Proposal submitted on 10/14/2020
Acceptance decision expected about 1/2021

SUBMISSION TYPE
Master Tutorial

TITLE
Police Officer Selection Is Broken and We Can Fix It

SHORTENED TITLE
Police Officer Selection Is Broken; How to Fix It

ABSTRACT
Psychometric analyses predict many new police officers will fail on the job with few minority candidates hired. Reasons for hiring errors and adverse impact and 6 psychometric and administrative approaches to help hire ethnically diverse academy classes while maintaining and even enhancing expected job performance will be covered, along with other IO approaches to improve police job performance.

PRESS PARAGRAPH
Many police managers are stymied in their attempts to hire black police officers due to the severe adverse impact that traditional employment tests have on black candidates. Many police departments and citizens are dismayed at the extent of poor job performance by police officers. Psychometric research and theory indicate that a high proportion of new police officers will fail on the job. This tutorial presents 6 psychometric and administrative approaches designed to help police departments hire ethnically/racially diverse academy classes while maintaining and even enhancing expected job performance. Real life examples are provided. A new, practical approach to the supervision of police officers designed to improve police job performance will be presented.

WORD COUNT
2,871

REPRODUCIBLE RESEARCH
The SPSS code and data file used for the simulations reported in Tables 1-4 can be made available.

TIME LIMIT REQUEST: 80 minutes

LEARNING OBJECTIVES

1. Describe one method to determine the proportion of new police officer hires that are expected to be false positives.

2. Describe two research-based approaches designed to improve both diversity in hiring and expected job performance of police officers.

3. Describe two research-based reasons for using tests of *g* on a pass-fail basis only.

4. Describe two administrative approaches designed to increase both expected job performance and diversity in hiring of police officers.

INTENDED AUDIENCE

This is intended for an audience schooled and experienced in employee selection testing, but all doctorate-level I/O psychologists should be able to understand much of the material.

BIO FOR JOEL P. WIESEN

Dr. Wiesen heads his own firm, Applied Personnel Research. He specializes in employee selection testing. In 1975, he was hired by Massachusetts to validate its civil service examinations. Since 1993, he has consulted on employee testing and developed written tests for government and business. He is approved by the American Psychological Association to sponsor continuing education courses for psychologists. He has served as an expert on testing matters for Attorneys General in two states, the US Department of Justice, other governmental entities, private law firms, and other organizations. He received a doctorate in psychology from Lehigh University and is licensed as a psychologist in three states.

Introduction

Many police managers are stymied in their attempts to hire black police officers and many citizens are dismayed at apparently clear examples of faulty job performance by police officers. This tutorial explores these topics anew using simulation methods and presents 3 psychometric and 3 approaches, to help police departments (PDs) hire ethnically diverse officers and enhance job performance. This tutorial focuses mainly on hiring and presents one new, practical approach to supervising police officers.

Many False Positive Hires, with Few Black Hires

A psychometric analysis reveals a high expected proportion of false positives among newly hired police officers (i.e., new hires who will fail on the job) and severe adverse impact on black candidates as a group. Consider this hiring scenario for a large jurisdiction: (a) corrected validity of general mental ability *(g)*, r=.24; (b) 10,000 candidates, including 1,000 black candidates; (c) half of the job candidates can do the job, (d) a standardized B-W mean score difference (*d*) of 1 on the test of *g*; and (e) a *d* of .24 for job performance (based on the regression formula y=.24x, where y is job performance, .24 is validity, and x is test score for *g*), as summarized in Table 1.

Table 1. Simulation Assumptions: Selection Using a Test of *g*

Validity: Test of *g*, r=.24

(r=.27 for supervisor ratings, which becomes .24 when the correction for unreliability of the predictor is removed, Aamodt, 2004, e.g., page 35)

Number of candidates: 10,000 candidates, including 1,000 black candidates

Number of hires: Hire 1,000 new officers (top 10% of scores)

Quality of candidates: Half of candidates have the ability to be successful officers

B-W difference on *g*: *d*=1.0

B-W difference on job: *d*=.24

This simulation shows that 33% of the hires are expected to fail on the job with only 13 black officers hired (rather than the expected 100). The expected adverse impact ratio is .12 (see

Table 2), consistent with previous simulations (e.g., Wiesen, 2018).

