing samples, and far less expensive than reviewing and scoring the recorded responses to a constructed response multimedia test. Therefore, the multiple hurdle selection method could maximize cost-savings in the selection process.

Multiple Hurdle Selection Process for Entry-Level Police Officers

The noncompensatory multiple hurdle selection process is the most common selection strategy for evaluating entry-level police officer applicants (DeCicco, 2000; Potter, 2013). Although the specific tests and combinations vary by public and private municipalities in the United States, most police forces follow the standards established by their state's civil service agency. The most common combination for entry-level police officers includes tests to determine a candidate's eligibility to meet the minimum requirements, physical and mental fitness, moral standards, and communication skills (Annell et al., 2015; DeCicco, 2000; Hoffman, 2018; Kringen, 2016; Potter, 2013; Riccucci & Riccardelli, 2015; Ryan et al., 2000). This combination of tests could assist hiring departments when identifying which candidates are most likely to be successful in the police academy and as police officers.

Minimum requirements. The most common minimum requirements for an entry-level police officer are citizenship, education, age, and a driver license (Potter, 2013). Requiring an applicant to be a citizen in the United States is dependent on the local or state agency (Go Law Enforcement, 2019). The minimum education requirements are usually a high school diploma or equivalent certification (Potter, 2013). However, some cities expect applicants to have a degree or certification in law enforcement or criminal justice (Hilal et al., 2017; Park, 2013).

Physical fitness. Some of the physical abilities listed in the job summary of a police officer include running, jumping, explosive strength, extent flexibility, and dexterity (National Center for O*NET Development, 2018). Physical fitness tests are a subject of several legal cases that resulted in a court decision of disparate impact because the job analysis did not demonstrate the need for physical abilities, there was adverse impact in the assessment outcomes, and the standards present in the assessment were not enforced for existing police officers (Arthur et al., 2013; Barrett et al., 2011; DeCicco, 2000; Highhouse et al., 2016; Potter, 2013; Riccucci & Saldivar, 2014). In response to litigation, many police agencies have established different guidelines for males and females, tests that align with the job requirements, and methods of reinforcing standards with sworn officers (Potter, 2013).

Mental fitness. The primary purpose of conducting the mental fitness assessment is to obtain the candidate's "clinical symptoms, personality characteristics, behavioral tendencies, interpersonal functioning, and interests" (Ben-Porath & Tellegen, 2011). A study conducted by the Bureau of Justice showed that nearly 100% of departments that serve 25,000 or more citizens utilize psychological evaluation as a standard protocol in the assessment and selection process for entry-level police officers (Roberts, Tarescavage, Ben-Porath, & Roberts, 2018). The most common psychological test is the

Minnesota Multiphasic Personality Inventory–2, and/or the Minnesota Multiphasic Personality Inventory–2–Restructured Form. The Inwald Personality Inventory is also very common in police selection, as is the 16PF, the PAI, and the CPI (Weiss & Inwald, 2018).

Moral standards. A three-prong approach is a common method of evaluating a candidate's moral standards through a background investigation, drug testing, and a polygraph examination (Potter, 2013). The purpose of the background check is to validate the information provided by the applicant during the application process, check their references, and explore the candidate's legal, financial, employment, and public record history. Drug testing can be used to detect the use of illegal and controlled substances. A lie-detector (polygraph) examination is also administered to deter a candidate from falsifying information when replying to structured interview questions that relate to the background check, psychological testing, and information disclosed during the screening process (DeCicco, 2000). Although the polygraph has not been shown to be a predictor of performance in a police academy, the test was shown to be a statistically significant predictor of academy completion (Park, 2013).

Communication skills. Strong communication skills are essential to the job of a police officer. The National Center for O*NET Development (2018) include active listening, speaking, negotiation, persuasion, knowledge of the English language, and oral expression in the job summary report for a police officer. A candidate's communication skills can be assessed through a written test and verbal responses in structured interviewing (DeCicco, 2000; Potter, 2013). The use of written tests is the subject of controversy in the literature because this method is only required by eight U.S. states (Riccucci & Riccardelli, 2015). However, evidence provided by Riccucci and Riccardelli shows the method of written tests is utilized by almost all CSCs in large U.S. cities.

City of Columbus Entry-Level Police Officer Testing

The Uniform Testing Unit of the Columbus CSC utilizes a noncompensatory multiple hurdle selection process for the selection of their police recruits (Columbus, 2019d). Once an applicant provides evidence to meet the minimum requirements and standards of an abbreviated background questionnaire, they move onto the initial testing process (see Appendix A; Columbus, 2019d). The testing process in Columbus occurs at the second selection hurdle and includes four examinations:

1. a multiple-choice test;
2. a written work sample;
3. COPE;
4. a physical fitness test.

The results of the second hurdle are used to determine which candidates are eligible to participate in the remaining eight hurdles that precede a notification of appointment for the Columbus Police Academy (Columbus, 2019c). All of the examinations are pass/fail except for COPE, which is scored using BARS (Pulakos, 2007). If a candidate passes all four of the examinations, then they are placed into one of three bands based upon their COPE score and credit for military service. These types of tests are consistent with industry practices (Annell et al., 2015; DeCicco, 2000; Hoffman, 2018; Kringen, 2016; Potter, 2013; Riccucci & Riccardelli, 2015; Ryan et al., 2000) and precede the remaining eight-phases of the multiple hurdle selection process that occurs before a candidate becomes eligible to be a recruit in the Academy (Columbus, 2019d).

The remaining eight steps of the multiple hurdle selection process that determines admission into the Academy and offer of employment begins with candidates who were placed in the highest band based upon their performance in Phase 3 of the testing process plus an eligibility-based military credit of 10-points (Columbus, 2019c; Columbus, 2019d). The noncompensatory selection process for the Academy is consistent with the common assessment selection and scoring strategy for evaluating entry-level police officer applicants (DeCicco, 2000; Potter, 2013). The steps following the second hurdle are:

1. self-reported background information;
2. a polygraph examination;
3. a review of background information and results of the polygraph exam;
4. a background investigation that includes employment history, criminal record, and references;
5. a panel interview;
6. conditional appointment as determined by the City of Columbus Public Safety Director;
7. a medical examination that includes vision, physical, and psychological components and the potential for a second polygraph;
8. acceptance into the academy and offer of employment.

As discussed in the limitations section, there are multiple variables that could influence attrition in this process that are unrelated to successful performance in each step. Reasons for attrition could include time, communication of progress between the city and the applicant, and applicant perspectives (McCarthy et al., 2017; Ryan et al., 2000).

Constructed Response Multimedia Test

The constructed response multimedia test generally consists of video-based, job-related scenarios that are presented to applicants who respond to a camera that records their response (see Corey et al., 1996; Cucina et al., 2015; De Soete et al., 2013; Kroll & Ziegler, 2016; Norton, McCloskey, & Hudson, 2011). The job-related scenarios are designed to replicate situations the applicant should expect to experience on the job. The job-related scenarios contribute to the face validity of the assessment while also providing a realistic job preview (Breaugh & Billings, 1988; Cucina et al., 2015; De Soete et al., 2013; Tuzinkski, 2013). The applicant receives instructions to reply to the screen/video image as though they are responding to a real-life situation. The applicant's responses are then reviewed by a panel of raters who utilize BARS to score the applicant on one or more criterion.

Behaviorally Anchored Rating Scales

BARS is one of the multiple tools that exist for assessing specific performance. The ratings incorporated into BARS should be defined by subject matter experts (Pulakos, 2007). When developing BARS, industry standards for job analysis techniques should be used to identify scenarios, often referred to as critical incidents (Harvey et al., 2007; Society of Industrial and Organizational Psychology, 2018). Subject matter experts contribute to defining the rating criteria for BARS because of their work experience or familiarity with the job requirements (Pulakos, 2007). When contributing to BARS development, subject matter experts could be people who have experience working in or supervising the job.

Ratings for BARS usually range from 1 to 5, or from 1 to 7, where the higher number correlates with highly effective performance (Pulakos, 2007). Two benefits of using BARS are the quantitative nature of the ratings and the consistency of the method (Dessler, 2011). Based on this approach, BARS could be used to score multiple scenarios that contribute to an average score for the person undergoing evaluation (Dessler, 2011; Pulakos, 2007). The quantitative score could then be used to compare individual performance and assign competency levels for multiple people working the same job in a department or organization.

Another feature of BARS is the consistency of the rating method. While some researchers claim BARS are a consistent measurement tool (Dessler, 2011; Pulakos, 2007), others have identified the negative effect BARS can have on an individual's appraisal (Tziner, Joanis, & Murphy, 2000). One way to ensure consistency with the BARS is to develop scenarios and scales thoroughly (Dessler, 2011; Pulakos, 2007). An example of thorough development of BARS is when multiple reviews with subject matter experts occur and confirmation is obtained that the behavioral statements are a consistent measure through one or more pilot tests (Pulakos, 2007).

The interpretation of BARS is essential to the correct use of the rating method. Dessler (2011) and Pulakos (2007) provided examples that went beyond three tiers, expanding to a Likert scale of 5 to 7. Both Dessler and Pulakos explained that an applicant might demonstrate certain behaviors that are listed in different categories. Therefore, the rater needs to be able to distinguish how effective, or ineffective, the applicant performs and score the behavior using the appropriate construct. If the rater is unable to make this interpretation, the result could be subjectivity, indicators of bias, and rater disagreement.

The Validity of Constructed Response Multimedia Testing

The constructed response multimedia test is a standard method in many entry-level police officer assessment and selection processes (Aamodt, 2004; Corey et al., 1996; Cucina et al., 2015; DeCicco, 1999; Doerner & Nowell, 1999). This high-fidelity test (Cucina et al., 2015) is often administered in an assessment center approach where applicants participate in multiple exercises that do not require knowledge or training in police officer policies and procedures (DeCicco, 2000). When BARS include measurements for problem-solving skills and effective interpersonal responses rather than consideration for specific knowledge of police officer policies and procedures, then this rating method could improve the fairness of the assessment (see Arthur & Villado, 2008; Wolgast et al., 2017).

The measure of criterion-related validity is how well the test predicts job performance (Cook, 2016). In multiple studies, the constructed response multimedia tests were strong predictors of candidate performance in a police academy (Corey et al., 1996; Cucina et al., 2015; Doerner & Nowell, 1999). The criterion-related validity in the studies on police academy recruits did not vary based upon subgroup differences, which is an indicator that using this approach can mitigate the risk of adverse impact. However, in each of the criterion-validity studies (Corey et al., 1996; Cucina et al., 2015; Doerner & Nowell, 1999), the raters had been police officers themselves, worked for a CSC, or had long-term experience in assessment and selection.

Differential validity is when an evaluation method is a better predictor of performance for one group than another (Berry et al., 2013; Cook, 2016). Although Schmidt and Hunter (1986) were adamant that differential validity was not present in their review of 85 years of research, differential validity is identified in several recent studies (Berry, Cullen, & Meyer, 2014; Berry et al., 2013; Rayson, Holliman, & Belyavin, 2000; Roth et al., 2017). When differential validity occurs, one group is outperforming another on the job even though both groups were tested similarly using the same method. Differential validity is not the same as subgroup differences in performance on the assessment. Subgroup differences in performance on an assessment can be calculated to determine if there are indicators of adverse impact in the test outcomes. However, differential validity should also be calculated to determine whether scores on the tests are better predictors of performance in the academy based on the evaluator method and applicant characteristics.

Incremental validity is when a predictor can explain a measurable outcome such as performance on a test or a job (Cook, 2016; Hunsley & Meyer, 2003; Meyer, 2007). In addition to measuring the applicant demographic characteristics as predictors of performance on the assessment and the academy, the focus is on the value of adding the community evaluator as a new method of evaluation. Calculating incremental validity contributes to understanding the amount of variance that each predictor variable contributes to the outcome when measured separately and together (Hunsley & Meyer, 2003; Meyer, 2007). The result could contribute to understanding if the new evaluation method of a community evaluator results in incremental validity for applicant testing outcomes and recruit performance in the police academy.

Reliability of Constructed Response Multimedia Testing

Reliability is the term used to describe the level of consistency of a test, method, or instrument (Cook, 2016). Cook presented an extensive overview of reliability in personnel selection research that included retest reliability, internal consistency reliability, and interrater agreement. Retest reliability is the comparison of scores that are obtained from people on two different occasions using the same test, method, or instrument (Cook, 2016). Internal consistency is an evaluation of the items in a test to ensure that each item is appropriate (Cook, 2016). Interrater agreement is the level of agreement between raters who assess the same people (Cook, 2016).

Evaluating interrater agreement based on the panel of evaluators contributes to understanding the level of agreement among the assessors (Cook, 2016). Individuals who have experience in a position may have different interpretations, expectations, and perceptions of job requirements (Conley & Sacket, 1987; Sacket & Laczo, 2003). Because of the opportunity for variability among raters, statistically analyzing the results is one way to measure the reliability of the evaluations (Fleiss, 1971; Shrout & Fleiss, 1979). A reliability index can be useful when evaluating the level of agreement, or variance, among raters. This statistical analysis can also contribute to identifying rater qualifications, bias, and comprehension of the rating method (Dierdoff & Wilson, 2003; Morgeson, Delaney-Klinger, Mayfield, Ferrara, & Campion, 2004).

The measurement of rater agreement has been calculated in police officer selection studies that use a constructed response multimedia test and BARS as the rating method (see Cucina et al., 2016; De Soete et al., 2013; Doerner & Nowell, 1999). Doerner and Nowell (1999) evaluated the reliability of a behavioral-personnel assessment device (B-PAD) and found that rater and ratee characteristics, as defined by race and sex, did not have a statistically significant effect on applicant scores. Intraclass correlations in two studies demonstrated consistent and statistically significant rater agreement (Cucina et al., 2016; De Soete et al., 2013). Although the researchers (Cucina et al., 2016; De Soete et al., 2013) adhered to the guidelines for selecting and reporting intraclass correlations (see Koo & Li, 2016), the raters in these two studies were referred to as trained reviewers and the researchers did not provide any additional demographic details.

Community Involvement

In many cities of the United States, the relationship between the public and the police is strained (Bergman, 2018; Chatterjee, 2016; Gould, 2017; Ruggs et al., 2016; Todak, 2017). Some researchers call for methods to improve community relations through hiring procedures, public engagement initiatives, surveys, and policy changes (Bergman, 2016; Chatterjee, 2016; Gould, 2017; Ruggs et al., 2016; Todak, 2017). Herndon (2016) was the only researcher to respond to Ruggs et al. (2016) with an explanation of the use of force in law enforcement. Herndon's research also included a suggestion about how changes in the community could improve the relationship between citizens and police. However, in *Public Management*, a trade journal, Gould (2017) provided suggestions for both law enforcement and the community to improve engagement between the two. Gould's research included a recommendation to involve citizens in the assessment and selection process for police officers.

Communication and personality trait studies in police officers have demonstrated the importance of measuring the communication style, and personality dimension constructs as a method of predicting performance (Lawrence, Christoff, & Escamilla, 2017). Lawrence et al. found that the evaluation of communication style and psychological characteristics of police officer applicants are a predictor of police-community interactions. Lawrence et al. also explained how evaluating communication styles and personality dimensions in the assessment and selection process are important constructs when measuring the predictive validity of a method.

The President's Task Force on 21st Century Policing

In 2015, the President of the United States commissioned a task force to "build trust between citizens and their peace officers" (The President's Task Force on 21st Century Policing [Task Force], 2015). The Task Force established six pillars for building this relationship: (1) Building Trust and Legitimacy; (2) Policy and Oversight; (3) Technology and Social Media; (4) Community Policing and Crime Reduction; (5) Training and Education; and (6) Officer Wellness and Safety. However, there is limited evidence in the peer-reviewed literature on the outcomes of the Task Force initiative.

Community Oriented Policing

Prior to 2015, government-funded projects that provided funding to law enforcement agencies included Columbus Law Enforcement Block Grants (Lilley & Boba, 2008), Community Oriented Policing Services (Lilley & Boba, 2008; Simmons, 2012) and Hiring in the Spirit of Service (Simmons, 2012). The three U.S. government-funded projects were intended to promote community involvement in the recruitment, selection, and development of police officers. Although Simmons (2012) listed five cities in the country that engaged members of the community in their process, only community members in Detroit, Michigan were provided the opportunity to vote as a stakeholder in the selection phase. The effect of the community members as participants in the rating process in Detroit is unknown.

The recommendations by DeCicco (2000) and research by Simmons (2012) that occurred before the Task Force (2015) demonstrate that some of the ideas and tactics suggested in 2015 to support the six pillars are not new to the field of police officer assessment and selection. The evidence by Simmons (2012) and DeCicco (2000) is supported in the Task Force (2015) report, where research commissioned by U.S. President Lyndon Johnson in 1967, *The Challenge of Crime in a Free Society*, is cited. As previously discussed, the importance of creating a valid selection process, engaging the community, and ensuring the approaches are legally defensible are recurring themes in entry-level police officer selection literature.

Importance of Understanding the Effect of the Rater on Selection Outcomes

Evaluations of assessment and selection methods for entry-level police officers include reliability and validity (Lievens et al., 2015; Sackett, Shewach, & Keiser, 2017); the effect of technology (Cucina et al., 2015); and adverse or disparate impact (De Soete et al., 2017; Guajardo, 2014; Highhouse et al., 2016; Hilal et al., 2017; Kringen, 2016; Riccucci & Riccardelli, 2015). However, none of these studies measured the effects of community member participation in the rating process. As calls for community involvement in the selection process become more prevalent (DeCico, 2000; Simmons, 2012), and cities begin to implement this method (see Ferrell, 2017; Rouan, 2017; Simmons, 2012), evaluating the results is necessary to determine the effectiveness of this alternative approach.

Summary and Transition

This literature review is evidence that there are a significant number of studies on the importance of entry-level police officer selection, recommendations for the use of alternative methods, disagreements on interpretations of *The Guidelines*, and calls for investigation of CSC methods. However, none of the studies included an investigation of the combination of a constructed response multimedia test and community evaluators as raters for a large and diverse group of entry-level police officer applicants. Additionally, none of the studies reviewed assessed the effect of a community member as an assessor on the assessment and selection outcomes for entry-level police officers.

This chapter demonstrates how literature supports the theoretical framework of MRT and GMA. MRT has been employed when measuring rater agreement, subgroup differences, and the predictive validity of a constructed response multimedia test (Corey et al., 1996; Cucina et al., 2015; Doerner & Nowell, 1999; Wolgast et al., 2017). The constructed response multimedia test in Phase 3 of the second step in Columbus' multiple hurdle selection process aligns with the method used in the research because applicants watch job-related scenarios and are then required to demonstrate problem-solving and interpersonal skills (Columbus, 2019d). GMA applies to the predictive validity component of this study because problem-solving, a proxy for cognitive ability (Arthur et al., 2013), is measured in the alternative selection method of a constructed response multimedia test (Columbus, 2019d).

This review of empirical studies supports the need for research on the predictors of applicant performance on the constructed response multimedia test and candidate performance in the police academy. Measuring subgroup differences of a large and diverse population of applicants addresses the limitation identified in a similar study (Lievens et al., 2015) and goes further to explore the alternative method of a community evaluator. As part of the City of Columbus' multiple hurdle selection process, a constructed response multimedia test (COPE) is used to measure the constructs of problem-solving and interpersonal skills that are scored by raters who utilize BARS in their evaluation of applicants (Columbus, 2019d).

Chapter 3 is a presentation of the research design for this dissertation, definition of the sample, and statistical methodology used to analyze indicators of adverse impact in the selection process, rater agreement, and predictors of performance on the constructed response multimedia test and in the police academy. Chapter 4 includes an analysis of the data and results of the statistical methodology. Chapter 5 consists of an interpretation of the results, limitations, recommendations for future research, social change implications of this study, and conclusions.

Chapter 3: Research Method

Introduction

The purpose of this quantitative study was to determine the effect of integrating community evaluators as an alternative selection device for the selection process of entry-level police officers in Columbus, Ohio. The City of Columbus CSC used a 10-step noncompensatory multiple hurdle selection process for the assessment and selection of entry-level police officers to determine eligibility for acceptance into the Academy (Columbus, 2019c; Columbus, 2019d). In the second step, a testing process consisting of four phases were:

1. a multiple-choice test;
2. a written work sample;
3. a constructed response multimedia test (COPE);
4. a physical fitness test.

