ADVERSE ACTION IN THE SELECTION PROCESS: BEYOND THE UNIFORM GUIDELINES

A Thesis

Presented to the faculty of the Department of Psychology

California State University, Sacramento

Submitted in partial satisfaction of
the requirements for the degree of

MASTER OF ARTS

in

Psychology

(Industrial/Organizational)

by

Michelle Gabrielle Avalos

SPRING
2019

ADVERSE ACTION IN THE SELECTION PROCESS: BEYOND THE UNIFORM GUIDELINES

A Thesis

by

Michelle Gabrielle Avalos

Approved by:

_____, Committee Chair
Dr. Lawrence Meyers

_____, Second Reader
Dr. Jianjian Qin

_____, Third Reader
Dr. Tim Gaffney

_____
Date

Student: Michelle Gabrielle Avalos

I certify that this student has met the requirements for format contained in the University format manual, and that this thesis is suitable for shelving in the Library and credit is to be awarded for the thesis.

_____, Graduate Coordinator _____
Dr. Lisa M. Bohon Date

Department of Psychology

Abstract

of

ADVERSE ACTION IN THE SELECTION PROCESS: BEYOND THE UNIFORM GUIDELINES

by

Michelle Gabrielle Avalos

The goal of this thesis is to supplement the analyses of a previously administered state-level multiple-choice examination using the 80 Percent Rule with Differential Item Functioning (DIF) analyses. DIF analysis addresses group differences when candidates are equated for level of ability (a more direct way to address test fairness) in an attempt to identify potentially biased exam items. By using a hierarchal logistic regression approach, the study evaluates Non-Uniform and Uniform DIF alongside the 80 Percent Rule assessment in an effort to address limitations identified in the 80 percent rule statistic as introduced in Section 4D of the Uniform Guidelines. The archival data utilized in this study were provided by a State Agency located on the West Coast of the United States. Job applicants (N = 1,517) across the state participated in the selection measure for a Cashier-based Classification and 8 demographics were utilized in the assessments (Male, Female, Over 40, Under 40, Caucasian, Hispanic, African-American, and Asian). It was concluded that the use of hierarchal logistic regression analysis in this study provided (1) clarity on violations identified through the 80 Percent Rule assessment and (2) evidence of item bias for several protected groups.

Conclusions and recommendations were also made regarding residual DIF and future studies.

_____, Committee Chair
Dr. Lawrence Meyers

_____
Date

ACKNOWLEDGEMENTS

I would first like to thank Dr. Lawrence Meyers, my advisor and committee chair, who supported me through innumerable edits, meetings, and advice throughout my time as an undergraduate and graduate student. He provided countless resources, opportunities, and pathways for me and for that I will always be truly grateful. I would also like to thank Lisa Tavano-Hall, my early statistics professor at the University. She was the first professor to acknowledge and fuel my passion for statistics and always recalled when I was absent in an auditorium full of students. Although she is no longer with us, her teachings will remain with me for a lifetime.

To my family, friends, and loved ones, I thank you for your support during this period in my life. Whether it was missed events, late arrivals, or early departures to work on my studies, I always felt a sense of understanding and support. Most of all, I'd like to thank my husband, Jorge, and the newest member of our little family, our daughter Emma. Jorge, I couldn't have done this without you. You inspired me to pursue my degree in Industrial Organizational Psychology and supported me every step of the way. Your unwavering patience, help, and love will always be remembered. Emma, my little miracle, you are the reason for everything I do and want to succeed in – thank you for being my inspiration.

TABLE OF CONTENTS

Page

LIST OF TABLES

Tables Page

LIST OF FIGURES

Figures Page

Chapter 1

INTRODUCTION

Historical Introduction

The Keju System is often recognized as the first formal Civil Service examination system, originating in China in 606 A.D. during the Han Dynasty. The purpose of this system was to formalize an appointment process for the hiring of grand councilors in various government departments within China. Established during the T'ang dynasty, Chinese emperors sought to appoint candidates through a series of rigorous exam phases, where passing candidates would obtain titles such as "budding scholar" during each of the progressive stages of the exam (Elman, 1991; Suen & Vu, 2006). Each examination stage spanned from one to nine days, containing content on various forms of knowledge, skills, and abilities deemed necessary for civil service, such as Confucian philosophy, national history, poetry, and written word/essay writing. Candidates who passed each of the phases were not only awarded the eventual position of grand councilor (equivalent to a prime administer), but bestowed financial benefits, power, and fame (Suen, 2006). This examination system lasted over 1,298 years, ending in the year 1905 as one of the most sophisticated testing systems of its time.

In the United States, civil service examinations originated from a series of political reactions that took place from 1789 to 1883. In early U.S. history (the first six presidencies), a spoils system began to develop in the public sector as various political parties and affiliations began to grow (U.S. Office of Personnel Management, 2003). Appointments in the government were seen as a cyclical process; each president would

place as many officers into political power as possible prior to resignation, resulting in turmoil and the reactions of the successor in reversing the order and doing the same. From 1801-1809, Thomas Jefferson was the first to acknowledge the impact of this process and claimed "partisan political considerations," enforcing the removal of unfair appointments based on nepotism and political favor. Some of his most iconic removals were that of John Adam's "midnight appointments," lifetime appointments given to officials during Adams presidency, and various executive appointments (excluding judges determined to be on good behavior; Peterson, 1970).

In an attempt to eliminate the possibility of future "midnight appointments" or irreversible long-term appointments, James Monroe's treasurer, William Crawford, authored the Tenure of Office Act or "The Four Year Bill" in 1820, allowing those in office to only hold a position for a 4-year term, similar to the presidency (Fish, 1905). Although this Act eliminated the issues of lengthy and undeserving tenures for certain officials, it added fuel to the ever expanding spoil system, in a sense encouraging presidential candidates to adjust the government officials in each new term.

From Andrew Jackson's presidency to Chester Arthur's presidency, the spoils system continued to grow and drive political appointments. However, awareness of qualifications and meeting standards for appointment were beginning to be introduced into the government. During Jackson's presidency, meritocracy (i.e., the belief that goods and services should be awarded on the basis of merit) was introduced (Hartigan & Wigdor, 1989). In 1853, nearing the end of Millard Fillmore's presidency, newly passed legislation designated certain compensation salaries, based on appointment, for specific

civil service employees; a first of its kind. The legislation even included an examination process, although the questions were rarely job related (e.g., “What did you have for breakfast this morning?) (U.S. Office of Personnel Management, 2003). Shortly after, during Ulysses Grant’s presidency in 1871, Grant mandated that regulations be enforced for the conduct and receipt of appointments. Grant also established the Advisory Board of Civil Service, now known as the Civil Service Commission, to accomplish this. Although it was eventually dismantled, it remained an important attempt and symbol of a national desire to standardize qualification regulations for civil service appointments.

President Garfield’s death was an event that significantly influenced the future of civil service examinations. Prior to his presidency, James Garfield spoke highly of Grant’s Advisory Board attempt, promising to instill the same principles in his presidency. Charles Guiteau, one of his political affiliates, spoke highly of Garfield and his view points. This aided his campaign and bolstered his slim standing in the polls that led to Garfield’s eventual election (Editor, 2015). Assuming Garfield would honor Grant’s principles and Guiteau’s work with an office appointment, Guiteau met with the president several times, only to be disregarded and eventually dismissed from the White House completely. Garfield’s disregard of the spoils system infuriated Guiteau which drove him to madness and the eventual assassination of the President.

The nation reacted strongly to President Garfield’s death, acknowledging the severity and distortion that the spoils system brought to White House. In 1883 the Pendleton Civil Service Reform Act was approved, creating a regulation that all federal job qualifications be based on merit, an idea modeled after British and Weberian Prussian

systems (McGrath, 2013). In the same year the Civil Service Commission was re-established to replace the spoils system, bringing about over 105,000 employees into the government and minimally qualifying them through an open examination system (Gifford 2012).

The re-establishment of the Civil Service Commission served several purposes. First and foremost, it sought to fairly employ candidates through a regulated merit system. In order to do this, the Commission needed to redefine the examination process, focusing on the assurance that all examinations be job-related and not just a barrier for employment (Guion, 1998). This would also accomplish the Commission's second goal - obtaining the best individuals possible for employment. By obtaining the best, the Government would not only raise its appearance of professionalism and power, but better serve the people of the United States of America.

From 1883 to 1964, the Commission's focus was on the growth of the government and the reform of current regulations. During President Roosevelt's administrations, the President pushed for the growth of the Federal labor force. Being a Civil Service Commissioner himself from 1889 to 1995, Roosevelt aimed to enforce the policies of the Commission, leading a large growth in available jobs that were in accordance to these regulations (Haynes, 1998). The growth in the workforce expanded to State Service, creating the need for individual States to develop their own merit-based policies and procedures; New York and Massachusetts were the first states to have Merit laws in 1906 beyond the regulations of the Federal Government; California established its first commission in 1913 (King, 1978).

Several decades passed and various laws and systems were developed: the Retirement Act of 1920 (implementing retirement and benefits), Classification Act of 1923 (defining positions and salaries), and the Veterans preference act of 1944 (U.S. Office of Personnel Management, 2003). During John F. Kennedy's presidency, however, the focus shifted from growth and reform to an emphasis on employment rights and equality, reacting to various political protests and social justice reforms. Early in his presidency, Kennedy began to focus on Women's rights and developed the Presidential Commission on the Status of Women (PCSW). The goals of the commission were (1) to ensure women were respected in the workforce and (2) to create positions for women in the workforce. On March 5th, 1961 the President formally issued Executive Order 10925, requesting that the Civil Service Commission develop a program to ensure future and current employees are treated fairly without regard to their race or national origin (A Brief History, 2018). It was this order that set the tone for the next several decades in Federal civil service history.

Fairness in Testing: Legal Implications

The Executive Order 10925 was quickly followed by the Civil Rights Act of 1964, during President Johnson's administration. Within the Act, Title VII (Equal Employment Opportunity) proclaimed that any form of discrimination against race, color, religion, sex, or national origin within the workplace was unlawful (Meyers, 2006). Published August 24, 1966 in the Federal Register by the Equal Employment Opportunity Commission (EEOC), Title 29-Labor, Chapter XIV, Part 1607.11 expanded upon Title VII's reference of "within the workplace":

> *"A test or other employee selection standard—even though validated against job performance in accordance with the guidelines in this part—cannot be imposed upon an individual or class protected by Title VII where other employees, applicants, or members of a minority or a sex group have been denied the same employment, promotion, transfer, or membership opportunities as have been made available to other employees or applicants".*

This introduction to Equal Employment Opportunity was unprecedented in the workplace and established standards for future hiring practices in an effort to ensure fairness of pre-employment testing. Nevertheless, issues in the courtroom made it apparent that the definition of testing equality in Title VII needed to be expanded upon, beyond the expansion of the 1966 Federal Registrar.

Griggs v. Duke Power (1971) was one of the first court cases that brought attention to Title VII's lack of clarification in what constitutes discrimination. The claims brought against Duke Power Company concerned entrance requirements that (a) did not intend to measure knowledge, skills, or abilities necessary for specific jobs or promotional opportunities and; (b) discriminated against African Americans, a protected class. In 1955, the company introduced the requirement of a high school diploma to job assignments (excluding labor-based jobs) and transfers from jobs related coal handling to operational, laboratory, or maintenance positions. After the introduction of Title VII, the company produced an additional entrance hurdle and mandated that all new employees pass aptitude "intelligence tests" prior to appointment. Beyond the aptitude tests (i.e., the Bennett Mechanical Comprehension Test and the Wonderlic Personnel Test) and the high

school entrance requirement being unrelated to the company's job characteristics, it was also argued by Willie S. Griggs and various petitioners that the requirements hindered the advancement and employment of African Americans (Employment Testing, 1972).

Due to these allegations, the courts evaluated the company's practices and found that the selection results were in violation of Title VII. In Chief Justice Burger's majority opinion on the case, the following was stated:

> "*Whites register far better on the Company's alternative requirements than Negroes…both requirements operate to disqualify Negroes at a substantially higher rate than white applicants"*

The Chief Justice's statement established two key interpretations for the once-ambiguous Title VII regulations. First, it recognized the violation of equal opportunity, regardless of intention. Although Duke Power's aim in the implementation of the requirements were to standardize the appointment process across ethnicities, the courts determined the results still led to substantial inequality in selection. Second, the Chief Justice's statement defined the violation of equality by providing statistics on African Americans being disqualified at a "substantially higher rate" than Caucasians (Columbia Review Journal, 1972). During the case, the Wonderlic and Bennet test pass rates were evaluated and results portrayed a higher pass rate for Caucasians (58%) than African-Americans (6%). As the intelligence tests pass rates were based on average pass rates of high school students, the courts also reviewed the 1960 census and found the pass rate of Caucasians (34%) to be much higher than that of African Americans (12%) as well. In Chief Justice Burger's address on the case, both of these factors created unnecessary barriers for the

African American workforce at the Duke Power Plant. It was also determined that the intelligence tests did not pertain to the applicant's ability to do the job. The courts ruled in favor of Griggs and the petitioners on March 8th, 1971.

In the court case of *Hazelwood School District v. United States* (1978), the District Court used Chief Justice Burger's interpretation of Title VII as reference for a determination of adverse impact. Essentially, the case involved a disparate impact allegation set forth by the United States on the Hazelwood School District for hiring a disproportionate ratio of African American teachers as compared to Caucasian teachers (Hazelwood School District v. United States, 1978). On April 27th, 1978, the case began in District Court where Hazelwood provided evidence that, although the ratio of African American teachers to Caucasian teachers was substantially different, it matched the student population of the area. The courts ruled in favor of Hazelwood and the case was appealed to the Federal Courts. Upon further review, the Federal Courts determined the District courts wrongfully associated the statistics of the population of teachers to the population of students within the School District; the teacher ratio was not generalizable to the student ratio. A re-evaluation of the teaching ratios took place by comparing census figures of similar school districts in the area and it was found, on average, that other school districts maintained a ratio of African-American teachers to Caucasian teachers at 15.4% (e.g., for every 100 Caucasian teachers, roughly 15 were African-American), while Hazelwood maintained a ratio of 1.8%. On June 27th, 1979, the Federal Court overturned the District's decision, ruling in favor of the United States (Meier, Sacks, Zabell, 1984).

Establishing the Guidelines for Adverse Impact

In 1968, the Office of Federal Contract Compliance (OFCC) began work to develop a definition of adverse impact. By 1970, EEOC and OFCC collaborated and established a set of guidelines that expanded further on the 1966 Federal Registrar expansion of Title VII. The definition of adverse impact still maintained that if adverse action is determined, then validation of the selection effort should be assessed. The new definition also referenced the newly published *Standards for Education and Psychological Tests and Manuals* (1966) for a means of assessment; however, the publisher's intent at the time was not to serve as a standard for employment testing.

Regardless of its intent, the 1970 definition began to introduce mandates of validity that were recognized by practitioners as unclear and impossible, such as the validity needing to result in a "high degree of utility" and evidence that attempts were made to demonstrate no other alternatives for testing were possible (Guion, 1998). Minor alterations to this rule were made and in 1971, provisions of Title VII, as well as this definition, were extended to the public sector (i.e., State Service) (Biddle, 2011; Selection Manual, 1979).

In the same year as the *Griggs v. Duke Power* (1971) case, the Technical Advisory Committee on Testing (TACT) was assembled by the State of California Fair Employment Practice Commission (FEPC). This group, consisting of 32 specialists, was entrusted with the responsibility to develop a set of California Guidelines on Employee Selection Procedures, which included the re-definition of adverse impact (Biddle, 2005). They began their research by reviewing the precedential proceedings listed earlier in this

paper. Although the extant court cases acknowledged the need of clarification for adverse impact to determine court rulings and interpreted data by means of "substantially higher rates," the commission agreed the phrase was not transferable across cases and was not sufficiently standardized in a mathematical sense.

The insufficient standardization was mainly attributable to the absence of an established threshold. In the *Griggs v. Dukepower* (1971) case, Chief Justice and the courts determined that a fifty-four percentage point difference between Caucasian pass rates (56%) and African American pass rates (2%) was large enough to warrant adverse action. This judgement was ruled without any clarification as to what degree the breadth of percentage was in meeting criteria of the violation, as no threshold at the time existed. However, the commission did find evidence that the decision of the violation fell within previous court case determinations, such as the *United States v. HK Porter Company* (1968; five to ten percent breadth equated to violation) and *Arrington v. Massachusetts Bay Transportation Authority* (1969; fifty-five percent equated to violation) cases.