Table 2. Simulation Outcomes for Selection Based on *g*

Hires who **cannot** do job: **329 (33%,** chance=50%)

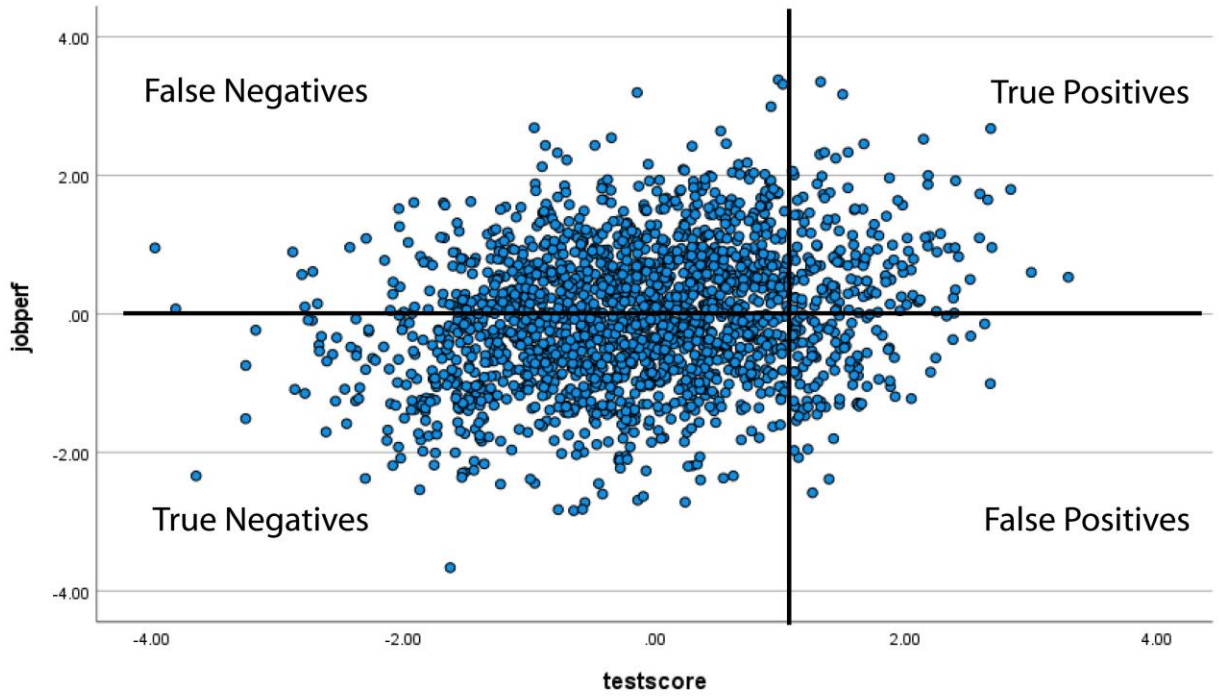
Black candidates selected: **13**

White candidates selected: **987**

Adverse impact: **.12** (severe)

A (reduced n) scatter plot shows a considerable number of false positives (in the lower right quadrant of Figure 1).

