2012

Employment Testing for Selection and Promotion Post Ricci and Lewis Decisions

William Maccarone
University of Rhode Island

EMPLOYMENT TESTING FOR SELECTION AND PROMOTION POST RICCI AND LEWIS DECISIONS

WILLIAM MACCARONE

University of Rhode Island

In this paper the history and development of disparate treatment as a result of standardized testing in both selection and promotion will be analyzed. Historical trends and litigation will be examined with particular emphasis on the validity and utility of utilizing standardized tests. In particular the recent *Ricci* and *Lewis* decisions will be scrutinized with an eye towards the influence those landmark cases may or may not have over future hiring and promotion practices.

In *Ricci*, the Supreme Court struck down a decision of a municipality to not use a promotional process that had a disparate impact upon minorities, while in *Lewis*, the Supreme Court struck down a decision of a municipality to utilize hiring practice with disparate impact upon minorities. There is an inherent question to be asked. Why was the testing practices of one organization upheld, the testing practices of a similar organization struck down, and why are so many of these tests so problematic to begin with?

In *Ricci*, the court utilized desperate treatment theory of Title VII, discrimination protection to "protect" a "non-suspect" class of individuals against racially motivated employment decisions following race based statistical testing bias in promotional testing. In *Lewis*, the Court utilized the disparate impact provisions of Title VII to find cause of action for a "protected class of individuals" following statistical race based testing bias in employment testing. These decisions seem diametrically opposed. How can any municipality move forward with any future testing policy, when they are open to liability regardless of their actions if there is any race based bias?

WHY STANDARDIZED TESTS?

The challenge of finding productive employees is not new. Predicting methods to determine which applicants will be productive employees has evolved from generation to generation. (Hunter & Schmidt, 1998: 262) A recent survey of employment recruiters indicate that employers have five ranked goals when looking for new employees; "(1) generating high-quality employment applications, (2) generating the best possible return on investment, (3) stimulating a desire to work for the organization, (4) filling specific positions, and (5) generating diversity." (Mello, 2006: 347)

While employers have a vested economic interest in selecting and promoting productive employees there are often economic interests in minimizing the costs associated with selection and promotions. Numerous studies have been undertaken to predict job performance given different job requirements. (Murphy, 1989)(Cited in Sackett & Lievens, 2008: 423) Invariably, mental ability testing is one of the best single predictors of future job performance while costing the least amount to administer. (Schmidt & Hunter, 1998)

How are tests used in selection and promotion?

Organizations that desire to hire or promote personnel need to develop some method of achieving this requirement of accurately making selection and promotion decisions. Many organizations utilize applications and an initial interview to provide opportunity to take a large group of applicants and reduce it down to a more manageable subset. (Mello, 2006) Often employers will utilize word of mouth or internal job postings to create a pool of potential candidates when an organization is looking to expand. (Mello, 2006) Informal networks of relatives and friends are a source of new candidates. Since this can be a highly problematic process for public employers, especially in the arena of public safety where more formal hiring

and promotional systems were created. (Lasky, 1997)

Civil Service laws, championed under the Pendleton Act, were developed following the attempted assassination of President James Garfield by a man disgruntled for failing to land a job in government service. (Lasky, 1999) Civil Service laws in Pennsylvania date back to 1919, and were designed to attract competent employees free from any religious or political affiliations. (Lasky, 1997) The 1919 Pennsylvania legislation also included merit testing for new employees and the "just-cause" requirement for current employees. (Lasky, 1997)

These laws created civil service commissions that were tasked with developing both personnel rules but also placing employees within classifications and often utilizing "ranked eligibility list[s]" for competitive job openings. (Lasky, 1997) These ranked lists were usually created following a written test administration. (Lasky, 1997) From these ranked lists various selection processes evolved. Some selection processes allowed the hiring authority to select one from the top two ("Rule of Two") or three ("Rule of Three") top scoring individuals to fill vacancies, while other more stringent rules required the top scorer ("Rule of One") to be selected and offered the hiring authority very little latitude in the selection process. (Lasky, 1997) Having such a highly structured selection process can present great difficulties for public employers finding the best possible candidates while assuring fairness and transparency, especially when one of the top concerns of human resource professional include diversity in the workplace. This is the conundrum many public employers faced and still face today.

Can General Mental Ability Tests Predict Performance?

General mental ability (GMA) (also known as "general cognitive ability and general intelligence") tests have long been considered one of the most valid single predictors of both "future performance and learning." (Hunter & Hunter, 1984; Ree & Earles, 1992)(Cited in Schmidt & Hunter, 1998, p. 262) GMA tests typically have the highest validity and one of the lowest administration costs. (Schmidt & Hunter, 1998) Theorists estimate that the Federal government could realize a $13 billion increase in productivity if it utilized strict rank ordering of selection and promotional GMA test scores without other factors. (Rynes, Brown, & Colbert, 2002) On a smaller scale, the Philadelphia Police Department could realize an estimated $12 million in additional productivity utilizing solely GMA testing alone. (Rynes, Brown, & Colbert, 2002)

GMA tests are very similar to general intelligence tests. Scientists are split in determining if GMA type tests effectively measure for predicting future performance, whether knowledge is the necessary prerequisite for performance, whether cognitive type tests measure all kinds of knowledge that may be necessary for future performance predictions, and whether any unitary factor can effectively measure performance. (Rominger & Sandoval, 1998) Despite these differences GMA testing is prevalent in selection and promotion, especially in civil service.

Employers began standardized testing to both screen potential employees and determine current employees eligibility for promotion in the 1950's. (Rominger & Sandoval, 1998) These tests are not necessarily meant to "measure intelligence itself, but a related construct: i.e. future job performance." (Rominger & Sandoval, 1998: 335) Employers turned to standardized tests because they were considered reliable "predictors of job performance" and they allowed employers to rank test takers based upon "level of performance." (Rominger & Sandoval, 1998, p. 302) Unfortunately, reliance upon standardized testing scores may create and perpetuate a systematic and chronic underrepresentation of women and minorities in the workplace. (Rominger & Sandoval, 1998) This underrepresentation is often evidenced when testing employee's cognitive abilities in comparison to actual job performance. (Rominger & Sandoval, 1998) Despite being one of the best available predictors of job performance, GMA's are not at all absolute, and they may fail to

predict job performance as well. (Rominger & Sandoval, 1998)

Employers have attempted to mitigate these disparities by including other predictors within the selection or promotion process; however these attempts often continue to perpetuate the bias. (Chung-Yang & Cronshaw, 2002) In some circumstances judicial and legislative decrees were instituted to require employers to accommodate women and minorities following GMA testing. (Chung-Yang & Cronshaw, 2002) Despite over thirty years of perceived fair employment efforts, litigation in both discrimination and reverse discrimination has continued to grow in an employment context. (Rominger & Sandoval, 1998)

What are the Costs Associated with Testing/Adverse Impact

One of the biggest concerns surrounding GMA testing relates to its impact on minority test-takers. Many civil service employers have struggled with methods to hire and promote the best suited employees, while balancing genuine societal goals of diverse workplaces. Higher test scores are directly correlated with the test-takers "economic, social, and educational status" which many believe creates a repetitive process of underrepresentation of minorities within employment opportunities and promotion. (Widgor & Sackett, 1993, p. 184) However, research soundly indicates that cognitive ability tests are one of the best indicators for future job performance. (Gardner & Deadrick, 2008) This leaves employers with a difficult choice, by-pass a valid selection method, look to alter testing methods to reduce racially biased results, or ignore the racial bias to achieve the best possible candidate.