A secondary issue involved in the standardization of adverse impact was found in ensuring statistical significance was considered in the final determination of adverse impact. Court hearings were beginning to provide statistics in their assessment of adverse impact beyond a standard percentage comparison, such as in the *Casteneda v. Partida* (1977) and *Hazelwood School District v. United States* (1978) cases. In both proceedings, the courts used a "Standard Deviation Analysis" between the focus group and reference group. Specifically, if the focus group deviated by more than two or three standard deviations as compared to the reference group in a selection procedure, the

results were said to indicate precise statistical disparities between the two groups (Meier, Sacks, Zabell, 1984). Although these analyses were a step in right direction, the Supreme Court recognized that an investigation of standard deviations did not officially constitute statistical significance assessment. Agreeing that a form of statistics should be utilized to further determine a violation of adverse impact, but wanting to include a test of statistical *significance,* the commission determined a form of the Pearson correlation statistic would be utilized (Biddle, 2011).

When the TACT group worked on the final negotiations of the rule, several concerns were addressed. First was the establishment of the percentage threshold. TACT referred to The Office of Federal Contract Compliance Testing Order of 1968, which first introduced the conceptualization of a "rule of thumb" statistic (Guion, 1998). The committee split into two teams, arguing between a liberal adoption of 70% and a stringent adoption of 90%, eventually leading to a compromise of 80%. Second was the feasibility of the statistical analysis. Although the team agreed on the importance of statistical procedures, they were aware of the lack of familiarity that practitioners had with complex mathematics, especially considering the limitations of technology and computer software at the time. Therefore, the committee agreed that (a) if the 80 percent threshold was not breached, no further analysis was necessary and; (b) if a violation of the 80 percent threshold was found, the implementation of a statistical analysis would need to occur. In 1972, the Equal Employment Opportunity Act was passed, appointing the Civil Service Commission power and authority over the enforcement of equal opportunity in Federal Civil Service (Adverse Impact, 2015). The Uniform Guidelines

was also published this year, containing the definition of adverse impact and various human resource testing practices:

> *Adverse effect refers to a total employment process which results in a significantly higher percentage of a protected group in the candidate population being rejected for employment, placement, or promotion. The difference between the rejection rates for a protected group and the remaining group must be statistically significant at the .05 level. In addition, if the acceptance rate of the protected group is greater than or equal to 80% of the acceptance rate of the remaining group, then adverse effect is said to be not present by the definition* (Section 7.1)

The 80 Percent Rule

In 1973, the Uniform Guidelines were re-reviewed and edited until a final revision was determined, resulting in the 1978 *Uniform Guidelines* as it is known today, endorsed by the Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. The adjustments made did little to impact the original definition of Adverse Impact, however, the updated version provided additional interpretations to the 80 Percent Rule, also known as the 4/5^{ths} rule of thumb.

The definition of the 80 Percent Rule states that if a selection ratio between a focus group's pass rate and a reference group's pass rate (the focus group normally being a protected class and the reference group being the majority of the testing group) is below 80 percent, then adverse impact is present and an investigation is warranted; if the opposite occurs, then adverse impact is not present and an investigation is not warranted.

The 80 Percent Rule is recommended to be implemented at each stage of the selection process to ensure no violations occur. This includes application acceptance, examination pass rates, hiring approvals, and anything else involved in assessing candidates for employment opportunities. To achieve the results in each of these stages, the following impact ratio formula would be utilized, where P_{Foc} represents the pass rate of the focus group, T_{Foc} represents the total individuals of the focus group, P_{Ref} represents the pass rate of the reference group, and T_{Ref} represents the total individuals of the reference group:

$$\frac{P_{Foc}}{P_{Foc}} \Big/ \frac{P_{Ref}}{T_{Ref}} = \text{Impact Ratio of Selection Stage}$$

The result of the assessment is considered the impact ratio of the selection stage (Collins & Morris, 2008; Morris 2001) and describes the magnitude of the selection rate difference between the focus and reference group. If a violation of the rule occurs, further assessment of the data by use of a statistical test is warranted (this concept is further explicated in Chapter 2).

As the 1978 version of the *Guidelines* indicate, the rule must be interpreted in consideration of other supplemental information. If a small difference occurs (within 80%) between the selection rate of a focus group and reference group, normally adverse impact is not warranted. However, according to the *Guidelines*, adverse impact may still exist in practical terms when the rates are statistically significant, or if the organization purposefully discouraged applicants based on race, sex, or ethnic group. Conversely, large differences may not necessarily warrant adverse impact if the differences are based on small numbers, are not statistically significant, or if the recruitment/program needs

create atypical scenarios of protected classes being small compared to the total applicant pool.

Adoption of the 80 Percent Rule in California State Service

In 1971, provisions of Title VII of the Civil Rights Act were extended to the public sector, including California State Civil Service. When the *Guidelines* were established, California's State Personnel Board (SPB) – an overseeing agency of civil service in California – quickly adopted Section IV's regulation of the *Guidelines* to maintain records of information regarding adverse impact pertaining to ethnicity, disability, and sex. Government Code 19232 was enacted at that time to ensure adherence to the regulation and is still enforced today for each of the Departments within the State of California.

Beginning at the examination phase of selection, applicants are requested, on a voluntary basis, to identify sex, age, ethnicity, and employment-related disability. To ensure anonymity, the State Agency database randomly assigns candidates with a six-digit identification number when collecting Equal Employment Opportunity (EEO) information. When data is collected, it is mandated that documentation containing EEO data be removed from the selection process prior to determining applicant qualifications to ensure judgements related to employment are free from discrimination.

The SPB requires that each State Agency utilize the "Rule of Thumb" statistic known as the 80 Percent rule when assessing adverse impact in the selection processes and defines this regulation in their California Selection Manual, published in 1979. Following the *Guidelines,* SPB developed a system database utilizing the same impact

ratio formula mentioned earlier, with minor adjustments. In the *Guidelines*, the formula refers to the reference group as the majority group (numerically) in the testing phase of the selection process. Depending on the candidate pool, this could vary greatly, although most often the reference group tends to be Caucasian or male (Biddle, 2015). To adhere to the Affirmative Action and EEO standards of California and to ensure record-keeping standardization, the SPB uses Caucasians as the base group for calculating the 80 percent rule for ethnicity, regardless of the majority size of the groups.

The SPB also mandates that if the numbers of candidate rates are small (i.e., occurred by chance alone) the agencies would not assume the existence of adverse impact (Selection Manual, 1979, Section 4415, p. 9). The term "small" is defined by SPB as the difference being so small that if one person were to shift, the result of adverse impact would differ. If the selection rate was continuous in nature and a pattern emerged from the selection rates over time, adverse impact would then be identified. The Selection Manual cites section IV, D of the *Guidelines,* acknowledging that the determination of adverse impact is not purely based on the arrhythmic 80 Percent rule. In an updated section of the manual, SPB emphasizes a need to calculate the statistical significance of differences in selection rates to verify the existence of adverse impact (Selection Manual, 1979, Section 4420, p. 4, 1979). The manual refers to the same guidance as the *Guidelines,* stating that larger groups may still need a statistical test if adverse impact is identified just over 80 Percent. If smaller groups occur in a selection process, and adverse impact exists through the rule, statistical significance tests may be warranted as well.

Beyond the Selection Manual and various trainings offered by the California Department of Human Resources (CalHR) – an overseeing state agency lateral with SPB – very little guidance is given regarding when a statistical test is warranted and what statistical test is preferred. In 2003, SPB submitted an updated Selection Manual to all California agencies, clarifying the points addressed in the *Guidelines*, but brought no attention to the statistics mentioned after the 4/5ths rule was established (Merit Selection Manual Appendix D, 2003). Instead, the board provided a list of references in the manual to guide those involved in testing, which contained various adverse impact statistical assessment suggestions, such as the *Assessment, Measurement, and Prediction of Personnel Decisions* (Guion, 1998) and *Standards for Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985).

Records pertaining to adverse impact claims were also void of any mention of statistical tests on the SPB website. For example, in 2001, quality assurance reviews were conducted for both the Department of Veterans Affairs and the Department of General Services. Issues were identified in the proper assessment of adverse impact, but the departments were only notified to implement a "re-evaluation of selection procedures" prior to results being published or to identify the job related-ness of the selection processes by a supporting job analysis if adverse impact exists in testing results (Quality Assurance Review of Veterans Affairs, 2001; Quality Assurance Review of General Services, 2001). In 2006, an audit was performed on the Department of Transportation concerning its selection practices and compliance to various state policies.

Regarding adverse impact, the Department reported on their adverse impact protocol, specifying that “an analysis of statistical data” is performed on all testing data to determine whether adverse impact exists. Again, no reference to the form of statistical test used was noted. In December of 2011, an assessment funded through SPB was performed to determine an evaluation between current ranking systems employed in California and a three-rank system pilot program. A goal of the pilot program was to determine whether adverse impact would be minimized with the new ranking system. Although the board reached out to a private consultant that had the abilities to employ statistical testing (Donnoe & Associates, Inc.), the assessment of adverse impact only focused on base percentages between various minority groups, similar to the 80 percent rule of thumb.

The SPB’s policies on adverse impact are vague and the agency recommends that each of the Departments develop a specific program designed to evaluate adversity and refer to the Uniform Guidelines as a source of guidance in the selection process. According to the Report on the Status of the State Discrimination Complaint Process (2002), only 46 (51.7%) of 89 State Departments in California evaluated have a formal discrimination assessment process in place. However, many of these processes are not publicized nor clarify whether statistics are utilized in the determination of Adverse Impact beyond the 80 Percent Rule. In a recent SPB evaluation report, over 900 discrimination reports were filed within Departments across the State of California within the 2011 calendar year. With the lack of uniformity and standardization across agencies regarding the interpretation of adverse impact, any of these reports could breed issues

similar to those discovered during the casework of *Griggs v. Dukepower* (1971) and *Hazelwood School District v. United States* (1978). Thus, to ensure best practices and allow for a higher standard within State Civil Service, a uniform statistical supplement to the rule of thumb should be adopted when determining evidence of adverse impact in the selection process within State agencies.

Chapter 2

BACKGROUND OF THE STUDY

Importance of Validity

As mentioned in Chapter 1, a uniform set of statistics should be established for the State of California in regard to the assessment of adverse impact. Before investigating the various statistical analyses available as options, it is important to expand on the definition of validity in the context of adverse impact. The concept of validity was first introduced into the realm of adverse impact in the late 1960s and has since been fundamental in understanding the determination of discrimination in selection (Anderson & Rogers, 1971). It is presented in the new *Standards* (2014) as a source of evidence and theory that is represented in a set of test scores. From a legal perspective, a selection procedure is said to be valid if it is job related, is consistent with business necessity (Civil Rights Act, 1991), and the proceedings fall in accordance to the *Uniform Guidelines,* various professional standards, and previous precedential court rulings. Section 3A of the *Guidelines* state that if adverse impact is present, discrimination exists, unless a validation study is available for the specific selection effort being assessed. A key aspect of validity in selection is that it is a culminated effort or aggregation of evidence that a particular selection measure indicates job relatedness (Society for Industrial and Organizational Psychology, 2003). Therefore, all feasible forms of validity should be assessed when determining the overall soundness of a selection measure. Currently, the *Standards* expand on five main forms of validity evidence; Evidence Based on Test Content, Evidence Based on Response Processes, Evidence Based on Internal Structure,

Evidence Based on Relations to Other Variables, and Evidence for Validity and Consequences for Testing.

Evidence Based on Test Content

This form of evidence suggests that a strong connection between the content on a test and the construct desired to be represented, must exist. This is also known as a content strategy (Uniform Guidelines: Section 15C1, 2015). To ensure content validity, several requirements must be met. Most selection test developers have a content specification that expands how a specific job or set of jobs should be portrayed (Standards, 2014). This specification is often referred to as a work or job analysis. This form of analysis ensures that the appropriate themes, formatting, etc. are presented on in the selection procedure. It would also entail an evaluation of both Knowledge, Skills, Abilities, Other Characteristics (KSAOS) and tasks be made and incorporated into the selection procedure (Society for Industrial and Organizational Psychology, 2003).

When conducting an analysis, job experts (known as Subject Matter Experts or SMES) are referenced to ensure that the content domain is properly represented and structured on an exam. A practitioner ensures this in the preliminary stages by either observing the experts portraying the construct or by interviewing them in a conference setting. The content received from the sample should be derived from a representative sample and should present knowledge, skills, abilities, tasks, equipment, and so on for the specified job. The analysis should also include measures of importance, frequencies, and other selection scales in order for the knowledge, skills, and abilities to be rated by need of representation for the specified selection measure (Standards, 2014). To ensure that

the appropriate construct is measured through content, construct irrelevance (irrelevant distractions from the original construct) is avoided and alignment, or matching capabilities/traits to a test assessing capabilities/traits, is implemented. Low rated characteristics (i.e. those not needed upon entry) should not be part of the content for a selection procedure and would not be considered "valid."

Mental resources, including intelligence, aptitude, etc., should not be supported by content validity (alone). For resources that are supported, the knowledge, skills, and abilities should be included and operationally defined and to approximate observable work behaviors as compared to an observable work product (Uniform Guidelines: Section 15C1, 2015). The closer the content validity measure (job analysis) is to the content of the selection procedure, the greater the evidence of validity. Additionally, if a higher score on a selection procedure is a result of better job performance, as deemed by a content valid job analysis, final scores are allowed to be ranked by proficiency.

Evidence Based on Response Process

Evidence based on response process suggests that validity is enhanced when confounding measures are removed from a selection measure. This can be minimized by assessing individual responses and determining what confounds exist. Some confounds relate to the text material and whether the language, vocabulary, and presentation is appropriate for the construct being represented (Standards, 2014). If words are too general or abstract, or if the language is confusing, confounds may be introduced. With this type of evidence, raters must also be taken into consideration. In this regard, errors (known as rater errors) could impact the manner in which an examinees response is rated

(e.g., the response style). These errors include but are not limited to Halo effects, Priming effects, and centralized scoring (Meyers, 2008). By assessing these confounds, researchers and test developers can ensure a sense of validity for the response process.

Evidence Based on Internal Structure

Validity evidence based on internal structure provides insight on how test items conform to the construct that a test developer expects to measure. A construct is conceptualized by the developer and is measured by a given exam or measurement tool.

For example, time management, a construct, can be captured through the criterion of an in-basket exercise measuring the ability to prioritize tasks by importance. Because it is generally agreed upon that task prioritization is an acceptable predictor of time management, it can be said that evidence based on internal structure is present when time management is measured through an in-basket exercise. Essentially, in this example, the in-basket exercise is measuring what "it is intended to measure."

Samuel Messick (1989), an American psychologist whose research focused heavily on validity evidence based on internal structure, defines the concept as a unified evaluative assessment of how empirical evidence and theoretical rationale support inferences and actions based on selection scores. To ensure that this inference can occur, statistical diagnostic tools such as Item Characteristic Curves (ICCs), inter-item correlations, and factor analysis are often employed. ICCs are often investigated to ensure that ability is taken into account when measuring performance on a test. The S-shaped curve – known as a sigmoidal curve – identifies performance on a continuum of examinee ability. Inter-item correlations provide information on the amount of explained

variance accounted for and shared between the items on a test. Factor analysis is employed to investigate factors of dimensionality represented by items in an exam—which affirms whether the intended constructs for measurement are empirically represented in an exam. These three examples of statistical methods all allow for the establishment of validity evidence based on internal structure.

Evidence Based on Relations to Other Variables

As noted by Campbell & Fiske (1959), three methods exist to establish validity based on evidence found in tests relating to other variables: convergent and divergent evidence, test criterion correlations, and validity generalization.

A construct represented in a selection measure is assessed through partly by correlating the measure to other measures. If a strong correlation exists between a measurement tool in question and a pre-established measure representing the same construct, the evidence suggests that the measure is convergent.

Conversely, if no correlation is said to exist between a measurement tool in question and a pre-established measure representing a vastly different construct, the evidence suggests that the measurement tool in question has divergent validity. An example of this would be a comparison of a supervisory measurement to a technical skills measurement. The constructs vary greatly (one comprised of leadership, guidance, and management skills while the other incorporates technical knowledge of the classification) and a relation between the two should not exist.