In Phase 3, COPE was designed to measure the problem-solving and interpersonal skills of the candidate (Columbus, 2019d; Corey et al., 1996; Cucina et al., 2015; Doerner & Nowell, 1999; De Soete et al., 2013; Lievens, 2015). In 2017, the City of Columbus modified Phase 3 of the four-phase process when they introduced the alternative approach of integrating community evaluators as raters (Ferrell, 2017; Rouan, 2017). Therefore, the focus on Phase 3 of the process administered by the CSC (Columbus, 2019d) was essential to this study.

In this chapter, I describe the quantitative approach and nonexperimental design for this study that includes a discussion of the variables. Definition of the population, data sources, collection, and assessment methods are also explained. The chapter will conclude with the steps to mitigate the risk of internal and external validity as well as the ethical procedures and research principles associated with this study.

Research Design and Rationale

I used a nonexperimental design for this quantitative study. Candidate demographics, candidate performance on the constructed response multimedia test measured by rater scores, and recruit performance in the Academy were collected from archival sources. Because I used archival data, there were no known participant time or resource constraints consistent with the design choice. Data analyzed in this study were not generated and collected for research purposes. Instead, the data for this study came from the results of the third phase of testing by the Columbus CSC from 2015–2017, and Academy results from 2015–2018.

The rating method in this study was the composition of rating panels who scored candidates on their performance on COPE. The rating panels consisted of uniform and CSC raters from 2015-2016 or uniform and community evaluators in 2017. An examination of rater agreement and subgroup differences was conducted. In the first research question, the independent variables were rating method and applicant demographics as predictors of scores on the assessment. For Research Question 2, the independent variables were rating method, applicant demographics, and score on the assessment as predictors of performance in the Academy as measured by the recruits' final score. For Research Question 3, the independent variables were rating method, applicant demographics, and score on the assessment as predictors of graduation from the

Academy. The variables were consistent with other studies where subgroup differences and predictive validity of a constructed response multimedia test used in entry-level police officer assessment were measured (Aamodt, 2004; Corey et al., 1996; Cucina et al., 2015; DeCicco, 1999; Doerner & Nowell, 1999).

Several researchers have conducted quantitative analyses on constructed response multimedia tests used in a multiple hurdle selection process for entry-level police officers that included interrater reliability (see Doerner & Nowell, 1999), or indicators of adverse impact (see De Soete et al., 2013), and predictive validity (see Corey et al., 1996; Cucina et al., 2015). Although this study is different from previous research because of the unique composition of this constructed response multimedia test (COPE) and the rating method (absence or presence of community evaluators), similar quantitative methods were used to conduct this analysis. Quantitative methods are appropriate for measuring rater reliability, subgroup differences, and predictive validity (see Corey et al., 1996; Cucina et al., 2015).

Methodology

Population and Sampling Procedures

The population in this study includes all the adults who participated in the assessment and selection process for police officers as applicants, candidates, and recruits in the Academy, or as raters who participated in the scoring of candidates on COPE from 2015–2017. The City of Columbus uses a banded approach to grouping candidates based on their performance in the third phase of the testing hurdle; thus, Academy recruits can be chosen from previous testing years. Data for Academy recruits and graduates were collected from 2015–2018.

The size of the population for this study was dependent on the number of police officer applicants, candidates, and recruits in Columbus during 2015–2018. However, calculating a power analysis to determine the minimum sample size was necessary. One of the recommendations from research is a minimum *N* of 100 for multiple regression exercises that use two variables (Warner, 2013). In addition, the use of a statistical program is a more accurate method of calculating a research sample size. The power analysis to calculate sample and effect size was facilitated using the G*Power program, which requires the researcher to input effect size and error probability of the study (Faul, Erdfelder, Buchner, & Lang, 2009; G*Power, 2014). Effect size (0.5), error probability ($\alpha = .05$), and a confidence interval of (.95) are commonly accepted values effective for reducing Type I and Type II errors in research (Frankfort-Nachmias & Leon-Guerrero, 2015). Using the recommended parameters, identifying multiple linear regression, and a three predictor variables, the recommended sample size from the G*Power program was 119 (G*Power, 2014).

Meeting and exceeding the sample size of candidates was not identified as a limitation before collecting data because, based on communication with the public safety test team manager at the CSC, the applications for the entry-level police officer position in Columbus, Ohio exceeded 1,000 per year and, on average, 800 became eligible for the third phase of the testing hurdle in previous years. However, the number of candidates appointed to the Columbus Police Academy as recruits was dependent on the results of the 10 steps in the multiple hurdle selection process. The number of recruits could not be, and was not, known until the data were collected (see Appendices C & D).

Instrument

COPE was developed in-house using a job analysis and by conducting several critical incident exercises with subject matter experts (City of Columbus CSC, 2012; see Harvey et al., 2007; Society of Industrial and Organizational Psychology, 2018). Scenarios were developed in close partnership with more than one group of subject matter experts (City of Columbus CSC, 2012). A primary objective for developing COPE was to ensure consistency and alignment between the three test versions that included eight different scenarios.

The BARS used to score each candidate consisted of behavior statements that align with a 5-point Likert scale (1 = *unacceptable*) to (5 = *excellent).* Therefore, the maximum raw score that could be earned for each scenario was 10 points. With eight scenarios, the maximum score that could be assigned by a rater was 80. The highest raw score an applicant could earn was 240 because there were three raters on each panel. The CSC then calculated *z* scores, by board, to determine an applicant's final score. If a candidate was eligible for veteran's preference points, these 5 or 10 points were applied to a passing score but were not used to move an applicant's score into the passing range (City of Columbus CSC, 2012).

The *z* score is then used to place the candidates who passed all four tests into one of three bands (90, 80, 70), which results in grouping scores that are within the same range (Murphy & Myors, 1995; Schmidt & Hunter, 1995). Candidates placed in the 90 band are the first to begin the following eight steps of the multiple hurdle selection process that precede the Academy (City of Columbus CSC, 2012). Once all the candidates in the 90 band have been approached to continue the multiple hurdle selection process, the candidates in the 80 band become eligible to continue to selection process followed by those in the 70 band. This approach could benefit candidates with lower scores that may have otherwise resulted in rejection from the multiple hurdle selection process (Murphy & Myors, 1995).

The community evaluators participated in 2 days (16 hours) of instruction with the Columbus Division of Police that included a job shadowing period with a police officer. The goal of the 2 days (16 hours) of instruction was to ensure the community evaluators had a basic understanding of the job duties of a police officer. Before scoring candidates, the raters in this study (uniform and community evaluators) received 1 day (8 hours) of training administered by the Columbus CSC. Training included (a) the purpose of COPE, (b) information and exercises on applying BARS when scoring candidates, and (c) how to identify and avoid several types of rater bias. The goal of the training was to ensure all raters were proficient with the evaluation and scoring process.

Data Collection

This study was conducted using archival data. Permissions necessary to gain access to the data were approval from the Walden University IRB (approval #12-20-18-0601405) and the City of Columbus public records request in accordance with the City of Columbus Public Records Policy (City of Columbus, 2019e). Historical or legal documents were not requested as sources of data.

I filed a public records request to the Columbus CSC Uniform Testing Unit that included (a) applicant nonpersonally identifiable information (numeric) that was assigned as a candidate identifier during the assessment and selection process with demographics that included name, race/ethnicity, and gender; (b) applicant scores for the third phase of the multiple hurdle selection process; (c) results of the multiple hurdle selection process that included final selection outcomes for each phase; (d) evaluator names and the nonpersonally identifiable information (numeric) assigned for the rating process with demographics that include role (Columbus CSC employee, uniform, or community); (e) ratings assigned by evaluators to applicants by scenario; and (f) documents used to report statistics for each of the testing phases and the 10-phase multiple hurdle process that included, but were not limited to, attrition and costs to administer.

The public records request to the Academy included (a) employment records that identify which applicants were accepted into the police academy as candidates, (b) class test score charts that include recruits' grades or GPA as a measure of performance in the academy and graduation status, and (c) class seniority worksheets that include evaluations of recruit performance and graduation status. The request for these records was for 2015–2018 and include Academy recruits who were applicants in the 2015–2017 testing cycles.

Data were received from the CSC on December 21, 2018, and January 7, 2019. Applicant counts for each phase of the police officer selection process was provided for 2015–2017, and the numbers associated with each phase of the selection process are current through December 31, 2018. The counts for each phase of the initial testing and multiple hurdle selection processes were provided for gender and race/ethnicity (see Appendix C). The CSC changed their reporting method for candidates by groups in 2017 and provided a different format for that year (see Appendix D). Costs for each phase of the multiple hurdle selection were not provided. When providing the requested data, the Columbus CSC (2012) included a redacted version of the *2012 Entry-Level Police Officer COPE Development Report* to provide additional background on Phase 3 of the testing process. There were no discrepancies identified in the data collected from the CSC.

Data from the Academy were received on January 8, 2019. The Academy provided final academy score and graduation status for all recruits ranging from 2015–2018. The Academy did not provide candidate identification number, race, or gender for the recruits. In some cases, only recruit last names were provided. To correct for this, the final disposition report provided by the CSC was used to match Academy recruits who completed the 10-step multiple hurdle selection process to their COPE score. The Academy also trained people from other municipalities who did not participate in the multiple hurdle selection process administered by the CSC. Therefore, these cases were excluded from the analysis. Of the 286 recruit names provided by the Academy, 162 were matched with COPE scores.

Operationalization of Constructs

Evaluation method. The evaluation method under investigation in this study was the change to the rating panel for the third phase of testing. In years 2015 and 2016, each panel was comprised of uniform evaluators and a CSC employee. In 2017, a community evaluator was introduced to the panel, and the CSC employee served as a moderator. The investigation into this evaluation method entailed an examination of interrater agreement. Because there are three or more raters on a panel and the BARS are an ordinal scale, the appropriate test for measuring interrater agreement is Kendall's *W* (Field, 2013; Gisev, Bell, & Chen, 2013; Lund Research, 2018). The statistical significance of Kendall's *W* is designated by a value ranging from 0 (no agreement between raters) to 1 (absolute agreement between raters), where .976 would explain 97.6% of variability among raters and demonstrate strong agreement (Field, 2013; Lund Research, 2019a). However, the statistical significance of Kendall's *W* is also indicated by a *p*-value, where $p < .05$ is deemed to be statistically significant (Lund Research, 2019a).

Subgroup differences. Subgroup differences were measured by gender (male/female) and race/ethnicity (majority/minority). Because the test method under investigation is a simulation exercise, there is a strong probability that subgroup differences may be low (De Soete et al., 2013). Subgroup differences in selection outcomes were measured using a *t* test and Cohen's *d* to estimate the effect size (see Arthur et al., 2002; Cucina et al., 2015; DeSoete et al., 2014; Field, 2013). The *t* test was used to determine the ratio of explained and unexplained variance between the gender (male or female) and race/ethnicity (majority or minority) groups individually (see Field, 2013). The calculation of *d* was dependent on the standard deviations that are identified between the groups (see Field, 2013). The groups did not have equal standard deviations for performance on the assessment, so the standard deviations for each group were pooled according to community evaluator presence, gender, and race/ethnicity. The formula used for calculating effect size was d = $(M_1\text{-}M_2)/SD_{pooled}$ where $SD_{pooled} = ((SD_1 \text{ x } N_1) + (SD_2 \text{ x } N_2))/(N_1+N_2)$.

Predictive and incremental validity. The predictive validity for the assessment and rating method in this study is an indicator of how well the test predicts performance in the Academy (see Corey et al., 1996; Cucina et al., 2015). Incremental validity was measured to determine if the modification of an existing rating method effected the predictive validity of this assessment. A multiple linear regression model was used to calculate validity because there were more than two independent variables in the equation (Lund Research, 2019b). The corrected criterion-validity coefficients are reported for each rating method and range from 0 to 1, and the results are compared to those published in Schmidt and Hunter's (1998) meta-analytic validity summaries. As identified in the limitations section, restriction of range was taken into consideration (see Berry et al., 2013) when comparing the results to Schmidt and Hunter's (1998) summaries. The Thorndike formula (as cited in Wiberg & Sundström, 2009) was integrated into the tests for this study prior to comparing the results to the findings by Schmidt and Hunter (1998).

Research Questions and Hypotheses

Quantitative methods were used to answer the following research questions to determine the effect of the community evaluator on the assessment and selection of police officers in Columbus, Ohio. The three questions were intended to measure selection outcomes and performance in the Academy.

Research Question 1: Does evaluation method type and/or candidate demographic characteristics predict the score on the Columbus Civil Service Commission constructed response multimedia test for candidates between 2015–2017?

$H_0$1: Evaluation method type (community evaluator presence or absence) and/or candidate demographic characteristics (gender and race/ethnicity) do not significantly predict the score on the Columbus Civil Service Commission constructed response multimedia test for candidates between 2015–2017.

H_a1: Evaluation method type (community evaluator presence or absence) and/or candidate demographic characteristics (gender and race/ethnicity) significantly predict the score on the Columbus Civil Service Commission constructed response multimedia test for candidates between 2015–2017.

Research Question 2: Does evaluation method type, candidate demographic characteristics, and/or score on the Columbus Civil Service Commission constructed response multimedia test predict Academy performance for recruits who were candidates between 2015–2017?

$H_0$2: Evaluation method type (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus CSC constructed response multimedia test do not significantly predict Academy performance for recruits who were candidates between 2015–2017.

H_a2: Evaluation method type (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the

Columbus CSC constructed response multimedia test significantly predict Academy performance for recruits who were candidates between 2015–2017.

Research Question 3: Does evaluation method type (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test predict Academy graduation for recruits who were candidates between 2015–2017?

$H_0$3: Evaluation method (community evaluator presence or absence of), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test do not significantly predict Academy graduation for recruits who were candidates between 2015–2017.

H_a3: Evaluation method (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test significantly predict Academy graduation for recruits who were candidates between 2015–2017.

Data Analysis

A *t* test and effect size (Cohen's *d*) was calculated to measure subgroup differences based on rating method (absence or presence of a community evaluator), gender (male or female), and race/ethnicity (majority or minority) groups individually (DeSoete et al., 2014; Field, 2013; Frankfort-Nachmias & Leon-Guerrero, 2015). Because the White group was the largest group of candidates and recruits, these cases were classified into the majority group. Any participant who identified as non-White or did not provide a race/ethnicity were classified into the minority group. Interrater agreement was calculated to measure the degree of consistency and agreement among raters using Kendall's *W* because the three raters who were randomly assigned to each rating panel then assigned ordinal values to the constructs being measured (Field, 2013; Gisev et al., 2013; Lund Research, 2019a).

To determine if the evaluation method and/or demographic differences significantly predicted an applicant's score on the Columbus CSC constructed response multimedia test (COPE), the predictor variables were (a) evaluation method (community evaluator presence or absence); (b) gender (male or female); and (c) race/ethnicity (majority or minority). The outcome (criterion) variable was the candidate's COPE score (ordinal). Descriptive statistics and intercorrelations were calculated to identify the relationships between the variables (see DeSoete et al., 2014; Doerner & Nowell, 1999). Linear regression was used to determine whether a statistically significant relationship existed between the predictors and the outcome variable (see Doerner & Nowell, 1999; Field, 2013; Warner, 2013).

To determine if the type of evaluation method, candidate's demographic characteristics, and/or COPE score significantly predicted a recruit's performance in the Academy, the predictor variables were (a) evaluation method (community evaluator presence or absence); (b) gender (male or female); (c) race/ethnicity (majority or minority), and COPE score (ordinal). The outcome (criterion) variable was the recruit's final performance score in the Academy (ordinal). Linear regression was used to determine whether a statistically significant relationship existed between the predictors and the outcome variable (see Doerner & Nowell, 1999; Field, 2013; Warner, 2013).

To determine if the type of evaluation method, candidate's demographic characteristics, and/or COPE score significantly predicted Academy graduation, the predictor variables were (a) evaluation method (community evaluator presence or absence); (b) gender (male or female); (c) race/ethnicity (majority or minority); and (d) COPE score (ordinal). The outcome (criterion) variable was recruit graduation from the Academy (did not graduate = 0, graduate = 1). Binary logistic regression was appropriate because there were only two possible outcomes (Field, 2013; Warner, 2013).

SPSS (2017) software was used to categorize and analyze the data received from the City of Columbus. Before analyzing any of these data, an exploratory analysis was conducted using SPSS to test that assumptions for regression were met (see Field, 2013; Frankfort-Nachmias & Leon-Guerrero, 2015; Warner, 2013). Tests for assumptions included linearity, independence of error, homoscedasticity, multicollinearity, undue influence, and normal distribution of errors (see Field, 2013; Frankfort-Nachmias & Leon-Guerrero, 2015). Each assumption was reviewed for the respective research questions where linear regression was used to determine whether the variables were statistically significant predictors of subgroup differences, performance on COPE, and performance in the Academy.

Threats to Validity

The reliability of the data in this study is dependent upon the local CSC responsible for collecting and reporting statistics associated with the Columbus, Ohio police officer exam and the affiliated Academy. As mentioned in the assumptions section, the Columbus CSC and Academy are Commission on Accreditation for Law Enforcement Agencies accredited agencies and are subject to audit via internal controls and requests for data from external sources. In addition, the data utilized in this study is subject to review of the Walden University Research Reviewer, Internal Review Board, and Dissertation Committee.

As previously mentioned, I have worked for the City of Columbus CSC Public Safety Test Team. In this role, my responsibilities have included designing, editing, and administering entry-level and promotional examinations for police and fire personnel, writing technical reports, and conducting data analysis. I was not part of the design or administration of COPE but did work as a substitute moderator for two days on one rating panel in 2017. This information is disclosed to alleviate any assumptions of bias, ethical issues, or impropriety.

External validity is a measure of how well a study can be generalized to a population with respect to the study participants, materials, and environment (Warner, 2013). In this study, the population is limited to raters, applicants, and candidates who participated in the Columbus, Ohio entry-level police officer assessment and selection process from 2015–2017 and Academy recruits from 2016–2018. The situations being tested in this study are not artificial, data are not being manipulated, and experiments are not being conducted. Because the constructed response multimedia test and Academy requirements are specific to Columbus, generalization to the population of police officer applicants is limited.

Threats to internal validity include, but are not limited to, history, maturation, testing, instrumentation, and statistical regression (Leighton, 2010). In this study, the time sequence associated with the four-year span is subject to events during that period that may have influenced applicant behavior (see Bergman, 2018; Todak, 2017), recruiting methods (Hilal et al., 2017; Newman & Lyons, 2009), and civil service cut off scores (Hoffman, 2018). Applicants are evaluated using one of three versions of the same constructed response multimedia test and scoring BARS throughout the three-year period, which mitigates the risk for familiarity with the instrument.

Ethical Procedures

The data collection includes demographic information for applicants and raters involved in the selection process. However, to protect the anonymity of those who participated in the selection process, unique identifiers and names are not published.
I obtained acknowledgment in writing from the Columbus CSC public safety test team manager and the commander of the police academy regarding the use of data. The data used in this study are available through a City of Columbus Public Records Request and are in accordance with the City of Columbus Public Records Policy (Columbus, 2019e). The employees of the Columbus CSC and respective police Academy did not collect data from participants on my behalf.

A vulnerable population is defined as a group that is one or more of the following: (a) chronically unhealthy, (b) underage, (c) incarcerated, (d) racial minorities, and (e) ethnic minorities (National Academy of Science, 2014). Evaluators in a selection process, police officer applicants, candidates, and recruits are not considered to belong to a vulnerable population. However, the records in this study include assessment ratings by the evaluator, race and gender for applicants, candidate scores on the assessment, and Academy outcomes. Although these data are available through a public records request, I have tried to ensure the participants remain anonymous and the data are protected. Protection of the data includes storing the records on a password protected computer.

The research study was reviewed by the Walden University IRB for compliance with human research and ethical standards. It was determined to meet institutional standards. Permission to conduct this study was granted by the Walden Institutional Review Board (IRB), approval #12-20-18-0601405.

Summary

In Chapter 3, I have identified and provided justification for the quantitative approach and methods that were used in this study. Explanation of the design included definition and operationalization of variables, the methodology that corresponds with the literature presented in Chapter 2, and a discussion of the instrument (COPE). The population, data collection, ethical procedures, and threats to external and internal validity were discussed and applied to the study. Data collection procedures adhered to the Walden University IRB and City of Columbus Public Records Policy.