What makes a test biased

There are numerous ways to determine if a test is biased. Some rely on simple mathematical percentages of pass rates among protected groups, while industrial/organizational psychologists utilize objective factors in determining possible bias. One of these models is called differential prediction or the Cleary Model. (Chung-Yan & Cronshaw, 2002) "Predictive bias is found when mean criterion (e.g. job performance) predictions for groups differentiated on some other basis than criterion performance are systematically too high or too low relative to mean criterion performance of the groups." (Society for Industrial Organization Psychology, Inc. (SIOP), 1987, p. 18)(Cited in Chung-Yan & Cronshaw, 2002, p. 491)

"A test is biased for members of a subgroup of the population, if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. (Chung-Yan & Cronshaw, 2002) In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of 'unfair,' particularly if the test produces a prediction that is too low. If the test is used for selection, members of a subgroup may be rejected when they were capable of adequate performance." (Ching-Yan & Cronshaw, 2002, p. 491)

A test is fair if both groups tested have the same relationship with independent and dependent variables being analyzed. (Chung-Yan & Cronshaw, 2002) Another manner of determining test bias is to determine: "when the difference between the mean test scores of two groups is greater relative to the difference between their mean job performance ratings." (Chung-Yan & Cronshaw, 2002, p. 491) Utilizing this process is probably the simplest way to determine if a test is biased. The test score difference should be correlated and proportional to job performance; this is called the Thorndike model. (Chung-Yan & Cronshaw, 2002)

Utilizing the common regression line, under the Cleary model, allows for the best selection of candidates for job performance because it is the best way to determine individual performance; however it tends to potentially leave many candidates that would have performed well underrepresented due to poor test scores. (Chung-Yan & Cronshaw, 2002) These Type II

errors, often called false-negative errors are a major concern for organizations seeking a diversified workplace, when the errors are over-representative of a protected class. Traditional human resource theorists are not as concerned with Type II errors, since there is little detrimental effect upon the organizations selection and promotion if there is an ample hiring pool available to compensate for the lower overall number of applicants. (Chung-Yan & Cronshaw, 2002) Using the Thorndike model candidates that could have performed well are also identified, however there may be more false-positives in utilizing the Thorndike method were individual scores will falsely indicate superior performance. (Chung-Yan & Cronshaw, 2002) False positive, or Type I errors are more of a concern for human resource theorists since the goals of the organization can be detrimentally effected if enough poorly chosen individuals are selected or promoted.

An important factor that requires discussion is that is that all of these models require a valid test. The job performance measure needs to be accurate. (Chung-Yan & Cronshaw, 2002) Objective measures of job performance need to be utilized to avoid any possible rater bias. (Ching-Yan & Cronshaw, 2002) This has been proven necessary by research. When subjective indices are utilized to determine job performance (I.e. supervisor ratings) blacks tend to perform markedly lower than white counterparts. (Chung-Yan & Cronshaw, 2002) Additionally, GMA tests predict objective measures of performance better than subjective measures. (Chung-Yan & Cronshaw, 2002)

PART I - LEGAL ANALYSIS

1964 Civil Rights Act and Title VII

Congress passed the 1964 Civil Rights Act in an effort to prohibit employer discrimination based upon sex, religion, color, race, or national origin. (Canton, 1987) Following the passage Title VII of the Civil Rights Act of 1964, Congress sought "to achieve equality of employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees." (Canton, 1987, p. 684) The Civil Rights Act also created the Equal Employment Opportunity Commission (EEOC) to help implement the Act and achieve the Act's intent, by changing employment practices and allowing women and minorities' equal employment and economic opportunity. (Rominger & Sandoval, 1998)

The Act itself "forbids employers from engaging in 'employment practices', including the use of employment tests that are designed to discriminate on the basis of proscribed factors." (Rominger & Sandoval, 1998, p. 306) The Act defines unlawful employment in Figure 1.

FIGURE 1
STATUE

(a) It shall be an unlawful employment practice for an employer:

(1) To fail or refuse to hire or discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions or privileges of employment, because of such individuals' race, color, religion, sex, or national origin; or

(2) To limit, segregate or classify his employees or applicants for employment in any way which would deprive any individual of employment opportunities or to otherwise adversely affect his status as an employee because of such individual's race, color, religion, sex or national origin. *42 U.S.C. § 2000e-2*(a)(1)-(2) (as cited in Rominger & Sandoval, 306-07).

Additionally, the Act further provides:

(h) Notwithstanding any other provision of this subchapter, it shall be an unlawful employment practice for an employer to give or to act upon the results of any professionally developed ability test, provided such test, its administration or action upon the results is not designed, intended or used to discriminate because of race, color, religion, sex, or national origin.

Source: *42 U.S.C. § 2000e-2*(e)(2) (as cited in Rominger & Sandoval, 1998, p. 307).

While the Civil Rights Act sought to protect workers from obvious invidious forms of discrimination, it failed to address facially neutral employment practices which had discriminatory effects. (Arakawa & Park Sonen, 2010, p. 469) In the myriad of litigation that stemmed following the passage of the Act, four theories of employment discrimination developed. These theories are: "disparate treatment, policies or practices that perpetuate the effects of past discrimination, adverse impact, and failure to accommodate an employee's religious observance or practices." (Canton, 1987, p. 684)

Disparate Treatment

The Supreme Court established a process for plaintiffs to claim employment actions were made based on race. In *McDonnell Douglas*, the Court proffered plaintiffs need to show (1) they are a racial minority, (2) they applied for and were qualified for a position employer was hiring for, (3) despite qualifications the applicant was rejected, and (4) the employer never hired anyone, and continued to seek applicants with similar qualifications. (*McDonnell Douglas Corp., v. Greene*, 411 U.S. 792, 802 (1973)) In response, the burden then shifts to the employer to show that the decision was based upon a "legitimate [] [and] nondiscriminatory" reason. (Id. at 802) Once the employer makes that showing, the burden then shifts back to the plaintiff to illustrate that the reason was not indeed legitimate and nondiscriminatory, but was a mere "pre-text" designed to mask the employers illegitimate and discriminatory intent. (*Id.* at 804) While this process provided a mechanism to fight discriminatory actions, it did not capture all forms of possible discrimination in the employment arena, especially in the realm of employment testing.

Disparate Impact

The Supreme Court recognized this exclusion and responded in 1971, with the *Griggs Power* decision, and in 1991 Congress amended Title VII, to explicitly include disparate impact within the purview of the Act. (Arakawa & Park Sonen, 2010) Under the revision any employment practice which results in a ***disparate impact*** "based on race, color, religion, sex or national origin" is unlawful unless the employer can "validate its test as job-related consistent with business necessity." *Griggs v. Duke*

Power Co., 401 U.S. 424 (1971) (cited in; Hoodhood, 2010: 112) (emphasis added)

In *Griggs*, the employer required employees to either pass a standardized general intelligence test or possess a high school diploma before hiring or promoting workers. (Canton, 1987, p. 685-86) While *Duke Power* employed ninety five workers, only fourteen were African American, and none of the African Americans worked outside of the "labor department." (Rominger & Sandoval, 1998, p. 309) These jobs were the lowest paid in the entire plant, and the only way of working in other more high paying positions required passing an aptitude test. (Rominger & Sandoval, 1998, p. 309-10) The Company initiated the testing program after Title VII was instituted. (Rominger & Sandoval, 1998, p. 310)

The Supreme Court made several key finding in *Griggs*. Firstly, the Court stated that any employment practice "which operates to exclude Negroes" is prohibited unless; the practice can be shown to be "job-related". *Griggs*, 401 U.S. 424 (Cited in Canton, 1987: 685) This is often referred to as the "business necessity" exception. (Canton, 1987: 685) Secondly, the Court concluded employer intent was irrelevant. (Canton, 1987) Thirdly, the burden of proof in disparate impact cases lie with the employer. (Rominger & Sandoval, 1998)

In 1975, the Court again addressed this issue in *Albemarle Paper Co. v. Moody*, when the Court developed a "three-part analysis" in determining whether adverse impact resulted from employment hiring practices. *Albemarle Paper Co. v. Moody* 422 U.S. 405 (1975) (cited in Canton, 1987, p. 686) The initial part of the analysis is for the employee to illustrate a prima facie case of discrimination, by showing a significantly different racial pattern between the applicant pool and the applicants selected for hire or promotion. (Canton, 1987) This illustration causes the burden to then shift to the employer to prove the test is job-related and that no other means of selection are available for the employer to meet their business needs. (Canton, 1987) Finally, after this illustration the burden shifts back to the employee to prove that the test is not a "business necessity" and is actually just a "pre-text" for discrimination. (Canton, 1987, p. 686)

In *Albermarle*, the Court, for the first time looked to psychometric evidence in determining whether a test can actually predict employee future performance. (Rominger & Sandoval, 1998) The Company in this case contracted with industrial psychologists to develop a job-relatedness analysis four months before trial. (Rominger & Sandoval, 1998) The psychologists utilized statistical correlation between the test scores and average supervisory rankings (subjective criterion) to determine the tests were job-related. (Rominger & Sandoval, 1998) The Supreme Court overruled this analysis and instead utilized validity standards developed by the American Psychological Association. (Rominger & Sandoval, 1998) This was the first of many times to come the Court, test-takers, and employers would utilize experts in seeking to validate testing processes and justify race based employment decisions.