Another statistical method of establishing validity based on the correlation coefficient is criterion-related validity evidence. The two approaches normally employed

to establish criterion related validity evidence are predictive and concurrent methods. Concurrent validity refers to the comparison between the measure in question and an associated outcome assessed at the same time. According to the *Standards* (2014), concurrent validity establishes the "status quo" of a specific period of time.

An example of a concurrent validity assessment would be if a group of Subject Matter Experts (SMES) were attempting to pilot a new examination for a specific classification, and had a group of individuals' test the old examination and new examination in the same period of time. If developed properly, the SMES would expect that the results of the new examination would reflect similar results of the old examination, as indicated by a significant correlation coefficient between the two tests.

Predictive validity is normally assessed by comparing scores of a specific selection procedure (e.g., an entrance examination) to future performance reviews. If the two occurrences are strongly correlated (i.e., if high scores on the examination correlate positively with high ratings on the performance review and vice versa), predictive validity is said to exist. This type of assessment is also known as a criterion correlation. Issues with this type of criterion related validity evidence includes range restriction (i.e., only those that passed the selection hurdle and obtained positions would be assessed on future performance) and lack of standardization, as performance reviews are normally subjective and based on personal experiences or encounters with the employee. Therefore, predictive studies are only feasible based on the dependent measure or outcomes being justified (Standards, 2014).

Lastly, validity evidence can be established by comparing results in a selection measure or scale to previous selection measure or scale results (also known as validity generalization). This can be done in both a quantitative and qualitative manner. Literature reviews would be considered qualitative and statistical comparisons would be considered quantitative. Having an item become "transportable" is also key. This identifies that the item can be interpreted as valid in multiple settings. Synthetic validity is another form of generalization. This assumes that elements can be grouped and synthesized together by association of comparable items.

Evidence of Validity and Consequences of Testing

Evidence based on consequences of testing can be found when one considers the impact that the testing environment has in testing results. Often, this is related to the soundness and interpretation of selection scores. Even the awareness of being assessed may impact the readiness of a candidate and thus, could impact the scoring. An example of this would be if a job series obtained a promotional pattern and the supervisors prepared materials, lessons, or training to prepare for a selection measure (beyond the standard scope of training for the current job series). If potential candidates are preparing for a selection measure, the knowledge assessed in the selection measure may not accurately portray the candidate's readiness for the position. Rather, it may instead measure the ability for the candidate to memorize the material.

Importance of Reliability

One of the prerequisites of a test's validity is reliability. Reliability specifies the precision of a test. A measure is said to be reliable if it produces similar results under

similar conditions. For example, if an organization develops an examination meant to measure certain criterion (e.g., supervisory skills), the organization would hope that the measure produces similar results under similar circumstances. In testing, similar circumstances could refer to the examination (i.e., the examination may be identical or conceptually the same) or the candidates taking the examination (i.e. familiarity with the job, exposure to the questions, or the candidate pool is the same or similar in background). If an individual takes an examination twice and receives considerably different scores, it can be said that the examination is low in reliability. Organizations would be ill advised to make selection decisions based on unreliable examinations such as this, as true performance is unclear from the scores that are observed.

Ideally, the goal in testing (such as pre-employment testing) is to develop a measure that assesses an individual's knowledge of a specific content area, without error. A measure of an individual's score is known as a true score or (T) in Classical Test Theory (CTT; Guion 2011). However, in some form, error is always introduced into the equation as it is nearly impossible to eliminate all error from a measure. This is known as error variance (E). The result of these factors (true score and error variance) leads to an observed score (X), expressed in the following formula below:

$$X = T + E$$

Error variance (E), also expressed as (V_O), is calculated by taking the sum of squared deviations between the two sets of data of equivalent forms, divided by the degrees of freedom in the model. The more the data sets deviate from each other, the more error variance is said to exist and the less reliable the measurement is. If an

examination is said to deviate greatly between administrations, a valid inference cannot be made as to how the test would perform in future conditions (Guion, 1998).

Additionally, the difference between a true score and an observed score can be due to natural error variance and is important to distinguish from other forms of error. The mental and physiological state of a candidate or the environment of the assessment may adjust between administrations and influence variance to occur naturally between scores. This is known as variance due to systematic causes or true score variance (V_T). The remaining error, known as variance due to random error (V_E), is derived from other factors that impact the reliability of the examination. This can include poorly developed questions, trick responses, unclear distractors, and miskeyed responses. To ensure proper interpretation of an observed score, error obtained from a measure should be partitioned into these two categories, expressed in the following formula below (Meyers, 2009):

$$V_E + V_T = V_O$$

Reliability Coefficients

As described above, reliability is the precision of a measurement. In CTT, test scores and their consistency across replications are mainly evaluated through reliability coefficients. Reliability coefficients can be conceptualized in the following formula:

$$r_x = 1 - \frac{V_E}{V_O}$$

The ratio $\frac{V_E}{V_O}$ represents non-systematic error obtained in a set of observed scores; the more non-systematic error present, the less precise an inference one can make from the results of a measure. Reliability coefficients are represented by a value derived from the subtraction of this error. This coefficient value, ranging between 0 and 1, represents

the amount of unique variance found in the observed set of scores. When less error is present, the reliability is closer to 1, indicating a more precise measurement. For example, if a measure yielded a reliability coefficient of .95, the measure can be said to have good reliability.

According to the *Standards* (2014), three categories of reliability are recognized: (a) coefficients derived from the administration of alternate forms in independent testing sessions (*alternate-form coefficients*), (b) coefficients obtained by an administration of the same form on separate occasions (*test-retest coefficients),* and (c) coefficients based on the relationships and interactions among scores derived from individual items or subsets within a test, all data accruing from a single administration (*internal-consistency coefficients*).

Alternate-forms reliability, also known as parallel-forms reliability, maintains that two tests are considered parallel if they measure the same constructs, but display the content in "alternative" manners (i.e., different versions of the same test). To assess this, the two forms are correlated and an alternate-forms reliability coefficient is obtained. If the true scores and error variance between the tests are closely related, the reliability coefficient would be near a value of 1, indicating that the tests are reliable comparisons of one another (Furr & Bacharach, 2008). Issues with this type of reliability include uncertainty that the true scores on one test equate to the true scores on the other test, as content might differ to some extent between the forms. Additionally, because it is ordinarily the case that the same individuals take both forms of the test so that a correlation between forms can be computed, recall and practice effects between

administrations can impact examinee performance, which in turn can impact the true score variance differences between the forms. If this occurs, a valid inference can also not be made from the reliability coefficient achieved.

As with alternate-forms reliability, test-retest reliability is also assessed by comparing test scores from two independent administrations. The main difference between the two forms of reliability is that test-retest reliability assesses the consistency of the same test over an interval of time, rather than alternative forms. In this form of reliability, the "same test" means that the tests used in both administrations are exactly the same, without any alterations or adjustments to the items as would be the case in alternate forms testing. The ultimate goal of test-retest reliability is to ensure that the test performs the same across multiple administrations across time. If true scores and the amount of error variance differ across time, the reliability coefficient is adversely affected. A benefit of this type of reliability over alternate-forms reliability is the minimization of error due to content adjustments, as the items stay the same (Furr & Bacharach, 2008). However, concerns with this method used to establish a reliability index still arise, as recall and practice effects for specific test constructs can still occur.

Internal Consistency

Of particular relevance to this study is the form of reliability concerning internal consistency. In pre-employment testing within the State of California, adverse impact concerns the performance of examinees within a single test administration. A single-administration measure is said to be internally consistent when individuals are performing consistently throughout the domains of, or items contained in, the measure in

one test administration (e.g. those who perform poorly on one question should perform poorly across other similar questions). The term "domain" is key in this definition, as it is common that measures are comprised of more than one content domain with performance varying between content areas.

To assess internal consistency at the item level with State-level testing, item analyses can be employed to evaluate frequencies, percentages, and item difficulties. An item difficulty identifies the percentage of individuals that correctly respond to a specific item; the higher the percentage, the easier the item. Values that perform below .20 would be considered too difficult and values above .90 would be considered too easy and should be considered for removal from the test (Interpreting Test Results, 2018).

Additionally, the relationship between how well an individual did on an item as compared to their total examination score can be assessed, and is known as item discrimination. One means of assessing this is through a Point Biserial correlation or Item Total correlation statistic. This statistic assesses how well a candidate's performance on an individual item mirrors their overall performance on the test and can be computed for both the keyed (correct) response and the distractors (incorrect responses) in the test.

For keyed responses, the item-total correlation should be positive, preferably .20 or above (Interpreting Test Results, 2018). The positive value demonstrates that candidates who scored higher on the test overall are choosing the correct response for the item and those who did poorer on the test are not. Distractors (incorrect responses) should be indicated by values closer to zero, or negative. Negative values demonstrate

that the lower scoring candidates are selecting the incorrect choices over the correct response. Items with positive distractor-total correlations should be deleted or revised.

To assess overall internal consistency on a set of items, several strategies can be employed, but the most widely recognized are: (a) Split-half reliability, (b) Kuder-Richardson 20 (KR-20), (c) and Cronbach's Alpha. Split-half reliability, introduced in the early 1900s, divides a data set into two equal portions and compares the item total scores for each examinee. Once the scores are methodically split, the subsets of the scores are correlated (Gamst, Guarino, Meyers, 2013). Values are indexed between 0 and 1; higher values indicative of higher reliability.

One concern associated with the split-half formula is the alternative methods available in splitting data, resulting in varied correlation results. In 1937, Kuder and Richardson developed the KR-20 coefficient, which is an evolution of the split-half reliability method. Rather than splitting the data once and comparing portions to achieve a correlation, the formula estimates the average of all possible split-half correlations and is equated using the following formula (Gravetter & Forzano, 2009):

$$KR20 = \left(\frac{n}{n-1}\right)\left(\frac{SD^2 - \sum pq}{SD^2}\right)$$

As with the split-half reliability formula, the KR-20 coefficients are indexed from 0 to 1, where higher values represent more reliability.

Although an improvement of the split-half formula, the KR-20 equation has a limitation in which only dichotomous variables (values of 0 and 1) can be utilized to calculate the reliability index. In 1951, Cronbach developed an adaption of the KR-20 using the following formula:

$$Cronbach's\ alpha = \left(\frac{n}{n-1}\right)\left(\frac{SD^2 - \sum variance}{SD^2}\right)$$

Essentially, the two formulas are the same, except in regards to the subtraction of the sum of variances ($\sum variance$), rather than the subtraction of $\sum pq$, as indicated in KR20 formula. The $\sum pq$ represents a summation of the products of *p* multiplied by *q*, where *p* represents the proportion of individuals choosing 0 per item, and *q* represents the proportion of individuals choosing 1 per item (Gravetter & Forzano, 2009). To allow for a greater range of responses (i.e., Polytomous items), such as a 5-point Likert scale, Cronbach altered this portion of the equation to instead represent the culmination of variances across all of the items. By doing so, Cronbach's alpha eliminates the limitation of using only dichotomous data and is the most recognized reliability statistic used today.

As with the previous reliability coefficients, values for Cronbach's alpha are indexed between 0 and 1; higher values indicative of higher reliability. General guidelines suggest acceptable reliability indices should be at a minimum of .70 to consider a measure reliable. Specifically, values between .70 and .79 are adequate, .80 to .89 are good, and .90 and above is excellent (Gamst, Guarino, Meyers, 2013).

If overall internal consistency is poor and/or questions have been determined to perform poorly within the test, test items may need to be removed. Post-hoc analyses can be generated to determine how the data set would perform (i.e., how the mean and variance would adjust) if an item were deleted. However, caution should be taken in the interpretation of these analyses, as various factors can impact the results of a reliability assessment. This includes but is not limited to the number of candidates, length of test or test segment, candidate exposure to test materials (training/studying), and test content

(Mehrens & Lehmann, 1973). Because of these concerns, it is highly recommended at the State-level that the test booklet and other test-relevant information be available when interpreting the reliability of an exam.

Standard Error of Measurement

The Standard Error of Measurement (SEM) is another means to assess the reliability and precision of a measure. It is defined as the standard deviation of the sample means if one were to draw an infinite number of testing samples from the same group of individuals and is represented by the following equation:

$$S_e = S_x\sqrt{1 - r_{xx}}$$

Essentially, if a group of individuals took a test multiple times, under the same conditions, it is likely that their scores would vary in some degree (either being increased or decreased between administrations, assuming recall effects were not a concern). This creates a distribution of possible scores, as well as a group of sample means, that can be referenced to estimate how a test sample may perform in future administrations. SEM can be computed through reference of this distribution. This parameter can be used to create a confidence interval around the mean to predict the likelihood that a future sample mean may fall within a given range (Standards, 2014). To compute a 95% confidence interval, or a chance that a sample mean will fall 95 times out of 100 in a given range, one would multiply the SEM by 1.96 and add and subtract this product from the mean, identifying the upper and lower bounds of the confidence interval (Gamst, Guarino, & Meyers 2013). The larger the range is between the bounds, the more difficult it is to

predict where the sample mean would fall in future administrations, and the less precise the test would be altogether.

Differential Item Functioning Analysis

As mentioned in Chapter 1, adverse impact refers to the disproportionate selection of a group of individuals (a focus group; normally a protected class) over another group of individuals (a reference group; normally Caucasian or male) in a selection decision. In the most simplistic terms, differences in selection rates between groups can be referred to as a difference of means between the groups in a testing measure. Assuming all groups within a selection measure have equal ability (a core assumption underlying the notion of adverse impact), if a difference of means occurs in a selection measure between groups, it may be due to an overall testing bias.

Where adverse impact analysis falls short is its inability to address the issue of whether group differences in a selection decision are due to bias attributable to the items on the test or from the varying abilities of candidates that coincides with members of different groups. Differential Item Functioning (DIF) analyses can be employed to statistically control for ability, which addresses this concern (Hurtz & Meyers 2007). The benefit of DIF analysis over adverse impact analysis is that it addresses biases at the item level, not the test level. By evaluating bias at the item level, we can determine whether groups are performing significantly differently from one another for each item (after correcting for alpha inflation) when controlling for ability level.

DIF analyses assist in the identification of true testing bias. This is important as some differences in overall test performance between groups are not due to bias and may

be improperly interpreted as such by an adverse impact analysis. For example, if a set of students attended a voluntary study session for an upcoming test and performed better on the test than the students who did not attend, this would indicate differences in ability, not testing bias. Testing bias refers to systematic and clear differences between groups in a measure that are presumably not due to the overall abilities of the test takers.

To determine whether DIF has occurred, a statistical test must be performed. If an item is found to perform significantly different between groups, the size of the effect of the difference should then be evaluated. Evaluation of the effect will allow for proper interpretation of the magnitude of the differences between the groups seen in the item. In certain conditions, an item may produce a significant difference but the effect of the difference (i.e., Effect Size) is too small to warrant concern. In other conditions, an effect may be large, and if disregarded, could contribute to overall testing bias (Sumizu & Zumbo, 2005).

Uniform and Non-Uniform DIF

DIF can be exhibited in two ways: uniform and non-uniform DIF. Uniform DIF occurs when one group of candidates score comparatively higher than another group, consistently and uniformly, across the entire range of ability on a given item. When this occurs, the item is said to perform for one group uniformly differently than the other when matching for ability. This concept is illustrated in Figure 1. Note that the Reference Group's performance on the item was consistently higher than the Focus Group's performance in both the low ability and high ability spectrum. This trend is an indication of Uniform DIF.