Chapter 4 includes a presentation of the statistical test results and analysis of the data that aligns with the statistical methodology discussed in Chapter 3. Beginning with demographic information and descriptive statistics, Chapter 4 also includes the results of each statistical test conducted in this study. An evaluation of rater agreement, subgroup differences in the selection outcomes, acceptance or rejection of the null hypotheses for the three research questions, and an overview of the results is also provided. Chapter 5 consists of an interpretation of the results, limitations, recommendations for future research, potential social change implications, and conclusions of this study.

Chapter 4: Results

Introduction

The purpose of this quantitative study was to determine the effect of integrating community evaluators as an alternative selection device into the Columbus CSC assessment and selection process for entry-level police officers. The examination of scores on the constructed response multimedia test, also referred to as the COPE, included an evaluation of scores assigned by each panel of raters as measured by Kendall's *W* to evaluate the level of agreement of the rating panels for 2015–2017. Subgroup differences in selection outcomes were evaluated for indicators of adverse impact. The first part of the statistical analysis included an examination of whether the rating method, candidate gender, and race/ethnicity were statistically significant predictors of performance on COPE as measured by candidates' scores from 2015–2017. The second part of the statistical analysis involved whether the rating method, gender, race/ethnicity, and COPE score were statistically significant predictors of recruit performance in, and graduation from, the Academy.

Data Collection

As discussed in Chapter 3, I obtained and analyzed archival data from the City of Columbus Uniform Testing Unit (CSC) and Police Academy after receiving approval from Walden University's IRB. These data contained personally identifiable information (names and demographic characteristics) that were necessary for this study. Questions regarding the data and materials provided by the CSC were discussed with the public safety test team manager. My questions pertained to the assessment, selection, and training of the community evaluators, development procedures for COPE, scoring methods, and rating panel constructs. Other than the discrepancy identified with the data from the Academy, which resulted in reducing the dataset from 286 cases to 162, there were no other known issues with these data. Attempts to obtain additional information on the 124 removed cases from the Academy dataset were unsuccessful.

Sample Demographics

This study encompassed three subsets of data: (a) data received from the Columbus CSC that included adjusted final scores for candidates on COPE, (b) data received from the Columbus Police Academy for recruit performance in 2016–2018, and (c) data received from the Columbus CSC that included raw scores for candidates as determined by a three-person rating panel from 2015–2017. The first data subset included the population of all candidates who were scored on COPE for 2015–2017. Table 1 provides the frequency distribution for the applicants. Gender data for the 2,510 candidates scored on COPE was 2,080 (82.9%) male, 419 (16.7%) female, and 11 (0.4%) did not provide gender information. Most of the ethnic distribution of the 2,510 candidates were White 1,892 (75.4%). The remaining race/ethnic groups included 314 (12.51%) Black or African American; 122 (4.9%) Two or More races; and 100 (4.0%) Hispanic or Latino; 43 (1.7%) Asian; 13 (0.5%) Missing/Blank; 12 (0.5%) American Indian or Alaskan Native; eight (0.3%) Prefer Not to Answer; and six (0.2%) Native Hawaiian or Pacific Islander.

Table 1

Frequency Distribution of COPE Candidates

	Gender			
	Frequency	%	Valid %	Cumulative %
Male	2,080	82.9	83.1	83.1
Female	419	16.7	16.7	99.8
Missing	11	0.4	0.2	100
Total	2,510	100	100	
	Race/Ethnicity			
	Frequency	%	Valid %	Cumulative %
Majority	1,892	75.4	75.4	75.4
Minority	618	24.6	24.6	100
Total	2,510	100		
White	1,892	75.4	75.8	75.8
2 or More Races	122	4.9	4.9	80.7
American Indian or Alaskan	12	0.5	0.5	81.1
Asian	43	1.7	1.7	82.9
Black or African American	314	12.5	12.6	95.4
Hispanic or Latino	100	4.0	4.0	99.4
Native Hawaiian or Pacific Islander	6	0.2	0.2	99.7
Prefer Not to Answer	8	0.3	0.3	100
Missing	13	0.5		
Total	2,510	100		

For this study, the White candidates 1,892 (75.4%) were coded into the majority group and all 618 (24.6%) applicants who identified as a race other than White, or did not provide an answer, were coded into the minority group.

The second data subset includes the population of Academy recruits from 2015–2018, their COPE scores, Academy score, and graduation status. Table 2 provides the detailed summary frequency distribution for Academy recruits. COPE scores were available for a total of 162 recruits who participated in, or graduated from, the Academy from 2016–2018. Of the 162 recruits, 137 (84.6%) were male; 23 (14.2%) were female; 128 (79%) were White (Majority); and 34 (21%) were non-White (Minority). The demographics of these datasets are representative of the findings from a recent study by Meier et al. (2018) that included recruits in police academies (85% male).

The 11 candidates and two recruits who did not provide gender information were removed from the subgroup differences and regression analyses where gender was used as a predictor. The justification for this approach is based upon research by Arthur et al. (2013) that addresses how the differences in subgroups can include psychological and physical assumptions. There is a lack of evidence to support whether the candidates withheld demographic data based upon perceptions (Ryan et al., 2000), or because the group (gender, race/ethnicity) that candidates identified with was not listed as an option.

Table 2

Frequency Distribution of Academy Recruits

	Gender			
	Frequency	%	Valid %	Cumulative %
Male	137	84.6	85.6	85.6
Female	23	14.2	14.4	100.0
Missing	2	1.2		
Total	162	100		
	Race/Ethnicity			
	Frequency	%	Valid %	Cumulative %
Majority	128	79.0	79.0	79.0
Minority	34	21.0	21.0	100.0
Total	162	100.0		
White	128	79.0	80.0	80.0
2 or More Races	9	5.6	5.6	85.6
Black or African American	17	10.5	10.6	96.3
Hispanic or Latino	6	3.7	3.8	100.0
Missing	2	1.2		
Total	162	100.0		

Table 3 provides the frequency distribution of candidates based on the presence of a community evaluator from the first and third datasets. Among the 2,510 candidates who were scored on COPE from 2015–2017, a community evaluator participated in scoring 831 (33.1%). Of the 162 recruits, community evaluators participated in scoring 53 (32.7%) on COPE because they were part of the evaluation process in 2017, whereas

109 (67.3%) were evaluated by panels that did not include a community evaluator in 2015 and 2016.

Table 3

Frequency Distribution of COPE Candidates and Academy Recruits and Presence of a Community Evaluator During COPE

	COPE Candidates			
	Frequency	%	Valid %	Cumulative %
Community Evaluator Not Present	1,679	66.9	66.9	66.9
Community Evaluator Present	831	33.1	33.1	100
Total	2,510	100	100	
	Academy Recruits			
	Frequency	%	Valid %	Cumulative %
Community Evaluator Not Present	109	67.3	67.3	67.3
Community Evaluator Present	53	32.7	32.7	100
Total	162	100	100	

The third dataset also included the raw scores for 2,510 candidates that were assigned by each of the three-person rating panels from 2015–2017. Table 4 includes the frequency distribution for the number of rating boards and candidates scored from 2015–2017. The number of boards was increased from seven in 2015 and 2016 to eight in 2017 to accommodate the number of community evaluators hired by the City of Columbus. Prior to the addition of the eighth rating board in 2017, the number of candidates scored per board ranged from 105 to 134 in 2015, and 113 to 119 in 2016. In 2017, the number of candidates scored per board ranged from 91 to 116.

Table 4

Frequency Distribution of Boards and Candidates Scored from 2015–2017

2015 (*n* = 865)			
Board	Frequency	%	Valid %
1	134	15.5%	15.5%
2	132	15.3%	15.3%
3	105	12.1%	12.1%
4	124	14.3%	14.3%
5	125	14.5%	14.5%
6	124	14.3%	14.3%
7	121	14.0%	14.0%
2016 (*n* = 814)			
Board	Frequency	%	Valid %
1	113	13.9%	13.9%
2	117	14.4%	14.4%
3	115	14.1%	14.1%
4	117	14.4%	14.4%
5	117	14.4%	14.4%
6	116	14.3%	14.3%
7	119	14.6%	14.6%
2017 (*n* = 831)			
Board	Frequency	%	Valid %
1	91	11.0%	11.0%
2	108	13.0%	13.0%
3	105	12.6%	12.6%
4	104	12.5%	12.5%
5	98	11.8%	11.8%
6	98	11.8%	11.8%
7	116	14.0%	14.0%
8	111	13.4%	13.4%

Note. *n* = number of applicants evaluated.

Interrater Agreement

The evaluation of rater agreement in this study included the third dataset from 2015–2017. In 2015 and 2016, each panel consisted of two uniform evaluators and one CSC employee. In 2017, when the community evaluator was introduced to the panel, the CSC employee served as a moderator. The community evaluators were chosen as a method of providing residents of Columbus the opportunity to have a voice in the selection process (Ferrell, 2017; Rouan, 2017). The three raters were randomly assigned to a panel and used the ordinal scale of the BARS to score candidates on problem solving and interpersonal skills (City of Columbus, 2012).

Kendall's *W* was run to determine whether there was a statistically significant level of agreement between the rater's judgment for each panel in 2015–2017. The overall total raw score assigned by each rater was used to measure agreement. In all 3 years, the rater agreement in their assessments was statistically significant, $W = .950$ to $.969$, $p < .01$ (see Table 5). The interpretation of this range is that Kendall's *W* explains 95% to 96.9% of variability among raters and demonstrates strong agreement (see Lund Research, 2019a). Although the level of agreement in 2017 was slightly lower than the rating panels without community evaluators, the difference is not statistically significant.

Table 5

Rater Agreement for 2015–2017 as Measured by Kendall's W

2015 Boards (Total *n* = 865)			
	n	*W*	*p*-value
1	134	0.970	0.000
2	132	0.970	0.000
3	105	0.964	0.000
4	124	0.952	0.000
5	125	0.981	0.000
6	124	0.985	0.000
7	121	0.960	0.000
Mean	124	0.969	
2016 Boards (Total *n* = 814)			
	n	*W*	*p*-value
1	113	0.974	0.000
2	117	0.946	0.000
3	115	0.982	0.000
4	117	0.949	0.000
5	117	0.961	0.000
6	116	0.983	0.000
7	119	0.982	0.000
Mean	117	0.968	
2017 Boards (Total *n* = 831)			
	n	*W*	*p*-value
1	91	0.961	0.000
2	108	0.966	0.000
3	105	0.951	0.000
4	104	0.939	0.000
5	98	0.951	0.000
6	98	0.944	0.000
7	116	0.931	0.000
8	111	0.959	0.000
Mean	104	0.950	

Note. n = number of applicants evaluated; *W* = Kendall's W.

Descriptive Statistics

A total of 2,510 candidates completed and were scored on COPE. The mean (and standard deviation) of the final scores reported was 80.3% (9.954), and the range was 43 to 108. Academy scores were reported for all participants who attended the Academy, regardless of their graduation status. Of the 162 candidates accepted to the Academy as recruits between 2015–2018, the mean (and standard deviation) Academy score was 90.83% (9.29), and the range was 0 to 97.43. One recruit's Academy score was reported as 0 and there was no evidence to support the recruit participated in the Academy. After removing the participant with a score of 0, the adjusted mean (and standard deviation) Academy score was 90.83% (9.29), and the range was 35.43 to 97.43 (see Table 6).

Table 6

Descriptive Statistics for COPE and Academy Scores for COPE Candidates and Academy Recruits

	COPE Candidates				
	N	Min	Max	*M*	*SD*
COPE Score	2,510	43	108	80.03	9.954
	Academy Recruits				
	N	Min	Max	*M*	*SD*
COPE Score	162	59	106	85.42	10.380
Academy Score	162	0.00*	97.43	89.08	12.786
	Academy Recruits (Revised)				
	N	Min	Max	*M*	*SD*
COPE Score	161	59	106	85.42	10.380
Academy Score	161	35.43	97.43	91.39	5.909

Note. *N* = number of participants; Min = minimum score; Max = maximum score; *M* = mean score; *SD* = standard deviation; * = Removed from the analysis.

Table 7 includes the breakout of the 2,510 COPE Candidate Scores by gender and race/ethnicity. Although the range of scores was lower for males (43 to 107) than females (53 to 108), the means (and standard deviation) for males = 79.88 (9.991) and females = 80.84 (9.733) were similar. The means (and standard deviation) were also similar for the majority = 79.96 (9.971) and minority = 80.25 (9.909) groups. The minority groups with a mean score higher than the White (majority) group included Two or More Races (80.57), American Indian or American Native (80.33), Black or African American (80.99), and Prefer Not to Answer (84.88).

Table 7

Descriptive Statistics for Candidate COPE Score by Gender and Race/Ethnicity

	Gender				
	N	Min	Max	*M*	*SD*
Male	2,080	43	107	79.88	9.991
Female	419	53	108	80.84	9.733
Missing	11	59	96	76.73	10.189
Total	2,510				
	Race/Ethnicity				
	N	Min	Max	*M*	*SD*
Majority	1,892	43	108	79.96	9.971
Minority	618	43	107	80.25	9.909
Total	2,510				
White	1,892	43	108	79.96	9.971
Two or More Races	122	55	103	80.57	9.393
American Indian or American Indian	12	67	94	80.33	8.316
Asian	43	56	97	77.09	10.033
Black or African American	314	43	104	80.99	9.820
Hispanic or Latino	100	53	107	79.34	10.880
Native Hawaiian or Pacific Islander	6	65	83	72.67	6.713
Prefer Not to Answer	8	74	97	84.88	6.792
Missing	13	63	96	77.31	9.776
Total	2,510				

Note. N = number of participants; Min = minimum score; Max = maximum score; *M* = mean score; *SD* = standard deviation.

Table 8 includes the breakout of COPE scores and Academy scores by gender and race/ethnicity for the sample of 161 Academy recruits. Males and females accepted to the Academy performed similarly on COPE (85.54 and 84.61 respectively) and in the Academy (91.49 and 90.48 respectively). However, the Gender Missing group (*N* = 2) performed the highest on COPE (92) and in the Academy (95.15). The performance of the majority and minority groups was similar on COPE (85.03 and 87.24 respectively) and in the Academy (91.57 and 90.72 respectively). The Race/Ethnicity Missing group performed the highest on COPE (92) and in the Academy (95.15).

Table 8

Descriptive Statistics for Candidate COPE Score and Academy Score by Gender and Race/Ethnicity

			Gender			
	Score	*N*	Min	Max	*M*	*SD*
Male	COPE	136	59.00	106.00	85.54	10.082
	Academy	136	35.43	97.43	91.49	6.209
Female	COPE	23	59.00	104.00	84.61	12.398
	Academy	23	77.50	94.07	90.48	3.833
Missing	COPE	2	88.00	96.00	92.00	5.657
	Academy	2	92.90	97.40	95.15	3.182
Total		161				
			Race/Ethnicity			
	Score	*N*	Min	Max	*M*	*SD*
Majority	COPE	128	59.00	106.00	85.03	10.038
	Academy	128	35.43	97.43	91.57	6.375
Minority	COPE	33	59.00	104.00	87.24	11.608
	Academy	33	83.14	97.40	90.72	3.561
Total		161				
White	COPE	128	59.00	106.00	85.03	10.04
	Academy	128	35.43	97.43	91.57	6.38
Two or More Races	COPE	9	67.00	103.00	89.56	10.83
	Academy	9	87.50	94.80	91.74	2.14
Black or African American	COPE	17	59.00	104.00	85.76	12.23
	Academy	17	83.14	94.60	88.85	3.60
Hispanic or Latino	COPE	5	62.00	97.00	86.20	14.25
	Academy	5	92.10	94.80	93.49	1.17
Missing	COPE	2	88.00	96.00	92.00	5.66
	Academy	2	92.90	97.40	95.15	3.18
Total		161				

Note. N = number of participants; Min = minimum score; Max = maximum score; *M* = mean score; *SD* = standard deviation.

Subgroup Differences

Group differences based on the rating method were measured by Community Evaluator (present/not present) using a *t* test and Cohen's *d* to estimate the effect size (see Arthur et al., 2002; DeSoete et al., 2014; Cucina et al., 2015; Field, 2013). An independent samples *t* test was performed utilizing the first dataset to assess whether mean COPE score differed significantly for a group based upon the presence of a community evaluator on the rating panel. For both groups, the assumption of homogeneity of variance was assessed by the Levene test (*F*). Because the value of *F* was small and not statistically significant (*p*-value > .05), no significant violation of the equal variance assumption was indicated. Therefore, the pooled variances version of the *t* test was used. The mean scores of each group were not statistically significant (*p*-value > .05) and the *d*-value (-0.005) is unlikely to yield adverse impact (see Table 9).

Table 9

t-Test Results and Effect Sizes for Rating Method

Group	*n*	*M*	*SD*	*F*	*p*	*t*	*p*	*d*
CE Not Present	1,679	80.01	10.014					
CE Present	831	80.06	9.838	0.788	0.375	-0.128	0.898	-0.005

Note. n = 2,510 for both groups; *M* = mean; *SD* = standard deviation; *F* = Levene F; *p* = probability value*;* *t* = t-ratio; *d* = Cohen's d; *CE* = community evaluator.

Table 10 includes the sample size, COPE score means (and standard deviations) for rating method (community evaluator present/not present), gender (male/female), and race/ethnicity (majority/minority). COPE scores from a total of 2,499 candidates were used for this research question (the 11 cases where gender was missing were removed from the sample of 2,510). A total of 1,673 (67%) candidates were scored without a community evaluator, and 826 (33%) were scored with a community evaluator. The distribution of scores for the sample size was roughly normal, the variances of scores were not significantly different for males/females or majority/minority, and scatterplots did not indicate nonlinear relations or bivariate outliers.

Table 10

COPE Score Means and Standard Deviations by Rating Method, Gender, and Race/Ethnicity

Community Evaluator			*n*	*M*	*SD*
Not Present	Male	Majority	1,077	80.00	10.140
		Minority	307	79.89	9.888
	Subtotal		1,384		
	Female	Majority	204	80.41	9.750
		Minority	85	80.09	9.558
	Subtotal		289		
	Total		1,673		
Present	Male	Majority	520	79.37	9.624
		Minority	176	80.64	10.330
	Subtotal		696		
	Female	Majority	87	82.22	10.045
		Minority	43	81.53	9.356
	Subtotal		130		
	Total		826		

Note. n = 2,499 for all groups; *M* = mean; *SD* = standard deviation.

Subgroup differences were measured by gender (male/female *N* = 2,499) and race/ethnicity (majority/minority *N* = 2,510) using a *t* test and Cohen's *d* to estimate the effect size (Arthur et al., 2002; Cucina et al., 2015; DeSoete et al., 2014; Field, 2013) for each group. An independent samples *t*-test was performed utilizing the first dataset to assess whether mean COPE score differed significantly for a group based upon gender or race/ethnicity. For both groups, the assumption of homogeneity of variance was assessed by the Levene test (*F*). In each case, the value of *F* is small and not statistically significant (*p*-value > .05), indicating no significant violation of the equal variance assumption. Therefore, the pooled variances version of the *t*-test was used for each group. The mean scores of each group were not statistically significant (*p*-value > .05) and the *d*-values ranging from -0.0291 to -0.0965 are unlikely to yield adverse impact (see Table 11).

Table 11

t-Test Results and Effect Sizes for Gender and Race/Ethnicity

				Gender				
	n	*M*	*SD*	*F*	*p*	*t*	*p*	*d*
Male	2,080	79.88	9.991					
Female	419	80.84	9.733	0.167	0.683	-1.789	0.074	-0.0965
Subtotal	2,499							
				Race/Ethnicity				
	n	*M*	*SD*	*F*	*p*	*t*	*p*	*d*
Majority	1,892	79.96	9.971					
Minority	618	80.25	9.909	0.049	0.824	-0.633	0.527	-0.0291
Subtotal	2,510							

Note. n = sample size; *M* = mean; *SD* = standard deviation; *F* = Levene F; *p* = probability value*;* *t* = t-ratio; *d* = Cohen's d.

Table 12 includes the sample size, COPE score means (and standard deviations) for rating method (community evaluator present/not present), gender (male/female), and race/ethnicity (majority/minority). COPE scores from a total of 159 recruits were used for this research question (one recruit with a score of zero and two recruits who did not provide gender were excluded from the analysis). A total of 108 (68%) recruits were scored on COPE without a community evaluator and 51 (32%) were scored with a community evaluator. For the 108 recruits who were not rated by a community evaluator, the mean COPE score was similar for females in the majority (81.0%) and minority (81.6%) group and for males in the majority (84.27%) and minority (84.12%) group. Whereas the 51 COPE scores where a community evaluator was present were higher for females (91.00%) when compared to males (86.58%) but were the highest overall for minority males (93.20%). There were no minority female recruits in the Academy who were rated by a community evaluator. For the 109 recruits who were not rated by a community evaluator, the mean Academy score for females in the majority group (91.15%) were similar to males in the majority group (91.51%) and higher than minority group females (89.75%) and males (84.60%). For the 51 Academy scores where a community evaluator was present, the mean for majority males was the highest (92.17%), followed by minority males (91.65%) and females (90.10%).