EEOC

Since its creation in 1964, the EEOC has also evolved into more than just an investigatory agency. The EEOC not only investigates complaints of disparate impact, treatment, and discrimination, it also produces guidelines to help employers avoid liability and develop fair employment standards. (Hoodhood, 2010) The EEOC's Uniform Guidelines on Employee Selection Procedures, is designed to accomplish just that goal. One of the Guides requirements is that employers validate any "selection procedure" including employment examinations that result in an adverse impact. (Hoodhood, 2010, p. 124-25) The EEOC also advises all employers to validate any examination regardless of potential adverse impact. (Hoodhood, 2010) The validation procedures recognized by the EEOC are "criterion-related, content related, and construct-related." (Hoodhood, 2010, p. 125-26)

One of the more controversial procedures utilized by the EEOC, seeks to set a method to calculate when an employment practice is determined to produce adverse impact. The

Commission has adopted a "four-fifths" rule. Under the "four-fifths" rule plaintiffs can make a prima facie showing of disparate impact if the pass rate of one particular group is less than "four-fifths" the pass rate of another's. (Winrow, & Schieber, 2010) The court system has typically given EEOC recommendations great deference and many courts have enforced EEOC recommendations, however they have never formulated a bright line rule codifying the "four-fifths" rule in law. (Winrow & Schieber, 2010)

Additionally, employers can use EEOC guidelines as a defense when facing possible civil actions for unfair employment practices. (Hoodhood, 2010, p. 124) Employers typically look to these guidelines to make race based employment decisions, however the Supreme Courts unwillingness to completely endorse many of these guidelines has created potential difficult scenarios for employers. In particular, if an employer has a statistical disparity within a promotional or selection examination, and fails to utilize the results, claiming the EEOC "four-fifths" rule as a defense, how can they overcome the inherent conflict between this guideline and the clear language of Title VII, disparate treatment for race based employment decisions? This is exactly what the Courts had to decide in *Ricci*.

PART II – HUMAN RESOURCE ASPECTS

SELECTION – THE BASICS –

Validity/Reliability

The selection process is one of the most important aspects of any human resource function. Often selection errors are the root of legal action taken against employers, as was the case in *Lewis*. Employers are often making informed determinations of a candidate's future job performance, longevity, and ability to function as productive members of the organization. It is almost impossible for all selection methods to be 100% valid and reliable. (Mello, 2006) Both validity and reliability are interdependent upon each other and are necessary to defend employers from potential discrimination suits. (Mello, 2006)

Validity can be measured in three ways, content, criterion, and construct-related validity. Content validity relates to the actual knowledge necessary for job performance. (Mello, 2006) This will often require a thorough job analysis in which critical knowledge, skills, and abilities are identified as necessary to perform the required job. (Mello, 2006) Criterion; or empirical validity is when job performance is analyzed in relation to the screening process. This is advantageous over content validity because it predicts the candidate's job performance. (Mello, 2006) Construct-related validity utilizes job analysis to create individual traits necessary for successful job performance and those traits are then tested. (Rominger & Sandoval, 1998) These tests are often referred to as personality of behavioral test and require complex analysis in comparison to content and criterion validation. (Rominger & Sandoval, 1998)

Reliability is the "consistency of the measurement being taken." (Mello, 2006: 353) Reliability should be consistent "across time and across evaluators", which means the candidate should receive similar results on repeat evaluations and evaluators should reach similar conclusions following repeated evaluations. There are typically two types of errors associated with low reliability. (Mello, 2006: 353) The first is when an important criterion necessary for job performance is missing; this is referred to as deficiency error. (Mello, 2006) The second is called contamination error, in which the "unwanted influences" detrimentally effect the selection or if the knowledge, skills, and abilities utilized as criterion are not required for successful job performance. (Mello, 2006: 354)

Selection processes often include any combination of interviews, testing, and reference checks, physical examinations and even abilities tests under certain circumstances. While all processes are of legitimate concern, here we will focus on the testing process. Job selection testing can take various forms. Depending upon the needs of the organization and the structure of the job, testing can include technical, interpersonal, or problem solving abilities, or even personality traits

or any "other ***job-related performance indicators***." (Mello, 2006: 356)(Emphasis added)

Work sample and trainability testing are two of the more common forms of tests. The work sample test seeks to create sampling of actual work involved in the particular job being tested for. (Mello, 2006) Where, trainability tests seek to identify the candidates that have the aptitude to learn the important functions of the job. (Mello, 2006)

Personality testing utilizes the "Big Five" personality dimensions to create image of candidate in which to best judge job suitability. (Mello, 2006: 357) The "Big Five" include the following traits; "sociability, agreeableness, conscientiousness, emotional stability, and intellectual openness." (Mello, 2006: 357) Personality tests have fallen into disfavor with many professionals due to social concerns and lack of job relatedness, in short personality tests are not the best indicator of job performance and as such they often fail judicial scrutiny in adverse impact challenges.

Single vs. Multiple Criteria Selection Systems

Due to the overwhelming evidence supporting adverse impact on minorities from traditional mental ability testing employers have sought to reduce the overall negative impact on minorities; through avoiding the traditional rank-order hiring system often associated with these tests. They have utilized several methods to meet the staffing goals in a manner that is more consistent with societal interests in a diverse workplace. Some of these have proven successful, while others have drawn substantial scrutiny. These include banding test scores in larger groups, utilizing other measures to gauge performance, such as work sample testing, assessment centers, and non-traditional written examinations. Perhaps some or all of these selection systems could be utilized at differing points of the selection or promotion process to produce the best possible results, while not only minimizing the disparate impact but also choosing the best possible candidate.

Banding test scores into hiring pools

One of the most controversial options available is test score banding. Proponents of banding argue that banding can be very useful to minimize disparate impact among minorities. These advocates argue that banding already exists in typical strict rank-order testing since the candidates are banded together in a very narrow band typically consisting of one point each. (Campion, Outz, Zedeck, Schmidt, et al, 2001: 152) The traditional rank-order system inherently implies that the statistical difference between each number within the ranking is relevant to performance and selection. (Campion, Outz, Zedeck, Schmidt, et al, 2001: 152) However, this statistical difference may not always be relevant.

There are two types of banding. The first and often least controversial called "traditional banding" is developed using expectancy charts that indicate projected levels of job performance based upon the individuals test score grouping. (Campion, Outz, Zedeck, Schmidt, et al, 2001: 153) The band group is often based upon administrative conveniences and professional judgment, which can be at best subjectively based and at worst malignant. The second banding type utilizes objective scientific data to determine the exact boundaries of the groupings and can be supported by mathematical and scientific justification.

In a perfect world all tests will be 100% valid and all scores will be a 100% reliable measurement of future performance; however this is often never a possibility. The most reliable and valid test will still contain a margin of error at the very least. Banding of this nature is called "standard error of the difference (SED) between scores". (Campion, Outz, Zedeck, Schmidt, et al, 2001: 153) Proponents argue that SED banding allows the hiring authority the ability to scientifically determine band sizing, which will eventually lead to objective results. (Campion, Outz, Zedeck, Schmidt, et al, 2001) The premise behind this process asserts the difference between the top and bottom of the banded scores is "psychometrically indistinguishable." (Campion, Outz, Zedeck, Schmidt, et al, 2001: 155) The degree of banding is directly related to the

reliability of the test. If the test is proven very reliable, than a very narrow band will most likely result. (Campion, Outz, Zedeck, Schmidt, et al, 2001)

Opponents counter creating "psychometrically indistinguishable" scores within a larger subset of scores, creates great concern for the entire testing process including validity and reliability. (Campion, Outz, Zedeck, Schmidt, et al, 2001: 156-7) If the test is proven to be reliable, than how can a scientist then counter, that certain percentages of the scores are unreliable and should be "banded" based upon "indistinguishable" characteristics of the scoring. (Campion, Outz, Zedeck, Schmidt, et al, 2001: 158) Additionally, the differences that scientist can utilize to create a band of scores, may be "large and important" in the selection process. (Campion, Outz, Zedeck, Schmidt, et al, 2001: 159)

Still others posit that banding, as with many business choices, is simply a matter of efficiency and value judgments. The employer is making a value judgment that utilizing a less expensive test at the potential cost of a less qualified employee is weighed against the use of more expensive testing criterion designed at providing the employer a higher quality candidate. (Campion, Outz, Zedeck, Schmidt, et al, 2001) Or even worse, employers utilize banding to allow selection of certain groups within an organization without outwardly admitting the rationale. (Campion, Outz, Zedeck, Schmidt, et al, 2001)

Courts have generally allowed the use of banding in selection procedures as long as the selection from within the band is not based upon illegal criteria often consisting of race, sex, etc. (Campion, Outz, Zedeck, Schmidt, et al, 2001) Even opponents of banding agree that comparing some very close scores and utilizing this grouping in conjunction with other secondary criteria for eventual candidate selection. (Campion, Outz, Zedeck, Schmidt, et al, 2001) The real issue is how to develop objective criteria for setting bands and then developing additional criteria for eventual selection. In general, banding can be a useful tool in personnel selection when used appropriately.