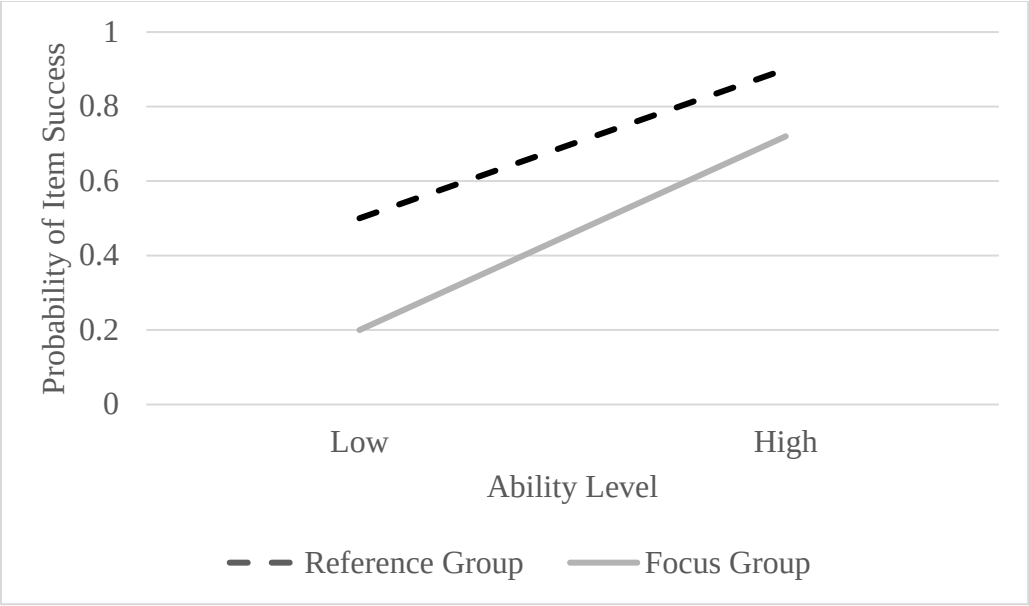

Figure 1. Uniform Differential Item Functioning.

Non-Uniform DIF is exhibited when an interaction occurs between groups and ability level as illustrated in Figure 2. In this example, individuals in the Focus Group with low ability seemed to perform worse than high ability individuals in the Focus Group. However, individuals in the Reference Group with low ability seemed to perform better than high ability individuals in Reference Group.

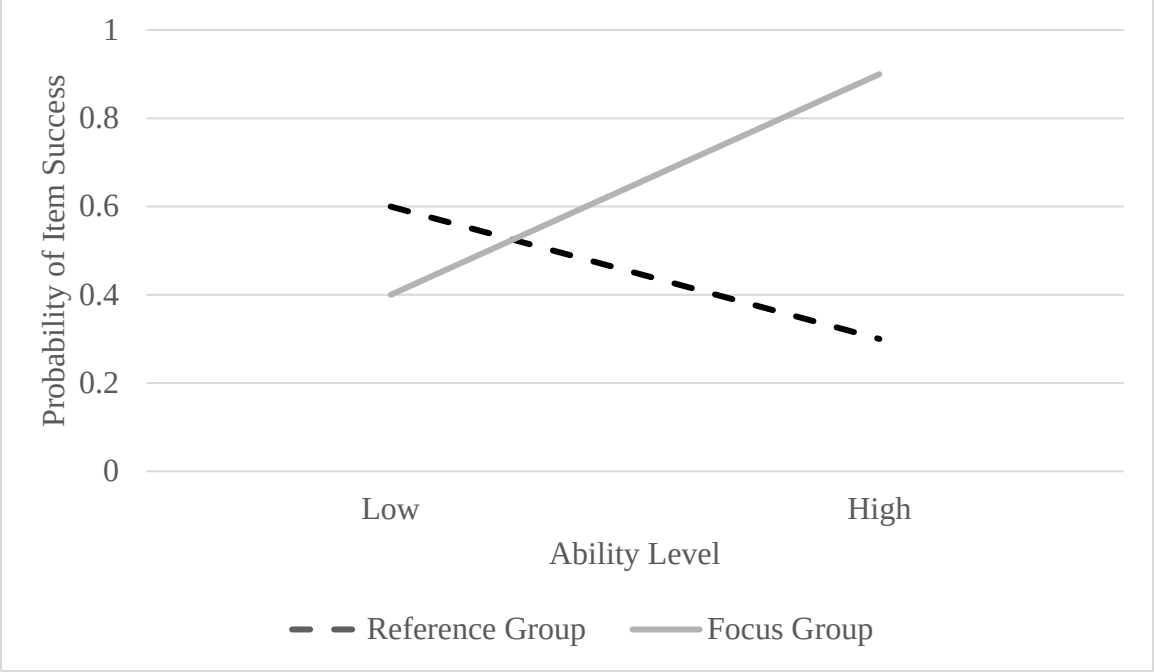

Figure 2. Non-Uniform Differential Item Functioning.

The Ability Measure

Since true ability is not directly observable, a measure of performance is required to group candidates into these high and low ability groups (Scheuneman & Slaughter, 1991). With respect to Classical Test Theory (CTT), ability level in DIF is normally evaluated as the total test score one receives after completing a measure. Total test score is used in this context because it is the only available criterion measure to represent an individual's ability level on the specific knowledge base being tested.

Another indicator available to estimate the construct of ability in DIF is Theta. Theta is an estimate of ability found in Item Response Theory (IRT). IRT (also known as the latent trait theory) assumes that each item, depending on its difficulty level, will provide information for varying examinee ability levels. Therefore, ability in IRT is addressed at the item level instead of the total test level, as contrasted by CTT. Theta is a

logarithmic odds transformation of the ratio of the probability of success over the probability of failure of an individual item. The value derived is known as a logarithmic odds (log odds) transformation of the ratio, also known known as a logit. Similar to test scores in CTT, logits are referenced when determining whether an item should be deleted or kept in a measure. Although IRT is supported substantially throughout literature as an effective testing paradigm in measuring ability, the purpose of this paper is to address ability thorough the lens of CTT. Those interested in further exploration of IRT are encouraged to read Lord and Novick's *Statistical Theories of Mental Test Score's* (1968).

Considering the Ability Measure

Several considerations should be made when referencing test score as a representation of ability level. First, the test must be a valid measure of the construct being assessed (e.g., reading comprehension). By avoiding confounding items, conducting a job analysis, and ensuring the items properly measure the construct, it can be presumed that total scores best reflect an individual's ability as it relates to the construct of interest. Secondly, the measure should be reliable. High reliability in a measure allows a test developer to trust that an examinee's test score reflects the ability of a test taker with relatively little measurement error, regardless of the number of administrations conducted; as reliability increases, the ratio of error variance to true score variance decreases, resulting in a more accurate and trustworthy test score.

Another consideration to be made when using test scores to represent ability is the dimensionality of the items being measured. A single total test score should reflect a single ability level. When two or more abilities are reflected in the total test score, DIF

can be falsely detected due to inconsistencies in performance between the sets of items. For example, a group of individuals may take a single test on mathematical principles, but the content may focus on algebraic and geometric principles (i.e., two dimensions). It is possible that test takers might vary in performance between the two constructs, thus rendering the total test score less meaningful than subscores representing each dimension. Best practice suggests researchers should split the items into the number of dimensions that exist in the test and assess each dimension separately, ultimately ensuring unidimensional ability level is assumed in each of the subtotal scores.

The last consideration that could impact the integrity of the total score is whether DIF is present in any of the items that make up the total test score. When DIF exists in a set of items, items containing DIF should be removed and a new total test score should be derived with the remaining items. Then, DIF analyses should be performed a second time, using a revised and more error-free total test score, ensuring any residual DIF has been identified (Zieky, 2003).

Considering Additional DIF Detection Factors

Various factors leading to Type I and Type II errors can also impact the detection of DIF. A Type I error is commonly referred to as a "false positive," or an incorrect rejection of a null hypothesis (i.e., the claim that DIF exists when it does not). A Type II error is known as a "false negative," or an incorrect acceptance of a null hypothesis (i.e., the claim that DIF does not exist when it in fact does).

Sample size is one such factor to take into consideration. For instance, the larger the sample size, the less likely a Type I and II error would occur (Acar, 2011). This is

because more information is being provided in the analysis and the sample becomes more representative of the population, mitigating the effects of outliers. To lower the likelihood of error, over 500 individuals in a DIF analysis is preferred. Less than 100 individuals in an analysis greatly increases the chance of committing Type I and II errors (Biddle, 2006).

The existence of extraneous variables is another factor to consider when conducting DIF analyses. Extraneous variables are undetected variables that influence the effects of another variable. In a DIF analysis, an extraneous variable is any variable that influences the independent variable of group membership (e.g., focus group versus reference group). For instance, organizations that intend to measure specific knowledge of a group of internal and external exam candidates should take precautions to ensure the knowledge being tested is equally accessible all groups being tested. Testing conducted without such precautionary efforts cannot affirm whether item performance differed due to group membership or due to knowledge accessibility.

A third factor to consider is whether or not deletion of items displaying DIF is a proper course of action. The mere presence of DIF in an item does not justify the deletion of an item, nor does it indicate item bias. Further, it is possible that items displaying DIF may be valid and useful in the measure, and the deletion of the item could decrease a measure's effectiveness. According to the *Guidelines*:

> *The use of any selection procedure which has an adverse impact on the hiring, promotion, or other employment or membership opportunities of members of any race, sex, or ethnic group will be considered to be discriminatory and inconsistent*

with these guidelines, unless the procedure has been validated in accordance with these guidelines.

Often, it is unclear whether practitioners should delete, alter, or retain items exhibiting DIF but nevertheless contributing to the validity of a selection measure (Scheuneman & Slaughter, 1991). When items displaying DIF require a background knowledge in a specialized focus or degree to properly make a determination, Subject Matter Experts (SMEs) should be consulted to consider the appropriateness of keeping the item. Otherwise, test developers may consider the items in question as they relate to the overall content of the exam. For example, developers may opt to remove a poor-performing item measuring writing comprehension instead of working to modify it, especially if there are similar items that do not exhibit DIF.

Available DIF Detection Methods

Currently, five DIF detection procedures outside of the IRT realm exist (ANOVA, two-way ANOVA, analysis of covariance (ANCOVA), Mantel-Haenszel chi-square, and logistic regression; Hurtz & Meyers, 2007). In a one-way Analysis of Variance (ANOVA) test, the independent variable, group association (protected group or reference group), is compared on the dependent variable, item performance (correct or incorrect). Although the simplicity of the method makes it an attractive option for practitioners, it brings about a few concerns. First, ability is not considered in the method, making it impossible to distinguish the effects of group association from the effects of ability level on item performance. Second, only uniform DIF can be detected as non-uniform DIF requires a comparison between ability and group.

To correct for the missing measure of ability, a two-way ANOVA can be performed (Meyers, Gamst, Guarino 2013). In this method, group association is still compared on item performance, however, the data is split between an additional independent variable, ability (i.e., total score). This allows to test for both uniform (main effect) and non-uniform (interaction) DIF. Although this is an improvement from a one-way ANOVA, several issues still exist. Ability level is still not considered when testing uniform DIF, making it difficult to determine whether the effects of group association are truly influencing item performance. In addition, splitting ability into only two groups (high and low ability) limits the amount of information provided in the test design, increasing the chance of Type I and Type II error.

An Analysis of Covariance (ANCOVA) resolves these issues by identifying ability as a covariate, rather than an independent variable. A covariate is a continuous variable that acts as a statistical control between the independent and dependent variables. By using this model, it is assumed that a linear relationship exists between ability level and the dependent variable, test performance. This is also known as the assumption of homogeneity of regression. If the relationship significantly differs between group memberships (i.e., the independent variable) further evaluation of the item is warranted. A fallback from this analysis, however, is that once the assumption of homogeneity of regression is violated, the analysis must stop and further evaluation of the main effect (uniform DIF) cannot occur (Hurtz & Meyers, 2007).

A larger issue that ANOVA, two-way ANOVA, and ANCOVA designs face when detecting DIF is the inability to handle binary data properly (which is how data is

portrayed in a DIF analysis). Each of these analyses fall under the scope of the General Linear Model (GLM) and are not meant to process binary values. An assumption of the GLM is homoscedasticity or equal error variance (Meyers, Gamst, Guarino 2013). DIF data cannot meet this assumption as group membership in DIF analyses do not have comparable variability on item performance, as item performance is identified in a categorical manner. Additionally, the GLM uses a method of least squares to approximate an equation based on a set of data that can reach values much greater than one, much less than zero.

Mantel-Haenszel-Chi Square

One method in detecting DIF beyond the GLM is the Mantel-Haenszel-Chi Sqaure (MH; 1959), a method structured to process binary data. In this method, each item is evaluated to determine if item performance differs between group memberships. This is assessed in a 2x2 array (known as a contingency table method), where one dimension represents item performance (correct or incorrect) and the other assesses group membership (i.e., female versus male, Caucasian versus Asian, etc.). Each of the arrays are then split by *X* number of ability levels (total test scores). A significant MH test indicates that a relationship exists between item performance and group membership, while controlling for ability level.

The simplicity of the MH method makes it the most popular procedure utilized in detecting DIF (Clauser & Mazor, 1998; Zwick, 1990). However, by categorizing ability into *X* number of ability levels (total test scores), MH provides very limited information on high and low ability test takers. As a result, the ability to detect non-uniform DIF

using the MH method is limited to the number of interactions between the categories created for ability level (Holland & Wainer, 1993).

Logistic Regression

A second method in detecting DIF beyond the GLM, and the focus of the study, is Logistic Regression (LR). The purpose of LR is to predict group membership on the dependent variable by calculating the likelihood that a specific item response type will belong to either the target group (i.e., correct item response) or reference group (incorrect item response) group (Gamst, Meyers, Guarino 2013) while taking into account ability level. The dependent variable, item performance, can be binary or multinomial (consisting of multiple categories). The current study focuses on binary LR where incorrect items are marked as 0 and correct items are marked as 1. Due to these parameters, rather than fitting a linear equation (found in ordinary least squares), LR fits a sigmoidal 'S' shaped curve to predict item performance, shown in Figure 3 below.

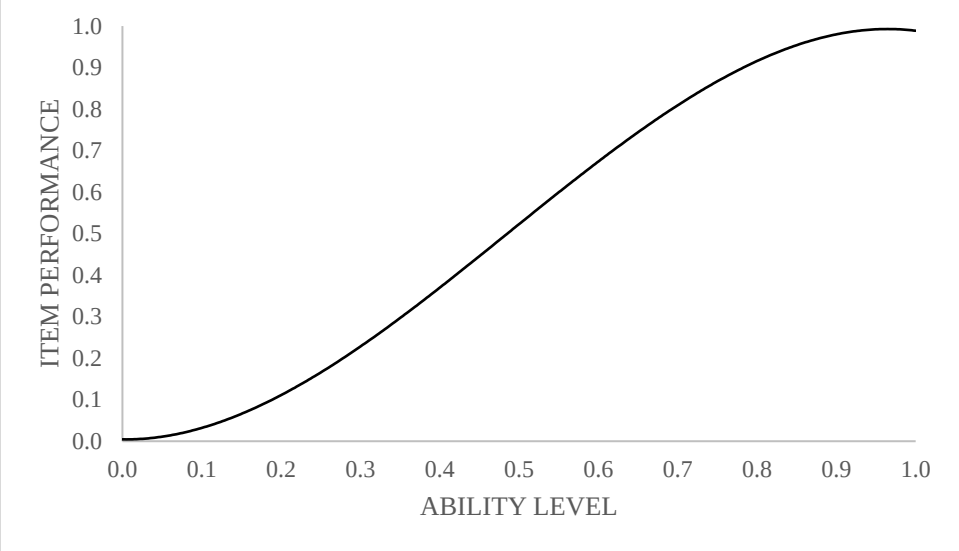

Figure 3. Sigmoidal 'S' Curve.

LR can be performed hierarchically in steps, also known as blocks. In the first block, ability level (total test score, a continuous variable) is entered into the model. By being entered into the first block, ability acts as a covariate (similar to an ANCOVA method) to group membership and item performance. The second block entered into the model is group membership. In this step, if a significant relationship is observed between group membership and item response type while controlling for ability level, it can be said that uniform DIF (the main effect) has been detected in the item. The third, and final, block entered into the model is the interaction between ability and group membership (this value is created by multiplying the two variables together). If this step is significant (with ability and group membership controlled for) then in can be said that non-uniform DIF (the interaction effect) has been detected in an item.