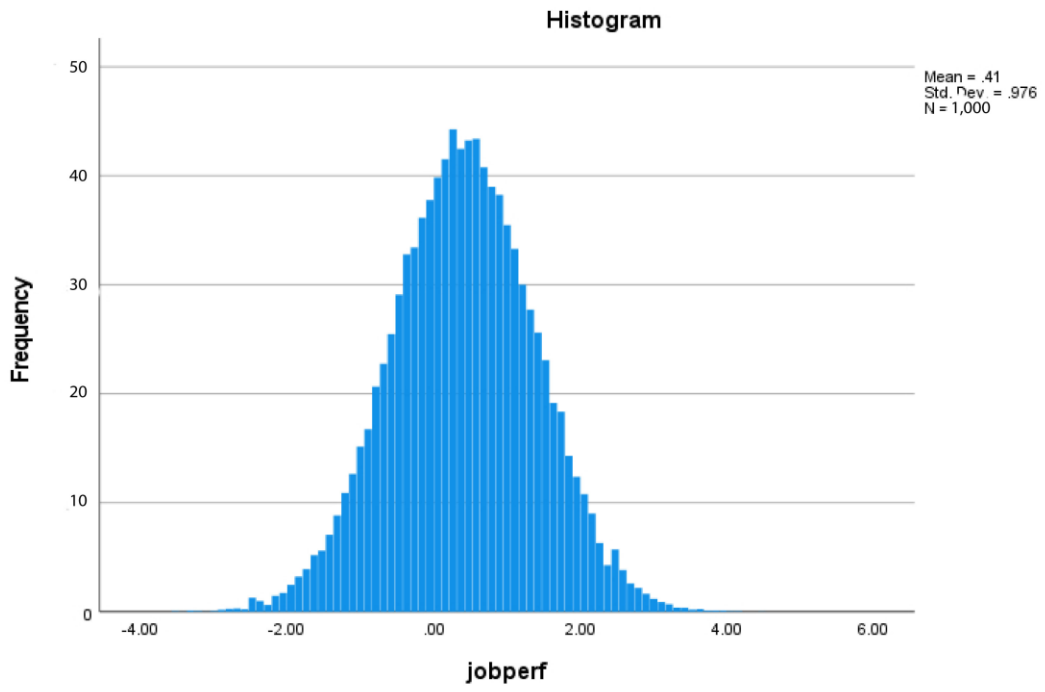
Figure 1. Scatterplot of Test Score and Job Performance Simulation



The proportion of new hires expected to be able to perform the job successfully can be expressed as an accuracy statistic known as the Positive Predictive Value (PPV), in this case .67. The PPV is used in other fields (e.g., medicine) as one of several prediction accuracy statistics. PPV is rarely used by IO psychologists; we prefer to use r. Hiring managers want all new hires to be capable of job success. Thus, hiring managers envision the world in terms of PPV rather than r. I recommend that the profession of IO psychology develop guidelines for acceptable levels of PPV. PPV is partly a function of the quality of the candidates (see below).

Mean expected job performance reveals a similar disturbing pattern of poor job performance. The simulation results in 7 hires 2 or more standard deviations below the candidate mean in job performance, 74 hires 1 or more standard deviations below the mean, and 337 hires at or below the mean. (See Figure 2.) The mean job performance among those hired is .41, a substantial improvement over a the z of 0 that would result from chance selection. Yet, 74 of those hired are well below the mean of all **candidates** in job performance.

Figure 2. Frequency Distribution of Job Performance for Those Hired



I take the results of this simulation as renewed impetus for searching for better ways to

select police officers. In fact, the actual PPV of police hiring is probably below 50% based on the following logic. Police officers need characteristics beyond *g*. Surely some of the non-measured, non-*g* characteristics are crucial and some job candidates lack those characteristics. If we assume that 50% of the candidates lack sufficient levels of these other characteristics to perform the job, then 50% of the 67% of new hires who were identified by the written test as true positives actually are expected to fail on the job. This leaves us with a PPV of 34%, meaning 2/3 of the new hires are expected to fail on the job. I take the results of this extension of the simulation as even greater impetus for searching for better ways to select new police officers.

The Main Cause for Adverse Impact on Minority Candidates

The main reason for the frequent failure to hire a diverse workforce is the adverse impact on minority candidates resulting from non-zero weighting of traditional tests of *g*. (Possible reasons for the black-white difference in mean test score are considered elsewhere, e.g., Wiesen, 2001).

Ranking Based on Composites of *g* and Personality

Composites of measures of *g* and personality are (increasingly) used rather than measures of *g* alone. Such composites also are expected to result in high numbers of false positive hires and adverse impact, although somewhat less severe than hiring based on *g* alone, as shown by the next simulation.

Let us modify the simulation above as follows: (a) select based on a composite of *g* and a test of personality, (b) assume r=.25 for the personality test, (c) equally weight *g* and personality, and (d) assume B-W *d*=0 on the personality test (see Table 3).

Table 3. Additional Simulation Assumptions, Including a Personality Test

Validity: Personality test, r=.25

B-W difference: zero for personality

Composite: Equally weight *g* and personality

The outcome of this modified simulation is that 26% of the hires are expected to fail on the job and the expected adverse impact ratio is .22 (see Table 4).