Alternative Paper/Pencil Fill In Format Tests

Another relatively new and innovative method for personnel selection utilizing a "paper and pencil test format" is called "constructed based response" testing. (Winfred, Edwards, & Barrett, 2002) This type testing is designed to mirror the results of the traditional cognitive ability test with considerably less adverse impact. (Id.) This is accomplished by changing the mode of testing. The new mode of testing seeks to eliminate or reduce the test takers advantage due to "testwiseness", test taking strategy, motivation, or anxiety levels. (Id. at 988) The process seeks to utilize many of the same questions posed on GMA tests, often associated with high levels of validity, utility, and adverse impact, yet replacing the traditional multiple choice answer selections with fill in the blank, "write-in" or "mark-in" responses. (Id. at 996)

Some limited studies have shown this testing to be an acceptable alternative to traditional multiple choice tests. The levels of adverse impact in one study were less for the constructed based response style exam; however, the sample was very small, as was the differences. (Winfred, Edwards, & Barrett, 2002) However, psychologists suggest that the reduced adverse impact is most likely correlated to reduced levels of reading comprehension that is required for this type test. (Id.) This could be very problematic if reading comprehension is viewed as an important KSA for the tested position. Additionally, there is a substantial increase in the costs associated with administration of the construct based exam in comparison to traditional multiple choice. (Id.)

Assessment Centers

Assessment centers utilize real work simulations rated by multiple graders. (Rominger & Sandoval, 1998) One of the advantages to utilizing assessment centers as a selection and promotion tool is that they are highly predictive and legally defensible. (Eurich, Krause, Cigularov, & Thornton, 2009) Successful usage of assessment centers requires extensive job analysis. In particular organizations utilizing centers often employ a variety of methods to perform a valid

analysis including interviewing current workers, supervisors, written questionnaires to current employees and supervisors, and even employing teams of workers to develop appropriate job analysis. (Eurich, Krause, Cigularov, & Thornton, 2009) Experts recommend individualized job analysis as the best manner to achieve valid predictive results.

Recent trends show that increasingly employers are not spending the requisite time and money in developing individualized analysis and are utilizing "off the shelf" type of "wholesale and/or adapted use of standard assessment centers". (Eurich, Krause, Cigularov, & Thornton, 2009, 396) However the number and frequency of assessment center usage is increasing, which many argue is a positive sign. (Eurich, Krause, Cigularov, & Thornton, 2009) One very costly yet highly predictive method utilizes "multiple simulations to imitate a typical work day an assessee may encounter". (Eurich, Krause, Cigularov, & Thornton, 2009, 397) This in particular is deemed a highly valid and realistic predictive exercise. (Eurich, Krause, Cigularov, & Thornton, 2009) It has always been believed that assessment centers are not as susceptible to disparate impact and even if there is a disparity in the result, if the center is created appropriately it will be narrowly tailored towards job-relatedness.

Work Sample Testing

Another manner for employers to avoid or minimize adverse impact litigation following cognitive ability tests or GMA testing, would be to incorporate different components that traditionally result in less racial bias and a higher degree of work relatedness. One type option is a work sample test. Work sample testing differs from assessment centers, in that work sample testing may include a traditional pencil and paper type as one of many exercises all designed to mirror actual tasks that are required for the particular job. (Roth, Bobko, McFarland, & Buster, 2008) Work sample tests often require the applicant possess the training required for adequate performance at the time of the test, which significantly questions its ability to be implemented in firefighter entrance exams, however in promotional exams its use may prove promising. (Hunter & Schmidt, 1998)

Early research indicated a significant reduction in adverse impact for work sample tests, especially when they contain "hands-on performance tasks" as a component of selection. (Roth, Bobko, McFarland, & Buster, 2008) More recent research has indicated there is still a measurable adverse impact. In fact, one recent study indicated that the adverse impact for moderately complex jobs was almost identical when work samples and traditional cognitive ability tests were used. (Id.)

The cause appears to be a direct correlation between cognitive abilities and performance of work sample tests. (Roth, Bobko, McFarland, & Buster, 2008) The higher an individual's cognitive ability, the better performance on work sample tests. In fact work sample exams typically consist of "bundling" important KSA's that are believed to be job-related. Therefore, requirements that are determined to be job-related do often require knowledge, skills, and other cognitive abilities. (Id. at p. 645) However, this data is especially important given government regulations relating to hiring (Uniform Guidelines of 1978), when more than one method of selection is sufficiently valid predictor of job performance, the employer must use the method with the least adverse impact.

ST. LOUIS, RICCI, LEWIS, AND POLITICS

When asked, most firefighters will cite tradition as one of the top reasons for choosing their career. Steeped within this tradition is also legacy, the legacy of fathers and uncles passing on jobs to their sons and nephews, the legacy of relatives occupying high ranking roles within fire departments for generations. Historically the firefighters throughout the United States have been white males. (Brodin, 2011) While municipal and societal efforts to bring fire department demographics in line with the community often legacy gets in the way. In the 1960's and 70's open discrimination and racism was rampant in some jurisdictions. (Id.) Many pundits argue that legacy and racism in the fire service is one in the same, while others point to past judicial decrees

mandating certain percentages of minorities at the expense of performance indicators as political correctness at the expense of life safety.

Firefighters have a unique job. They are not only required to possess a great deal of knowledge, skills, and abilities, they are also required to live and work in extremely close confines with each other for extended periods of time. (Id.) They are required to eat, sleep, and work together, often for shifts exceeding twenty-four hours in duration. Societal problems are not left at the firehouse door-step; they are as prevalent inside the firehouse as any other public or private venue of the time. As culture evolves over time, firefighting is often stuck in the past steeped by the tradition and legacy and unable or even incapable of dynamic change. (Id.)

For years the fire service utilized any means necessary to preserve the tradition and legacy, even if at the expense of the service. The Supreme Court has held that fire departments across the country have "pervasively discriminated against minorities" and once minority firefighters gain entrance they are often met with "silent treatment", harassment, and at times even physical harm. (Brodin, 197-98) Often examinations given by municipal fire departments utilize the "business necessity" defense allowable under the EEOC Guidelines to avoid making wholesale modifications in hiring and promotion in response to disparate treatment. In October of 2011, a U.S. District Court in New York found that the New York Fire Department (FDNY) was systematically and deliberately "segregated" for "over forty years" often utilizing business necessity as a tool of exclusion. *Vulcan Soc., Inc. v. City of New York,* U.S. Dist. LEXIS 115074, (Oct. 5, 2011) at 4-5.

The St. Louis Experience

There seems to be little argument that standardized tests consisting of GMA written examination tend to disproportionately affect minority candidates despite efforts to remediate the issues. One possible solution to reducing this likely scenario was explored thirty years ago by the St. Louis Missouri Fire Department. St. Louis did not easily change its hiring and promotional procedures; unfortunately the court system forced the City to take action on several different occasions. A U.S. Court of Appeals ordered the City, Firefighters union, Firefighters Institute for Racial Equality (FIRE), and the U.S. Department of Justice to form a Test Development Committee. (Duffe, Gebhart, & McCurley, 1998) The City had a long history of minority underrepresentation within the fire department and following several disparate impact law suits over promotional exams the court ordered a new valid examination in 1979. (See id.)