Logistic Regression as the Preferred Method of DIF Detection

Regarding the detection of DIF, both MH and LR have their respective advantages and disadvantages (Rogers & Swaminathan, 1993; Ferne & Rupp 2007). LR is heavily dependent on the amount of data received. Too few scores on the sigmoidal regression curve may adversely affect the ability to predict an individuals' score effectively. This is not such an issue with the MH method, as a researcher can determine the most appropriate ability parameters for contingency tables based on the information provided. Another consideration is statistical power. Because LR is structured to assess both uniform and non-uniform DIF, it may be less powerful in detecting uniform DIF than the MH method (Navas-Ara & Gomez-Benito, 2002; Rogers & Swaminathan 1993). Nevertheless, LR is more able than MH to detect non-uniform DIF, as LR receives more

information from the continuous ability measure than MH receives from the categorical ability measure. A third and final consideration is the complexity of the method in question. Up until recently, LR has been seen as more computationally complex than MH, making it a less-desirable option for practitioners to use. However, with an increased accessibility to statistical data software packages (e.g., IBM SPSS), LR is becoming a more feasible option. Modern statistical software can compute LR statistics by referencing a data set, where MH still requires the researcher to manually determine and categorize the number of ability parameters for the study (Mazor, Kanjee, & Clauser 1995).

For the purposes of this study, LR was the preferred method of DIF detection. It should be noted that many studies include multiple methods in detecting DIF (both CTT-based and IRT) and best practices suggests utilizing multiple methods to detect DIF, given the various limitations and considerations mentioned above.

Purpose of the Present Study

The wide-scale adoption of the 80% Rule in examination units across the State can in part be attributed to the simplicity of the analysis. While the advantages of the 80% Rule's simplicity are readily observable, the straightforwardness of the analysis lends itself to some glaring limitations. Specifically, the 80% Rule does not provide test developers insight on potential bias at the item-level. The rule employs a comparison of group total success rates and overlooks potential sources of biases within the items themselves. It also presumes that the groups are comparable in their levels of ability measured by the test, and this assumption may not be warranted in some instances.

DIF analysis addresses group differences when candidates are equated for level of ability (a more direct way to address test fairness) in an attempt to identify potentially biased exam items. Many DIF analyses can be used to identify these biases, however, LR has seen an increase in use in recent years (Hosmer & Lemeshow, 2000) and is noted in literature as one of the preferred methods in detecting DIF (Duncan 2009; Wang & Bian, 2014). The purpose of this study is to evaluate the effects of supplementing the analyses of previously-administered State-level examinations using the 80% Rule with Uniform and Non-Uniform DIF indices through use of LR analysis. Sample data provided by a large western State agency will be referenced during this study. It is expected that, through a consideration of current and available diagnostic methods, a more informed decision can be made with respect to the issue of adverse impact.

CHAPTER 3

METHOD

Sample Description

The archival data utilized in this study were provided by a State Agency located on the West Coast of the United States. A total of 1,517 job applicants across the State participated in the selection measure. To ensure confidentially, the testing measure name, candidate names, and State Agency name and location have been excluded from this study. However, the classification was identified as the Cashier Classification to allow readers a general understanding of the work produced by the position type. Demographic information was collected on a voluntary basis for all candidates who participated in the study. As shown in Table 1, 32.7 % males and 66.5% females participated in the study, 0.8% declined to state. In regard to ethnicity, individuals self-identified as Black/African-American (56.4%), Asian (13.5%), Hispanic (7.7%), Caucasian (7.4%), other (4.7%), Filipino (3.0%), Pacific Islander (1.9%), and American Indian/Alaskan Native (0.9%), and 4.5 % chose "Decline to state." In regard to age group, individuals self-identified as between the ages of 22-39 years (62.7%), 40-69 years (30.2%), less than 21 years (6.2%), over 70 years (0.1%), or declined to state (0.8%).

Instrument

Individuals interested in applying to vacancies associated with the Cashier classification (a classification solely utilized within a single State department) are required to participate in a multiple-choice selection measure. This test was designed to

measure mathematical skills as it pertains to cashiering, a core-competency identified through several knowledge, skills, and abilities found with a content-validated job analysis for the specified classification. There were a total of 25 multiple-choice items. Test items were scored 1 for the correct answer and 0 for an incorrect answer. No partial points were awarded. The mean test score was 22.241 with a standard deviation of 3.476. The reliability coefficient (internal consistency) Cronbach's Alpha was .842. Individual item means and standard deviations were also collected and can be found in Appendix A. Table 2 describes the mean test scores and standard deviations by demographic group.

Table 1

Test Taker Sample Size by Demographic Group

Demographic Group	*N*	%
Gender		
Male	496	32.7
Female	1009	66.5
Ethnicity		
American Indian/Alaskan	14	0.9
Asian	205	13.5
Black/African-American	855	56.4
Caucasian	112	7.4
Filipino	44	3.0
Hispanic	117	7.7
Pacific Islander	29	1.9
Other	72	4.7
Age Group		
Less than 21 years old	94	6.2
22-39 years old	951	62.7
40-69 years old	458	30.2
Over 70 years old	2	0.1

Table 2

Test Taker Mean Score & Standard Deviation by Demographic Group

Demographic Group	*N*	*M*	*SD*
Gender			
Male	496	22.635	3.199
Female	1009	22.053	3.564
Ethnicity			
American Indian/Alaskan	14	20.571	4.452
Asian	205	23.722	1.619
Black/African-American	855	21.713	3.736
Caucasian	112	23.089	2.336
Filipino	44	23.045	3.289
Hispanic	117	22.256	3.518
Pacific Islander	29	22.793	2.857
Other	72	22.300	3.871
Age Group			
Less than 21 years old	94	22.074	3.139
22-39 years old	951	22.043	3.667
40-69 years old	458	22.668	3.075
Over 70 years old	2	24.000	0.000

Procedure

Two analyses were performed for the Cashiering examination: 80 percent rule analysis and LR analysis method. It is acknowledged that the 80 percent rule is neither an assessment of Uniform DIF nor Non-Uniform DIF. Rather, it is a ratio comparison between the reference group and the focal group. Therefore, the purpose of this study is to evaluate how the LR method supplements the 80 percent rule in the interpretation of adverse impact.

80 Percent Rule Evaluation

Demographic data were grouped into gender, ethnicity, and age group. Within each of the groupings, test scores from the focal groups were compared to the respective reference group to evaluate whether a violation of the 80 percent rule had occurred. For gender, the Male group was considered the reference group (Biddle, 2015) and the data were compared to the focal group, Females. For ethnicity, the Caucasian group was considered the reference group (Biddle, 2015) and the data were compared to the Asian, Black/African American, and Hispanic focal groups. The American Indian/Alaskan, Filipino, and Pacific Islander focal groups were not compared to the reference group as the data sets were less than 100 participants and any percentage violations found would likely be due to chance alone. For age group, the less than 21 years old group and 22-39 year group were combined to create the reference group, which was compared to the 40-69 group and 70+ group, combined to create the 40+ focal group. The choice to utilize the groupings in this manner was based on the Age Discrimination in Employment Act of 1967, identifying individuals 40+ years as a protected class.

Within each of the groupings, if a difference greater than 80 percent occurred between the pass rate of the focal group and reference group, the focal group was flagged for a violation. The following formula was utilized to determine this difference:

$$\frac{Passed\ exams\ of\ focal\ group}{Total\ exams\ (N)\ of\ focal\ group} \Big/ \frac{Passed\ exams\ of\ reference\ group}{Total\ exams\ (N) of\ reference\ group}$$

Analysis for Item Removal

For each of the demographic subgroups (gender, ethnicity, and age group) producing a sufficient subject size, the LR method was performed on all available reference/focal comparisons, resulting in a total of five distinct evaluations of DIF per item (a total of 125 separate DIF assessments at 25 items with five evaluations each). For each assessment, a three-step LR method was performed. In the first step, ability level (represented as total test score) was entered into the model as a continuous variable. The second step entered into the model was group membership. The third and final step entered into the model was the interaction between ability and group membership.

For each of the steps in the LR Method, a Chi-square significance test was performed. In the first step, an evaluation of test score (ability level) and item performance was assessed. In the second step, uniform DIF was evaluated by reviewing whether the addition of the group membership variable yielded in a significant Chi-Square. In the third step, Non-Uniform DIF was evaluated by reviewing whether the addition of the interaction between group membership and test score variable yielded a significant Chi-square.

If statistical significance was found in the third step, the effect of Non-Uniform DIF was further evaluated by reviewing the variance value difference between step one

and step three. If significance was not found in the third step, the effect of Uniform DIF was evaluated by reviewing the estimated explained variance value differences between step one and step two. For the LR Method, the Nagelkerke R^2 represented estimated explained variance. The parameters determined for Nagelkerke R^2 were supported by previous research on DIF analysis with the LR method (Shimizu and Zumbo 2005) and are as follows:

1. Values between the effect sizes less than .035 was considered a negligible effect.
2. Values between the effect sizes less than .07, greater than .035, with a Chi square of .025 were considered a moderate effect.
3. Values between the effect sizes greater than .07, with a Chi square of .025 was considered a large effect.

Chapter 4

ANALYSIS OF THE DATA

Descriptive Statistics

Descriptive statistics for the 25 item test were reviewed. The mean test score for all items was 22.241 with a standard deviation of 3.477. The reliability coefficient (internal consistency) Cronbach's Alpha was .842.

Descriptive statistics (subject size, mean score, and standard deviation) and reliability coefficients were provided for each of the comparison groups in Table 3 (Gender [*Male* vs. *Female*], Ethnicity [*Caucasian* vs. *African-American, Caucasian* vs. *Hispanic*, and *Caucasian* vs. *Asian*], and Age Group [*Over 40* and *Under 40*]. All groups met a reliability coefficient above the .700 cut-off except for the *Asian* demographic group (α = .617). According to California Department of Human Resources (CalHR), a minimal reliability coefficient of .700 is required for a scale to be considered reliable and consistent (Developing, Using, and Evaluating Written Examinations, 2017).

Table 3

Test Taker Mean Score, Standard Deviation, and Reliability Coefficient by Demographic Group for 25-Item Test

Demographic Group	*N*	*M*	*SD*	*A*
Gender				
Male	496	22.635	3.202	.837
Female	1009	22.053	3.566	.841
Ethnicity				
Asian	205	23.722	1.622	.617
African-American	855	21.713	3.738	.841
Caucasian	112	23.089	2.346	.730
Hispanic	117	22.256	3.533	.849
Age Group				
Under 40 (Combined)	1045	22.046	3.625	.847
Over 40 (Combined)	460	22.670	3.066	.824

Adverse Impact Analyses

In order to perform the 80 Percent Rule evaluation on each of the five comparison groups, a pass rate cut-off score must ordinarily be identified for the Cashier examination. As the purpose of this study is not to determine the minimally acceptable competency of candidates for the Cashiering position, all possible cut-off scores were

explored to perform adverse impact assessments for each of the five comparison groups. A display of counts for each of the comparison groups' cut-off pass rates are available in Appendix B.

Of the pass rates tested for each comparison group at each possible cut-off score, only two comparison groups (African-American/Caucasian and Hispanic/Caucasian) were found to contain 80 percent rule violations for the 25 item test, as presented in Table 4. For the African-American/Caucasian comparison group, violations were found at the 23 and 24 point cut-off. For the Hispanic/Caucasian group, one violation was found at the 24 point cut-off.

Table 4

Adverse Impact Table for 25-Item Test

Cut Score	*Group Comparison*	*FG Pass*	*RF Pass*	*FG Pass/Fail*	*RF Pass/Fail*	*FG/RF*
23	African-American/ Caucasian	500	82	.585	.732	.799
24	African-American/ Caucasian	373	67	.436	.598	.729
24	Hispanic/ Caucasian	55	67	.470	.598	.786

Note: *FG Pass* = Number of individuals in focal group who passed, *RF Pass* = Number of individuals in reference group who passed. *FG Pass/Fail*= Pass rate of focus group. *RG Pass/Fail*= Pass rate of reference group. *FG/RF* = Pass rate of focus group over pass rate of reference group (80 percent rule violation check).

Logistic Regression

The 25 items in the test were assessed for DIF using the Logistic Regression procedure, using the five comparison groups: *Male* vs. *Female*, *Caucasian* vs. *African-*

American, *Caucasian* vs. *Hispanic*, and *Caucasian* vs. *Asian*, and *Under 40* vs. *Over 40.* Of the 125 assessments, 6.4% (8) were classified as indicating Non-Uniform DIF and 10.4% (13) were classified as indicating Uniform DIF. Of the assessments indicating Non-Uniform DIF, 62.5% (5) were classified as negligible effects, 12.5% (1) classified as a moderate effect, and 25.0% (2) classified as large effects, as indicated in Table 5. Of the assessments indicating Uniform DIF, 61.5% (8) were classified as negligible effects, 7.7% (1) was classified as a moderate effects, and 30.8% (4) were considered large effects, as indicated in Table 6. The Nagelkerke R^2 values and DIF classification category of each item and comparison group are displayed in Appendix C.

Table 5

Logistic Regression Non-Uniform DIF Classifications by Comparison Group for 25-Item Test

Comparison Group	Classification Category		
	Negligible	*Moderate*	*Large*
Caucasian/Asian	0	1	0
Caucasian/African-American	3	0	0
Caucasian/Hispanic	0	0	2
Under 40/Over 40	1	0	0
Male/Female	1	0	0
Grand Total	5	1	2

Table 6

Logistic Regression Uniform DIF Classifications by Comparison Group for 25-Item Test

Comparison Group	Classification Category		
	Negligible	*Moderate*	*Large*
Caucasian/Asian	0	1	3
Caucasian/African-American	2	0	0
Caucasian/Hispanic	2	0	1
Under 40/Over 40	4	0	0
Male/Female	0	0	0
Grand Total	8	1	4

Of the 25 items in the test, 12 items displayed Non-Uniform or Uniform DIF (items 4, 6, 8, 9, 11, 13, 14, 18, 19, 23, 24, and 25). Of these items, only six items displayed moderate to large DIF effect sizes (items 6, 8, 9, 11, 19, and 24), as presented in Table 7.

Table 7

Logistic Regression Uniform and Non-Uniform DIF Classifications by Item Number for 25-Item Test

Comparison Group	Classification Category					
	Uniform DIF			Non-Uniform DIF		
	Negligible	*Moderate*	*Large*	*Negligible*	*Moderate*	*Large*
Item 4	1	0	0	0	0	0
Item 6	1	0	2	0	0	1
Item 8	0	0	1	1	0	0
Item 9	0	0	1	0	0	0
Item 11	1	0	0	0	1	0
Item 13	2	0	0	0	0	0
Item 14	0	0	0	1	0	0
Item 18	0	0	0	1	0	0
Item 19	1	0	0	2	0	1
Item 23	1	0	0	0	0	0
Item 24	0	1	0	0	0	0
Item 25	1	0	0	0	0	0
Grand Total	8	1	4	5	1	2

Descriptive Statistics for Shortened Test

Results of the LR test indicated moderate to large DIF in nine items. According *to ETS Standards for Quality of Fairness* (2002), it is stated that items indicating moderate to large levels of DIF impact the integrity of an examination and should be removed when possible. After removing the six items displaying moderate to large DIF effect sizes (Items 6, 8, 9, 11, 19, and 24), descriptive analyses were performed to reassess the comparison groups for the remaining 19 items in the test (Items 1, 2, 3, 4, 5, 7, 10, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 25). This resulted in a mean of 17.026, a standard deviation of 2.780, and a reliability coefficient of .823 for the 1,517 test takers.

Descriptive statistics (subject size, mean score, and standard deviation) and reliability coefficients were provided for each of the comparison groups in Table 8 (Gender [*Male* vs. *Female*], Ethnicity [*Caucasian* vs. *African-American*, *Caucasian* vs. *Hispanic*, and *Caucasian* vs. *Asian*], and Age Group [*Over 40* and *Under 40*]. All groups met the reliability coefficient criteria for the .700 cut-off except the *Asian* demographic group (α = .555).

Table 8

Test Taker Mean Score, Standard Deviation, and Reliability Coefficient by Demographic Group for Shortened Test

Demographic Group	*N*	*M*	*SD*	*A*
Gender				
Male	496	17.347	2.524	.813
Female	1009	16.874	2.854	.821
Ethnicity				
Asian	205	18.288	1.188	.555
African-American	855	16.593	3.005	.822
Caucasian	112	17.741	1.864	.706
Hispanic	117	16.966	2.776	.816
Age Group				
Under 40 (Combined)	1045	16.864	2.899	.828
Over 40 (Combined)	460	17.389	2.430	.799

Logistic Regression of Shortened Test

Of the 125 assessments for the shortened test, 4.0% (5) were classified as indicating Non-Uniform DIF and 3.2% (4) were classified as indicating Uniform DIF. Of the assessments indicating Non-Uniform DIF, 60.0% (3) were classified as negligible effects, 40.0% (2) classified as moderate effects, and 0.0% (0) classified as large effects, as indicated in Table 9. Of the items indicating Uniform DIF, 100.0% (4) were classified as negligible effects and none were classified as moderate or large effects, as indicated in

Table 10. The Nagelkerke R^1 values and DIF classification category of each item and comparison group are displayed in Appendix D.

