Tools to Increase Diversity and Validity in Hiring Police Officers

Joel P. Wiesen, Ph.D.

Hiring an ethnically/racially diverse police force can be essential to maintain, strengthen and, too often, rebuild the relationship between a police department and the community it serves. Unfortunately, many police managers are largely stymied in their efforts to hire black police officers due to the pervasive adverse impact (AI) that traditional employment selection procedures have on black candidates as a group. The societal impact of this AI has spurred much thought and research by many I/O psychologists over at least the past 25 years (e.g., Ployhart & Holtz, 2008), but now the societal impact seems more pressing than ever before. As a result I undertook a fresh, critical look at the selection tools that are or could be used to hire police officers. This led me to conclude that there are practical, professionally acceptable selection tools or approaches, most novel or little used, that may be used to support the hiring of a diverse police force while substantially maintaining or even improving test validity. This column is the first of a three-part series that summarizes the fruits of my multi-year quest for such tools. In this column I will introduce the topic and present a few tools. The later columns will describe tests with little or reverse impact, provide some more innovative tools and approaches, give one real-life and a few hypothetical examples of using these tools, and discuss some legal considerations. My ultimate goal is to reinvigorate efforts to hire a higher, more representative proportion of minority police officers while substantially maintaining or even improving the levels of job performance.

Each of the columns will present several tools that are, or have the potential of being, practical, effective, and acceptable to users and other concerned parties. For each tool, a small number of supportive references or a logical rationale for its use is presented. The few citations to the literature are not meant to be exhaustive, but only illustrative. The main focus of these columns is on the hiring of black police applicants, but the tools and principles often apply to Hispanic applicants as well. I use the term “minority” to refer to both black and Hispanic applicants.

The Main Cause for Adverse Impact on Minority Applicants

To restate the obvious, the main reason for the frequent failure to hire a diverse academy class is the large AI on minority applicants that pervasively results from ranking candidates based, even in part, on traditional multiple choice (MC)

tests. This AI is largely driven by the mean score difference between minority and white applicants. There are many and varied reasons for this mean score difference (e.g., different educational experiences or opportunities related to schools, family economic resources, or geographic areas where people live or work) and these multiple reasons or causes cannot all be addressed directly by an employer's hiring process.

Tool 1:
Measure the Ability to Remember and Identify Faces that Mirror the Community

Remembering and identifying minority faces is easier for members of that minority group (e.g., Levin, 2000). So, it is reasonable to expect a measure of this ability to have reverse impact, that is, to favor black applicants. Being able to remember and recognize faces of the racial/ethnic groups found in the community is likely to be a predictor of good job performance. The level of validity of this ability may be evaluated using content and/or criterion validation.

Tool 2:
Rely on Content Validation

There is a long history of our profession giving more credence to criterion-related validation research than to content approaches. This is seen clearly in the federal Uniform Guidelines on Employee Selection Procedures (UGESP, EEOC, 1978). Yet this preference is at odds with the current and the previous two editions of the Standards for Educational and Psychological Tests. The current edition clearly states, "Validity is a unitary concept. ... Like the 1999 Standards, this edition refers to types of validity evidence, rather than distinct types of validity" (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014, page 13). We have to go back to the 1974 Standards to find a statement in support of the primacy of criterion validation, such as, "Other forms of validity are not substitutes for criterion-related validity" (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1974, page 27). Given the relative dearth of criterion-related research on the selection of police officers and the difficulty of conducting such research, we will be very slow to adopt new testing tools and approaches unless we rely on content validation research. I hope the UGESP will not serve as a barrier to the reduction of AI by the adoption of non-traditional, content valid selection tools.

Tool 3:
Empower the PD to Help Guide the Design of the Selection System

There are many decisions to be made when designing and implementing an employee selection system. To empower a PD to contribute to such decision making, we can and should provide projections of the expected level of AI and expected level of job performance of any proposed selection system, both based on plausible assumptions concerning the numbers of applicants and hires, and validity. Depending on these projections, a PD might request its testing consultant to provide one or more alternative selection system designs, with associated projections of expected levels of AI and job performance. Decisions

about making tradeoffs between expected levels of AI and job performance, or between screening costs and AI, may best be made by the hiring officials rather than testing experts.