In 1974 the City's Fire Captain promotional exam consisted of a multiple choice written test that measured technical knowledge, a seniority component, and job performance rating. The written exam and seniority equated to forty-five percent (45%) of the overall score each and an additional ten percent (10%) was based on the job performance rating. (Duffe et al., 1998: 449) The test produced a disparate impact on black firefighters and the Eighth (8th) Circuit Court of Appeals ruled the test was not job-related and failed to test major components of fire captain's duties including supervisory skills. (Duffe et al., 1998) In response the City sought to include an "assessment center" component within the promotional testing to accurately test supervisory duties. (Duffe et al., 1998,: 449)

The City developed a new fire captain promotional testing process which included an "assessment center" in which candidates were given three different scenarios likely to confront a fire captain. (Duffe et al.: 449) The scenarios included a "fire scene simulation" in which the candidate viewed and actual fire video and made recommendations of their orders and tactics to handle the fire. (Duffe et al., 1998: 449) The second scenario required the candidate to prepare and deliver a "training simulation" which a fire captain would regularly perform as part of their daily supervisory functions. (Duffe et al., 1998: 449) Finally, the assessment included an "interview simulation" in which the candidate was faced with a personal problem between two firefighters and the perspective "Captain" was

called upon to council his subordinates. (Duffe et al., 1998: 449)

The new assessment component accounted for seventy percent (70%) of the promotional score with the multiple choice "technical firefighting knowledge" portions accounting for the remaining thirty percent (30%). (Duffe et al., 1998: 449) The results showed no significant differences between white and minority candidates following the assessment portion, however, the multiple-choice exam still illustrated a significant difference between whites and minorities. (Duffe et al., 1998) The scores were combined to reveal disparate impact, again.

As a result, the Court ordered the City to create a Test Development Committee tasked with creating a new innovative testing procedure and in the interim rank order the candidates based upon the assessment center results and promote one minority firefighter, for every two white firefighters. (Duffe et al., 1998) The Committee developed a multiple hurdles testing program that incorporated a content-valid multiple choice-exam that tested basic knowledge, assessment center component designed to evaluate supervisory and administrative skills, and a fire scene simulation to evaluate technical knowledge. The significant difference in the new scheme allocated one hundred percent (100%) of the final score based upon the assessment center simulation. (Duffe et al., 1998) The initial two portions, the written examination and fire scene evaluation were simply pass/fail. Candidates needed to pass both to proceed onto the assessment portion. The most costly aspect of the testing procedure lies with the individualized assessment center portion.

The new examination process resulted in continued statistical differences between white and minority passage rates for both the written and assessment portions. Over a period of four exams (11 years) minority candidates passed at rates between eighty percent (80%) to eighty-nine percent (89%) as compared to whites on the written pass/fail portion. (Duffe et al., 1998: 455) This is contrasted with the fire scene pass/fail component where white candidates failed at a much higher rate than minority candidates. On two examinations white candidates passed at rates of eighty-nine (89) and seventy-eight (78) percent when compared to minorities, however on two occasion's whites passed at rates of forty-eight and forty-six percent. (Id. p. 455) There was no statistical difference between minority and white candidates passage rates following the assessment center portion of the testing process. (Duffe et al., 1998)

St Louis was able to overcome any statistical differences between blacks and whites by utilizing different procedures and pooling the data together to make promotions in a manner that limited disparate impact. However, there does not appear to be any objective criteria for determining the weights of the different procedures based on validity or utility. This data would seem essential to any valid promotional procedure. The St. Louis "solution" appears to be more political and less scientific. Great caution needs to be taken when politics and testing combine, as we see below.

The Ricci case

New Haven was no stranger to litigation over minorities in the fire service. In 1973, the New Haven Fire Department had only eighteen black firefighters in the ranks, no Hispanics, and four hundred and eighty-six whites. (Brodin, 2011) Only one out of one hundred and seven officers was black. (Id.) Three separate times between 1973 and 2002 the courts weighed in requiring the City to change its hiring and promotion practices to better effectuate valid testing and minority representation. (Id.) The Court has even appointed Special Master's to oversee the hiring and promotion practices of the fire department to assure the court's rulings were implemented.

In the fall of 2003, the City of New Haven sought to administer promotional examinations for the positions of Fire Captain and Fire Lieutenant. The process was defined by the City Charter (which required a "merit system" consisting of job-related examinations administered by Civil Service Board (CSB) pursuant to the "rule of three") and the collective bargaining agreement (CBA) between the City and firefighters union (which weighted examinations of written

exam (60%) and oral exam (40%)). (*Frank Ricci, et al., v. John DeStefano*, 554, F. Supp. 2d 142 (D. Conn., Sept. 28, 2006) p. 146) The City hired an experienced consultant Industrial/Organized Solutions, Inc. (IOS) to "develop and administer the examinations." (Id. p. 146) As a result IOS created a written and oral examination based upon the "knowledge, skills, and abilities necessary for the lieutenant and captain positions". (Brodin, 2011, p. 167) The City paid ISO, ***$100,000*** for the test. (*Ricci v. DeStefano*, 2009 U.S. LEXIS 4945, 16)(Emphasis added)

The examinations were administered in November and December of 2003. Forty-four percent (44%) of the firefighters taking the Lieutenants exam and thirty-nine percent (39%) of the Lieutenants taking the Captains exam were minorities. (Brodin, 2011, p. 168) Following the ranking of candidates utilizing the CSB and CBA only ten white test-takers were eligible for Fire Lieutenant promotions, while seven out of nine whites were eligible for promotion to Fire Captain. (Id.) In sum, none of the thirty-four minorities qualified for promotion to Lieutenant, while only two of sixteen minority Lieutenants qualified for promotion to Fire Captain. (Id.)

The CSB conducted five public hearings in which numerous witnesses advocated both for and against certification. (*Ricci*, 554 F. Supp. 2d, pp. 146-47) The plaintiff firefighters union president Patrick Egan, sought to have the test validated (as "job-related") by IOS, which could allow the test to be utilized even if there was an adverse impact as a result. (Id. p. 147) The City's Corporate Counsel, Thomas Ude, openly expressed concern over the possible legal issues facing the city, in addition to moral concerns, based upon societal goals of a diverse work-force. (Id. pp. 146-47) Explicitly Ude was concerned that the City would most definitely face legal challenges if they certified the exam results, even if the test was validated as "job-related" since the City would have "less discriminatory alternatives for the selection process" available to it. (Id. p. 147)

The "lead test developer" for the consultant hired by the City, Chad Legal Department testified that the examination was developed following numerous interviews with current New Haven fire officers and questionnaires designed to create an adequate job analysis which then could be used to develop a realistic and pertinent test. (*Ricci*, 554 F. Supp. 2d, p. 148) The consultants even "rode-along" with new Haven Fire units all in an effort to "generate a list of tasks, knowledge, skills, and abilities that are considered essential to performance" as a New Haven fire officer. (*Ricci*, US LEXIS 4945, p. 23) The test was then reviewed by two independent high ranking national fire officers in an effort to further bolster credibility. (*Ricci*, 554 F. Supp. 2d, p. 148) Finally, all materials tested on the exam were given to all candidates on a syllabus attached to the promotional application. (Id. p. 148) The exam also contained an oral component, in which the candidates were asked the same questions, generated from the same source materials as the written exam, by a panel of three fire independent out-of-state professionals. (*Ricci*, US LEXIS 4945, p. 17) The panel of three included one white, one black, and one Hispanic member. (Id. p. 24) Legel testified that in his professional opinion the test was "facially-neutral". (*Ricci*, 554 F. Supp. 2d, p. 148)

Even one of IOS's competitors testified at one of the hearings. Dr. Christopher Hornick, (a consultant that competes with IOS for testing contracts) testified that although he had "not had time to study the test at length or in detail" nor did he review any past New Haven Fire Department promotional exams results, he was able to claim his company would have developed a test that would have "significantly and dramatically less adverse impact" than the IOS test. (*Ricci*, 554 F. Supp. 2d, p. 149) Another witness, Dr. Janet Helms who is a professor of counseling psychology at Boston College testified that while she couldn't point to any one aspect of the exam that created the disparate impact, since she did not examine the test, most written exams will create disparate impact on "under-represented groups". (Id. p. 150) She also opined that since 67% of the survey respondents (current New Haven firefighters) were white the exam questions could be biased to their particular job-knowledge, thus creating a racially biased exam. (Id. p. 150)

The CSB ultimately decided to not certify the examination results following the hearings. The board split 2-2, with one member abstaining, since her brother was one of the candidates seeking promotion. (*Ricci*, 554 F. Supp. 2d, p. 151) As a result eighteen New Haven firefighters (seventeen white, one Hispanic) sued the City alleging violation of Title VII, disparate treatment and equal protection under the Constitution. (Id. p. 152) They claimed that political pressure from supporters of the sitting mayor led to the decision. (Id. p. 151) The City countered that they had an obligation to not certify the exam results as a result of Title VII, and cannot be liable under anti-discrimination laws for complying with current anti-discrimination statutes. (Id. p. 152)

This "reverse discrimination" suit placed the City in the unusual position of claiming the test had discriminated against minorities through disparate impact and the plaintiff firefighters sought to prove that "business necessity" dictated the results utilized despite the adverse result. (Brodin, 2011, p. 171) The District Court dismissed the initial law suit against the City claiming the City had no racial animus and was simply acting to "to remedy the disparate impact" created by the examinations, the First Circuit Court of Appeals affirmed the lower courts holding. (Brodin, 2011, p. 172) The Supreme Court however, overturned this ruling in a 5-4 decision that some have argued has turned disparate impact on its head in today's litigious environment.