Table 9

Logistic Regression Non-Uniform DIF Classifications by Comparison Group for Shortened Test

Comparison Group	Classification Category		
	Negligible	*Moderate*	*Large*
Caucasian/Asian	0	2	0
Caucasian/African-American	1	0	0
Caucasian/Hispanic	0	0	0
Under 40/Over 40	1	0	0
Male/Female	1	0	0
Grand Total	3	2	0

Table 10

Logistic Regression Uniform DIF Classifications by Comparison Group for Shortened Test

Comparison Group	Classification Category		
	Negligible	*Moderate*	*Large*
Caucasian/Asian	0	0	0
Caucasian/African-American	0	0	0
Caucasian/Hispanic	0	0	0
Under 40/Over 40	4	0	0
Male/Female	0	0	0
Grand Total	4	0	0

Of the 19 items in the shortened test, 9 items displayed Non-Uniform or Uniform DIF (items 4, 12, 13, 14, 18, 23, 25). Of these items, only two items displayed moderate to large DIF effect sizes (items 13 and 23), as presented in Table 11.

Table 11

Logistic Regression Uniform and Non-Uniform DIF Classifications by Item Number for Shortened Test

Comparison Group	Classification Category					
	Uniform DIF			Non-Uniform DIF		
	Negligible	*Moderate*	*Large*	*Negligible*	*Moderate*	*Large*
Item 4	1	0	0	0	0	0
Item 12	1	0	0	0	0	0
Item 13	0	0	0	1	1	0
Item 14	0	0	0	1	0	0
Item 18	0	0	0	1	0	0
Item 23	1	0	0	0	1	0
Item 25	1	0	0	0	0	0
Grand Total	4	0	0	3	2	0

Adverse Impact Analyses for Shortened Test

A display of counts for each of the comparison groups' cut-off pass rates for the shortened test are available in Appendix E. Of the pass rates tested for each comparison group at each possible cut-off score, only one comparison group (African-American/Caucasian) at one item (cut-score 18) was found to obtain an 80 percent rule violation in the 19-item test, as presented in Table 12.

Table 12

Adverse Impact Table for Shortened Test

Cut Score	*Group Comparison*	*FG Pass*	*RF Pass*	*FG Pass/Fail*	*RF Pass/Fail*	*FG/RF*
18	African-American/ Caucasian	463	77	.54	.69	.79

Note: *FG Pass* = Number of individuals in focal group who passed, *RF Pass* = Number of individuals in reference group who passed. *FG Pass/Fail*= Pass rate of focus group. *RG Pass/Fail*= Pass rate of reference group. *FG/RF* = Pass rate of focus group over pass rate of reference group (80 percent rule violation check).

Chapter 5

FINDINGS AND INTERPRETATION

Findings and Conclusion

The Adverse Impact analysis of the 25-item test indicated that two comparison groups exhibited 80 Percent Rule violations: the Hispanic group compared to the Caucasian group and the African-American group compared to the Caucasian group. When the same set of items were assessed with the LR analysis, violations at the item level were found with the Hispanic group compared to the Caucasian group, but no violations (beyond negligible violations) were found with the African-American group compared to the Caucasian group. Additionally, the LR analysis found multiple violations with the Asian group compared to the Caucasian group in the 25-item test, whereas the Adverse Impact analysis did not. Differences between the two analyses also occurred in the shortened 19-item test: the Adverse Impact analysis indicated a violation with the African-American group compared to the Caucasian group, where the LR analysis indicated a violation with the Asian group compared to the Caucasian group.

It is important to note the differences between the results of the Adverse Impact analyses and the results of the LR analyses. Both procedures aim to detect performance differences between the comparison groups, however, LR analysis provides an additional frame of reference through statistical significance and the control of ability level to detect issues identified at the item level. This alternative perspective created an expanded assessment of violations that the Adverse Impact analysis could not fulfill (due to the simplicity of the analysis). As such, the LR analysis allowed for (1) a confirmation of

violations found in the Adverse Impact assessment, (2) the ability to negate violations identified in the Adverse Impact assessment (attributable to non-significant or negligible effects), and (3) the identification of alternative violations not found in the initial analysis (Roth, Bobko, and Switzer 2006).

Residual DIF after Item Removal

Overall, the findings in this study match those found in similar studies utilizing LR to detect DIF (Cuevas & Cervantes, 2012). DIF analyses assist in the identification of true testing bias and the current study provides evidence that the elimination of items displaying DIF in the 25-item test minimized the amount of error (violations) found in the 19-item shortened test.

However, it is important to acknowledge the residual violations found in the shortened test as indicated in both the LR analysis and Adverse Impact analysis. Specifically, after six items were identified as displaying significant moderate to large effects of DIF, the items were removed from the test and the shortened test was reassessed through both an LR analysis and Adverse Impact analysis to confirm the effectiveness of the parameters for item elimination, as suggested by Shimizu and Zumbo (2005). After removal of the items displaying DIF, two additional items displayed DIF in the LR analysis (the original test displaying six violations) for the shortened test and the Adverse Impact Assessment indicated one violation for one comparison group (the original test displaying three violations for two comparison groups).

After a thorough review of the relevant literature, no studies approached the reassessment of items displaying DIF after an initial removal. However, several studies

suggest applying multiple rounds of DIF analysis to ensure the proper elimination of items displaying DIF have occurred (Penfield & Lee, 2010). Simultaneously, efforts to eliminate DIF must represent a compromise with the length of the exam, as content covered by the exam items needs to be sufficient to ensure proper coverage of topics tested as well as ensuring the overall validity of the examination. However, no specific standard exists with regard to the number of items required for a particular scale (some constructs can even be effectively identified with as few as 3 items; Netemeyer, Sharma & Bearden 2003). The current study had 19 items remaining after the initial DIF removal and, although two items displayed residual DIF, no further items were considered for removal as it was determined that any additional items removed would adversely impact thereliability and construct validity of the test.

Limitations

The first limitation identified in the study was the varying sizes of the 16 demographic groups. Prior to conducting each of the analyses, groups with fewer than 100 candidates were eliminated from the study to avoid Type I (the claim that DIF exists when it does not) and Type II (the claim that DIF does not exist when it does) errors. This led to the immediate elimination of three ethnic groups: Filipino, Pacific Islander, and American Indian/Alaskan Native. Groups identified as “Other” or “Declined to state” were also removed from the study due to small group sizes and the inability to categorize the groups into a specific demographics (as ethnicity, age, and gender were not specified in the two selections).

After eliminating the groups with less than 100 candidates, ten demographic groups remained, varying between 112 and 1009 subjects. In addition to small subject sizes, groups varying greatly in size in a particular analysis can also lead to Type I error (Rusticus & Lovato, 2014). Therefore, when possible, similar groups were evaluated to see if they were eligible to be dummy coded into new focus/reference groups to allow for more even group sizes within a particular analysis (Schenker & Raghunathan, 2007). For this study, age was combined from four groups (less than 21, 22-39, 40-69, and over 70 years) that varied in size to two more similar sized groups (under 40, over 40). No other groups were deemed appropriate to combine in the study.

The result of eliminating and combining the demographic groups led to a smaller window (eight demographic groups) for analyzing the data. Additionally, several of the remaining groups still varied greatly in size in comparison to one another. The variance in group size and limited data can be in part attributed to how the examination was advertised. The positions advertised for this classification were only available within a particular State County. Additionally, the classification was defined as entry level. As both location and job type are factors found to attract specific demographic groups (May 2016 Occupation Profiles, 2017), it can be said that the external validity of the study may have been impacted by the available candidate pool.

A second limitation of the study was the absence evaluating why items displayed DIF, beyond the statistical analysis. According to Shimzu and Zumbo (2005), items with effect size differences greater than .35 (with a Chi Square less than .025), between either Block 1 and Block 2 (Uniform DIF) or Block 1 and Block 3 (Non-Uniform DIF) were

identified as indicating moderate to large DIF and were removed from the original 25-item examination. However, no additional review occurred to determine whether the items displaying DIF were truly detecting differences in performance while controlling for ability, or if the significant finding was due to an extraneous factor. For example, it is possible that DIF can be detected in an item, but the performance differences might have been due to variability in the style of how the question was structured, rather than an actual performance difference between the demographic groups (Longford, Holland, & Thayer 1993). As noted above, residual significant DIF was found in the shortened test, even after item removal – this may be partially due to unnecessary removal of items in the original test, as the test questions were not further evaluated by Subject Matter Experts for a true need for deletion based on performance differences alone (Scheuneman & Slaughter, 1991).

Another limitation of the study was the lack of identifying a cut-score prior to performing the Adverse Impact assessment. When developing an examination, a thorough assessment is normally done to determine a proper pass point cut-off score for the measure. For this study, a cut-off score was not determined, making it difficult to identify whether an Adverse Impact assessment violation through the 80 Percent Rule was a concern. For example, in the shortened test (19 questions), an 80 Percent Rule Violation occurred at the 18-point pass point cut-off (95 percent pass rate). The higher the pass point cut-off, the more difficult the examination is to pass, creating a smaller subject pool. Since the pass rate was not identified, it is unknown whether a violation at the 18-point cut-off is justified (i.e., individuals should have, on average, received at least

18 points on the examination) or whether the cut-off was too high, creating an unrealistic reference point for determining adverse impact (although it can be assumed that a pass rate of 95 percent is rather high and likely not a realistic cut-off point for a standard examination).

Implications for Future Studies

The current study compared adverse impact analysis through the 80 Percent Rule Assessment to LR. However, as identified in Chapter 2, several alternatives such as Item Response Theory (IRT) and Mantel-Haenszel Chi Square could also provide useful information, in addition to the LR analysis on the interpretation of Adverse Impact Assessments. Future studies should explore these alternative methods of DIF on interpreting the 80 Percent Rule and evaluate the benefits that each method provides in interpreting a potential violation.

Additionally, future studies should expand on whether residual DIF occurs similarly throughout the various methods of DIF, after items have been removed initially. Most studies comparing the methods of DIF on a data set focus on the identification of DIF, rather than comparing the effectiveness of removing poorly performing items through a second assessment (shortened test), as seen in this study.

The strategies involved in the removal of the items based on DIF parameters should also be explored, beyond the criteria presented by Shimzu and Zumbo (2005). In this particular study, only 25 items were present in the test, limiting the ability to explore the effects of DIF removal to a minimum (as each DIF deletion led to a substantial loss in the test content and meaning of the construct; Hambleton, 2006). Future studies should

explore a longer test, where options to evaluate the need to delete items performing differentially can be explored.

Appendix A

Item Means and Standard Deviations

Item #	Mean	SD
Item 1	.963	.189
Item 2	.950	.218
Item 3	.875	.330
Item 4	.943	.233
Item 5	.925	.264
Item 6	.984	.125
Item 7	.980	.139
Item 8	.926	.262
Item 9	.933	.251
Item 10	.825	.380
Item 11	.645	.479
Item 12	.974	.160
Item 13	.794	.404
Item 14	.886	.318
Item 15	.970	.173
Item 16	.898	.303
Item 17	.877	.328
Item 18	.890	.313
Item 19	.883	.321
Item 20	.904	.294
Item 21	.914	.281
Item 22	.875	.330
Item 23	.778	.416
Item 24	.843	.364
Item 25	.805	.400

Appendix B

Number of Candidates Passing at Cut-Off Score Level by Test and Comparison Group

Original Test Cut-off Score	Male	Female	Age (Under 40)	Age (40+)	Caucasian	Hispanic	African American	Asian
6	495	1007	1044	458	112	117	852	205
7	495	1007	1044	458	112	117	852	205
8	494	1005	1041	458	112	116	850	205
9	492	1000	1035	456	112	115	844	205
10	492	997	1033	455	112	114	843	205
11	491	992	1027	455	112	114	839	205
12	488	986	1020	453	112	114	832	205
13	486	982	1015	452	112	114	829	205
14	481	969	999	450	111	112	817	205
15	476	953	985	443	111	110	801	205
16	469	934	963	439	110	109	780	204
17	465	917	946	435	109	109	765	203
18	459	892	921	429	107	108	740	202
19	448	871	897	422	105	105	717	202
20	433	829	848	414	103	99	675	200
21	417	785	808	395	98	96	632	198
22	390	730	754	367	93	88	581	190
23	355	628	667	316	82	80	500*	168
24	281	485	507	258	67	55*	373*	146

Note: No candidates obtained a failing score below the 6 point cut-off for the 25 item test. Focus group counts marked with * indicate a group violating the 80 percent rule when compared to its respective reference group.

Appendix C

Nagelkerke R^2 Values and DIF Classification Category by Item and Comparison Group

Item #	Comparison Group	Block 1 Nagelkerke R^2	Block 2 Nagelkerke R^2	Block 3 Nagelkerke R^2	Block 2 Significance Level	Block 3 Significance Level	Classification Category
1	African American/ Caucasian	0.0740	0.0740	0.0740	0.745	0.874	N/N
1	Asian/ Caucasian	0.0820	0.0840	0.0980	0.730	0.329	N/N
1	Hispanic/ Caucasian	0.0050	0.0070	0.0180	0.805	0.433	N/N
1	Female/ Male	0.0760	0.0760	0.0780	0.998	0.443	N/N
1	Over 40/ Under 40	0.0740	0.0740	0.0760	0.694	0.371	N/N
2	African American/ Caucasian	0.2000	0.2000	0.2020	0.993	0.437	N/N
2	Asian/ Caucasian	0.1850	0.1900	0.1930	0.602	0.665	N/N
2	Hispanic/ Caucasian	0.1910	0.1910	0.1980	0.843	0.475	N/N
2	Female/ Male	0.2150	0.2150	0.2160	0.645	0.383	N/N
2	Over 40/ Under 40	0.2180	0.2210	0.2210	0.200	0.892	N/N
3	African American/ Caucasian	0.1760	0.1760	0.1800	0.831	0.172	N/N
3	Asian/ Caucasian	0.2000	0.2000	0.2050	0.896	0.388	N/N
3	Hispanic/ Caucasian	0.2430	0.2430	0.2460	0.991	0.555	N/N
3	Female/ Male	0.2050	0.2050	0.2050	0.892	0.882	N/N
3	Over 40/ Under 40	0.2080	0.2130	0.2140	0.039	0.321	N/N
4	African American/ Caucasian	0.1330	0.1330	0.1330	0.871	0.790	N/N
4	Asian/ Caucasian	0.1520	0.1520	0.1790	0.981	0.128	N/N
4	Hispanic/ Caucasian	0.1940	0.1990	0.2000	0.546	0.737	N/N
4	Female/ Male	0.1560	0.1590	0.1600	0.157	0.708	N/N
4	Over 40/ Under 40	0.1630	0.1790	0.1840	0.002*	0.100	N/N
5	African American/ Caucasian	0.2850	0.2870	0.2890	0.359	0.303	N/N
5	Asian/ Caucasian	0.1500	0.1510	0.1600	0.825	0.444	N/N
5	Hispanic/ Caucasian	0.4070	0.4070	0.4080	0.941	0.802	N/N
5	Female/ Male	0.2880	0.2880	0.2890	0.766	0.531	N/N
5	Over 40/ Under 40	0.2940	0.2950	0.2950	0.382	0.540	N/N
6	African American/ Caucasian	0.2190	0.2490	0.2490	0.019*	0.781	N/N