Table 4. Simulation Outcomes for Selection Based on *g* and Personality

Hires who **cannot** do job: **257** (**26%)**

Black candidates selected: **25**

White candidates selected: **975**

Adverse impact: **.23** (severe)

The high number of false positives in this modified simulation is due to: (1) modest validity of the composite (r=.35, based on the simulation) and (2) the assumed quality of the candidate group. Any composite that gives considerable weight to *g* is expected to result in large

values for the B-W *d* and severe adverse impact (Sackett & Ellingson, 1997).

Test Batteries Will Differ if Selected Based on Utility Rather than Validity

Test utility is **low** when the candidate pool is highly qualified. For example, if 95% of the candidates have the cognitive ability to do the job, a test of *g* can have utility of no more than 5%. (I ignore cost of recruitment, testing, training, etc., and focus on job performance for the sake of this tutorial.) Non-cognitive abilities with lower validity can have higher utility when the candidate pool is relatively weak in measured areas. Some large PDs require candidates to have a college degree. Arguably, a very high proportion of candidates with a college degree have the cognitive ability required to be a police officer, but a much smaller proportion of college graduates have the personality characteristics needed to be successful officers. The utility for a non-cognitive test will be higher than the utility for a test of *g* in an candidate group of college graduates.

Six Approaches to Improve Expected Job Performance and Increase Diversity in Hiring

This tutorial will describe and provide psychometric support for 6 approaches to hiring police officers. The approaches fall in two general categories, as follows:

- Psychometric

  (1) Use tests of general cognitive ability (*g*) on a pass/fail basis only, if at all.

  (2) Select tests based on utility, not validity.

  (3) Test specific job-related cognitive abilities and other characteristics that show small ethnic group differences (i.e., *d* near zero).

- Administrative

  (4) Have testing consultants project the number of diversity hires and expected job performance.

  (5) Focus on quality over quantity in recruiting, especially minority candidates.

  (6) Grant residency preference in hiring.

(1) Use Tests of General Cognitive Ability (*g*) on a Pass/Fail Basis Only, If at All.

Use a traditional multiple choice (M/C), entry-level test of *g* on a P/F basis, with a job-related passing point. Even weighting the traditional M/C test at less than 50% is likely to result in severe adverse impact in hiring (Sackett & Ellingson, 1997, Tables 1-3, 5). This is a major reason why many PDs that have moved away from total reliance on the traditional M/C test still have difficulty hiring a diverse workforce (e.g., Diversity on the Force, 2015). By setting a suitable passing point on a traditional M/C test and ranking on test components that have little or no adverse impact, we can expect the new hires to have sufficient intelligence and the requisite personal and interpersonal skills and other abilities to do the work, yet with less adverse impact.

This P/F use of *g* is contrary to some apparently obvious implications of meta-analysis research, yet there are many strong arguments for using a M/C test of *g* P/F, if at all. First, some PDs require a college degree, vitiating the validity of *g* among the candidates. Second, the validity of *g* for police is lower than for other jobs (r=.24, Aamodt, 2004). Third, the historic meta-analytic findings in favor of *g* are tempered by more recent research. When the validity of

assessment centers and tests of *g* are compared directly, the validity of *g* is much lower than that of broader assessment center exercises (Sackett, Shewach and Keiser, 2017). Additionally, the correlation between leadership and intelligence is dramatically higher for observational than for M/C or short answer measures of intelligence, .60 vs .19 (Judge, Colbert & Ilies, 2004, Table 2).

There is reason to think that job criteria are biased. The B-W mean difference in job performance on many jobs is only .5 sd, while the difference in test performance often is 1 s.d. The regression formula y=rx (where y is job performance and x is test performance) would only explain the .5 versus 1 sd difference in job and test performance if employers hired randomly or from the whole range of test performance, which is not the case. This suggests that the job criteria may be biased. Various lines of research support such bias. Studies of salaries show tall people are paid more than short (Judge & Cable, 2004), men are paid more than women both within and between occupations (Hegewisch & Barsi, 2020), and physically attractive people of both genders are paid more than unattractive (Marlowe, Schneider & Nelson, 1996). Further, claims of lack of differential validity have been challenged (e.g., Aguinis, Culpepper & Pierce, 2010 and 2016).