Table 12

COPE and Academy Score Means and Standard Deviations by Rating Method, Gender, and Race/Ethnicity

Community Evaluator				*n*	*M*	*SD*
Not Present	Female	Majority	COPE Score	10	81.00	11.738
			Academy Score	10	91.15	2.833
		Minority	COPE Score	5	81.60	17.953
			Academy Score	5	89.75	2.769
	Subtotal			15		
	Male	Majority	COPE Score	77	84.27	10.514
			Academy Score	77	91.51	7.637
		Minority	COPE Score	16	84.12	10.959
			Academy Score	16	89.89	3.642
	Subtotal			93		
	Total			108		
Present	Female	Majority	COPE Score	8	91.00	7.071
			Academy Score	8	90.10	5.502
		Minority	COPE Score	-	-	-
			Academy Score	-	-	-
	Subtotal			8		
	Male	Majority	COPE Score	33	86.58	8.359
			Academy Score	33	92.17	3.596
		Minority	COPE Score	10	93.20	7.406
			Academy Score	10	91.65	3.370
	Subtotal			43		
	Total			51		

Note. n = 159 for all groups; *M* = mean; *SD* = standard deviation.

Correlational Analysis

The correlation matrix for COPE Scores by rating method, gender, and race/ethnicity is provided in Table 13. There were only two correlations with statistical significance, and both of these were very small (| *r* | < .1). There was a very small positive correlation between gender and score *r*(2,497) = .04, *p* < .05. There was a very small positive correlation between race/ethnicity and gender *r*(2,497) = .06, *p* < .01. Because none of the correlations between the predictor variables in Table 13 are greater than .70, there is no evidence of multicollinearity.

Table 13

Results of the Pearson Correlation for COPE Score, Rating Method, Gender and Race/Ethnicity

(*n* = 2,499)		Score	Community Evaluator	Gender	Race/ Ethnicity
Pearson Correlation	Score	-			
	Comm. Eval.	0.00	-		
	Gender	0.04	-0.02	-	
	Race/Ethnicity	0.01	0.03	0.06	-
Sig. (1-tailed)	Score	-			
	Comm. Eval.	0.486	-		
	Gender	0.037	0.167	-	
	Race/Ethnicity	0.276	0.046	0.001	-

Note: n = sample size.

Academy Scores by rating method, gender, race/ethnicity, and COPE Score is provided in Table 14. The strength of the correlations were very small (| *r* | < .1) for all variables with the exception of a small positive correlation between COPE Score and Community Evaluator (.1 < | *r* | < .3). The correlation between Community Evaluator and COPE Score was the only statistically significant correlation, *r*(157) = .210, *p* < .01.

Because none of the correlations between the predictor variables in Table 14 are greater than .70, there is no evidence of multicollinearity. The Thorndike Model (as cited by Wiberg & Sundström, 2009) was used to correct for restriction of range for the validity of COPE score as a predictor of performance in the Academy. The uncorrected value *r*(157) = -.025, *p* = .375. The corrected value *r*(157) = -.024. The corrected *r* value is much lower than those included in Schmidt and Hunter's (1998) summaries for corrected verbal work samples *r*(3,159) = .44 and cognitive ability (*r* = .56).

Table 14

Results of the Pearson Correlation for Academy Score, Community Evaluator, Gender, Race/Ethnicity, and COPE Score

(*n* = 159)		Academy Score	Community Evaluator	Gender	Race/ Ethnicity	COPE Score
Pearson Correlation	Academy Score	-				
	Comm. Eval.	0.05	-			
	Gender	-0.06	0.02	-		
	Race/Ethnicity	-0.07	0.00	0.02	-	
	COPE Score	-0.03	0.21	-0.03	0.07	-
Sig. (1-tailed)	Academy Score	-				
	Comm. Eval.	0.279	-			
	Gender	0.226	0.383	-		
	Race/Ethnicity	0.171	0.490	0.385	-	
	COPE Score	0.375	0.004	0.347	0.181	-

Test of the Assumptions

Testing assumptions of regression before interpreting the output is a component of validity (Field, 2013). A test of assumptions in this study included normal distribution of the outcome (criterion) variable with no outliers, a linear relationship between the predictor and outcome variables, homoscedasticity, and absence of multicollinearity.

Normal Distribution

The curve imposed on a histogram by SPSS can be interpreted for normal distribution (Warner, 2013). Figure 1 includes a normal distribution of COPE Scores for the amended sample of COPE Candidates (*N* = 2,499). Figure 2 includes a normal distribution of COPE Scores for Academy Recruits (*N* = 159). Figure 3 includes a normal distribution for Academy Score for Academy Recruits (*N* = 159).

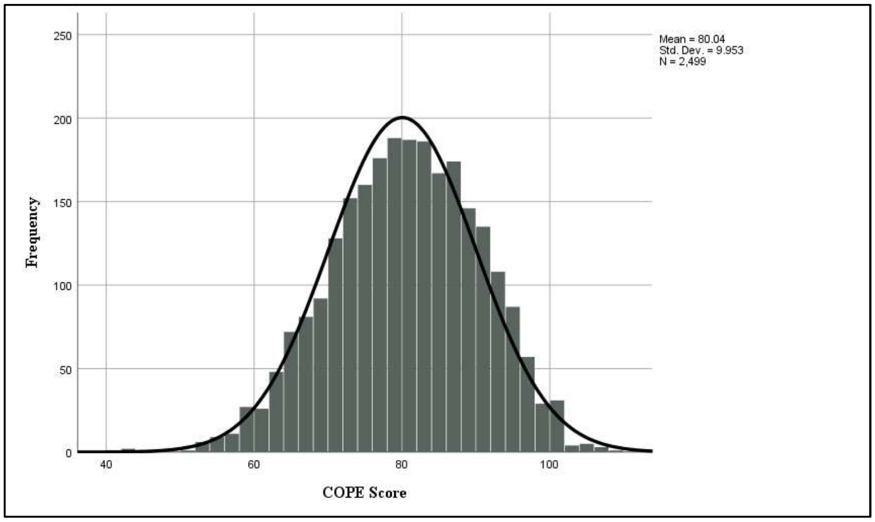

Figure 1. Test of normal distribution for COPE score.

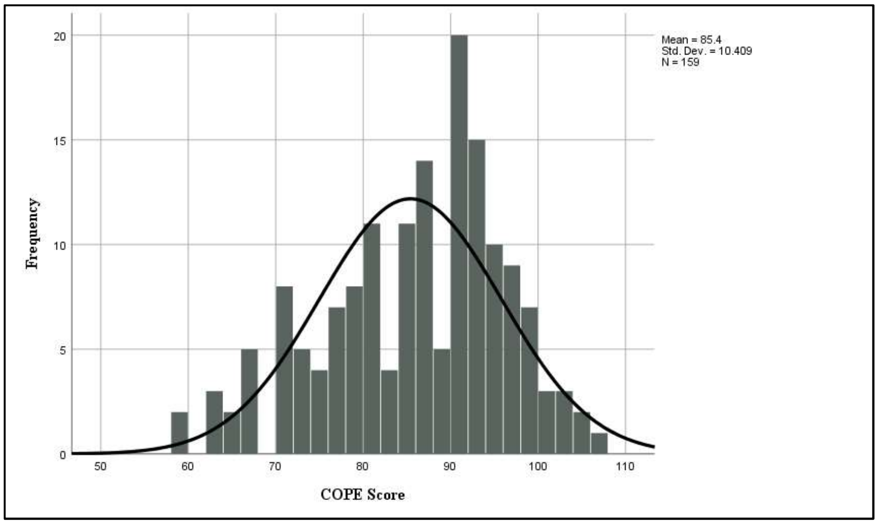

Figure 2. Test of normal distribution for COPE score for Academy recruits.

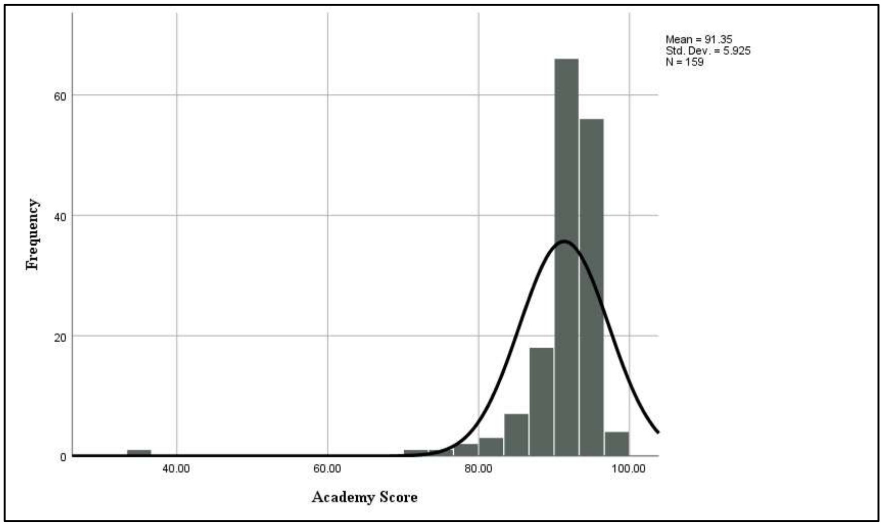

Figure 3. Test of normal distribution of Academy score for Academy recruits.

Linearity

A linear relationship should exist between the dependent and independent variables, which is discernable through the evaluation of a regression plot or a fit line (Warner, 2013). A linear relationship was examined using a Normal Probability – Probability Plot (Normal P-Plot of Regression) for Research Questions 1 and 2. The equation line in Figure 4 shows evidence of a linear relationship with a skew value between the predictor variables (rating method, gender, and race/ethnicity) with the criterion variable (COPE Score) that is close to zero.

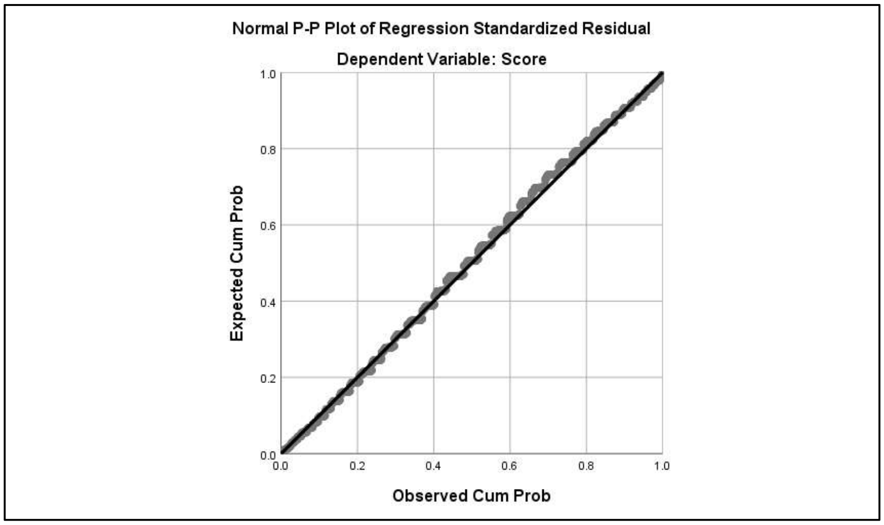

Figure 4. Normal P-Plot of regression for COPE score.

The equation line in Figure 5 shows evidence of a linear relationship between the predictor variables (rating method, gender, race/ethnicity, and COPE Score) with the criterion variable (Academy Score). However, there is evidence of skewness in the Normal P-Plot of Regression for Academy Score.

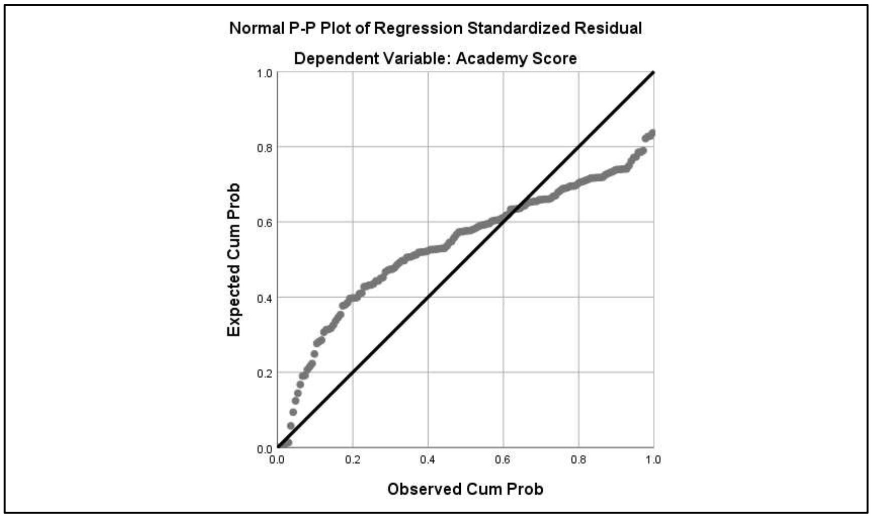

Figure 5. Normal P-Plot of regression for Academy score.

Homoscedasticity

Homoscedasticity exists when residuals are equally distributed along a regression line (Warner, 2013). A scatter plot provides a method for determining if linear relationship exists and whether regression is an appropriate method of analysis. A scatter plot was created using the predictor variables (rating method, gender, and race/ethnicity) with the criterion variable (COPE Score). A discernable pattern is not obvious in Figure 6 and the range does not exceed -3 and 3; therefore, the data meets the assumption of homoscedasticity.

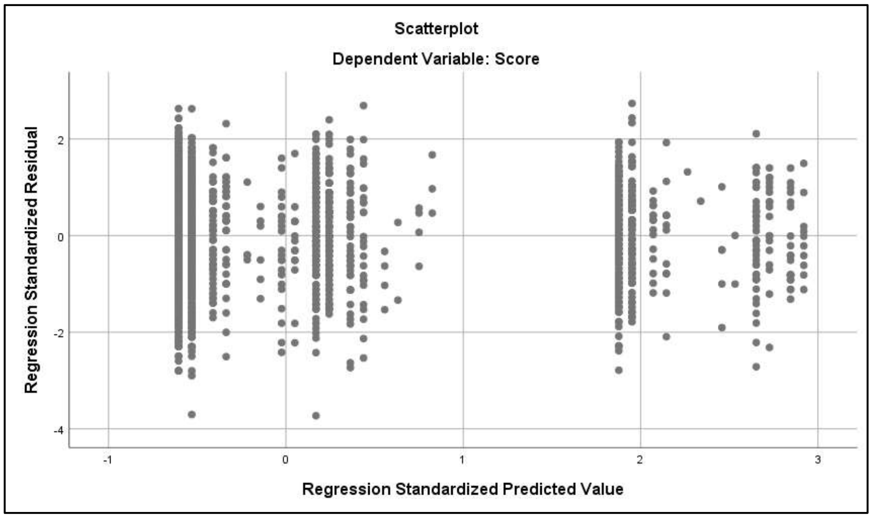

Figure 6. Scatter plot of regression for COPE score.

A scatter plot was created using the predictor variables (rating method, gender, race/ethnicity, and COPE Score) with the criterion variable (Academy Score). A discernable pattern is apparent in Figure 7 because the points are not equally distributed above and below zero on the X axis, or to the left and right of 0 on the Y axis. Therefore, the data does not meet the assumption of homoscedasticity.

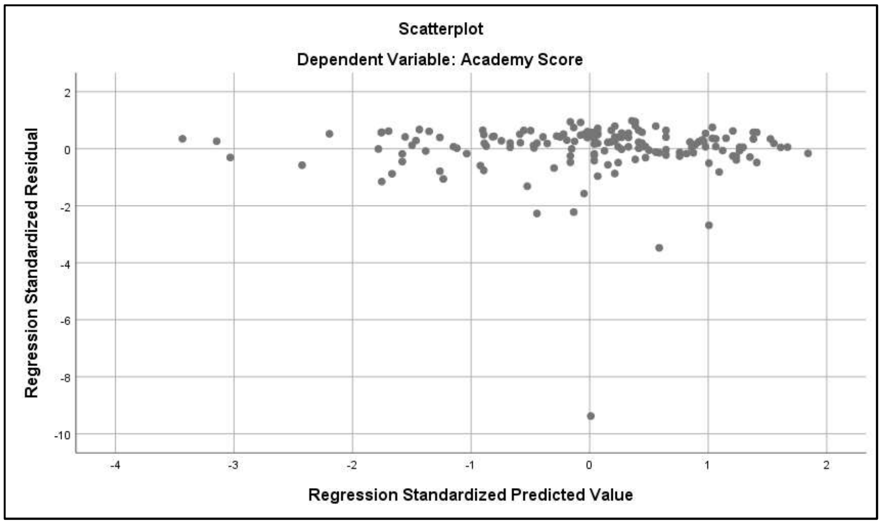

Figure 7. Scatter plot of regression for Academy score.

Multicollinearity

Multicollinearity is when predictor variables are highly correlated (Warner, 2013). The variance inflation factor (VIF) indicates that values close to, or above 10, indicate high levels of multicollinearity (Warner, 2013). Because the VIF indicators in Table 15 for the predictor variables of rating method, gender, and race/ethnicity and the criterion variable COPE Score are greater than 1 and less than 5, the assumption can be made that the variables are moderately correlated.

Table 15

Collinearity Statistics for COPE Score

Model		VIF
1	(Constant)	
	Community Evaluator	1.001
	Gender	1.004
	Race/Ethnicity	1.004

Note. VIF = variable inflation factor.
a. Dependent Variable: Score

The VIF indicators in Table 16 for the predictor variables of rating method, gender, race/ethnicity, and COPE Score on the criterion variable Academy Score are greater than 1 and less than 5. Therefore, the assumption can be made that the variables are moderately correlated.

Table 16

Collinearity Statistics for Academy Score

Model		VIF
1	(Constant)	
	Score	1.053
	Community Evaluator	1.047
	Gender Recoded	1.003
	Majority Minority	1.006

Note. a. Dependent Variable: Academy Score

Statistical Analysis

This section includes the statistical findings relative to research question. Each research question and hypothesis are presented in alignment with the methods presented in Chapter 3. Additional analysis is included based upon the initial findings for each research question.

Research Question 1

Does evaluation method type and/or candidate demographic characteristics predict the score on the Columbus Civil Service Commission constructed response multimedia test for candidates between 2015–2017?

H_{01}: Evaluation method type (community evaluator presence or absence) and/or candidate demographic characteristics (gender and race/ethnicity) do not significantly predict the score on the Columbus Civil Service Commission constructed response multimedia test for candidates between 2015–2017.

H_{a1}: Evaluation method type (community evaluator presence or absence) and/or candidate demographic characteristics (gender and race/ethnicity) significantly predict the score on the Columbus Civil Service Commission constructed response multimedia test for candidates between 2015–2017.

The first data subset was used to assess whether evaluation method and/or demographic differences significantly predict a candidate's score on the Columbus CSC constructed response multimedia test (COPE) using a multiple linear regression analysis. The outcome (criterion) variable was applicant's score on COPE (ordinal). The predictor variables in the equation were coded (a) whether a community evaluator was present or absent (0 = no; 1 = yes), (b) applicant gender (0 = male; 1 = female), and (c) applicant race/ethnicity (0 = majority; 1 = minority). Cases where the race/ethnicity was not identified were placed in the minority group. Eleven cases where gender was not provided were excluded from the analysis.

Table 17 includes the results of the first multiple linear regression analysis to determine whether the presence of a community evaluator and candidate gender were statistically significant predictors of a candidate's COPE score. This multiple regression analysis was not found to be statistically significant ($p > .05$).

Table 17

Results of Multiple Linear Regression Analysis (Rating Method and Gender as Predictor Variables)

Variable	*B*	95% CI	β	*sr*	*p*
(Constant)	79.872	[79.362, 80.382]			
Community Evaluator	0.030	[-0.800, 0.860]	0.001	0.001	0.944
Gender	0.954	[-0.091, 1.999]	0.036	0.036	0.074

Note. CI = confidence intervals for *B; sr* = semipartial correlation.
Dependent Variable: COPE Score.