Item #	Comparison Group	Block 1 Nagelkerke R^2	Block 2 Nagelkerke R^2	Block 3 Nagelkerke R^2	Block 2 Significance Level	Block 3 Significance Level	Classification Category
6	Asian/ Caucasian	0.1670	0.3190	0.3190	0.012*	1.000	L/N
6	Hispanic/ Caucasian	0.2990	0.3990	0.4930	0.021*	0.023*	L/L
6	Female/ Male	0.2860	0.2930	0.3030	0.204	0.130	N/N
6	Over 40/ Under 40	0.2670	0.2680	0.2790	0.581	0.112	N/N
7	African American/ Caucasian	0.1210	0.1370	0.1410	0.080	0.417	N/N
7	Asian/ Caucasian	0.1310	0.1860	0.1880	0.102	0.706	N/N
7	Hispanic/ Caucasian	0.2300	0.3060	0.3100	0.045	0.682	N/N
7	Female/ Male	0.1530	0.1530	0.1610	0.911	0.127	N/N
7	Over 40/ Under 40	0.1550	0.1570	0.1590	0.463	0.384	N/N
8	African American/ Caucasian	0.2850	0.2900	0.3010	0.104	0.025*	N/N
8	Asian/ Caucasian	0.2520	0.3760	0.3790	0.000*	0.538	L/N
8	Hispanic/ Caucasian	0.3860	0.4030	0.4480	0.223	0.051	N/N
8	Female/ Male	0.2740	0.2740	0.2760	0.990	0.230	N/N
8	Over 40/ Under 40	0.2700	0.2700	0.2700	0.735	0.781	N/N
9	African American/ Caucasian	0.3070	0.3170	0.3190	0.030	0.390	N/N
9	Asian/ Caucasian	0.1540	0.2430	0.2860	0.009*	0.071	L/N
9	Hispanic/ Caucasian	0.3900	0.3990	0.4220	0.488	0.265	N/N
9	Female/ Male	0.3100	0.3110	0.3120	0.352	0.350	N/N
9	Over 40/ Under 40	0.3160	0.3220	0.3220	0.044	0.956	N/N
10	African American/ Caucasian	0.3080	0.3090	0.3090	0.407	0.455	N/N
10	Asian/ Caucasian	0.2050	0.2110	0.2120	0.294	0.686	N/N
10	Hispanic/ Caucasian	0.2600	0.2600	0.2620	0.868	0.549	N/N
10	Female/ Male	0.3210	0.3220	0.3220	0.189	0.729	N/N
10	Over 40/ Under 40	0.3260	0.3280	0.3290	0.169	0.224	N/N
11	African American/ Caucasian	0.0850	0.0880	0.0900	0.168	0.181	N/N
11	Asian/ Caucasian	0.1930	0.2030	0.2540	0.086	0.000*	N/M
11	Hispanic/ Caucasian	0.0950	0.1020	0.1070	0.268	0.370	N/N
11	Female/ Male	0.0920	0.0930	0.0940	0.157	0.380	N/N
11	Over 40/ Under 40	0.0920	0.1050	0.1090	0.000*	0.033	N/N
12	African American/ Caucasian	0.3460	0.3460	0.3470	0.731	0.695	N/N

Item #	Comparison Group	Block 1 Nagelkerke R^2	Block 2 Nagelkerke R^2	Block 3 Nagelkerke R^2	Block 2 Significance Level	Block 3 Significance Level	Classification Category
12	Asian/ Caucasian	0.1090	0.1170	0.1200	0.629	0.726	N/N
12	Hispanic/ Caucasian	0.3820	0.3820	0.3850	0.977	0.767	N/N
12	Female/ Male	0.3710	0.3720	0.3720	0.772	0.743	N/N
12	Over 40/ Under 40	0.3670	0.3830	0.3850	0.017	0.448	N/N
13	African American/ Caucasian	0.3250	0.3300	0.3310	0.044*	0.706	N/N
13	Asian/ Caucasian	0.1660	0.1830	0.2020	0.092	0.064	N/N
13	Hispanic/ Caucasian	0.2640	0.2940	0.2970	0.026*	0.533	N/N
13	Female/ Male	0.3340	0.3340	0.3340	0.448	0.774	N/N
13	Over 40/ Under 40	0.3350	0.3350	0.3370	0.616	0.091	N/N
14	African American/ Caucasian	0.1570	0.1570	0.1610	0.873	0.135	N/N
14	Asian/ Caucasian	0.1210	0.1220	0.1230	0.711	0.715	N/N
14	Hispanic/ Caucasian	0.1950	0.1950	0.2030	0.915	0.296	N/N
14	Female/ Male	0.1580	0.1590	0.1650	0.452	0.031*	N/N
14	Over 40/ Under 40	0.1610	0.1620	0.1630	0.411	0.324	N/N
15	African American/ Caucasian	0.2280	0.2300	0.2330	0.466	0.375	N/N
15	Asian/ Caucasian	0.0110	0.0180	0.0180	0.535	0.988	N/N
15	Hispanic/ Caucasian	0.1430	0.1450	0.1550	0.770	0.455	N/N
15	Female/ Male	0.1670	0.1710	0.1710	0.222	0.972	N/N
15	Over 40/ Under 40	0.1780	0.1800	0.1880	0.413	0.088	N/N
16	African American/ Caucasian	0.5440	0.5450	0.5450	0.338	0.772	N/N
16	Asian/ Caucasian	0.3640	0.3770	0.3800	0.288	0.604	N/N
16	Hispanic/ Caucasian	0.4180	0.4500	0.4530	0.044	0.505	N/N
16	Female/ Male	0.5200	0.5200	0.5200	0.517	0.802	N/N
16	Over 40/ Under 40	0.5230	0.5260	0.5270	0.085	0.338	N/N
17	African American/ Caucasian	0.5070	0.5080	0.5090	0.581	0.315	N/N
17	Asian/ Caucasian	0.3490	0.3490	0.3510	0.876	0.678	N/N
17	Hispanic/ Caucasian	0.4280	0.4280	0.4430	0.943	0.177	N/N
17	Female/ Male	0.4850	0.4890	0.4890	0.051	0.448	N/N
17	Over 40/ Under 40	0.4860	0.4870	0.4900	0.256	0.072	N/N
18	African American/ Caucasian	0.4100	0.4100	0.4260	0.918	0.001*	N/N

Item #	Comparison Group	Block 1 Nagelkerke R^2	Block 2 Nagelkerke R^2	Block 3 Nagelkerke R^2	Block 2 Significance Level	Block 3 Significance Level	Classification Category
18	Asian/ Caucasian	0.6530	0.6580	0.6600	0.495	0.623	N/N
18	Hispanic/ Caucasian	0.5280	0.5300	0.5760	0.598	0.013	N/N
18	Female/ Male	0.4700	0.4700	0.4710	0.380	0.630	N/N
18	Over 40/ Under 40	0.4630	0.4640	0.4650	0.419	0.280	N/N
19	African American/ Caucasian	0.4460	0.4490	0.4600	0.125	0.005*	N/N
19	Asian/ Caucasian	0.4970	0.4980	0.5000	0.677	0.563	N/N
19	Hispanic/ Caucasian	0.3610	0.3920	0.4670	0.046*	0.001*	N/L
19	Female/ Male	0.4700	0.4700	0.4730	0.846	0.099	N/N
19	Over 40/ Under 40	0.4640	0.4640	0.4720	0.792	0.003*	N/N
20	African American/ Caucasian	0.4390	0.4400	0.4410	0.412	0.399	N/N
20	Asian/ Caucasian	0.4630	0.4770	0.4940	0.329	0.288	N/N
20	Hispanic/ Caucasian	0.3290	0.3340	0.3540	0.521	0.184	N/N
20	Female/ Male	0.4530	0.4540	0.4540	0.368	0.779	N/N
20	Over 40/ Under 40	0.4570	0.4600	0.4600	0.110	0.652	N/N
21	African American/ Caucasian	0.6000	0.6020	0.6020	0.326	0.986	N/N
21	Asian/ Caucasian	0.4200	0.4230	0.4270	0.695	0.629	N/N
21	Hispanic/ Caucasian	0.3510	0.3680	0.3830	0.200	0.224	N/N
21	Female/ Male	0.5360	0.5400	0.5410	0.056	0.448	N/N
21	Over 40/ Under 40	0.5400	0.5400	0.5410	0.703	0.333	N/N
22	African American/ Caucasian	0.5120	0.5120	0.5140	0.658	0.302	N/N
22	Asian/ Caucasian	0.3910	0.4410	0.4450	0.038	0.572	N/N
22	Hispanic/ Caucasian	0.4070	0.4080	0.4240	0.689	0.137	N/N
22	Female/ Male	0.4790	0.4800	0.4800	0.415	0.754	N/N
22	Over 40/ Under 40	0.4810	0.4830	0.4830	0.283	0.899	N/N
23	African American/ Caucasian	0.3890	0.3920	0.3920	0.121	0.808	N/N
23	Asian/ Caucasian	0.2550	0.2560	0.2730	0.724	0.084	N/N
23	Hispanic/ Caucasian	0.3370	0.3420	0.3440	0.359	0.514	N/N
23	Female/ Male	0.3760	0.3760	0.3770	0.470	0.629	N/N
23	Over 40/ Under 40	0.3760	0.3810	0.3820	0.011*	0.432	N/N
24	African American/ Caucasian	0.2840	0.2840	0.2840	0.584	0.898	N/N

Item #	Comparison Group	Block 1 Nagelkerke R^2	Block 2 Nagelkerke R^2	Block 3 Nagelkerke R^2	Block 2 Significance Level	Block 3 Significance Level	Classification Category
24	Asian/ Caucasian	0.1660	0.2220	0.2250	0.009*	0.525	M/N
24	Hispanic/ Caucasian	0.3730	0.3730	0.3930	0.806	0.076	N/N
24	Female/ Male	0.3410	0.3430	0.3440	0.085	0.364	N/N
24	Over 40/ Under 40	0.3360	0.3390	0.3400	0.065	0.424	N/N
25	African American/ Caucasian	0.5500	0.5500	0.5520	0.845	0.118	N/N
25	Asian/ Caucasian	0.4760	0.4860	0.4870	0.180	0.825	N/N
25	Hispanic/ Caucasian	0.4590	0.4600	0.4800	0.672	0.055	N/N
25	Female/ Male	0.5500	0.5500	0.5530	0.762	0.056	N/N
25	Over 40/ Under 40	0.5510	0.5570	0.5570	0.005*	0.806	N/N

Appendix B Heading Descriptions

Block 1 Nagelkerke R^2 – The amount of variance accounted for in the first block of the logistic regression procedure. Displays the amount of variance accounted for by candidates' total test score.

Block 2 Nagelkerke R^2 – The amount of variance accounted for in the second block of the logistic regression procedure. Displays the amount of variance accounted for by candidates' total test score and group membership.

Block 3 Nagelkerke R^2 – The amount of variance accounted for in the first stage of the logistic regression procedure. Displays the amount of variance accounted for by candidates' total test score, group membership, and the interaction of total test score and group membership.

Block 2 Significance Level – The significance level of the Block 2 Nagelkerke R^2 value. * Denotes that the value is significant at the .025 level.

Block 3 Significance Level – The significance level of the Block 3 Nagelkerke R^2 value. * Denotes that the value is significant at the .025 level.

Classification Category- The categorization of variance value differences (Shimizu and Zumbo 2005) between either (1) Block 2 Nagelkerke R^2 and Block 1 Nagelkerke R^2 or (2) between Block 3 Nagelkerke R^2 and Block 1 Nagelkerke R^2. Adjustments between the effect sizes less than .035 was considered a Negligible effect (N), adjustments between the effect sizes less than .07, greater than .035, with a Chi square of .025 was considered a Moderate effect (M), and adjustments between the effect sizes greater than .07, with a Chi square of .025 was considered a Large effect (L). The first letter

represents the classification category for Uniform DIF, the second letter represents the classification category for Non-Uniform DIF.

Appendix D

Nagelkerke R^2 Values and DIF Classification Category by Item and Comparison Group for Shortened Test

Item #	Comparison Group	Block 1 Nagelkerke R^2	Block 2 Nagelkerke R^2	Block 3 Nagelkerke R^2	Block 2 Significance Level	Block 3 Significance Level	Classification Category
1	African American/ Caucasian	0.0840	0.0840	0.0840	0.793	0.965	N/N
1	Asian/ Caucasian	0.0810	0.0840	0.1010	0.704	0.283	N/N
1	Hispanic/ Caucasian	0.0080	0.0090	0.0160	0.828	0.510	N/N
1	Female/ Male	0.0880	0.0880	0.0890	0.949	0.405	N/N
1	Over 40/ Under 40	0.0840	0.0840	0.0880	0.744	0.201	N/N
2	African American/ Caucasian	0.2170	0.2170	0.2180	0.921	0.500	N/N
2	Asian/ Caucasian	0.1960	0.1990	0.2030	0.644	0.601	N/N
2	Hispanic/ Caucasian	0.2040	0.2050	0.2090	0.788	0.572	N/N
2	Female/ Male	0.2240	0.2240	0.2260	0.590	0.311	N/N
2	Over 40/ Under 40	0.2290	0.2320	0.2320	0.168	0.920	N/N
3	African American/ Caucasian	0.1790	0.1790	0.1820	0.792	0.189	N/N
3	Asian/ Caucasian	0.2030	0.2030	0.2130	0.785	0.236	N/N
3	Hispanic/ Caucasian	0.2700	0.2700	0.2700	0.882	0.812	N/N
3	Female/ Male	0.2130	0.2130	0.2130	0.924	0.734	N/N
3	Over 40/ Under 40	0.2160	0.2210	0.2220	0.032	0.263	N/N
4	African American/ Caucasian	0.1290	0.1290	0.1290	0.867	0.784	N/N
4	Asian/ Caucasian	0.1150	0.1150	0.1290	0.964	0.287	N/N
4	Hispanic/ Caucasian	0.2020	0.2070	0.2100	0.506	0.632	N/N
4	Female/ Male	0.1510	0.1550	0.1550	0.169	0.656	N/N
4	Over 40/ Under 40	0.1590	0.1760	0.1800	0.002*	0.134	N/N
5	African American/ Caucasian	0.2880	0.2900	0.2940	0.382	0.131	N/N
5	Asian/ Caucasian	0.1950	0.1950	0.2110	0.949	0.306	N/N
5	Hispanic/ Caucasian	0.4640	0.4650	0.4670	0.823	0.657	N/N

Item #	Comparison Group	Block 1 Nagelkerke R^2	Block 2 Nagelkerke R^2	Block 3 Nagelkerke R^2	Block 2 Significance Level	Block 3 Significance Level	Classification Category
5	Female/ Male	0.2940	0.2940	0.2940	0.802	0.710	N/N
5	Over 40/ Under 40	0.3000	0.3010	0.3020	0.304	0.380	N/N
7	African American/ Caucasian	0.1390	0.1570	0.1610	0.064	0.407	N/N
7	Asian/ Caucasian	0.1440	0.1960	0.2020	0.112	0.594	N/N
7	Hispanic/ Caucasian	0.2370	0.3170	0.3200	0.042	0.665	N/N
7	Female/ Male	0.1750	0.1750	0.1790	0.829	0.284	N/N
7	Over 40/ Under 40	0.1750	0.1770	0.1810	0.519	0.328	N/N
10	African American/ Caucasian	0.3140	0.3150	0.3170	0.356	0.315	N/N
10	Asian/ Caucasian	0.2230	0.2270	0.2290	0.406	0.613	N/N
10	Hispanic/ Caucasian	0.2640	0.2640	0.2680	0.997	0.392	N/N
10	Female/ Male	0.3260	0.3270	0.3270	0.189	0.826	N/N
10	Over 40/ Under 40	0.3310	0.3330	0.3350	0.155	0.147	N/N
12	African American/ Caucasian	0.3360	0.3360	0.3390	0.721	0.405	N/N
12	Asian/ Caucasian	0.1760	0.1890	0.1980	0.516	0.583	N/N
12	Hispanic/ Caucasian	0.3880	0.3880	0.4000	0.975	0.492	N/N
12	Female/ Male	0.3680	0.3680	0.3680	0.863	0.980	N/N
12	Over 40/ Under 40	0.3610	0.3780	0.3800	0.015*	0.465	N/N
13	African American/ Caucasian	0.3340	0.3380	0.3390	0.054	0.730	N/N
13	Asian/ Caucasian	0.1790	0.1990	0.2360	0.055	0.011*	N/M
13	Hispanic/ Caucasian	0.3110	0.3360	0.3430	0.040	0.254	N/N
13	Female/ Male	0.3500	0.3500	0.3500	0.467	0.736	N/N
13	Over 40/ Under 40	0.3500	0.3500	0.3550	0.667	0.020*	N/N
14	African American/ Caucasian	0.1840	0.1840	0.1870	0.749	0.200	N/N
14	Asian/ Caucasian	0.1570	0.1590	0.1610	0.526	0.612	N/N
14	Hispanic/ Caucasian	0.2230	0.2230	0.2270	0.962	0.488	N/N
14	Female/ Male	0.1860	0.1870	0.1950	0.384	0.007*	N/N
14	Over 40/ Under 40	0.1890	0.1900	0.1900	0.321	0.418	N/N