Tool 4:
Shorten the Application Period

Simply allowing fewer applicants can be expected to increase the proportion of minority applicants hired. For this reason, long application periods can backfire in terms of AI. When smaller proportions of applicants are selected, we should expect more severe AI (e.g., Sackett & Ellingson, 1997, Table 1, and pg. 711, par. 2). This approach may be critiqued in terms of reduced test utility. Yet our profession focuses on validity almost to the exclusion of the consideration of the selection ratio and its impact on utility. (There is no caution against high selection ratios in SIOP, 2003.) It would be curious if one of the rare times the profession voices concern about a high selection ratio is when this is suggested in order to reduce AI. In any case, the effect of the number of applicants on AI and utility may be statistically estimated in advance, and that information can be provided to the PD.

Tool 5:
Do Not Rely on a Weighted Average of Low and High d to Cure the AI Problem

Do not rely on adding low d tests to a traditional MC test of g to solve the AI problem. (This is more of a caution than a tool.) Even weighting the traditional M/C test at much less than 50% and including other tests with low d in a weighted composite is likely to result in severe adverse impact in hiring (Sackett & Ellingson, 1997, Table 1, pg. 710). This is a major reason why the many PDs that have moved away from total reliance on the traditional M/C test and include other measures in their entry-level examinations (e.g., measures of work-style) still have difficulty hiring a diverse workforce (e.g., Diversity on the Force, 2015).

Parenthetically, I note that Sackett and Ellingson (1997) have been cited as sounding a strong warning of the danger of increasing d due to adding predictors to a test of g. This strong warning, not found in the article, is overstated. For example, Sackett and Ellingson report that a four predictor composite with a sum of d's of 1.5 is expected to have a maximum d of .75, and a d of .54 if the average intercorrelation of the predictors is .30 (their Table 4). More correctly stated, the practical guidance of Sackett & Ellingson is that adding predictors with small d's to a test of g is likely to yield a composite with lower d than the test of g alone, which may not be enough to reduce AI to acceptable levels (Sackett & Ellingston, 1997, p. 712).

Tool 6:
Use a Traditional Test of g on a Pass-Fail Basis or Not at All

This tool is so contrary to some apparently obvious implications of meta-analysis research that I will discuss it more fully than the tools mentioned above.

The widespread use of traditional MC tests to rank police applicants is based in large part on the meta-analytic findings of relatively high criterion-related validity of tests of g for hiring for a wide range of jobs (e.g., Hunter & Hunter, 1984) and the widespread finding of a linear relationship between test and job performance scores (e.g., SIOP, 2003, page 21). Yet there are many strong, logical arguments in favor of using tests of g on a pass-fail basis in the hiring of POs, if such tests are used at all. Let's look at a few such arguments (not in order of importance).

First, many PDs require a college degree. Only about 40% of people 25-34 years of age have any type of post-secondary school degree (US Department of Education, 2012). The criterion-related validity of M/C tests of g is surely much lower among college graduates than among the general population. But the professional literature does not caution against placing major reliance on g for jobs for which employers require a college degree (e.g., there is no mention of this in SIOP, 2003). If the proportion of the age appropriate population that completes a college degree is 40% and if the validity of g is .40 for the unrestricted population, then the validity of g may be expected to be only .23 among college graduates (based on a back of the napkin Monte Carlo study). It may well be that various alternative selection procedures have higher validities than .23. Stated differently, among college graduates it may well be that non-cognitive traits are more predictive of police officer job performance than cognitive

ability. The import of g as compared with non-cognitive abilities and characteristics may also be considered in terms of job analysis findings (which may find non-cognitive abilities to be of equal or more import than cognitive).