Table 18 includes the results of the second multiple linear regression analysis to determine whether the presence of a community evaluator and candidate race/ethnicity were statistically significant predictors of a candidate's COPE score. This multiple regression analysis was not found to be statistically significant (*p* > .05).

Table 18

Results of Multiple Linear Regression Analysis (Rating Method and Race/Ethnicity as Predictor Variables)

Variable	*B*	95% CI	β	*sr*	*p*
(Constant)	79.941	[79.420, 80.463]			
Community Evaluator	0.044	[-0.785, 0.873]	0.002	0.002	0.917
Race/Ethnicity	0.290	[-0.615, 1.195]	0.013	0.013	0.530

Note. CI = confidence intervals for *B; sr* = semipartial correlation.
Dependent Variable: COPE Score.

Table 19 includes the results of the third multiple linear regression analysis to determine whether the presence of a community evaluator, candidate gender, and candidate race/ethnicity were statistically significant predictors of an applicant's COPE score. This multiple regression analysis was not found to be statistically significant (*p* > .05).

Table 19

Results of Multiple Linear Regression Analysis (Rating Method, Gender, and Race/Ethnicity as Predictor Variables)

Variable	*B*	95% CI	β	*sr*	*p*
(Constant)	79.823	[79.274, 80.371]			
Community Evaluator	0.023	[-0.808, 0.853]	0.001	0.001	0.958
Gender	0.937	[-.110, 1.985]	0.035	0.035	0.079
Race/Ethnicity	0.223	[-0.688, 1.134]	0.010	0.010	0.631

Note. CI = confidence intervals for *B; sr* = semipartial correlation.
Dependent Variable: COPE Score.

Additional ANOVA tests were run combining Gender and Race/Ethnicity and the presence of a Community Evaluator as well as by each minority group with a population greater than 100, and all results lacked statistical significance (*p* > .05). Additional multiple linear regression analyses were run with interactions (Community Evaluator x Gender; Community Evaluator x Race/Ethnicity) and the significance for all predictor variables in both tests were not found to be statistically significant (*p* > .05). The results of the additional tests are displayed in Appendix E.

Based on the results of the regression and follow-up tests, the first null hypothesis, which stated “Evaluation method type (community evaluator presence or absence) and/or candidate demographic characteristics (gender and race/ethnicity) do not significantly predict the score on the Columbus Civil Service Commission constructed response multimedia test for applicants between 2015–2017,” was not rejected.

Research Question 2

Does evaluation method type, candidate demographic characteristics, and/or score on the Columbus Civil Service Commission constructed response multimedia test predict Academy performance for recruits who were candidates between 2015–2017?

H_{02}: Evaluation method type (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test do not significantly predict Academy performance for recruits who were candidates between 2015–2017.

H_{a2}: Evaluation method type (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test significantly predict Academy performance for recruits who were candidates between 2015–2017.

The second data subset was used to assess whether the type of evaluation method, candidate's demographic characteristics, and/or score on the Columbus CSC constructed response multimedia test (COPE) significantly predicted a recruit's performance in the academy, using a multiple linear regression equation. The predictor variables were coded (a) whether a community evaluator was present or absent (0 = no; 1 = yes), (b) candidate gender (0 = male; 1 = female), (c) candidate race/ethnicity (0 = majority; 1 = minority), and score on COPE (ordinal) was not recoded. The outcome (criterion) variable was the recruit's final performance score in the Columbus Police Academy (ordinal). Cases where race was not identified were placed in the minority group. Two cases where gender was not identified were excluded from the analysis. One candidate's score in the Academy was provided as a zero, which was an outlier in the distribution of scores. Therefore, this candidate was removed from the regression analysis. The distribution of scores for the sample size was roughly normal.

Table 20 includes the results of the first multiple linear regression analysis to determine if the predictor variables (a) whether a community evaluator was present or absent, (b) candidate gender, and (c) COPE score (ordinal) were statistically significant predictors of a recruit's score in the Academy. This multiple regression analysis was not found to be statistically significant (*p* > .05).

Table 20

Results of Multiple Linear Regression Analysis (Rating Method, Gender, and COPE Score as Predictor Variables)

Variable	*B*	95% CI	β	*sr*	*p*
(Constant)	93.174	[85.352, 101.006]			
Community Evaluator	0.714	[-1.333, 2.761]	0.056	0.055	0.492
Gender	-1.053	[-3.710, 1.604]	-0.063	-0.063	0.435
COPE Score	-0.022	[-0.114, 0.070]	-0.039	-0.038	0.633

Note. CI = confidence intervals for *B; sr* = semi partial correlation.
Dependent Variable: Academy Score.

Table 21 includes the results of the second multiple linear regression analysis to determine if the predictor variables (a) whether a community evaluator was present or absent, (b) candidate race/ethnicity, and (c) COPE score (ordinal) were a statistically significant predictor of a recruit's score in the Academy. This multiple regression analysis was not found to be statistically significant (*p* > .05).

Table 21

Results of Multiple Linear Regression Analysis (Rating Method, Race/Ethnicity, and COPE Score as Predictor Variables)

Variable	*B*	95% CI	β	*sr*	*p*
(Constant)	92.822	[85.054, 100.591]			
Community Evaluator	0.832	[-1.185, 2.849]	0.0664	0.065	0.416
Race/Ethnicity	-0.843	[-3.143, 1.458]	-0.0577	-0.06	0.47
COPE Score	-0.018	[-0.110, 0.074]	-0.031	-0.031	0.701

Note. CI = confidence intervals for *B; sr* = semi partial correlation.
Dependent Variable: Academy Score.

Table 22 includes the results of the third multiple linear regression analysis to determine if the predictor variables (a) whether a community evaluator was present or absent, (b) candidate gender, (c) race/ethnicity, and (d) COPE score (ordinal) were a statistically significant predictor of a recruit's score in the Academy. This multiple regression analysis was not found to be statistically significant (*p* > .05).

Table 22

Results of Multiple Linear Regression Analysis (Rating Method, Gender, Race/Ethnicity, and COPE Score as Predictor Variables)

Variable	*B*	95% CI	β	*sr*	*p*
(Constant)	93.115	[85.276, 100.953]			
Community Evaluator	0.700	[-1.348, 2.749]	0.055	0.054	0.500
Gender	-1.021	[-3.681, 1.638]	-0.061	-0.061	0.449
Race/Ethnicity	-1.076	[-3.442, 1.290]	-0.072	-0.072	0.370
COPE Score	-0.019	[-0.112, 0.073]	-0.034	-0.033	0.683

Note. CI = confidence intervals for *B; B* = Estimated values of raw (unstandardized) regression coefficients; CI = confidence interval for odds ratio; *sr* = semipartial correlation. Dependent Variable: Academy Score.

Additional regression equations were run with interactions (Community Evaluator x Gender x COPE Score; Community Evaluator x Race/Ethnicity x COPE Score). The results for both of these multiple regression analyses were not found to be statistically significant (*p* > .05). The results of these tests are included in Appendix F.

Because of the lack of statistical significance in the regression tests for Research Question 2, two additional tests were conducted. The first included revising Academy scores to weighted averages for recruits who did not graduate. The Academy provided a final score (Academy Score) for all recruits based on an average of their completed weeks and did not zero-fill for the week(s) after the recruit left the Academy. For example, one recruit had a reported score of 87.20, which was changed to 43.60 when weighing the averages. Changes to Academy Scores were titled Academy Score Revised. This adjustment to the Academy scores did not affect the statistical significance of the predictive validity (*p* = .459) or the full regression equation (*p* = .785). The results of this test are included in Appendix G.

The second test was implementing the recommendations from Goodwin and Leach (2006) to examine the variables for distribution and skewness. This resulted in identifying that both variables were negatively skewed (COPE Score = -0.513, Academy Score = -5.927) and were not normally distributed when using the Shapiro-Wilk Test of Normality (COPE Score *p* = .001, Academy Score *p* < .01) or when using the Kolmogorov-Smirnov Test of Normality (COPE Score *p* < .01, Academy Score *p* < .01). Even after conducting a Log Transformation (Log10) using SPSS, with the max scores for COPE (106) and Academy Score (97.43) for the respective reflections, these transformations did not result in a positive skewness for Log10COPE (-1.309) but did for Log10Academy Score (.098). Transforming the variables did not result in normal distribution for COPE Score or Academy Score. The results of the *ANOVA* test using Log10COPE and Log10AcademyScore did not result in statistical significance (*p* = .055). A linear regression test did not result in predictive validity for the correlation between Log10COPE and Log10AcademyScore (*p* = .338) or statistical significance for Gender (*p* = .066) or Log10Cope (*p* = .50) as predictors of Log10Academy Score. However, Race/Ethnicity (*p* = .019) was a significant predictor of Log10Academy Score. The results of these tests are included in Appendix H.

Based on the results of the regression tests and supplemental analysis, the null hypothesis, which stated "Evaluation method type (community evaluator presence or absence), recruit demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test do not significantly predict Academy performance for recruits who were candidates between 2015–2017" was not rejected.

Research Question 3

Does evaluation method type (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test predict Academy graduation for recruits who were candidates between 2015–2017?

> H_{03}: Evaluation method (community evaluator presence or absence of), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test do not

significantly predict Academy graduation for recruits who were candidates between 2015–2017.

H_{a3}: Evaluation method (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test significantly predict Academy graduation for recruits who were candidates between 2015–2017.

To assess whether the type of evaluation method, candidate's demographic characteristics, and/or score on the Columbus CSC constructed response multimedia test (COPE) significantly predict Academy graduation, a binary logistic regression test was performed using the second dataset. The predictor variables were (a) evaluation method (community evaluator presence or absence); (b) candidate gender (male or female); (c) race/ethnicity (majority or minority); and (d) COPE score (ordinal). The outcome (criterion) variable was recruit graduation from the Academy. The predictor variables were coded (a) whether a community evaluator was present or absent (0 = no; 1 = yes), (b) applicant gender (0 = male; 1 = female), (c) applicant race/ethnicity (0 = majority; 1 = minority), and score on COPE (ordinal) was not recoded. The outcome (criterion) variable was the recruit's graduation status (did not graduate = 0, graduate = 1).

Table 23 includes the results of the first binary logistic regression analysis using the predictors (a) community evaluator, (b) gender, and (c) COPE Score. The outcome (criterion) variable was recruit graduation from the Academy. A test of the full model compared with a constant-only or null model was not statistically significant ($p > .05$).

Table 23

Results of Binary Logistic Regression Analysis (Rating Method, Gender, and COPE Score as Predictor Variables for Academy Graduation)

							95% C.I.	
	B	S.E.	Wald	*df*	Sig.	Exp(B)	Lower	Upper
Constant	3.967	2.841	1.949	1	0.163	52.810		
Community Evaluator	-0.293	0.694	0.179	1	0.673	0.746	0.191	2.906
Gender	-1.015	0.733	1.917	1	0.166	0.362	0.086	1.525
COPE Score	-0.011	0.033	0.112	1	0.738	0.989	0.927	1.055

Note. *B* = Estimated values of raw (unstandardized) regression coefficients; SE = Standard Error; Wald = Wald Statistic; *df* = degrees of freedom; Sig. = Significance (*probability* value); Exp(B) = odds ratio; CI = confidence interval for odds ratio.

Table 24 includes the results of the second binary logistic regression analysis. The predictor variables were (a) evaluation method (community evaluator presence or absence); (b) race/ethnicity (majority or minority); and (c) COPE score (ordinal). The outcome (criterion) variable was recruit graduation from the Academy. A test of the full model compared with a constant-only or null model was not statistically significant (*p* > .05).

Table 24

Results of Binary Logistic Regression Analysis (Rating Method, Race/Ethnicity, and COPE Score as Predictor Variables for Academy Graduation)

							95% C.I.	
	B	S.E.	Wald	*df*	Sig.	Exp(B)	Lower	Upper
Constant	3.805	2.948	1.666	1	0.197	44.943		
Community Evaluator	-0.302	0.682	0.196	1	0.658	0.739	0.194	2.813
Race/Ethnicity	0.935	1.079	0.751	1	0.386	2.547	0.307	21.103
COPE Score	-0.013	0.034	0.144	1	0.704	0.987	0.923	1.056

Note. B = Estimated values of raw (unstandardized) regression coefficients; SE = standard Error; Wald = Wald Statistic; *df* = degrees of freedom; Sig. = Significance (*probability* value); Exp(B) = odds ratio; CI = confidence interval for odds ratio.

Table 25 includes the results of the third binary logistic regression analysis. The predictor variables were (a) evaluation method (community evaluator presence or absence); (b) candidate gender (male or female); (c) race/ethnicity (majority or minority); and (d) COPE score (ordinal). The outcome (criterion) variable was candidate graduation from the Academy. A test of the full model compared with a constant-only or null model was not statistically significant (*p* > .05).

Table 25

Results of Binary Logistic Regression Analysis (Rating Method, Gender, Race/Ethnicity, and COPE Score as Predictor Variables for Academy Graduation)

							95% C.I.	
	B	S.E.	Wald	*df*	Sig.	Exp(B)	Lower	Upper
Constant	4.019	2.910	1.908	1	0.167	55.655		
Community Evaluator	-0.252	0.696	0.131	1	0.717	0.777	0.199	3.041
Gender	-1.020	0.739	1.908	1	0.167	0.360	0.085	1.533
Race/Ethnicity	0.855	1.081	0.625	1	0.429	2.351	0.282	19.574
COPE Score	-0.013	0.034	0.153	1	0.695	0.987	0.923	1.055

Note. B = Estimated values of raw (unstandardized) regression coefficients; *SE* = Standard Error; *Wald* = Wald Statistic; *df* = degrees of freedom; *Sig.* = Significance (*probability* value); *Exp(B) = odds ratio;* CI = confidence interval for odds ratio.

Therefore, the third null hypothesis, which states, “Evaluation method (community evaluator presence or absence of), recruit demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test do not significantly predict Academy graduation for recruits who were candidates between 2015–2017,” was not rejected.

Summary and Transition

The data collected and analyzed for this study provides insight into the entry-level police officer assessment and selection process in Columbus, Ohio for the purpose of determining whether the introduction of a community evaluator effected selection outcomes. When community evaluators were part of the assessment process, rater agreement was slightly lower in 2017 than in 2015–2016. However, the difference was not statistically significant. Subgroup differences in the assessment and scoring process did not indicate adverse impact for 2015–2017. Although the frequency and descriptive statistics used to examine the data demonstrated a normal distribution of assessment and Academy scores for gender and race/ethnicity groups for 2015–2017, nonparametric tests resulted in evidence that both the COPE and Academy scores were not distributed normally. In addition, evidence of skewness in Academy scores could have contributed to the lack of statistical significance for Research Question 2.

In addition to a measurement of subgroup differences, an examination of the validity of the instrument was also conducted. Multiple regression was used to assess whether rating method, gender, and/or race/ethnicity, were predictors of candidate performance on COPE. The results of the regression tests did not provide evidence to suggest that the presence or absence of a community evaluator, candidate gender and/or race/ethnicity, did not predict an applicant’s score on the Columbus CSC constructed response multimedia test.

Multiple regression was used to assess whether rating method, gender, and/or race/ethnicity, and COPE Score were predictors of recruit performance in the Academy. Binary logistic regression was used to assess whether rating method, gender, and/or race/ethnicity, and COPE Score were predictors of recruit graduation from the Academy. There was no evidence to suggest that collectively, rating method, gender, race/ethnicity, and COPE Score predicted performance in, or graduation from, the Columbus Police Academy.

Chapter 5 will include an analysis of how these findings contribute to the literature on entry-level police officer selection through comparison with peer-reviewed literature. An interpretation of the findings in relationship to the theoretical framework of MRT and GMA will also be addressed. Additionally, limitations, trustworthiness, validity, and reliability will be described. Recommendations for future research in entry-level police officer selection and the integration of community evaluators into the process will be outlined, as well as implications for social change.

Chapter 5: Discussion, Conclusions, and Recommendations

Introduction

I designed this quantitative study to determine the effect of integrating community evaluators as an adjunct to the assessment and selection process for entry-level police officers. Although community members have been engaged as evaluators in the police officer assessment and selection process (Simmons, 2012), my attention was focused on evaluating the effectiveness of this approach. This required an examination of (a) the theoretical framework for this study; (b) personnel assessment and selection methods for entry-level police officers; (c) assessment and selection outcomes associated with the video-based constructed response multimedia test; (d) the effect of community involvement in the selection of police personnel prior to designing and developing the research methods; (e) research methods that align with measuring the effect of a rating method and predictive validity; (f) data collection; and (g) data analysis.

Quantitative research methods were used to analyze the data provided by the CSC and Academy to determine if rating method (absence or presence of a community evaluator), gender, and race/ethnicity effected (a) rater agreement and (b) candidate performance on assessment and selection outcomes associated with the video-based constructed response multimedia test (COPE). Quantitative methods were also used to determine if rating method, gender, race/ethnicity, and COPE score predicted performance in, and graduation from, the Academy. This chapter includes my interpretations of the findings, discussion of limitations encountered, recommendations for research, and implications for social change resulting from this study.

Summary of Key Findings

Following approval from the Walden IRB, I collected data from the City of Columbus CSC and Columbus Police Academy by submitting a public records request with each unit (see Columbus, 2019e). I obtained a total of 2,510 valid records from the CSC that were used to assess rater agreement, subgroup differences, and answer Research Question 1. I obtained a total of 162 valid records from the Academy that were used to answer Research Questions 2 and 3, which also required using the dataset from the CSC. The target populations for this study were the candidates and raters who participated in the entry-level police officer assessment and selection process from 2015–2017 and the recruits who participated in the Academy from 2015–2018. Assessing the gender and race/ethnicity for test candidates and Academy recruits was required for measuring subgroup differences. Therefore, only participants who identified as male or female were assigned to the gender group. The largest number of participants in both populations identified as White (1,892 candidates and 128 recruits) and were assigned to the majority group. The smaller groups that included Two or More Races, American Indian or Alaskan Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Pacific Islander, Prefer Not to Answer, and Missing/Blank (618 candidates and 34 recruits) were assigned to the minority group.

I conducted the data analysis for this study using SPSS version 25 (SPSS, 2018). Analysis included Kendall's *W* to measure rater agreement; *t* tests, effect sizes, and ANOVAs to measure subgroup differences; linear regression to test predictor and criterion variables (measured by performance); and binary logistic regression to test the predictors and criterion variable when only two possible outcomes existed (graduated or did not graduate). In the next section, I will discuss the results of these statistical tests in relation to the literature on entry-level police officer assessment and selection as well as the research questions and hypotheses for this study.

Interpretation of Findings

This study is expected to provide insights into the effect of evaluation methods on selection outcomes and effectiveness of an entry-level police officer assessment. The design of this study is consistent with the methodology in recent studies on the reliability and validity of entry-level police officer assessment and selection devices and outcomes by using the demographic characteristics, assessment scores, and performance in a police academy as variables (see Corey et al., 1996; Cucina et al., 2015; Doerner & Nowell, 1999; Lievens, 2015; Park, 2013).

Review of the Assessment

The constructed response multimedia test is a standard method of assessment in a multiple hurdle selection process for entry-level police officers (Aamodt, 2004; Corey et al., 1996; Cucina et al., 2015; DeCicco, 1999; De Soete et al., 2013; Doerner & Nowell, 1999). In this study, I reviewed the development of the assessment by evaluating the *2012 Entry Level Police Officer COPE Development Report* (City of Columbus CSC, 2012). The report demonstrated alignment with the *Principles for the Validation and Use of Personnel Selection Procedures 5th Edition* (Society of Industrial and Organizational Psychology, 2018). The high-fidelity method portrayed entry-level police officer environmental working conditions, visual and verbal clues, and the emotion of a situation to the candidate (Christian et al., 2010; Cucina et al., 2015; Tuzinski, 2013) that did not require reading or envisioning a scenario (Cucina et al., 2015). The situations presented in COPE were designed to provide a realistic job preview while contributing to a fair assessment of a candidate's competencies (Sinden et al., 2013; Society of Industrial and Organizational Psychology, 2018).