(2) Select Tests Based on Utility, Not Validity.

A relatively low validity test (e.g., personality with r=.15) can have higher utility than a higher validity test (e.g., *g* with r=.25) in hiring situations often encountered (Wiesen, 2017, 2018).

(3) Test Specific Job-Related Cognitive Abilities and Other Characteristics That Show Small Ethnic Group Differences (i.e., *d* near Zero).

Some examples of non-cognitive (or less, or differently cognitive) test areas and a test mode (some with high validity) that may have higher utility than a M/C test of *g* are: face recognition and memory; creative problem solving; oral communication; conscientiousness, integrity, etc.; new ways to measure intelligence; and structured oral exams. Recognition and memory for minority faces can be expected to have reverse impact because remembering and identifying minority faces is easier for members of that minority group (e.g., Levin, 2000). Creative problem solving is expected to have low *d* due to its low correlation with *g*. Oral communication has shown low *d* for law enforcement candidates (Hausknecht, Trevor & Farr, 2002). Integrity tests show *d* around zero for race (Ones & Viswesvaran, 1998) and high validity (r=.41), and the highest incremental validity over *g* (Schmidt & Hunter, 1998). Although personality has a reputation of having low validity, that validity has been found to increase with time (r=.18 to r=.45, year 1 to year 7 of med school, Lievens, Ones & Dilchert, 2009).

There are some newer ways to test intelligence that show lower *d* values (e.g., Agnello, Ryan, Yusko, 2015). There are various facets to *g*, some with smaller *d*s, and the facets of *g* are not equally valid for various jobs (e.g., Wee, Newman & Joseph, 2014). Structured oral exams have the highest validity of all tests, r=.57 (Aamodt, 2016, Table 5.2, page 194). In the past 20 years, the B-W *d* for structured interviews has been zero (Levashina, Hartwell, Morgeson & Campion, 2014, Table 3, page 254), perhaps due in part to the structured interviews assessing traits that go beyond *g* (e.g., conscientiousness and interpersonal).

(4) Have Testing Consultants Project the Number of Diversity Hires and Expected Job Performance.

Police managers often are shocked when a new selection system for entry-level police officer results in severe adverse impact. Testing consultants often (typically?) recommend using M/C tests of *g* for selecting police officers. When asked by municipal officials to design a selection battery to help the jurisdiction hire a diverse group of new police officers, consultants often respond by including a test with low B-W *d* as a weighted component in a test battery. However, this approach often results in severe adverse impact. By the time the police or municipal officials learn this, the exam has been given and graded and it is too late to make changes to the exam. The new police officer selection system becomes a political and, perhaps, a legal liability.

The solution I propose involves changing the role of consultants in designing the selection system for entry-level police officer. In the RFP process, require the consultants to give the municipal officials numeric predictions of the expected number of diversity hires and level of job performance (see Figure 3). The municipal officials can, in turn, use the information provided to decide which of various testing proposals to accept. That way, police chiefs and other municipal officials will be in a good position to make decisions concerning the relative managerial, legal, and political value of:

- diversity in newly hired police officers
- the cost and time required for developing and conducting the selection process
- the expected level of job performance
- practicality and transparency of the selection system

Municipal officials can help design the entry-level testing program if given numerical

estimates of the predicted diversity in hiring and the expected level of job performance for various testing approaches.

Pros:

- weighty decisions will be made by the responsible municipal officials, and
- municipal officials are less likely to be surprised after the fact by the level of adverse impact of a new hiring process.

Cons:

- municipal managers have limited expertise in psychometrics, so both they and their consultants will have to work hard to bring them up to speed on such concepts as: utility, expected level of job performance, expected level of adverse impact, and variance in these expected values,
- municipal officials may be hard pressed to provide some of the information that consultants need to make predictions of the number of diversity hires, and
- consultants will feel uncomfortable making predictions about the number of diversity hires.