Second, the validity of g for police may be lower than for other jobs. One meta-analysis found much higher validity for training than job performance: r = .76 versus r = .38 (Hirsh, Northrupt & Schmidt, 1986, as cited in Schmidt and Hunter, 2004, page 166, Table 3). (I could not find these exact corrected r values in the original article, Hirsh, Northrupt & Schmidt, 1986, Tables 8 and 11.) Other occupations show large differences in the same direction, but law enforcement showed the largest difference. The .38 is one of the lowest validities reported for a specific occupation. This large difference and the relatively low criterion-related validity coefficient indicate that much of the variance in job performance of police officers is due to factors other than cognitive ability. Further, the Hirsh et al. meta-analysis was based on studies done long ago, when it was less common to require entering police officers to have a college degree, so such correlations may be expected to be lower today. A more recent meta-analysis found the validity of g for police academy performance to be .44 and job performance .15 (Aamodt & Flink, 2000). A later literature review reported the validity for cognitive ability tests to be .27 for supervisor ratings and .62 for academy performance, after correction (Aamodt, 2004, page 35, and Table 3.1). This lower level of validity for job performance may well reflect the extent to which police training academies depend on written M/C test scores (McHenry, Hough, Toquam, Hanson & Ashworth, 1990, Table 8, p. 349). If we consider the correlation with academy performance to be inflated due to the use of multiple choice questions, we could conclude that the more defensible corrected correlation of .27 for job performance is no better than the validity of alternative selection procedures. Further, the corrected validity of .27 for job performance may be inappropriately inflated. Aamodt reports “correcting for attenuation in the predictor and criterion as well as range restriction” (Aamodt, 2004, page 35, and Table 2.5). Aamodt based the correction for predictor unreliability for tests of g on reliability of .82, a lower value than is often seen with 100 question police entrance exams, in my experience. In any case, when we use a test operationally, its validity reflects its actual reliability, not a perfect, corrected reliability. Further, calculated test reliability reflects the range of ability of the people tested. So, correcting for both unreliability and restriction of range may be double dipping, at least to some extent. (Of course, Aamodt is not unique in using both these corrections.)

Third, the meta-analytic findings in favor of g are tempered by other research findings. With respect to measuring cognitive ability, which is the main area that traditional M/C tests measure, not all measures are interchangeable (e.g., Hough, Oswald & Ployhart, 2001; Lang, Kersting & Lang, 2010; Schmitt, 2014). For example, the correlation between leadership and intelligence is dramatically higher for observational than for paper and pencil (M/C or short answer) measures of intelligence, .60 vs .19 (Judge, Colbert & Ilies, 2004, Table 2). This supports the use of measures of cognitive ability other than M/C tests. Importantly, different aspects of cognitive ability have different size B-W differences (e.g., Wee, Neuman & Joseph, 2014), suggesting that we carefully select the aspects of cognitive ability that are measured.

Fourth, there are murky issues concerning the fairness of traditional M/C tests. The B-W mean difference in job performance on many jobs is only .5 sd, while the difference in test performance often is 1 sd. To the statistically naive, the larger mean score difference in test performance than job performance is trou-

bling. Yet the seemingly logical statistical explanation of this .5 vs 1 sd difference is flawed. The statistical explanation relies on the (much) less than perfect relationship between test and job performance. The statistical relationship is simply the regression formula: y = r*x (where y is job performance, x is test performance, and r is the validity coefficient). If we let r = .5 and plug in the .5 and 1 sd differences in the place of x and y, we get this seemingly tidy relationship: .5 = .5 * 1. However, this tidy relationship 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. But if selection of applicants is based on test score (e.g., from the top 10%), the mean levels of job performance for the minority and non-minority applicants selected should be quite similar. That the job performance discrepancy is as large as would be expected based on random selection is puzzling and suggests to me that the job criteria may be biased.

Indeed, there are strong indications that the criteria used in many criterion validation studies may be unfair. Studies of salaries show that the workplace often makes apparently biased evaluations of job performance: tall people are paid more than short (Judge & Cable, 2004), men are paid more than women (Hegewisch, Williams & Henderson, 2011), and physically attractive people of both genders are paid more than unattractive (Marlowe, Schneider & Nelson, 1996). If these attributes are related to salary, presumably for spurious reasons, perhaps skin color may also be a source of bias in job criterion measures.

Although there is a body of research that suggests that traditional M/C tests do not show differential validity - that is, that tests in general are equally valid for blacks and whites - recent research by Herman Aguinis and his colleagues suggests otherwise, showing that tests are sometimes biased in favor of one group and sometimes in favor of the other group, with variance beyond what would be expected by chance (e.g., Aguinis, Culpepper & Pierce, 2010 and 2016).