COPE was administered using an assessment center approach where candidates participated in multiple exercises that did not require knowledge or training in police officer policies and procedures (City of Columbus CSC, 2012). The constructs measured by raters using BARS were problem-solving skills and effective behavioral and communication responses, and the BARS did not include consideration for specific knowledge of police officer policies and procedures. A full day of rater training was administered by the Columbus CSC to all raters on the business day before assessment scoring started. The Columbus CSC applied *z* scores by board and grouped the transformed assessment scores using banding, which is an appropriate approach to mitigate the risk of adverse impact (City of Columbus CSC, 2012; Murphy & Myors, 1995; Schmidt & Hunter, 1995).

Rater Agreement

Evaluating interrater agreement specific to a panel of evaluators contributes to understanding the level of agreement among the raters (Cook, 2016). Some studies have included reports that individuals who have experience in a position may have different interpretations, expectations, and perceptions of job requirements (Conley & Sacket, 1987; Sacket & Laczo, 2003). According to the public safety test team manager at the

Columbus CSC, experience in law enforcement was not a requirement for community evaluators. Because of the risk associated with variability among raters, performing a statistical analysis of the raw scores was an appropriate method to measure the reliability of the evaluations (Fleiss, 1971; Shrout & Fleiss, 1979). Interrater agreement can also be referenced when identifying rater qualifications, bias, and comprehension of the rating method (Dierdoff & Wilson, 2003; Morgeson, Delaney-Klinger, Mayfield, Ferrara, & Campion, 2004).

In all 3 years that COPE was administered, the rater agreement was statistically significant ($W = .950$ to $.969$, $p < .01$). A range of 95% to 96.9% of variability that is statistically significant ($p < .01$) demonstrates strong agreement among the raters (see Lund Research, 2019a). Although the level of agreement among raters in 2017, when community evaluators participated in the rating panels, was slightly lower than 2015–2016, the difference was not statistically significant and demonstrated strong agreement. This finding reinforces previous research that collectively the raters were trained appropriately, qualified, and demonstrated comprehension of the BARS associated with scoring candidates on the COPE (see Aamodt, 2004; Corey et al., 1996; Cucina et al., 2015; DeCicco, 1999; De Soete et al., 2013).

Subgroup Differences

One indicator of adverse or disparate impact used in litigation is the identification of subgroup differences in an assessment and selection process (De Soete et al., 2013; Highhouse et al., 2016; Wolgast et al., 2017). Subgroup differences, which are the mean differences between groups regarding psychological constructs (Arthur et al., 2013), is not the same as adverse impact, which is the effect of a decision or rule (Arthur et al., 2013; Lindsey et al., 2013). Selection strategies available to address the differentiation between subgroup differences and adverse impact are the design and scoring of an assessment (Arthur et al., 2013). In addition to the examination of subgroup differences for the purpose of indicators of adverse impact, this evaluation was also a response to the suggestion from De Soete et al. (2013) to replicate the methods of their study in a more diverse population.

In this study, group differences based on the rating method were measured by community evaluator (present/not present) using a *t* test and Cohen's *d* to estimate the effect size (see Arthur et al., 2002; Cucina et al., 2015; DeSoete et al., 2014; Field, 2013). The results of the pooled variances version of the *t* test indicated that the mean scores of each group were not statistically significant (*p*-value > .05) and the *d*-value (-0.005) is unlikely to yield adverse impact. Subgroup differences were measured by gender (male/female *N* = 2,499) and race/ethnicity (majority/minority *N* = 2,510) using a *t* test and Cohen's *d* to estimate the effect size (see Arthur et al., 2002; Cucina et al., 2015; DeSoete et al., 2014; Field, 2013) for each group. The results of the pooled variances version of the *t* test indicated that the mean scores of each group were not statistically significant (*p*-value > .05) and the *d*-values ranging from -0.0291 to -0.0965 are unlikely to yield adverse impact. Thus, the findings related to subgroup differences resulting from the administration of COPE in 2015–2017 did not indicate a statistically significant difference in group differences or indicators of adverse impact.

These were the three research questions and hypotheses for this study:

Research Question 1: Does evaluation method type and/or candidate demographic characteristics predict the score on the Columbus Civil Service Commission constructed response multimedia test for candidates between 2015–2017?

$H_0$1: Evaluation method type (community evaluator presence or absence) and/or candidate demographic characteristics (gender and race/ethnicity) do not significantly predict the score on the Columbus Civil Service Commission constructed response multimedia test for candidates between 2015–2017.

H_a1: Evaluation method type (community evaluator presence or absence) and/or candidate demographic characteristics (gender and race/ethnicity) significantly predict the score on the Columbus Civil Service Commission constructed response multimedia test for candidates between 2015–2017.

Research Question 2: Does evaluation method type, candidate demographic characteristics, and/or score on the Columbus Civil Service Commission constructed response multimedia test predict Academy performance for recruits who were candidates between 2015–2017?

$H_0$2: Evaluation method type (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test do not significantly predict Academy performance for recruits who were candidates between 2015–2017.

H_a2: Evaluation method type (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the

Columbus Civil Service Commission constructed response multimedia test significantly predict Academy performance for recruits who were candidates between 2015–2017.

Research Question 3: Does evaluation method type (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test predict Academy graduation for recruits who were candidates between 2015–2017?

$H_0$3: Evaluation method (community evaluator presence or absence of), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test do not significantly predict Academy graduation for recruits who were candidates between 2015–2017.

H_a3: Evaluation method (community evaluator presence or absence), candidate demographic characteristics (gender and race/ethnicity), and/or score on the Columbus Civil Service Commission constructed response multimedia test significantly predict Academy graduation for recruits who were candidates between 2015–2017.

Research Question 1

Using multiple regression analysis, the three independent variables (rating method, gender, and/or race/ethnicity) did not predict performance on COPE. For example, the *p*-values for each of the independent variables were .958 for rating method, .079 for gender, and .631 for race/ethnicity. This aligns with previous research on the fairness of a properly designed and administered constructed response multimedia test for the assessment and selection of entry-level police officers (Aamodt, 2004; Corey et al., 1996; Cucina et al., 2015; DeCicco, 1999; De Soete et al., 2013; Doerner & Nowell, 1999). There were no known studies on the effect of a community evaluator on assessment and selection outcomes. Therefore, the finding that rating method, the absence or presence of a community evaluator, is not a statistically significant predictor of selection outcomes can be interpreted to mean the reliability of this particular assessment was not effected by the integration of community evaluators into the selection process.

Research Question 2

Using multiple regression analysis, the four independent variables (rating method, gender, and/or race/ethnicity, and COPE score) failed to predict training performance as measured by a recruit's score in the Academy. For example, the *p*-values for each of the independent variables were .50 for rating method, .449 for gender, .370 for race/ethnicity, and .683 for COPE score. The Thorndike Model (as cited by Wiberg & Sundström, 2009) was used to correct for restriction of range for the validity of COPE score as a predictor of performance in the Academy. The uncorrected value was $r(157) = -.025$, $p = .375$ and the corrected value was $r(157) = -.024$. The corrected *r* value is much lower than those included in Schmidt and Hunter's (1998) summaries for corrected verbal work samples $r(3,159) = .44$ and cognitive ability ($r = .56$).

Research Question 3

Using binary logistic regression analysis, the four independent variables (rating method, gender, and/or race/ethnicity, and COPE score) failed to predict a recruit's graduation from the Academy. For example, the *p*-values for each of the independent variables were .717 for rating method, .167 for gender, .429 for race/ethnicity, and .695 for COPE score.

The lack of statistical significance for Research Question 1 aligned with previous research on this topic. However, the nonsignificant findings for Research Questions 2 and 3 were surprising. The predictive validity of this assessment and selection method, as determined by the analysis for Research Questions 2 and 3, does not align with previous research on constructed response multimedia testing for entry-level police officers, which demonstrated this method as a strong predictor of candidate performance in a police academy (Corey et al., 1996; Cucina et al., 2015; Doerner & Nowell, 1999). One explanation for the contradictory findings is the data for COPE and Academy scores did not meet the requirements for normal distribution based on results of the Shapiro-Wilk Test of Normality (COPE Score p = .001, Academy Score p < .01) or when using the Kolmogorov-Smirnov Test of Normality (COPE Score p < .01, Academy Score p < .01) and Academy scores did not meet the assumption for linearity because of skewness. An attempt to transform COPE and Academy scores using Log Transformation (Log10) to address the concerns with distribution were also unsuccessful. Results of these tests must be considered when evaluating the results of multiple linear regression (Field, 2013; Warner, 2013).

Theoretical Implications

The theoretical framework of MRT (Daft & Lengel, 1986) and GMA (Schmidt & Hunter, 1998, 2004) was used in this study to interpret the findings of the community evaluator on the assessment and selection of entry-level police officers. This framework aligns the consistent process associated with administering a media-rich assessment (Cucina et al., 2015; De Soete et al., 2013; Lievens et al., 2015), structured method of rating (Wolgast et al., 2017), and the predictive validity of similar assessments (Corey et al., 1996; Cucina et al., 2015; Doerner & Noell, 1999). The combination of these two theories as a framework was a unique component of this study.

Media Richness Theory

Researchers have used MRT to explain how different types of organizational communications can influence levels of uncertainty and equivocality (Daft & Lengel, 1986). The previous findings for selection process analysis where high-fidelity assessments were used have shown to contribute to smaller subgroup differences and better job performance than low-fidelity methods (Cucina et al., 2015; Kroll & Zeigler, 2016; Lievens et al., 2015). Lievens et al. (2015) also tested MRT to compare the predictive validity of verbal and written responses for police officer academy cadets using a constructed response multimedia test. The outcomes from the high-fidelity method of COPE included small subgroup differences but did not result in predictive validity based upon recruit performance in, or graduation from, the Academy. Therefore, only the findings for subgroup differences in assessment and selection support MRT and the measures of predictive validity failed to support MRT.

General Mental Ability

GMA was introduced by Spearman in 1904 and is also referred to as intelligence or cognitive ability (Schmidt & Hunter, 1998; 2004). In a meta-analysis of pre-employment methods, Schmidt and Hunter (1998) included summaries for corrected verbal work samples *r*(3,159) = .44 and cognitive ability (*r* = .56) as some of the strongest for predictive validity. Using the attributes of GMA discussed in Chapter 2 of this study, the constructed response multimedia test is expected to demonstrate statistically significant validity as measured by performance in a police academy when the combination of effective problem-solving and interpersonal skills are the constructs being measured (Corey et al., 1996; Cucina et al., 2015; Doerner & Nowell, 1999; Wolgast et al., 2017). Although COPE aligned with the job-related scenarios and requirement to demonstrate problem-solving and interpersonal skills (Columbus, 2019d), the measures of recruit performance in, and graduation from, the Academy failed to support the theory of GMA.

Limitations of the Study

Several limitations of this study were identified and discussed in Chapter 1. The limitations included: (a) a lack of predictive validity evidence for COPE, (b) the limited research on the reliability and agreement of evaluators with various amounts of job-related and rating experience using BARS to assess entry-level police officer candidates, (c) restriction of range because this study only explored one out of four phases of testing that occurred in the second of 10 steps in a multiple hurdle selection process, (d) the potential for confounding variables that influenced attrition throughout the multiple hurdle selection process, (e) the sample size of the study, and (f) the potential for researcher bias because I am employed by the Columbus CSC.

These limitations were addressed and mitigated throughout the research process. In addition, the following limitations could be addressed in future studies by modifying the research design and types of data collected.

Only data from one assessment and selection method were used in this study and the results could be unique to COPE. Although COPE and Academy scores appeared to be normally distributed, nonparametric tests proved otherwise. In addition, the demographic groups were not evenly distributed, and in some cases were very small.

Additional detail from the Academy regarding the data could have resulted in a larger sample. Because the data provided by the Academy did not include all of the full names or candidate identification numbers for recruits and did include recruits who were not part of the multiple hurdle selection process, the original sample of recruits was reduced 43% (from 286 to 162). Although this sample size met the minimum recommendation (*N* = 119) from G*Power (2014), the sample groups within the sample size and lack of variability in the scores could have contributed to the statistical significance of the results (Goodwin & Leach, 2006).

Additional detail from the Academy regarding the data could have resulted in a better understanding of how Academy scores were calculated. The Academy provided a final score (score) for all recruits based on an average of their completed weeks and did not zero-fill for the week(s) after the recruit left the Academy. For example, one recruit had a reported score of 87.20, which was changed to 43.60 when weighing the averages.

In addition to utilizing final Academy scores and graduation as an outcome (criterion) variable for COPE, recruit scores on specific phases of the Academy where interpersonal and problem-solving skills are evaluated could be considered for measuring predictive validity. If the recruit becomes a sworn-officer, performance evaluations and feedback from community members could also be used as criterion variables.

Recommendations for Future Research

The recommendations for future research include, but are not limited to, addressing the limitations I have previously identified. Although research on police officer assessment and selection dates back more than 100 years, there is an ongoing debate about the best methods to use in this process. Many city administrators and police forces are looking for opportunities to build relationships with their community.

The integration of community members into the assessment and selection process for entry-level police officers was suggested by Gould (2017), discussed at length by Simmons (2012), and was used as a new method in Columbus, Ohio (Ferrell, 2017; Rouan, 2017). This study is thought to be the first to include a measurement of the effect and is not intended to be generalized. However, as city administrators and police forces continue to explore methods of building engagement and relationships between the community and the police force, community evaluators may be a viable option. Although statistically significant evidence was not found to demonstrate that the integration of community evaluators as an adjunct to the assessment and selection process for entry-level police officers effected outcomes, additional studies should be conducted to measure this method.

Although this study focused on community evaluators in an entry-level police officer selection process, measurement of evaluator perceptions are necessary to further investigate the effect of this alternative method. Evaluator perceptions could include ideas about the responsibilities of a police officer or the transparency of the assessment and selection process. The measurement of pre- and post-assessment perceptions could contribute to determining whether the community evaluator experience contributed to building relationships with the community.

Only one CSC assessment and selection method was explored in this study, which contributed to the recommendations for additional research on the fairness of the selection process (McLarty & Whitman, 2016), and contributing factors to adverse or disparate impact, in the field of law enforcement (De Soete et al., 2013; Guarjado, 2014; Hilal et al., 2017; Kringen, 2016; Riccucci & Riccardelli, 2015; Riccucci & Sadivar, 2018). The inspection of selection procedures and alternative devices is one of the reasons for suggestions to include diverse members of the community in the assessment and selection process of police officers. Researchers have a responsibility to continue this investigation into assessment and selection methods, especially for law enforcement positions.

Implications for Social Change

The significance of this study was based on providing insights into the effect of evaluation methods on selection outcomes and the effectiveness of an entry-level police officer assessment. The findings from this study could benefit the City of Columbus and other police officer selection committees when identifying the best assessment and rating method for mitigating the risk of adverse impact. Selecting the most qualified candidates, while mitigating the risk of adverse or disparate impact, provides equal access to all applicants in the selection process and can reduce the costs and experiences resulting from poor performance, or attrition, in a police academy.

The findings from this study support positive social change by identifying that integrating the community into a structured assessment process did not have an impact on selection outcomes as measured by gender, race/ethnicity, or performance. This method could enable public safety departments to build relationships with the community by inviting members to participate in the assessment and selection process. Other potential social change may include the effect that the integration of community members could have on applicant and community perceptions of the assessment and selection process for entry-level police officers.

This study may have contributed to social change by taking the first approach to measuring the effect of the community evaluator on the assessment and selection of entry-level police officers. Researchers and practitioners can use this information when evaluating assessment and selection methods for people who interact with the public. Because of the high reliability identified in the agreement between raters, and the low subgroup differences associated with the evaluation method, the techniques utilized by the Columbus CSC could be useful to other city administrators and police forces who are considering this method of evaluation.

Conclusion

The importance of creating a valid selection process and improving engagement between law enforcement agencies and the community are recurring themes in the literature. The integration of community members as raters in an assessment and selection process is considered to be a method of improving relationships between a public safety division and the public. The goal of this study was to determine whether selection outcomes were influenced by the introduction of community evaluators into one phase of an assessment and selection process based on measurements of rater agreement, adverse impact indicators, and psychometric adequacy.

This study is believed to be the first to measure the effect of community participation in an entry-level police officer assessment and selection process. In this study, there was no evidence to suggest that integrating community evaluators into the assessment and selection process for entry-level police officers affected rater agreement or subgroup differences in selection outcomes. There was no evidence to suggest that candidate demographics were predictors of performance on the constructed response multimedia test, regardless of whether or not a community evaluator was present on a rating panel. Additionally, there was no evidence to suggest that the presence or absence of a community evaluator, candidate demographics, and score on the constructed response multimedia test predicted performance in, or graduation from, a police academy. The findings reported in this study were compared to empirical research and the similarities and differences were discussed.

Although the introduction of community evaluators as raters in a structured assessment test did not affect selection outcomes as measured by gender, race/ethnicity, or academy performance, there is significance in the findings. The results of this study can be interpreted to mean the reliability and validity of this structured assessment were not strengthened or weakened by the integration of community evaluators. The potential for social change that could result from this alternative method include increasing the transparency of a selection process, providing a voice for the community, and improving applicant perceptions. Integrating community evaluators when developing or administering structured assessment and selection processes may be a viable option for law enforcement agencies.

References

Aamodt, M. G. (2004). *Research in law enforcement selection*. Boca Raton, FL: Brown Walker Press.

Albrecht, J. F. (2017). *Police brutality, misconduct, and corruption.* https://doi.org/10.1007/978-3-319-64438-7

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Annell, S., Lindfors, P., & Sverke, M. (2015). Police selection - implications during training and early career. *Policing-An International Journal of Police Strategies & Management*, *38*(2), 221-238. https://doi.org/10.1108/PIJPSM-11-2014-0119

Arthur Jr, W., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology, 55*(4), 985-1008. https://doi.org/10.1111/j.1744-6570.2002.tb00138.x

Arthur, W. J., Doverspike, D., Barrett, G. V., & Miguel, R. (2013). Chasing the Title VII holy grail: The pitfalls of guaranteeing adverse impact elimination. *Journal of Business and Psychology, 28*(4), 473-485. https://doi.org/10.1007/s10869-013-9289-6

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*(2), 435-442. https://doi.org/10.1037/0021-9010.93.2.435

Arthur, W., & Woehr, D. (2013). No steps forward, two steps back: The fallacy of trying to “eradicate” adverse impact? *Industrial and Organizational Psychology, 6*(4), 438-442. https://doi.org/10.1111/iops.12081

Barrett, G. V., Miguel, R. F., & Doverspike, D. (2011). The uniform guidelines: Better the devil you know. *Industrial and Organizational Psychology, 4*(4), 534-536. https://doi.org/10.1111/j.1754-9434.2011.01386.x

Ben-Porath, Y. S., & Tellegen, A. (2011). Review of the Minnesota Multiphasic Personality Inventory–2–Restructured Form. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The twentieth mental measurements yearbook*. Minneapolis: University of Minnesota Press.

Bergman, M. E. (2018). Police shootings and race in the United States: Why the perpetrator predation perspective is essential to I-O Psychology’s role in ending this crisis. *Industrial and Organizational Psychology, 11*(1), 151-157. https://doi.org/10.1017/iop.2017.101

Berry, C. M., Cullen, M. J., & Meyer, J. M. (2014). Racial/ethnic subgroup differences in cognitive ability test range restriction: Implications for differential validity. *Journal of Applied Psychology, 99*(1), 21-37. https://doi.org/10.1037/a0034376

Berry, C. M., Sackett, P. R., & Sund, A. (2013). The role of range restriction and criterion contamination in assessing differential validity by race/ethnicity. *Journal of Business and Psychology, 28*(3), 345-359. https://doi.org/10.1007/s10869-012-9284-3

Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on

Black–White mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology, 66*(1), 91-126. https://doi.org/10.1111/peps.12007

Breaugh, J. A., & Billings, R. S. (1988). The realistic job preview: Five key elements and their importance for research and practice. *Journal of Business and Psychology*, *2*(4), 291-305. https://doi.org/10.1007/BF01013761

Bullock, V., Latham, A., & Aamodt, M. (2018, October). *Current trends in law enforcement research.* Poster session presented at the 44th Annual Conference of the Society for Police and Criminal Psychology, Sarasota, Fl.