Item #	Comparison Group	Block 1 Nagelkerke R^2	Block 2 Nagelkerke R^2	Block 3 Nagelkerke R^2	Block 2 Significance Level	Block 3 Significance Level	Classification Category
15	African American/ Caucasian	0.2570	0.2600	0.2650	0.390	0.239	N/N
15	Asian/ Caucasian	0.0210	0.0270	0.0340	0.585	0.537	N/N
15	Hispanic/ Caucasian	0.1340	0.1350	0.1510	0.774	0.354	N/N
15	Female/ Male	0.1930	0.1980	0.1980	0.178	0.863	N/N
15	Over 40/ Under 40	0.2050	0.2060	0.2130	0.481	0.108	N/N
16	African American/ Caucasian	0.5600	0.5610	0.5610	0.392	0.945	N/N
16	Asian/ Caucasian	0.3420	0.3570	0.3620	0.265	0.522	N/N
16	Hispanic/ Caucasian	0.4390	0.4670	0.4680	0.055	0.764	N/N
16	Female/ Male	0.5330	0.5330	0.5330	0.498	0.875	N/N
16	Over 40/ Under 40	0.5360	0.5390	0.5390	0.104	0.503	N/N
17	African American/ Caucasian	0.5130	0.5140	0.5150	0.645	0.233	N/N
17	Asian/ Caucasian	0.3420	0.3430	0.3480	0.782	0.492	N/N
17	Hispanic/ Caucasian	0.4500	0.4500	0.4650	0.835	0.176	N/N
17	Female/ Male	0.4920	0.4960	0.4960	0.053	0.723	N/N
17	Over 40/ Under 40	0.4920	0.4940	0.4960	0.279	0.122	N/N
18	African American/ Caucasian	0.4160	0.4160	0.4320	0.855	0.001*	N/N
18	Asian/ Caucasian	0.6500	0.6520	0.6540	0.669	0.665	N/N
18	Hispanic/ Caucasian	0.5620	0.5650	0.5980	0.470	0.034	N/N
18	Female/ Male	0.4800	0.4810	0.4810	0.369	0.611	N/N
18	Over 40/ Under 40	0.4720	0.4730	0.4730	0.466	0.482	N/N
20	African American/ Caucasian	0.4740	0.4750	0.4760	0.508	0.436	N/N
20	Asian/ Caucasian	0.4770	0.4890	0.4990	0.367	0.411	N/N
20	Hispanic/ Caucasian	0.3500	0.3560	0.3740	0.458	0.203	N/N
20	Female/ Male	0.4800	0.4810	0.4810	0.377	0.916	N/N
20	Over 40/ Under 40	0.4840	0.4860	0.4870	0.143	0.589	N/N
21	African American/ Caucasian	0.6240	0.6250	0.6250	0.386	0.934	N/N
21	Asian/ Caucasian	0.4390	0.4430	0.4440	0.624	0.790	N/N

Item #	Comparison Group	Block 1 Nagelkerke R^2	Block 2 Nagelkerke R^2	Block 3 Nagelkerke R^2	Block 2 Significance Level	Block 3 Significance Level	Classification Category
21	Hispanic/ Caucasian	0.3650	0.3800	0.3980	0.225	0.188	N/N
21	Female/ Male	0.5560	0.5600	0.5600	0.052	0.558	N/N
21	Over 40/ Under 40	0.5600	0.5600	0.5610	0.822	0.250	N/N
22	African American/ Caucasian	0.5280	0.5280	0.5300	0.752	0.169	N/N
22	Asian/ Caucasian	0.4460	0.4900	0.4970	0.052	0.414	N/N
22	Hispanic/ Caucasian	0.3890	0.3890	0.4220	0.772	0.036	N/N
22	Female/ Male	0.4880	0.4880	0.4880	0.396	0.910	N/N
22	Over 40/ Under 40	0.4900	0.4910	0.4910	0.255	0.910	N/N
23	African American/ Caucasian	0.4120	0.4140	0.4140	0.164	0.748	N/N
23	Asian/ Caucasian	0.2800	0.2800	0.3150	0.942	0.013*	N/M
23	Hispanic/ Caucasian	0.3450	0.3480	0.3500	0.481	0.539	N/N
23	Female/ Male	0.4010	0.4010	0.4010	0.509	0.572	N/N
23	Over 40/ Under 40	0.4010	0.4060	0.4060	0.013*	0.379	N/N
25	African American/ Caucasian	0.5440	0.5440	0.5460	0.768	0.157	N/N
25	Asian/ Caucasian	0.4700	0.4760	0.4770	0.298	0.749	N/N
25	Hispanic/ Caucasian	0.4580	0.4610	0.4740	0.483	0.113	N/N
25	Female/ Male	0.5490	0.5490	0.5510	0.777	0.063	N/N
25	Over 40/ Under 40	0.5510	0.5570	0.5570	0.003*	0.991	N/N

Appendix E Heading Descriptions

Block 1 Nagelkerke R^2 – The amount of variance accounted for in the first block of the logistic regression procedure. Displays the amount of variance accounted for by candidates' total test score.

Block 2 Nagelkerke R^2 – The amount of variance accounted for in the second block of the logistic regression procedure. Displays the amount of variance accounted for by candidates' total test score and group membership.

Block 3 Nagelkerke R^2 – The amount of variance accounted for in the first stage of the logistic regression procedure. Displays the amount of variance accounted for by candidates' total test score, group membership, and the interaction of total test score and group membership.

Block 2 Significance Level – The significance level of the Block 2 Nagelkerke R^2 value. * Denotes that the value is significant at the .025 level.

Block 3 Significance Level – The significance level of the Block 3 Nagelkerke R^2 value. * Denotes that the value is significant at the .025 level.

Classification Category- The categorization of variance value differences (Shimizu and Zumbo 2005) between either (1) Block 2 Nagelkerke R^2 and Block 1 Nagelkerke R^2 or (2) between Block 3 Nagelkerke R^2 and Block 1 Nagelkerke R^2. Adjustments between the effect sizes less than .035 was considered a Negligible effect (N), adjustments between the effect sizes less than .07, greater than .035, with a Chi square of .025 was considered a Moderate effect (M), and adjustments between the effect sizes greater than .07, with a Chi square of .025 was considered a Large effect (L). The first letter represents the classification category for Uniform DIF, the second letter represents the classification category for Non-Uniform DIF.

Appendix E

Number of Candidates Passing at Cut-Off Score Level by Test and Comparison Group for Shortened Test

Shortened Cut-off Score	Male	Female	Age (Under 40)	Age (40+)	Caucasian	Hispanic	African American	Asian
5	496	1005	1042	458	112	117	851	205
6	494	1003	1039	457	112	116	849	205
7	492	999	1034	456	112	116	843	205
8	492	994	1030	455	112	114	841	205
9	489	984	1019	453	112	113	831	205
10	482	975	1006	450	111	112	820	205
11	478	957	988	446	111	111	804	204
12	472	935	964	442	111	111	779	204
13	463	916	943	435	108	109	761	203
14	449	888	910	427	106	106	731	202
15	443	851	876	418	105	101	697	202
16	426	798	830	395	100	95	648	201
17	394	731	760	366	95	88	581	192
18	337	590	619	308	77	71	463*	174

Note: No candidates obtained a failing score below the 5 point cut-off for the 19 item test. Focus group counts marked with * indicate a group violating the 80 percent rule when compared to its respective reference group.

References

A Brief History of Affirmative Action. (2018). Retrieved from http://www.oeod.uci.edu/policies/aa_history.php

Acar, T. (2011). Sample size in differential item functioning: An application of hierarchical linear modeling. *Kuram ve Uygulamada Eğitim Bilimleri, 11*(1), 284-288.

Adverse impact - Society for Human Resource Management. (2015). Retrieved from https://www.shrm.org/ResourcesAndTools/tools-and-samples/toolkits/Pages/avoidingadverseimpact.aspx

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*.Washington, D. C.: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.Washington, D. C.: American Educational Research Association.

Anderson, B. R., & Rogers, M. P. (1971). *Personnel Testing and equal employment opportunity*. (Washington: U.S. Gov. Print. Off.).

Arrington v. Massachusetts Bay Transportation Authority, 306 F. Supp. 1355 (D. Mass. 1969)

Biddle, D. A. (2005). *Adverse Impact and Test Validation; A practitioner's Guide to Valid and Defensible Employment Testing* (2nd ed). Burlington, VT; Gower.

Biddle, D. A. (2011). *Adverse Impact and Test Validation: A Practitioner's Handbook* (3rd ed). Burlington, VT: Gower.

Campbell, D. T., & Fiske, D. W. (1959). "Convergent and Discriminant Validation by the Multitrait–multimethod matrix". *Psychological Bulletin, 56*(2), 81-105

Casteneda v. Partida, 430 U.S. 482 (1977)

Civil Rights Act of 1991 § 109, 42 U.S.C. § 2000e et seq (1991)

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.

Collins, M. W., & Morris, S. B. (2008). Testing for adverse impact when sample size is small. *Journal of Applied Psychology, 93*(2), 463-471. doi:10.1037/0021-9010.93.2.463

Cuevas, M., & Cervantes, V. H. (2012). Differential item functioning detection with logistic regression. *Mathématiques Et Sciences Humaines,* (199), 45-59. doi:10.4000/msh.12274

Developing, Using, and Evaluating Written Examinations. (2017). Sacramento, CA: California Department of Human Resources.

Editor, C. R. (2015). *The Assassination of President James Garfield: The History and Legacy of the President's Death*. CreateSpace Independent Publishing Platform. Gifford 2012).

Elman, B. A. (1991). Political, Social, and Cultural Reproduction via Civil Service Examinations in Late Imperial China. The Journal of Asian Studies, Vol. 50, No. 1. (Feb., 1991), pp. 7-28.

Employment Testing: The Aftermath of Griggs v. Duke Power Company. (1972). *Columbia Law Review, 72*(5), 900. doi:10.2307/1121428

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. Federal Register, 43(166), 38290-38309.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*(2), 113-148.

Fish, C.R. (1905). *The civil service and the patronage.* New York: Longmans, Green

Furr M.R., & Bacharach, V.R. (2008). *Psychometrics: An introduction.* Thousand Oaks, CA: Sage Publications, Inc.

Gravetter, F. J., & Forzano, L. B. (2009). *Research Methods for the Behavioral Sciences*. Stamford: Cengage Learning.

Griggs v. Duke Power Company, 401 U.S. 424 (1971).

Guion, R. M. (1998). Assessment, measurement, and prediction for personnel decisions. Mahwah, NJ; Lawrence Erlbaum Associates.

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions*. New York: Routledge.

Hambleton, R. K., (2006). Good practices for identifying differential item functioning. *Medical Care, 44* (11 Suppl. 3), S182-S186.

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the GATB*. Washington, D.C.: National Academy Press.

Haynes, J. E. (1998, October 19). Words and Deeds of America History; United States Civil Service Commission. Retrieved April 02, 2017, from http://memory.loc.gov/ammem/mcchtml/corhome.html

Hazelwood School District v. United States, 433 U.S. 299 (1978)

Holland, P. W. & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley.

Hurtz, G. M. & Meyers, L. S. (2007). *Differential Item Functioning*. Unpublished manuscript, Department of Psychology, California State University, Sacramento, California.

Interpreting Test Results. (2018). Retrieved from https://testingservices.utexas.edu/scanning/interpreting-test-results

King, G. (1978). The California State Civil Service System. *The Public Historian, 1*(1), 76-80. doi:10.2307/3377672

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171-196).

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* MA: Addison-Wesley

May 2016 Occupation Profiles. (2017, March 31). Retrieved from https://www.bls.gov/oes/2016/may/oes_stru.htm

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the mantel-haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*(2), 131-144.

McGrath, R.J. (2013). The rise and fall of radical civil service reform in the u.s. states. *73*(4), 638-649. Doi: 10.1111/puar.12075

Mehrens, W. A., & Lehmann, I. J. (1973). *Measurement and evaluation in education and psychology*. New York: Holt, Rinehart and Winston.

Meier, P., Sacks, J., & Zabell S. L. (1984). What Happened in Hazelwood: Statistics, Employment Discrimination, and the 80% Rule. *American Bar Foundation Research Journal,* Vol. 9, No. 1 (Winter, 1984), pp. 139-186.

Merit Selection Manual: Policy and Practices. (2003). Sacramento, CA: California State Personnel Board.

Messick, S. (*1989*). *Validity*. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-104). New York, NY American Council on education and Macmillan.

Meyers, L. S. (2006). *Civil Service, the Law, and Adverse Impact Analysis In Promoting Equal Employment Opportunity*. Unpublished manuscript, Department of Psychology, California State University, Sacramento, California.

Meyers, L. S. (2008). *Common Rater Errors*. Unpublished manuscript. Department of Psychology, California State University, Sacramento, California.

Meyers, L. S. (2009). *Reliability, Error, and Attenuation*. Unpublished manuscript. Department of Psychology, California State University, Sacramento, California.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2013). Applied multivariate research: Design and interpretation (2nd ed.). Thousand Oaks, CA, US: Sage Publications, Inc.

Morris, S. B. (2001). Sample size required for adverse impact analysis. *Applied H.R.M. Research, 6*(1-2), 13-32.

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of dif. *European Journal of Psychological Assessment, 18*(1), 9-15.

Netemeyer, R. G., Sharma, S., & Bearden, W. O. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage.

Penfield R. D., Lee O. Test-based accountability: potential benefits and pitfalls of science assessment with student diversity. *Journal of Research in Science Teaching*. 2010;47:6–24.

Peterson, M. D. (1970). Thomas Jefferson and the New Nation : A Biography. Cary, GB: Oxford University Press, USA. Retrieved from http://www.ebrary.com

Quality Assurance Review of the Department of General Services. (2001). Sacramento, CA: California State Personnel Board.

Quality Assurance Review of the Department of Veteran Affairs. (2001). Sacramento, CA: California State Personnel Board.

Report on the Status of the State Discrimination Complaint Process. (2002). Sacramento, CA: California State Personnel Board.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement, 17*, 105-116.

Roth, P. L., Bobko, P., & Switzer, F. S. (2006). Modeling the behavior of the 4/5ths rule for determining adverse impact: Reasons for caution. *Journal of Applied Psychology, 91*(3), 507-522. doi:10.1037/0021-9010.91.3.507

Rusticus, S. A. & Lovato C. Y. (2014). Impact of Sample Size and Variability on the Power and Type I Error Rates of Equivalence Tests: A Simulation Study. *Practical Assessment, Research & Evaluation.* Vol. 19. N. 11, 1-10. https://pareonline.net/getvn.asp?v=19&n=11

Schenker, N., & Raghunathan, T. E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine, 26*(8), 1802-1811. doi:10.1002/sim.2801

Scheuneman, J. D. & Slaughter, C. (1991). *Issues of test bias, item bias, and group differences and what to do while waiting for the answers*. Unpublished manuscript, Educational Testing Service.

Selection Manual. (1979). Sacramento, CA: California State Personnel Board.

Shimizu, Y., & Zumbo, B. D. (*2005*). A Logistic Regression for. *Differential* Item Functioning Primer. Japan Language Testing. Association Journal, 7, 110-124.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures*, (4th ed.). Bowling Green, OH: Society for Industrial and Organizational Psychology, Inc.

Suen, H. K., & Vu, L. (2006). Chronic Consequences of High-Stakes Testing? Lessons from the Chinese Civil Service Exam. Comparative Education Review, 50(1), 46–65.

U.S. Office of Personnel Management. (2003). *Our mission, role & history theodore roosevelt.* Retrieved from https://archive.opm.gov/about_opm/tr/history.asp

United States v. HK Porter Company, 296 F. Supp. 40 (N.D. Ala. 1968)

Zieky, M. (2003). A DIF Primer. Princeton, NJ: Educational Testing Service.

Zwick, R. (1990). When do item response function and Mantel-Haenszel Definitions of Differential Item Functioning Coincide? *Journal of Educational Statistics*, *15*(3), 185-197.