Figure 3. Form Proposed to Collect Psychometric Information from Potential Consultants

Police Officer Selection System Proposal Evaluation Form

Topic	Selection System Approach 1	Selection System Approach 2	Selection System Approach 3	Selection System Approach 4
1. Projected Number of Hires				
Projected number of whites hired				
Projected number of blacks hired				
Projected number of Hispanics hired				
Projected number of other minorities hired				
Projected number of men hired				
Projected number of women hired				
2. Projected Adverse Impact				
Projected Adverse Impact: Blacks				
Projected Adverse Impact: Hispanic				
Projected Adverse Impact: Other minorities				
Projected Adverse Impact: Women				
3. Projected Average Job Performance (fill in one of the three lines below)				
Average job performance using SAT scale (where average of all applicants is 500, standard deviation is 100)				
% of hire who will be successful on the job				
Other measure of projected job performance (described in narrative of proposal)				
4. What is the basis for ranking candidates? (narrative)				
5. Cost				
Total Cost for projected number of candidates				
Adjustment +/- for more or fewer candidates				
6. Time to implement full selection system				
7. Evaluation of Proposer (acceptable or not, plus narrative)				
Past experience				
Expertise				
References				
8. Quality of Proposal (acceptable or not, plus narrative)				
Understandable, logical, and complete (do I trust projections)				
Selection system practical and reasonable				
9. OVERALL EVALUATION (Rate on a 1-9 scale or rank from low to high)				

For more information on using this form contact:
Dr. Joel P. Wiesen, Director
Applied Personnel Research
62 Candlewood Road; Scarsdale, NY 10583
projections@appliedpersonnelresearch.com
(617) 244-8859

Instructions to Chief
Summarize 1-5 from proposals
Rate 6-8 numerically or with narrative
Rate 9, considering 1-8.

(5) Focus on Quality over Quantity in Recruiting, Especially Minority Candidates.

Tests of *g* alone result in high false positive rates in hiring decisions. To help reduce false positive rates, PDs should focus on quality over quantity in recruitment. Additionally, with smaller candidate groups, there would be less severe adverse impact. (A numeric example will be presented.)

With a high proportion of well-qualified candidates, the number of false positive police officer hires drops dramatically. In the extreme, if all candidates were highly qualified, there would be no false positive hires. In this respect, recruitment is more important than test validity in the selection of entry-level police officers. (Recruitment might be facilitated if high school rank is accepted as a substitute for passing a test of *g*.) Improving candidate recruitment is easier and faster than developing more valid tests.

(6) Grant Residency Preference in Hiring.

A disproportionate number of police officer jobs in cities with a high proportion of minority residents are filled by candidates from wealthy white suburbs with better schools. A possible way to address this is to grant residents some preference in hiring. There is a validity argument: Police officials report that in-depth knowledge of the community is very important for effective police patrolling (Blacks in IO Psyhology, 2020).

Real Life Examples (detail omitted due to SIOP proposal word limit)

- Bridgeport, CT
- Oklahoma City, OK
- US Army
- NYC 2012 Firefighter Exam
- Columbus, OH, Police Officer Exam

Other Methods to Improve Police Officer Job Performance

Police officers do much of their work with no supervision in real time. Body cams could be used much more effectively. Now videos for select incidents are reviewed after the fact. Police officers are often acting with no Sergeant (immediate supervisor) present. If sergeants viewed the body cam videos in real time, they could provide immediate feedback and guidance to officers. Real time oversight is now not possible until a Sergeant arrives on the scene or is queried by a police officer by radio. Vision is crucial to supervision.

References

Aamodt, M. G. (2004). *Research in Law Enforcement Selection*. Boca Raton, FL: Brown Walker Press.

Aamodt, M. G. (2016). *Industrial-Organizational Psychology: An Applied Approach* (8th ed.) Boston: Cengage Learning.

Aguinis, H., Culpepper, S. A. & Pierce, C. A. (2010). Revival of Test Bias Research in Preemployment Testing. *Journal of Applied Psychology, 95,* 648-680.

Aguinis, H., Culpepper, S.A., & Pierce, C.A. (2016). Differential prediction generalization in college admissions testing. *Journal of Educational Psychology, 108,* 1045-1059.

Agnello, P., Ryan, R. & Yusko, K. P. (2015) Implications of modern intelligence research for assessing intelligence in the workplace. Human Resource Management Review 25, 47–55.