Chatterjee, D. (2016). Approaching "Baltimore is burning" from a systems change perspective: Role of I-O psychologists as change agents. *Industrial and Organizational Psychology, 9*(3), 565-572. https://doi.org/10.1017/iop.2016.56

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*(1), 83-117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

City of Columbus, Ohio. (2019a). City of Columbus Charter Review. Retrieved from https://www.columbus.gov/council/Charter-Review/Charter-Review-Commission/

City of Columbus, Ohio. (2019b). City of Columbus Civil Service Commission. Retrieved from https://www.columbus.gov/civilservice/

City of Columbus, Ohio. (2019c). City of Columbus Police Academy. Retrieved from https://www.columbus.gov/copta/

City of Columbus, Ohio. (2019d). Police officer selection process. Retrieved from https://www.columbus.gov/police-officer/selection-process/

City of Columbus, Ohio. (2019e). Public records policy. Retrieved from https://www.columbus.gov/uploadedFiles/Human_Resources/Document_Library/Policies_and_Forms/POLICIES/PO17%20Public%20Records%20Policy.pdf

City of Columbus Civil Service Commission (2012). *2012 Entry level police officer COPE development report*.

Civil Rights Act of 1964, Public Law 88-352 (78 Stat. 241). Retrieved from https://www.eeoc.gov/laws/statutes/titlevii.cfm

Civil Rights Act of 1991, 42 U.S.C. CC 1981, 200e et seq. Retrieved from https://www.eeoc.gov/laws/statutes/cra-1991.cfm

Commission on Accreditation for Law Enforcement Agencies. (2010). Steps in the accreditation process. Retrieved from http://www.calea.org/content/steps-accreditation-process

Cook, M. (2016). *Personnel selection: Adding value through people – a changing picture.* Chichester, UK: Wiley-Blackwell.

Conley, P. R., & Sackett, P. R. (1987). Effects of using high- versus low-performing job incumbents as sources of job-analysis information. *Journal of Applied Psychology, 72*(3), 434-437. https://doi.org/10.1037/0021-9010.72.3.434

Corey, D., MacAlpine, D., Rand, D., Rand, R., Wolf, G. (1996). *B-PAD® Technical Reports, 4th Edition.* Napa, CA: The B-PAD Group, Inc.

Cucina, J. M., Su, C., Busciglio, H. H., Thomas, P. H., & Peyton, S. T. (2015). Video-based Testing: A high-fidelity job simulation that demonstrates reliability, validity, and utility. *International Journal of Selection and Assessment, 23*(3), 197-209. https://doi.org/10.1111/ijsa.12108

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554-571. https://doi.org/10.1287/mnsc.32.5.554

DeCicco, D. (2000). Police officer candidate assessment and selection. *FBI Law Enforcement Bulletin, 12*(69), 1-6. Retrieved from https://leb.fbi.gov/file-repository/archives/dec00leb.pdf

De Soete, B., Lievens, F., Oostrom, J., & Westerveld, L. (2013). Alternative predictors for dealing with the diversity-validity dilemma in personnel selection: The constructed response multimedia test. *International Journal of Selection and Assessment, 21*(3), 239-250. https://doi.org/10.1111/ijsa.12034

Dennis, A. R., & Kinney, S. T. (1998). Testing media richness theory in the new media: The effects of cues, feedback, and task equivocality. *Information systems research, 9*(3), 256-274. https://doi.org/10.1287/isre.9.3.256

Dessler, G. (2011). *Human resource management* (12th ed). Upper Saddle River: NJ: Prentice Hall.

Dierdorff, E. C., & Wilson, M. A. (2003). A meta-analysis of job analysis reliability. *Journal of Applied Psychology, 88*(4), 635-646. https://doi.org/10.1037/0021-9010.88.4.635

Doerner, W. G., & Nowell, T. M. (1999). The reliability of the behavioral-personnel assessment device (B-PAD) in selecting police recruits. *Policing: An International Journal of Police Strategies & Management*, *22*(3), 343-353. https://doi.org/10.1108/13639519910285099

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Ferrell, K. (2017). Community group helps select future Columbus police officers and firefighters [News Story, Columbus Dispatch]. Retrieved from http://bit.ly/2rjWEdE

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed). Thousand Oaks, CA: Sage Publications.

Finch, D. M., Edwards, B. D., & Wallace, J. C. (2009). Multistage selection strategies: Simulating the effects on adverse impact and expected performance for various predictor combinations. *Journal of Applied Psychology*, *94*(2), 318-340. https://doi.org/ 10.1037/a0013775

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382. https://doi.org/10.1037/h0031619

Frankfort-Nachmias, C., & Leon-Guerrero, A. (2015). *Social statistics for a diverse society* (7th ed). Thousand Oaks, CA: Sage Publications.

G*Power (Version 3.1.9.2) [Computer Software]. (2014). G*Power: Statistical power analyses for Windows and Mac. Retrieved from http://www.gpower.hhu.de/en.html

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social & Administrative Pharmacy*, *9*(3), 330-338. https://doi.org/10.1016/j.sapharm.2012.04.004

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of *r. The Journal of Experimental Education, 74*(3), 251-266. https://doi.org/10.3200/JEXE.74.3.249-266

Gosnell, H. F. (1923). Some practical applications of psychology in government. *American Journal of Sociology, 28*(6), 735-743. https://doi.org/10.1086/213551

Gould, R. (2017). Public engagement for police departments: How to build and regain trust. *Public Management, 6,* 26-27.

Guajardo, S. A. (2014). Workforce diversity: Ethnicity and gender diversity and disparity in the New York City Police Department. *Journal of Ethnicity in Criminal Justice*, *12*(2), 93-115. https://doi.org/10.1080/15377938.2013.837851

Gustafson, J. (2013). Diversity in municipal police agencies: A national examination of minority hiring and promotion. *Policing: An International Journal of Police Strategies & Management*, *36*(4), 719-736. https://doi.org/10.1108/PIJPSM-01-2013-0005

Harvey, J. L., Anderson, L. E., Baranowski, L. E., & Morath, R. A. (2007). Job analysis: Gathering job specific information. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 57–95). Mahwah, NJ: Lawrence Erlbaum Associates.

Herndon, J. S. (2016). A force for law and order. *Industrial and Organizational Psychology, 9*(3), 548-550. https://doi.org/10.1017/iop.2016.52

Highhouse, S., Doverspike, D., & Guion, R. M. (2016). *Essentials of personnel assessment & selection.* New York, NY: Routledge.

Hilal, S., Densley, J. A., & Jones, D. S. (2017). A signaling theory of law enforcement hiring. *Policing & Society*, *27*(5), 508-524. https://doi.org/10.1080/10439463.2015.1081388

Hoffman, C. C. (2018). Civil service mandated cutoff scores: Challenges and practice recommendations. *Industrial and Organizational Psychology, 11*(1), 158-172. https://doi.org/10.1017/iop.2017.102

Huffcutt, A. I., Van Iddekinge, C. H., & Roth, P. L. (2011). Understanding applicant behavior in employment interviews: A theoretical model of interviewee performance. *Human Resource Management Review*, *21*(4), 353-367. https://doi.org/10.1016/j.hrmr.2011.05.003

Huffcutt, A. I., & Youngcourt, S. S. (2007). Employment interviews. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 181-199). Mahwah, NJ: Lawrence Erlbaum Associates.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, *15*(4), 446-455. doi:10.1037/1040-3590.15.4.446

Kehoe, J. (2007). Selection strategies. In S. G. Rogelberg (Ed.), *Encyclopedia of* *industrial and organizational psychology* (Vol. 1, pp. 699-703). https://doi.org/10.4135/9781412952651.n268

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kringen, A. L. (2016). Examining the relationship between civil service commissions and municipal police diversity. *Criminal Justice Policy Review*, *27*(5), 480-497. https://doi.org/10.1177/0887403415612252

Kroll, E., & Ziegler, M. (2016). Discrimination in video-based job interviews. International *Journal of Selection and Assessment, 24*(6), 161-171. https://doi.org/10.1111/ijsa.12138

Lawrence, D. S., Christoff, T. E., & Escamilla, J. H. (2017). Predicting procedural justice behavior: Examining communication and personality. *Policing-An International Journal of Police Strategies & Management, 40*(1), 141-154. https://doi.org/10.1108/PIJPSM-07-2016-0107

Leighton, J. (2010). Internal validity. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 620-622). http://dx.doi.org/10.4135/9781412961288.n192

Lievens, F., De Corte, W., & Westerveld, L. (2015). Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. *Journal of Management, 41*(6), 1604-1627. https://doi.org/10.1177/0149206312463941

Lilley, D., & Boba, R. (2008). A comparison of outcomes associated with two key law-enforcement grant programs. *Criminal Justice Policy Review, 19*(4), 438-465. https://doi.org/10.1177/0887403408319886

Lindsey, A., King, E., McCausland, T., Jones, K., & Dunleavy, E. (2013). What we know and don’t: Eradicating employment discrimination 50 years after the civil rights act. *Industrial and Organizational Psychology, 6*(4), 391-413. https://doi.org/10.1111/iops.12075

Lund Research. (2019a). Kendall’s coefficient of concordance, *W* (Interrater reliability). Retrieved from https://statistics.laerd.com/premium/spss/kccir/kendalls-coefficient-of-concordance-in-spss.php

Lund Research. (2019b). Multiple linear regression. Retrieved from https://statistics.laerd.com/premium/sts/prediction-c-twoplus.php

Markus, K. & Lin, C. (2010). Construct validity. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 230-233). http://dx.doi.org/10.4135/9781412961288.n72

McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., & Ahmed, S. M. (2017). Applicant perspectives during selection: A review addressing “So what?,”“What’s new?,” and “Where to next?”. *Journal of Management, 43*(6), 1693-1725. https://doi.org/10.1177/0149206316681846

McDaniel, M. A., Kepes, S., & Banks, G. C. (2011). The Uniform Guidelines are a detriment to the field of personnel selection. *Industrial and Organizational Psychology, 4*(4), 494-514. https://doi.org/10.1111/j.1754-9434.2011.01382.x

McLarty, B. D., & Whitman, D. S. (2016). A dispositional approach to applicant reactions: Examining core self-evaluations, behavioral intentions, and fairness perceptions. *Journal of Business and Psychology, 31*(1), 141-153. https://doi.org/10.1007/s10869-015-9405-x

Meier, A. M., Arentsen, T. J., Pannell, L., & Putman, K. M. (2018). Attrition of police officers as predicted by peer evaluations during academy training. *Policing & Society*, *28*(1), 17–26. https://doi.org/10.1080/10439463.2015.1128904

Mendoza, J. L., Bard, D. E., Mumford, M. D., & Ang, S. C. (2004). Criterion-related validity in multiple hurdle designs: Estimation and bias. *Organizational Research Methods*, *7*(4), 418-441. https://doi.org/10.1177/1094428104268752

Meyer, R. D. (2007). Incremental validity. In S. G. Rogelberg (Ed.), *Encyclopedia of industrial and organizational psychology* (Vol. 1, pp. 341-342). https://doi.org/10.4135/9781412952651.n130

Morgeson, F. P., Delaney-Klinger, K., Mayfield, M. S., Ferrara, P., & Campion, M. A. (2004). Self-presentation processes in job analysis: A field experiment investigating inflation in abilities, tasks, and competencies. *Journal of Applied Psychology, 89*(4), 674–686. https://doi.org/10.1037/0021-9010.89.4.674

Murphy, K. R., & Myors, B. (1995). Evaluating the logical critique of banding. *Human Performance, 8*(3), 191-201. https://doi.org/10.1207/s15327043hup0803_5

National Academy of Science. (2014). *Proposed revisions to the common rule for the protection of human subjects in the behavioral and social sciences.* Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK217976/

National Center for O*NET Development. Police Patrol Officers. *33-3051.01*. Retrieved from https://www.onetonline.org/link/summary/33-3051.01

Norton, A., McCloskey, A., & Hudson, R. A. (2011). Prediction assessments: Using video-based predictions to assess prospective teachers' knowledge of students' mathematical thinking. *Journal of Mathematics Teacher Education, 14*(4), 305-325. https://doi.org/10.1007/s10857-011-9181-0

Omnibus Crime Control and Safe Streets Act. (1968). Public Law 90-351 (82 Stat. 197). Retrieved from https://www.justice.gov/crt/omnibus-crime-control-and-safe-streets-act-1968-42-usc-3789d

Park, J. A. (2013). *Cadet attrition and training performance at the Texas Department of Public Safety* (Order No. 3597641). Available from ProQuest Dissertations & Theses Global. (1459216961).

Perkins, D. D., & Zimmerman, M. A. (1995). Empowerment theory, research, and application. *American Journal of Community Psychology, 23*(5), 569-579. https://doi.org/10.1007/BF02506982

Ployhart, R. E., & MacKenzie, W. J. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology, Vol. 2: Selecting and developing members for the organization* (pp. 237-252). https://doi.org/10.1037/12170-008U.S

Ployhart, R. E., Schmitt, N., & Tippins, N. T. (2017). Solving the supreme problem: 100 years of selection and recruitment at the Journal of Applied Psychology. *Journal of Applied Psychology*, *102*(3), 291-304. https://doi.org/10.1037/apl0000081

Potter, G. (2013). The history of policing in the United States. Retrieved from https://plsonline.eku.edu/sites/plsonline.eku.edu/files/the-history-of-policing-in-us.pdf

President's Task Force on 21st Century Policing. (2015). *Final Report of the President's Task Force on 21st Century Policing*. Retrieved from http://www.theiacp.org/Portals/0/taskforce_finalreport.pdf

Pulakos, E. D. (2007). Performance measurement. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 293-317). Mahwah, NJ: Lawrence Erlbaum Associates.

Pynes, J. E. (2001). The triumph of techniques over purpose revisited: Evaluating police officer selection. *Review of Public Personnel Administration, 3*(21), 219-236. https://doi.org/10.1177/0734371X0102100304

Rayson, M., Holliman, D., & Belyavin, A. (2000). Development of physical selection procedures for the British Army. Phase 2: Relationship between physical performance tests and criterion tasks. *Ergonomics, 43*(1), 73-105. https://doi.org/10.1080/001401300184675

Riccucci, N. M., & Riccardelli, M. (2015). The use of written exams in police and fire departments: Implications for social diversity. *Review of Public Personnel Administration, 35*(4), 352-366. https://doi.org/10.1177/0734371X14540689

Riccucci, N. M., & Saldivar, K. (2014). The status of employment discrimination suits in police and fire departments across the United States. *Review of Public Personnel Administration, 34*(3), 263-288. https://doi.org/10.1177/0734371X12449839

Roberts, R. M., Tarescavage, A. M., Ben-Porath, Y. S., & Roberts, M. D. (2018). Predicting postprobationary job performance of police officers using CPI and

MMPI–2–RF test data obtained during preemployment psychological screening. *Journal of Personality Assessment,* 1-12. https://doi.org/10.1080/00223891.2018.1423990

Roth, P. L., Le, H., Oh, I., 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. https://doi.org/10.1037/a0034377

Rouan, R. (2017). Community gains voice in hiring [News Story, Columbus Dispatch]. Retrieved from https://www.govtech.com/em/disaster/Columbus-Hires-Civilian-Evaluators-to-Help-Pick-fFre-and-Police-Recruits.html

Ruggs, E. N., Hebl, M. R., Rabelo, V. C., Weaver, K. B., Kovacs, J., & Kemp, A. S. (2016). Baltimore is Burning: Can I-O Psychologists help extinguish the flames? *Industrial and Organizational Psychology, 9*(3), 525-547. https://doi.org/10.1017/iop.2016.5

Ryan, A. M., Sacco, J. M., McFarland, L. A., & Kriska, S. D. (2000). Applicant self-selection: Correlates of withdrawal from a multiple hurdle process. *Journal of Applied Psychology,85*(2), 163-179. https://doi.org/10.1037/0021-9010.85.2.163

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*(3), 707-721. https://doi.org/10.1111/j.1744-6570.1997.tb00711.x

Sackett, P. R., & Laczo, R. M. (2003). Job and work analysis. In W. C. Borman, D. R. Ilgen, R. J. Klimoski, & I. B. Weiner (Eds.), *Handbook of psychology: Vol. 12.*

Industrial and organizational psychology (pp. 21-37). Hoboken, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/0471264385.wei1202

Sackett, P. R., & Roth, L. (1991). A Monte Carlo Exanimation of banding and rank order methods of test score use in personnel selection. *Human Performance, 4*(4), 279-295. https://doi.org/10.1207/s15327043hup0404_3

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology, 102*(10), 1435-1447. https://doi.org/10.1037/apl0000236

Schmidt, F. L., & Hunter, J. E. (1995). The fatal internal contradiction in banding: Its statistical rationale is logically inconsistent with its operational procedures. *Human Performance, 8*(3), 203-214. https://doi.org/10.1207/s15327043hup0803_6

Schmidt, F.L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262-274. https://doi.org/10.1037/0033-2909.124.2.262

Schmidt, F. L. & Hunter, J.E. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*(1), 162-173. https://doi.org/10.1037/0022-3514.86.1.162

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing reliability. *Psychological Bulletin, 86*(2), 420-428. https://doi.org/10.1037/0033-

2909.86.2.420

Simmons, K. C. (2012). Stakeholder participation in the selection and recruitment of police: Democracy in action. *St. Louis University School of Law, Pub. L. Rev.*, *32*, 7-32. Retrieved from http://law.slu.edu/sites/default/files/Journals/kami_chavis_simmons_article.pdf

Society for Industrial and Organizational Psychology. (2018). Principles for the validation and use of personnel selection procedures (5th ed). Retrieved from http://www.siop.org/_Principles/principles.pdf

Spence, M. (1978). Job market signaling. *Uncertainty in Economics*, 281-306. https://doi.org/10.1016/B978-0-12-214850-7.50025-5

SPSS (Version 25) [Computer Software]. (2017). IBM Corporation.

Terman, L. M., Otis, A. S., Dickson, V., Hubbard, O. S., Norton, J. K., Howard, L., Flanders, J. K., Cassingham, C. C. (1917). A trial of mental and pedagogical tests in a civil service examination for policemen and firemen. *Journal of Applied Psychology, 1*(1), 17-29. https://doi.org/10.1037/h0073841

Todak, N. (2017). The decision to become a police officer in a legitimacy crisis. *Women & Criminal Justice, 27*(4), 250-270. https://doi.org/10.1080/08974454.2016.1256804

Tuzinski, K. (2013). Simulations for personnel selection: An introduction. In *Simulations for personnel selection* (pp. 1-13). Retrieved from https://www.researchgate.net/profile/Kathy_Tuzinski/publication/294283938_Simulations_for_Personnel_Selection_An_Introduction/links/58c6b75daca272e36dde

9752/Simulations-for-Personnel-Selection-An-Introduction.pdf

Tziner, A., Joanis, C., & Murphy, K. R. (2000). A comparison of three methods of performance appraisal with regard to goal properties, goal perception, and ratee satisfaction. *Group & Organization Management*, *25*(2), 175-190. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.821.456&rep=rep1&type=pdf

Uniform guidelines on employee selection procedure. (1978). 43 FR (August 25, 1978). Retrieved from https://www.gpo.gov/fdsys/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml

U.S. Office of Personnel Management. (2007). *Assessment decision guide*. Retrieved from http://apps.opm.gov/ADT/ContentFiles/AssessmentDecisionGuide071807.pdf

Warner, R. M. (2013). *Applied statistics: From bivariate through multivariate techniques* (2nd ed). Thousand Oaks, CA: Sage Publications.

Wee, S., Newman, D. A., & Joseph, D. L. (2014). More than g: Selection quality and adverse impact implications of considering second-stratum cognitive abilities. *Journal of Applied Psychology, 99*(4), 547-563. https://doi.org/10.1037/a0035183

Weiss, P. A., & Inwald, R. (2018). A brief history of personality assessment in police psychology: 1916–2008. *Journal of Police and Criminal Psychology, 33*(3), 189-200. https://doi.org/10.1007/s11896-018-9272-2

Wherry, R. J. S., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35,* 521-551. https://doi.org/10.1111/j.1744-

6570.1982.tb02208.x

Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research & Evaluation, 14*(5), 1-9. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.584.2121&rep=rep1&type=pdf

Wolgast, S., Backstrom, M., & Bjorklund, F. (2017). Tools for fairness: Increased structure in the selection process reduces discrimination. *Plos One*, *12*(12), 1-15. https://doi.org/10.1371/journal.pone.018951

Appendix A: Multiple Hurdle Selection Process for Police Officers in Columbus, Ohio

Contacting the Recruiting Unit

All information on becoming a officer with the Columbus, Ohio Division of Police is contained athttp://www.columbuspolice.org/default.htm. If you would like to be placed on a mailing list to receive an application, please contact the Civil Service Commission at [redacted] or the Minority Recruiting Unit at [redacted]

Article I. The Selection Process

Article II. *- Step One -*

The first step in the testing process is to file an application with the Columbus Civil Service Commission for the position of Police Officer. The requirements for filing an application are as follows:

1. You must be 20 years of age at the time of application and 21 years at the time of appointment
2. You must possess a valid driver's license and
3. You must have a high school diploma, or GED equivalent.
4. You must be a US Citizen.