The Supreme Court Weighs In (Ricci)

In a five to four decision the Supreme Court reversed the lower court and held that the City actually discriminated against the plaintiffs under disparate treatment grounds when it failed to certify the exam scores. (Brodin, 2011) The Court believed the "race-based decision" of the City to "reject [ ] the test results because 'too many whites and not enough minorities would be promoted were the lists to be certified'" violated the provisions of Title VII's disparate ***treatment*** prohibition. (*Ricci*, US LEXIS 4945.: 40)(Emphasis added) The Court then looked to determine if this violation was permissible in order to avoid future liability from minorities based upon the disparate results of the troubled test.

In proffering the *Ricci* decision the Court looked to past Equal Protection decisions where municipalities were tasked with formulating policies and making employment decisions that might not traditionally satisfy judicial scrutiny, in efforts to rid past prevalent race-based discrimination. (*Ricci*, LEXIS 4945, at 46; *citing Richmond v. J.A. Croson Co.,* 488 U.S. 469, 500 (1989); *Wygant v. Jackson Bd. of Ed.,* 476 U.S. 267, 277 (1986)) The Court adopted the "strong-basis in evidence" standard to "resolve any conflict between the disparate-treatment and disparate impact provisions of Title VII". (*Ricci*, LEXIS 4945, at 50) This standard was first developed in 1989, when the Court held certain "race-based" decisions are constitutional ***if*** there is a "strong basis in evidence" that such decisions were necessary to "remedy past racial discrimination". (Id. p. 45-46, citing *Wygant*: 277)(Emphasis added)

The Court acknowledges that the disparate impact from the exam was significant, especially in light of statistical evidence indicating minorities passage rates were approximately one-half white test-takers passage rates. (*Ricci*, LEXIS 4945, pp. 52-53) However, statistics taken alone will not suffice in making a disparate impact claim, in fact the law allows, following a "significant statistical disparity", the City to determine if the exam was "job-related" and "consistent with business necessity", or determination of other "equally valid, less-discriminatory alternative[s]" the City could have utilized to promote firefighters and failed to do so. (Id. p. 55)

In particular the Court took issue with the City's assertion the test was not "job-related" or "consistent with business necessity." (*Ricci*, US LEXIS 4945, p. 55) Evidence in the record from numerous experts advocating against certification indicated they, had either not thoroughly reviewed the exam, worked as competitors of IOS, or had no firefighting experience. Included in this group advocating against certification was the Mayor's office itself, whom had paid IOS a great deal of money to develop and administer a comprehensive examination process which

included a validation report to prove "job-relatedness" which the City refused to accept. (Id. pp. 56-57) The City's second argument that "less-discriminatory" alternatives means for existed, is also equally flawed according to the majority decision. The very prospect of changing scores or the weighting between oral and written examination scores, or banding competing scores together to alter civil-service "rule-of-three" requirements would also violate Title VII. (Id. pp. 58-60)[1]

One of the more controversial and compelling arguments cited in the Ricci case was Justice Alito's concurrence. Justice Alito utilizes transcripts from the CSB proceedings along with other extrinsic evidence to step beyond the direct issue at hand before the Court and espouse that even if the "strong basis in evidence" requirements were met by the City, its claim would fail based upon the "subjective question concern[ing] the employer's intent". (*Ricci*, US LEXIS 4945, at 69-70, *Alito Concurring*) Justice Alito, who was joined by Justices Scalia and Thomas, posit political pressure necessitated the Mayor of New Haven's decision to throw the test out, and fear of disparate impact liability was a mere pretext, hiding the true "illegitimate" purposes behind the CSB's actions. (See id. p. 72)

To support this approach Justice Alito, illustrated a pattern of back-room political pressure exerted by a local religious leader with personal and political ties to Mayor DeStefano. (*Ricci*, US LEXIS 4945 p. 73) This pressure was readily apparent throughout the case's history, the District Court even wrote "city officials worked behind the scenes to sabotage the promotional examinations because they knew that, were the exams certified, the mayor would incur the wrath of [Reverend Boise] Kimber" (Id. p. 73) Mayor DeStefano testified as a character witness for Kimber following his conviction for stealing from an elderly woman in 1996 and DeStefano appointed Kimber chair of the New Haven Fire Commissioners, until he resigned after telling firefighters new recruits with "too many vowels in their name[s]" would not be hired. (Id. pp. 73-74) This evidence in conjunction with the City's subsequent actions presents reasonable grounds for repudiating the City's decision.

While many human resource managers would prefer a bright-line rule for managing these type scenario's the Court has adopted this familiar standard in effort to give employers flexibility in making important decisions in light of competing interests among stakeholders the test takers, employers, and potentially aggrieved classes. The Court seems to struggle with any decision that is solely race-based once a selection process has been "clearly" "established" absent "very strong" evidence of liability if the race-based decision is not made.

Was it Really Such a Landmark Supreme Decision?

Pundits have argued the Ricci decision will impose peril beyond human imagination to public employers, minorities, and society. The titles of the various Law Reviews sum up the hysteria; "*Ricci v. DeStefano*: The New Haven Firefighters Case & The **Triumph of White Privilege**", "*Ricci v. DeStefano*: How the Supreme Court ***Muddled*** Employment Discrimination Law and ***Doomed Employers*** to Costly Litigation", "The Quintessential **Employer's Dilemma**: Combating Title VII Litigation by Meeting the Elusive Strong Basis in Evidence Standard", "***Damned*** If You Do and Damned If You Don't: Title VII and Public Employee Promotion Disparate Treatment and Disparate Impact Litigation." (Brodin, 2011; Hoodhood, 2010; Kormanyos, 2010; Roberts, 2010)(Emphasis added)

Similar to Ricci, the Court again had to recently decide how to interpret disparate impact following another firefighter case. This case did not involve promotion, but firefighter selection. One of the most interesting aspects of the decision is the great amount of time the legal system required in settling the eventual claim. The case originated

[1] The Court Specifically noted, Hornick, IOS's direct competitor that the City sought advice from, has since been hired by the City as a consultant. (Id. pp. 62-63)

from a 1995 entrance examination and was not ultimately decided until May of 2010. Again this case has been heralded as a landmark decision sure to alter the employment practices of municipal fire departments for years to come, however when scrutinizing the facts and following established human resource selection policies, one can't help but believe not much has changed, and that human resource professionals need to strictly adhere to established policies. This is what happened over a course of a decade and a half, when the Chicago Fire Department sought to find new recruits.

Lewis

The City of Chicago sought to create an eligibility list to hire future firefighters in the summer of 1995. (*Arthur L. Lewis v. City of Chicago*, 2005 U.S. Dist. LEXIS 42544, p. 5 (N.D. Ill. March 22, 2005)) The City required applicants to be residents of the city, at least 18 years old, and have a high school degree or equivalent. (Id.) The City conducted a "content orientated" examination designed to test "important aspects of performance" in which more than 26,000 people were tested. (Id. p. 7) The written exam was just the first step in a multi-step process that included physical abilities tests, background investigation, medical and drug screening, and eventual placement in the fire academy with successful completion and state board certification necessary to become a Chicago firefighter. (Id. p. 4)

The test included a written and video component based upon a 12th grade reading level. (*Lewis,* LEXIS 42544 p. 9) The written portion accounted for eighty-five percent of the total score, while the video portion accounted for the remaining fifteen percent. (Id. p. 10) The passing score of the exam was set at sixty-five (65); the average score attained was seventy-five (75). (Id. p. 11) The highest score was a ninety-eight (98) and the lowest was a twelve (12). (Id. p. 11) The test was developed by Dr. James Outtz, an industrial organizational psychologist, and was based upon knowledge, skills, and abilities termed "critical" or "essential" to be a Chicago firefighter, even before completing any required training. (Id. p. 9) Dr. Outtz created a list of forty-six skills, of which eighteen were "essential" and "needed on day one" from any firefighter candidate. (Id. p. 9) Of the eighteen seven were determined to be cognitive and four were tested on the exam. (Id.) These basic cognitive skills were:

> "(1) the ability to comprehend written information; (2) the ability to understand oral instructions; (3) the ability to take notes; (4) the ability to learn from or understand based on demonstration." (Id. p. 9)

Following the examination the City decided to create three pools of candidates, the first would be titled "well-qualified" and would include all candidates that scored 89 or above on the test, while the next pool, titled "qualified" would include all candidates that scored above the required 65 passing score and below the 89 required to be considered "well-qualified". (*Lewis*, 2005 LEXIS 42544 p. 6) The "pools" were utilized to cull candidates from the aggregate mass of passing scores and advance them to the next step in the hiring process. The City utilized this list and advanced candidates from the "well-qualified" pool from 1996 until 2001. (Id. p. 6) None of the candidates that scored between the passing score of 65 or the "well-qualified" arbitrary cut-off were initially hired or allowed opportunity to advance to the next step in the hiring process. (Id.)