Blacks in IO Psychology (2020). *Community Call Addressing Systemic Issues in Law Enforcement Workplaces* [Video]. YouTube https://www.youtube.com/watch?v=P2NfkdDNiBA&feature=youtu.be

Diversity on the Force: *Where Police Don't Mirror Communities; A Governing Special Report* (September, 2015). Washington: Governing.

Hausknecht, J. P., Trevor, C. T. & Farr, J. L. (2002). Retaking ability tests in a selection setting: implications for practice effects, training performance, and turnover. *Journal of Applied Psychology*, 87, 243–254.

Hegewisch, A. & Barsi, Z. (2020) *The Gender Wage Gap by Occupation*. Institute for Women's Policy Research. Downloaded 10/13/2020 from https://iwpr.org/wp©content/uploads/2020/07/2020©Occupational©wage©gap©FINAL.pdf

Judge, T. A. & Cable, D. M. (2004). The Effect of physical height on workplace success and income. *Journal of Applied Psychology, 89*, 428-441.

Judge, T. A., Colbert, A. E., & Ilies, R. (2004). Intelligence and leadership: A quantitative review and test of theoretical propositions. *Journal of Applied Psychology, 89*, 542-552.

Levashina, J., Hartwell, C. J., Morgeson, F. P. & Campion, M. A. (2014) The structured employment interview: narrative and quantitative review of the research literature. *Personnel Psychology, 67*, 241–293.

Levin, D. T. (2000) Race as a Visual Feature: Using Visual Search and Perceptual Discrimination Tasks to Understand Face Categories and the Cross-Race Recognition Deficit. *Journal of Experimental Psychology: General, 129*, 559-574.

Lievens, F., Ones, D. S., & Dilchert, S. (2009). Personality Scale Validities Increase Throughout Medical School. *Journal of Applied Psychology, 94*, 1514-1535.

Marlowe, C. M., Schneider, S. L., & Nelson, C. E. (1996). Gender and attractiveness biases in hiring decisions: Are more experienced managers less biased? *Journal of Applied Psychology, 81*, 11-21.

Ones, D. S. & Viswesvaran, C. (1998). Gender, Age, and Race Differences on Overt Integrity Tests: Results Across Four Large-Scale Job Applicant Data Sets. *Journal of Applied Psychology, 83*, 35-42.

Sackett, P. R. & Ellingson, J. E. (1997). The Effects of Forming Multi-Predictor Composites on Group Differences and Adverse Impact. *Personnel Psychology, 50*, 707-721.

Sackett, P. R., Shewach, O. R. & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology, 102*, 1435–1447.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Wee, S., Newman, D. A. & Joseph, D. L. (2014). More Than g: Selection quality and adverse impact implications of considering second-stratum cognitive abilities. *Journal of Applied Psychology, 99*, 547–563

Wiesen, J. P. (2001). *Some Possible Reasons for Adverse Impact.* Presentation at the 16th Annual Convention of the Society for Industrial and Organizational Psychology, San Diego, CA.

Wiesen, J. P. (2016, November). Tools to Increase Diversity and Validity in Hiring Police Officers. *The Personnel Testing Council of Metropolitan Washington Newsletter, XII (3),* 4-11.

Wiesen, J. P. (2017a, March). Tools to Increase Diversity and Validity in Hiring Police Officers – Part II. *The Personnel Testing Council of Metropolitan Washington Newsletter, XII(4),* 6-15.

Wiesen, J. P. (2017b, July). Tools to Increase Diversity and Validity in Hiring Police Officers – Part III. *The Personnel Testing Council of Metropolitan Washington Newsletter, XIII (1),* 6-17.

Wiesen, J. P. (2017c). *Quantitative Considerations in Balancing Validity, Utility, Fairness, and Adverse Impact*. Presentation at the 2017 Conference of the International Personnel Assessment Council, Birmingham, AL, July 19, 2016.

Wiesen, J. P. (2018). *Master Tutorial: Tools to Increase Diversity, Utility, and Validity in Hiring Police Officers*. Presented at the 33rd Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Chicago, IL.

Wiesen, J. P. & Aguinis, H. (2010). *New Methods for Reducing Adverse Impact and Preserving Validity*. Symposium presentation at the 25th Annual Convention of the Society for Industrial and Organizational Psychology, Atlanta, GA.