Article III. *- Step Two -*

If your application is approved, the Civil Service Commission will notify you of the time and place to report for the three-phase examination. The phases of the examination are:

I. Multiple Choice Examination (Pass/Fail)
II. Writing Sample (Pass/Fail)
III. Oral Exercise (B-Pad)
IV. Physical Capability

The phases of the examination are numbered according to the order in which the exam will be graded. The phases will not be administered in the order they are listed above. Phases I, II, and IV of the examination are administered as "hurdles". Candidates who fail to pass a phase will not receive a score for subsequent phases. A candidate must receive a passing score on all phases to be considered eligible. Passing scores from Phase III will determine the candidate's ranking on the eligible list.

In summary the standards for the Physical Capability will measure the following:

EVENTS	STANDARD
agility run - consists of sprinting and dodging around one-foot obstacles over a 60-yard course	21.0 second maximum time
1 bench press repetition	70% of candidate's body weight
vertical jump	*to be determined
300 meter run	**70 seconds
push-ups	23 repetitions
sit-ups	31 repetitions
1.5 mile run	**17 minutes, 53 seconds

* Will be assessed as part of the police officer exam, however, until a standard is determined for this event, all candidates who take this event will be given a passing grade on this event
**May not be tested as part of the police officer exams administered in 2002

Article IV. ***- Step Three -***
Candidates who score high enough on the Civil Service examination(s) to begin the selection process will be mailed a personal history questionnaire. A successful candidate will then be required to report to Police Headquarters for a pre-interview with a background investigator. Pictures, fingerprints, and waivers will be completed to assist in an extensive background investigation. A polygraph examination and oral interview will be scheduled at this time.

Article V. ***- Step Four -***
The candidate will be given a polygraph examination to verify all the information provided to the Background Investigator.

Article VI. ***- Step Five -***
The Civil Service Commission will review your entire package to made sure there are no violations of the background removal standards for Civil Service employment with the Division of Police

Article VII. ***- Step Six -***
A thorough investigation will be conducted by the background investigator including a visit to the candidate's residence.

- Step Seven -

All of the information compiled by the investigator is sent to the ORAL Review Board for review. The candidate will be required to interview with this Board and questions will be asked in regard to the background investigation.

Article VIII. ***- Step Eight -***

Oral review board recommendations and background investigations will be reviewed by the Police Administrative Subdivision chain of command. Summaries of each candidate will be forwarded to the City of Columbus Safety Director for consideration of a Conditional Letter of appointment.

Article IX. ***- Step Nine -***

This step will include a physical examination to include a cardiovascular stress test and a psychological evaluation to evaluate a candidate's overall fitness. NOTE: Vision requirements state that you must be **correctable to 20/20** and no more than **20/125 BINOCULAR** uncorrected, each eye. The Physical and Psychological must be passed before a final Offer of Employment is given.

Article X. ***- Step Ten -***

Candidates will be notified by letter of an appointment date for the Police Academy.

Appendix B: Minimum Qualifications for Police Officers in Columbus, Ohio

Minimum Qualifications

1. Must have a high school diploma or G.E.D.

2. Must be at least 20 years old to apply.

3. Must possess a valid driver's license.

4. Must be a U.S. citizen (permanent residency is not accepted).

Automatic Disqualifiers

1. Tried or purchased marijuana in the past 12 months.

2. Tried or purchased any other illegal drug(s) in the last 3 years *(EXCEPT Marijuana)*.

3. Been convicted while operating a motor vehicle (OVI, DUI, or OMVI) while under the influence of alcohol or drugs within the last five (5) years.

4. As an adult 18 or older:

   - Been ***convicted*** of a felony offense(s) (*Does not apply to misdemeanors (M1 - M4))*.

   - Verified, admitted or convicted of domestic violence within the last ten (10) years.

   - Intentional violation of any protection order or temporary restraining order within seven (7) years.

   - Non-compliance with court ordered child support, alimony or other financial responsibility within the preceding five (5) years.

   - Received four (4) or more moving violations in the past three (3) years (*Excluding parking tickets or seat belt violations*).

     Note: For a complete list of disqualifiers please read the entire Background Removal Standards for Police Officers and Police Communication Technicians provided by The City of Columbus Civil Service Commission.

Sworn personnel shall have no visible piercing (other than ears) or tattoos on head, neck or hands.

Appendix C: Police Officer Selection Statistics 2014-2016

		Gender		Race/Ethnicity					Diversity Counts	
Phase	*n*	Female	Male	Black	Other	White	Unknown	White Male + Unknown	*n*	Percentage
2014										
Applied	2712	434	2278	554	259	1899		1627	1085	40%
Passed MQ/ABQ	2521	397	2124	483	243	1795		1537	984	39%
Showed MC,WS,COPE	1320	199	1124	236	133	951		823	497	38%
Passed MC	1065	156	909	156	106	803		691	374	35%
Passed WS	923	144	779	112	91	720		610	313	34%
Passed COPE	776	127	649	97	71	608		511	265	34%
Showed Physical Fitness	647	99	545	82	58	507		433	214	33%
Eligible	546	75	471	77	47	422		369	177	32%
90 Band	161	28	133	26	19	116		96	65	40%
Showed PHQ	155	29	126	36	10	109		89	66	43%
Passed Background Standards	117	21	96	26	7	84		68	49	42%
Conditional Offer	86	17	69	19	3	64		52	34	40%
Passed Medical	64	13	51	11	2	51		42	22	34%
Appointed	58	12	46	10	2	46		37	21	36%
2015										
Applied	2859	468	2391	547	323	1989		1715	1144	40%
Passed MQ/ABQ	2761	450	2311	526	311	1924		1653	1108	40%
Showed MC,WS,COPE	1402	236	1166	261	160	981		833	569	41%
Passed MC	1034	171	863	153	103	778		662	372	36%
Passed WS	869	150	719	96	83	690		580	289	33%
Passed COPE	726	129	597	79	69	578		481	245	34%
Showed Physical Fitness	546	95	451	63	51	432		361	185	34%
Eligible	456	73	383	52	41	363		309	147	32%
90 Band	144	22	122	17	13	114		97	47	33%
Showed PHQ	259	47	212	39	7	212	1	176	83	32%
Passed Background Standards	109	18	91	13	6	90		76	33	30%
Conditional Offer	55	10	45	8	5	42		35	20	36%
Passed Medical	40	7	33	4	4	32		28	12	30%
Appointed	34	7	33	4	4	26		23	11	32%
2016										
Applied	2661	445	2216	603	325	1719	14	1495	1166	44%
Passed MQ/ABQ	2559	429	2130	563	313	1669	14	1449	1110	43%
Showed MC,WS,COPE	1231	206	1025	227	162	838	4	721	510	41%
Passed MC	943	157	786	158	109	673	3	575	368	39%
Passed WS	816	150	666	114	92	609	1	510	306	38%
Passed COPE	694	130	564	105	73	515	1	428	266	38%
Showed Physical Fitness	535	100	435	86	56	392	1	324	211	39%
Eligible	426	77	349	69	47	310	0	260	166	39%
Sent PHQ	318	59	260	52	36	230	0	230	123	39%
Showed PHQ	257	46	211	47	25	185	0	185	104	40%
Passed Background Standards	172	31	141	23	21	128	0	128	65	38%
Conditional Offer	107	26	81	12	11	84	0	84	41	38%
Passed Medical	95	20	75	10	7	78	0	95	32	34%
Appointed	92	19	73	10	6	76	0	76	31	34%

Note: MQ/ABQ = Minimum Qualifications/Abbreviated Background Questionnaire; MC = Multiple Choice; WS = Writing Sample; COPE = Columbus Oral Police Exam; PHQ = Personal History Questionnaire.

Appendix D: Police Officer Selection Statistics 2017

	Female									Male									Total			
Phase	AI/AN	Asian	Black	His/Lat	HA/PI	White	2 or More	Unknown	Total	AI/AN	Asian	Black	His/Lat	HA/PI	White	2 or More	Unknown	Total	Total All	Diversity All	Diversity Black	Diversity Female
Applied	4	4	115	24	0	233	37	1	418	14	47	406	96	6	1384	110	14	2077	2495	1097	521	418
Passed MQ/ABQ	4	4	111	24	0	227	35	1	406	14	46	390	93	6	1337	100	13	1999	2405	1055	501	406
Showed MC,WS,COPE	3	2	39	12	0	114	21	1	192	9	27	185	45	3	713	49	7	1038	1230	510	224	192
Passed MC	2	2	28	9	0	92	16	0	149	8	15	113	30	2	568	36	7	779	928	353	141	149
Passed WS	2	2	25	9	0	88	13	0	139	6	14	82	26	2	525	33	7	695	834	302	107	139
Passed COPE	2	1	23	9	0	81	13	0	129	6	13	71	19	2	443	28	5	587	716	268	94	129
Showed Physical Fitness	1	1	19	7	0	50	7	0	85	5	10	59	19	2	335	21	4	455	540	201	78	85
Eligible	1	0	13	4	0	37	6	0	61	4	7	54	17	1	283	19	4	389	450	163	67	61
Sent PHQ	1	0	9	3	0	31	4	0	48	2	5	43	13	2	203	17	4	288	336	130	52	48
Showed PHQ	1	0	8	2	0	26	4	0	41	2	5	32	12	1	174	14	4	244	285	107	40	41
Passed Background Standards	1	0	5	2	0	23	4	0	35	0	4	15	4	1	105	9	2	140	175	68	20	35
Conditional Offer*	1	0	1	1	0	14	2	0	19	0	3	10	3	0	64	7	0	87	106	42	11	19
Passed Medical*	1	0	1	1	0	10	1	0	7	0	3	7	3	0	53	4	0	70	77	24	8	7
Appointed*	1	0	1	1	0	10	1	0	14	0	3	7	2	0	47	4	0	63	77	30	8	14

Note: AI/AN = American Indian/Alaskan Native; His/Lat = Hispanic/Latino; HA/PI = Hawaiian/Pacific Islander; MQ/ABQ = Minimum Qualifications/Abbreviated Background Questionnaire; MC = Multiple Choice; WS = Writing Sample; COPE = Columbus Oral Police Exam; PHQ = Personal History Questionnaire.

Appendix E: Results of Supplemental Tests for Research Question 1

Table E1

ANOVA Results (Community Evaluator and Gender as Predictors for COPE Score)

ANOVA[a]

	Sum of Squares	df	Mean Square	F	Sig.
Regression	317.295	2	158.648	1.602	.202[b]
Residual	247135.293	2496	99.013		
Total	247452.588	2498			

a. Dependent Variable: Score

b. Predictors: (Constant), Gender, Community Evaluator

Table E2

Community Evaluator and Race/Ethnicity as Predictors for COPE Score

ANOVA[a]

	Sum of Squares	df	Mean Square	F	Sig.
Regression	40.792	2	20.396	0.206	.814[b]
Residual	248570.311	2507	99.151		
Total	248611.103	2509			

a. Dependent Variable: Score

b. Predictors: (Constant), Majority/Minority, Community Evaluator

Table E3

Community Evaluator and Two or More Races as Predictors for COPE Score
ANOVA[a]

	Sum of Squares	df	Mean Square	F	Sig.
Regression	56.853	2	28.427	0.288	.750[b]
Residual	198648.035	2011	98.781		
Total	198704.888	2013			

a. Dependent Variable: Score

b. Predictors: (Constant), Two or More Races/Ethnicity, Community Evaluator

Table E4

Community Evaluator and Hispanic as Predictors for COPE Score
ANOVA[a]

	Sum of Squares	df	Mean Square	F	Sig.
Regression	54.594	2	27.297	0.272	.762[b]
Residual	199686.111	1989	100.395		
Total	199740.705	1991			

a. Dependent Variable: Score

b. Predictors: (Constant), Hispanic, Community Evaluator

Table E5

Multiple Regression with Community Evaluator and Gender Interactions

Variable	*B*	95% CI	β	*sr*	*p*
(Constant)	79.979	[79.455, 80.503]			
Community Evaluator	-0.289	[-1.196, 0.617]	-0.014	-0.013	0.531
Gender	0.336	[-0.926, 1.597]	0.013	0.010	0.602
Eval_Gender	1.967	[-0.284, 4.217]	0.044	0.034	0.087

Note. CI = confidence intervals for *B; sr* = semipartial correlation.
Dependent Variable: COPE Score.

Table E6

Multiple Regression with Community Evaluator and Race/Ethnicity Interactions

Variable	*B*	95% CI	β	*sr*	*p*
(Constant)	80.040	[79.495, 80.584]			
Community Evaluator	-0.262	[-1.224, 0.699]	-0.012	-0.011	0.593
Race/Ethnicity	-0.129	[-1.253, 0.996]	-0.006	-0.004	0.823
Eval_Race/Ethnicity	1.190	[-0.705, 3.086]	0.034	0.025	0.218

Note. CI = confidence intervals for *B; sr* = semipartial correlation.
Dependent Variable: COPE Score.

Appendix F: Results of Supplemental Tests for Research Question 2

Table F1

Multiple Regression with Community Evaluator, Gender, and COPE Score Interactions

Variable	*B*	95% CI	β	*sr*	*p*
(Constant)	94.370	[84.373, 104.367]			
Community Evaluator	0.612	[-20.879, 22.102]	0.048	0.005	0.955
Gender	-7.347	[-29.220, 14.525]	-0.438	-0.054	0.508
COPE Score	-0.037	[-0.115, 0.080]	-0.065	-0.051	0.533
Score_CommunityEval	0.004	[-0.241, 0.249]	0.028	0.003	0.974
Score_Gender	0.082	[-0.182, 0.347]	0.420	0.050	0.539
CommunityEval_Gender	-8.377	[-73.252, 56.498]	-0.310	0.016	0.799
Score_CommunityEval_Gender	0.070	[-0.651, 0.7927]	0.237	-0.021	0.848

Note. CI = confidence intervals for *B; sr* = semipartial correlation.
Dependent Variable: Academy Score.

Table F2

Multiple Regression with Community Evaluator, Race, and COPE Score Interactions

Variable	*B*	95% CI	β	*sr*	*p*
(Constant)	92.574	[82.415, 102.733]			
Community Evaluator	3.540	[-18.882, 25.963]	0.282	0.025	0.756
Race/Ethnicity	0.336	[-20.240, 20.913]	0.023	0.003	0.974
COPE Score	-0.013	[-0.133, 0.107]	-0.023	-0.017	0.829
Score_CommunityEval	-0.037	[-0.294, 0.221]	-0.260	-0.023	0.780
Score_Race/Ethnicity	-0.023	[-0.266, 0.219]	-0.141	-0.015	0.850
CommunityEval_Race/Ethnicity	-7.194	[-63.198, 48.809]	-0.321	-0.021	0.800
Score_CommunityEval_Race/Ethnicity	0.105	[-0.509, 0.719]	0.437	0.027	0.736

Note. CI = confidence intervals for *B; sr* = semipartial correlation.
Dependent Variable: Academy Score.

Appendix G: Test Results of Revision to Academy Score for Research Question 2

Table G1

Correlations

(n = 159)		Academy Score Revised	Community Evaluator	Gender	Race/ Ethnicity	COPE Score
Pearson Correlation	Academy Score R	-				
	Community Eval.	-0.005	-			
	Gender	-0.105	0.024	-		
	Race/Ethnicity	0.003	0.002	0.023	-	
	COPE Score	-0.008	0.210	-0.031	0.073	-
Sig. (1-tailed)	Academy Score R	-				
	Community Eval.	0.477	-			
	Gender	0.095	0.383	-		
	Race/Ethnicity	0.485	0.490	0.385	-	
	COPE Score	0.459	0.004	0.347	0.181	-

Table G2

Results of Multiple Linear Regression Analysis (Rating Method, Gender, and COPE Score as Predictor Variables)

Variable	*B*	95% CI	β	*sr*	*p*
(Constant)	91.055	[76.822, 105.289]			
Community Evaluator	0.009	[-3.711, 3.729]	0.000	0.000	0.996
Gender	-3.205	[-8.035, 1.625]	-0.105	-0.105	0.192
Race/Ethnicity	0.174	[-4.122, 4.469]	0.006	0.006	0.936
COPE Score	-0.012	[-0.180, 0.155]	-0.012	-0.012	0.883

Note. CI = confidence intervals for *B; sr* = semipartial correlation.
Dependent Variable: Academy Score.

Appendix H: Results of Additional Tests of Normality and Variable Transformation for Research Question 2

Table H1

Descriptive Statistics for COPE Score and Transformed COPE Score

			Statistic	Std. Error
COPE Score	Mean		85.48	0.818
	95% Confidence Interval for Mean	Lower Bound	83.87	
		Upper Bound	87.10	
	5% Trimmed Mean		85.80	
	Median		87.00	
	Variance		107.739	
	Std. Deviation		10.380	
	Minimum		59	
	Maximum		106	
	Range		47	
	Interquartile Range		15	
	Skewness		-0.513	0.191
	Kurtosis		-0.357	0.380
Log10COPE	Mean		1.2718	0.02008
	95% Confidence Interval for Mean	Lower Bound	1.2322	
		Upper Bound	1.3115	
	5% Trimmed Mean		1.2897	
	Median		1.3010	
	Variance		0.065	
	Std. Deviation		0.25484	
	Minimum		0.00	
	Maximum		1.68	
	Range		1.68	
	Interquartile Range		0.31	
	Skewness		-1.309	0.191
	Kurtosis		3.770	0.380

Table H2

Descriptive Statistics for Academy Score and Transformed Academy Score

			Statistic	Std. Error
Academy Score	Mean		91.3930	0.46567
	95% Confidence Interval for Mean	Lower Bound	90.4733	
		Upper Bound	92.3126	
	5% Trimmed Mean		92.0904	
	Median		92.5000	
	Variance		34.913	
	Std. Deviation		5.90873	
	Minimum		35.43	
	Maximum		97.43	
	Range		62.00	
	Interquartile Range		3.50	
	Skewness		-5.927	0.191
	Kurtosis		50.966	0.380
Log10AcademyScore	Mean		0.7694	0.01969
	95% Confidence Interval for Mean	Lower Bound	0.7305	
		Upper Bound	0.8083	
	5% Trimmed Mean		0.7702	
	Median		0.7731	
	Variance		0.062	
	Std. Deviation		0.24986	
	Minimum		0.00	
	Maximum		1.80	
	Range		1.80	
	Interquartile Range		0.26	
	Skewness		0.098	0.191
	Kurtosis		2.408	0.380

Table H3

Correlations

(n = 159)		Log10 Academy Score	Community Evaluator	Gender	Race/ Ethnicity	Log10 COPE
Pearson Correlation	Log10AcademyScore	-				
	Community Eval.	-0.003	-			
	Gender	0.149	0.024	-		
	Race/Ethnicity	0.185	0.002	0.023	-	
	Log10COPE	0.033	-0.164	-0.007	-0.103	-
Sig. (1-tailed)	Log10AcademyScore	-				
	Community Eval.	0.485	-			
	Gender	0.030	0.383	-		
	Race/Ethnicity	0.010	0.490	0.385	-	
	Log10COPE	0.338	0.019	0.467	0.098	-

Table H4

Results of ANOVA (Rating Method, Gender, and Log10COPE Score as Predictor Variables)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	0.545	4	0.136	2.368	.055[b]
	Residual	8.867	154	0.058		
	Total	9.412	158			

a. Dependent Variable: Log10_AcademyScore

b. Predictors: (Constant), Log10_Score, Gender Recoded, Majority Minority, Community Evaluator

Table H5

Results of Multiple Linear Regression Analysis (Rating Method, Gender, and Log10COPE Score as Predictor Variables)

Variable	*B*	95% CI	β	*sr*	*p*
(Constant)	0.672	[.467, .876]			
Community Evaluator	0.001	[-0.081, 0.083]	0.002	0.002	0.979
Gender	0.100	[-0.007, 0.207]	0.145	0.148	0.066
Race/Ethnicity	0.115	[0.019, 4.469]	0.187	0.188	0.019
Log10COPE	0.051	[-0.099, 0.202]	0.054	0.054	0.500

Note. CI = confidence intervals for *B; sr* = semipartial correlation.
Dependent Variable: Log10Academy Score.