In 2001, the "well-qualified" pool was becoming increasingly exhausted and the City began advancing candidates from the "qualified" pool. (*Lewis*, 2005 LEXIS 42544 p. 6) No evidence was ever introduced to illustrate the "qualified" candidates did not possess the knowledge, skills, or abilities necessary to be a Chicago firefighter, in fact virtually all these candidates that entered the fire academy completed training and received "state certification". (Id. p. 7) The City acknowledged that any candidate receiving a 65 or higher possessed the "minimum level of cognitive ability" necessary to be a Chicago firefighter. (Id. p. 12)

The first perceived problem with the 1995 test was the disparate impact the test created on

minorities before the arbitrary qualified pools were created. Of the over 26,000 test takers, 11,649 (45%) were white while 9,497 (37%) were African American. (*Lewis*, 2005 LEXIS 42544 p. 6) Over 93% of all white test takers achieved a 65 or higher on the exam, compared to 72% of African Americans. (Id. p. 11) While the disparity is significant, the racial divide among the top scorers is even more evident. The "well-qualified" hiring pool of candidates scoring 89 or above consisted of 12.6% of white test takers compared to 2.2% of African American test takers. (Id. p. 7) This resulted in white candidates having five to one advantage over black candidates at the possibility to moving to the next step in the hiring process. Eventually a class of over 6,000 African American applicants filed suit in the Northern District of Illinois, arguing the City's actions were discriminatory. (Id. p. 1)

Dr. Outz, suggested banding the scores based upon the tests "internal 'reliability'" in which it was determined certain score ranges were statistically insignificant (*Lewis*, 2005 LEXIS 42544 p. 13) Based upon this "psychometric basis" Dr. Outtz suggested utilizing bands of thirteen points from the top-score of 98 to create classes within the test-takers. (Id. p. 14) Absent disparate impact analysis this would allow the City to hire any test-takers with a score of 85 or higher, or to even, mostly due to the extremely large standard error of thirteen, expand the hiring pool to include all candidates that achieved a passing score of 65 or higher. (Id.)

Chicago's Deputy Director of Personnel, after learning the disparate impact on minorities and hearing Dr. Outz's recommendations decided to set the cut-off score of 89. (*Lewis*, 2005 LEXIS 42544 p. 16) Joyce claimed that administrative convenience was utilized to set the cut-off score based upon the anticipated number of candidates required to meet the Chicago Fire Department's hiring needs. (Id.) The City sent letters to all test-takers informing them of their ranking (i.e. "well-qualified", "qualified", "fail"). The "qualified" pool were informed that because the exact number of candidates required was unknown and the high number of higher scoring candidates, it was "not likely" that they would be offered a job, however, they would be retained on the "eligibility list" until the next test was administered. (Id. p. 17)

Numerous suits were filed as a result of the 1995 examination. In one such action, a group of current white Chicago firefighters sought to overturn existing practices within the Chicago Fire Department designed to ameliorate the past racial discrimination within the department. (*Horan v. City of Chicago*, No. 98 C 2850, 2003 U.S. Dist. LEXIS 17173 (N.D. Ill. Sept. 30, 2003), p. 185) In that case, the group of firefighters asserted "that the 1995 entrance exam was content valid" and that there was a direct correlation between a higher score and superior job qualifications. (Id.) The City argued in its defense that the test was *NOT* necessarily "content valid" and that the test only tested a "narrow set of cognitive abilities" and "could not predict on-the-job performance". (*Lewis* U.S. Dist. LEXIS 42544, pp. 19-20)(Emphasis added)

In the disparate impact claim brought against the City by "qualified" candidates, the City argued that the test was merely designed to examine some cognitive aspects that are related to candidates "trainability". (*Lewis* U.S. Dist. LEXIS 4254 p. 23) This claim may have benefited the City under the facts and circumstances surrounding the *Horan* court case, however this rationale does not comport to the requirements of Title VII, in disparate treatment analysis. Title VII requires the City to illustrate a connection between firefighter performance and the arbitrary cut-off between the "well-qualified and "qualified" applicant pools. (Id. p. 26)

The Trial Court found the City liable for Title VII claims, of disparate impact from minority test-takers that had been determined "qualified" under the City's ranking scheme. (*Lewis* U.S. Dist. LEXIS 4254 pp. 25-6) In particular the Court found that the City's "business necessity" defense was fatally flawed. The evidence produced at trial cast wide doubt on whether the 1995 exam tested the four cognitive skills it sought to test accurately, whether the cut-off score of 89 was representative of a candidates "relative abilities", and most importantly the City completely failed to prove any

candidate scoring above 65 was less qualified than any "well-qualified" candidate. (Id. p. 27) The Court specifically looked to the Horan decision in emphasizing that the City readily admitted no real statistical difference related to job-performance existed between a candidate that scored 65 or 89. (Id. pp. 34-5) This factor has been further illustrated by hiring candidates that scored between 65 and 89, for the 2003 firefighter academy, with no reported impact on job or training performance. (Id. p. 36)

One of the plaintiff's witnesses, Dr. Charles Cranny, testified that the biggest problem with the exam was no direct correlation between test scores and job performance. (*Lewis* U.S. Dist. LEXIS 4254 p. 37) Absent this data, it is impossible to determine the correlation coefficient and determine how candidates actually performed in relation to job performance. (Id.) The City sought to refute Dr. Cranny's assertions by claiming that all cognitive tests will result in similar outcomes, and that utilizing past exams correlation coefficients and applying them to the 1995 test result will accomplish the same result. (Id. p. 41) While the Court did question this analysis, ultimately it was the Horan decision that again directly impeded the City's claim. In *Horan* the City argued the opposite, claiming "cognitive skills are varied and distinguishable and that the results - - and consequently the predictive value - - of a cognitive test can vary depending on which skills are tested." (Id. p. 42) Furthermore, the video component of the 1995 test was used for the first time, which undoubtedly questions how predictive past examinations correlated coefficient would be when applying to the 1995 test.

The District Court went on to predict that even if the "business necessity" defense was successful for the City, they still would have lost the Title VII claim. Under Title VII, disparate impact claims, the burden would have shifted back to the candidates to prove there was a less discriminatory method of selecting recruits. (*Lewis* U.S. Dist. LEXIS 4254 p. 45) By simply randomly electing candidates from all the passing applicants ("well-qualified" and "qualified") the City would have drastically reduced the disparate impact while still limiting the pool for "administrative convenience" purposes. (Id.) As a remedy the Court ordered that 132 candidates be hired from the class of 6,000 that initiated the action. (*Arthur L. Lewis, et al., v. City of Chicago*, 2007 U.S. Dist. LEXIS 24378 (N.D. Ill., March 20, 2007) pp. 3-4) The Court further ordered that back-pay with pre-judgment interest and seniority be awarded to all 132 candidates. (Id.)

The City promptly appealed to the 7th Circuit Court of Appeals. The 7th Circuit reversed the lower Court based upon a technicality. The City argued before the appeals court the plaintiffs were late in filing their complaint and missed the statutorily set statute of limitations of 300 days from the test date. (*Arthur L. Lewis et al., v. City of Chicago*, 528 F.3d 488 (7th Circ. 2008) p. 3) The plaintiffs contended that every time the City utilized the effected hiring list, a new cause of action was created, and while the actual examination date was beyond 300 days of the plaintiffs claim, the first usage of the hiring list was within the 300 day window. (Id. p. 3) The Court of Appeals construed the statute narrowly and failed to allow the plaintiffs claim forward. As with the Ricci decision, this was not the end of the story.

In May of 2010, the U.S. Supreme Court reversed the 7th Circuit clearing the way for some of the *Lewis* Plaintiffs to begin their careers as Chicago firefighter recruits fifteen years after the test. The Supreme Court did not enter the fray of determining whether the test was sufficiently job related, or whether the City satisfied the "business necessity" exception, but simply held that the intent and letter of Title VII required the Court to uphold the claim against the City. (*Arthur L. Lewis, et al., v. City of Chicago*, No. 08-974 (U.S. 2010))

EMPLOYMENT TESTING AND PUBLIC POLICY

Organizations require employment candidates that are best suited for their particular industry, however society requires employment practices free from discrimination, the real challenge lies in formulating a proper employment screening process while preserving society's goals. And secondly what are the costs associated with that endeavor? This paper began with several

questions regarding how public employers can move forward based upon recent Supreme Court decisions. There is good news for public employers. *Lewis* and *Ricci* are not as diametrically opposed as first thought. When looking at the detailed facts and circumstances of each decision it is easy to see how the Court ruled and the reasoning behind the decision.

Unfortunately, that does little to solve some of the bigger dilemmas facing municipal governments in selection and promotion decisions. This paper illustrates the necessity for utilizing objective measures in all hiring and promotional decisions. Developing a system that utilizes multiple hurdles to achieve eventual promotion or selection appears to be the best approach. This presents a potential huge cost for cash strapped city and state governments, at a time of fiscal uncertainty. Utilizing assessment centers in conjunction with traditional paper and pencil exams both based upon objective valid criteria is one possible solution. The validity of every portion of the examination needs to be clearly proven and potential cut-offs (to establish hiring pools) need to be established ***before*** the examination is administered. Finally, and possibly most importantly, the validity needs to be established with as much certainty as scientifically possible, since even with all these steps, disparate impact is still a possibility, and if individual test takers decide to pursue legal action the municipalities have a proven defense to support their actions.

WORKS CITED

Court Decisions

US Supreme Court Decisions

Arthur L. Lewis, et al. v. City of Chicago, 130 S. Ct. 2191 (2010).

Frank Ricci, et al. v. John DeStefano, 129 S. Ct. 2658 (2009).

City of Richmond v. J.A. Croson Co., 488 U.S. 469, 500 (1989).

Wygant v. Jackson Bd. of Ed., 476 U.S. 267, 277 (1986).

Albemarle Paper Co., v. Moody, 422 U.S. 405 (1975).

McDonnell Douglas Corp., v. Green, 411 U.S. 792 (1973).

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

Appellate Decisions

Frank Ricci, et al. v. John DeStefano, 530 F.3d 88 (2d Cir. 2008), rev'd 129 S. Ct. 2658 (2009).

Arthur L. Lewis, et al. v. City of Chicago, 528 F. 3d 488 (7th Cir. 2008), rev'd, 130 S. Ct. 2191 (2010).

Trial Court Decisions

Vulcan Soc. Inc. v. New York, U.S. Dist. LEXIS 115074 (E.D. N.Y. Oct. 5, 2011).

Frank Ricci, et al. v. John DeStefano, 554 F. Supp. 2d 142 (D. Conn. 2006), aff'd, 530 F.3d 88 (2d Cir. 2008), rev'd, 129 S. Ct. 2658 (2009).

Arthur L. Lewis, et al. v. City of Chicago, No. 98 C 5596, 2005 WL 693618 (N.D. Ill. Mar. 22, 2005), rev'd, 528 F. 3d 488 (7th Cir. 2008), rev'd, 130 S. Ct. 2191 (2010).

Horan v. City of Chicago, No. 98 C 2850, 2003 U.S. Dist. LEXIS 17173 (N.D. Ill. Sept. 30, 2003).

Scholarly Journals

Arakawa, Lynda L., and Michele Park Sonen. "Caught in the Backdraft: The Implications of Ricci v. DeStefano on Voluntary Compliance and Title VII." *University of Hawai'i Law Review* 32 (2010): 463-83. Print.

Brodin, Mark S. "Ricci v. DeStefano: The New Haven Firefighters Case & the Triumph of White Privilege." *Southern California Review of Law and Social Justice* (2011): 161-232. Print.

Campion, Michael A., James L. Outtz, Sheldon Zedeck, Frank L. Schmidt, Jerard F. Kehoe, Kevin R. Murphy, and Robert M. Guion. "The Controversy over Score Banding in Personnel Selection: Answers to 10 Key Questions." *Personnel Psychology* 54 (2001): 149-85. Print.

Canton, Doreen. "Adverse Impact Analysis of Public Sector Employment Tests; Can A City

Devise a Valid Test?" *University of Cincinnati Law Review* 56 (1987): 683-709. Print.

Chung-Yan, Greg A., and Steven F. Cronshaw. "A Critical Re-Examination and Analysis of Cognitive Ability Tests Using the Thorndike Model of Fairness." *Journal of Occupational and Organizational Psychology* 75 (2002): 489-509. Print.

Eurich, Tasha L., Diana E. Krause, Konstantin Cigularov, and George C. Thornton III. "Assessment Centers: Current Practices in the United States." *Business Psychology* 24 (2009): 387-407. Print.

Gardner, Donald, and Diana L. Deadrick. "Underprdiction of Performance for US Minorities Using Cognitive Ability Measures." *Equal Opportunities International* 27.5 (2008): 455-64. Print.

Gebhart, Gary M., William C. Duffe, and Roger A. McCurley. "Fire Service Testing in a Litigious Environment: A Case History." *Public Personnel Management* 27.4 (1998): 447-58. Print.

Hoffman, Calvin C., and George C. Thornton III. "Examining Selection Utility Where Competing Predictors Differ in Adverse Impact." *Personnel Psychology* 50.2 (1997): 455-69. Print.

Hoodhood, Erica E. "The Quintessential Employer's Dilemma: Combating Title VII Litigation by Meeting the Elusive Strong Basis in Evidence Standard." *Valparaiso University Law Review* (2010): 111-55. Print.

Kormanyos, Katie R. "Ricci v. DeStefano: How the Supreme Court Muddled Employment Discrimination Law and Doomed Employers to Costly Litigation." *University of Toledo Law Review* 41 (2010): 975-1004. Print.

Mello, Jeffrey A. *Strategic Human Resource Management*. 2nd ed. Mason: South-Western, 2006. Print.

Ree, M.J., and A.J. Earles. "Intelligence is the Best Predictor of Job Performance" *Current Directions in Psychological Science*, 1 (1992): 86-89. Print.

Roberts, Robert N. ""Damned If You Do and Damned If You Don't: Title VII and Public Employee Promotion Disparate Treatment and Disparate Impact Litigation." *Public Administration Review* 70 (4) (2010): 582-590. Print.

Rominger, Anna S., and Pamela Sandoval. "Employee Testing: Reconciling the Twin Goals of Productivity and Fairness." *DePaul Business Law Journal* 10 (1998): 301-47. Print.

Roth, Phillip, Phillip Bobko, Lynn McFarland, and Maury Buster. "Work Sample Tests in Personnel Selection: A Meta-Analysis of Black-White Differences in Overall and Exercise Scores." *Personnel Psychology* 61 (2008): 637-62. Print.

Rynes, Sara L., Kenneth G. Brown, and Amy E. Colbert. "Seven Common Misconceptions about Human Resource Practices: Research Findings versus Practitioner Beliefs." *Academy of Management Executive* 16.3 (2002): 92-103. Print.

Sackett, Paul R. and Filip Lievens. "Personnel Selection" *Annual Review of Psychology* 59 (2008): 419-450. Print.

Schmidt, Frank L., and John E. Hunter. "The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings." *Psychological Bulletin* 124.2 (1998): 262-74. Print.

Wigdor, Alexandra K., and Paul R. Sackett. "Employment Testing and Public Policy: The Case of the General Aptitude Test Battery." *Personnel Selection and Assessment; Individual and Organizational Perspectives* (1993): 183-204. Print.

Winfred, Arthur, Bryan D. Edwards, and Gerald V. Barrett. "Multiple-Choice and Constructed Response Tests of Ability: Race-Based Subgroup Performance Differences on

Alternative Paper and Pencil Test Formats." *Personnel Psychology* 55 (4): 985-1008. Print.

Winrow, Brian P., and Christen Schieber. "The Diparity Between Disparate Treatment and Disparate Impact: An Analysis of the Ricci Case." *Journal of Legal, Ethical, and Regulatory Issues* 13 (2010): 45-54. Print.