Even if we accept the position that our tests are unbiased predictors of job performance, there is reason to question their fairness. There is research that indicates that black and Hispanic employees do not encounter a level playing field at work. For example, more black and Hispanic employees than white report being the target of derogatory and exclusionary behaviors, with d of .49 for black and .42 for Hispanic employees (Bergman, Palmieri, Drasgow & Ormerod, 2007, Tables 2, 5). To the extent that our tests accurately predict biased job criterion data, the tests are biased.

In any case, it seems obvious that police work calls on many diverse abilities, including oral communication, interpersonal skill, and honesty, as well as innovative problem solving. If large differences in g drive the ranking of job applicants, candidates who score high on these other abilities are not likely to be hired.

For these reasons, it seems appropriate to either try to use a traditional M/C test in ways that minimize its AI or to replace such a test. Possible ways to select police officers without relying on the traditional M/C test for ranking purposes will be discussed in the next column.

Author Bio:

Dr. Wiesen has served as an expert in testing-related employment discrimination litigation, both for defense and plaintiff. He is a long time (long-distance) member of PTC/MW. For over 15 years he headed the Massachusetts civil service test development and validation program. Now he is owner and Direc-

tor of the consulting firm named Applied Personnel Research located in Scarsdale, NY. He is a published test author. He is licensed as a psychologist in three states.

Questions and comments on this column are welcome.
Write Dr. Wiesen at: j@jpwphd.com.

References

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

Aamodt, M. G. & Flink, W. (2000). Predictors of Police Academy Performance. Paper presented at the Annual Conference of the International Personnel Association Assessment Council.

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.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1974). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Bergman, M. E., Palmieri, P. A., Drasgow, F. & Ormerod, A. J. (2007) Racial and Ethnic Harassment and Discrimination: In the Eye of the Beholder? Journal of Occupational Health Psychology, 12, 144–160.

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

Equal Employment Opportunity Commission (EEOC), Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. Federal Register, 43, 38290-39315.

Hegewisch, A., Williams, C. & Henderson, A. (2011) The Gender Wage Gap by Occupation, Updated April 2011. Institute for Women's Policy Research Fact Sheet #C350a. Downloaded 5/23/2011 from http://www.iwpr.org/publications/pubs/the-gender-wage-gap-by-occupation-updated-april-2011/at_download/file

Hirsh, H. R., Northrop, L. C. & Schmidt, F. L. (1986) Validity Generalization Results for Law Enforcement Occupations. Personnel Psychology, 39, 399-420.

Hough, L. M., Oswald, F. L. & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. International Journal of Selection and Assess-

ment, 9, 152-194.

Hunter, J. E. & Hunter, R. F. (1984). Validity and Utility of Alternative Predictors of Job Performance. Psychological Bulletin, 96,72-98.

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.

Lang, J. W. B., Kersting, M. & Lang, J. (2010). General mental ability, narrower cognitive abilities, and job performance: the perspective of the nested-factors model of cognitive abilities. Personnel Psychology, 63, 595-640.

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.

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.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A. & Ashworth, S. (1990). Project A Validity Results: the Relationship Between Predictor and Criterion Domains. Personnel Psychology, 43, 335-354.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. Personnel Psychology, 61, 153-172.

Sackett and Ellingston (1997). The effects of forming multi-predictor composites on group differences and adverse impact. Personnel Psychology, 50, 707-721.

Schmidt, F. L. & Hunter, J. (2004). General Mental Ability in the World of Work: Occupational Attainment and Job Performance. Journal of Personality and Social Psychology, 86, 162-173.

Schmitt, N. (2014) Personality and Cognitive Ability as Predictors of Effective Performance at Work. Annual Review of Organizational Psychology and Organizational Behavior, 1, 45-65.

Society for Industrial and Organizational Psychology, Inc. (2003). Principles for the Validation and Use of Personnel Selection Procedures (4th ed.) Bowling Green, OH: Author.

US Department of Education, 2012. New State-by-State College Attainment Numbers Show Progress Toward 2020 Goal. http://www.ed.gov/news/press-releases/new-state-state-college-attainment-numbers-show-progress-toward-2020-goal (Retrieved February 2, 2016)

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.

Volume XII Issue 3

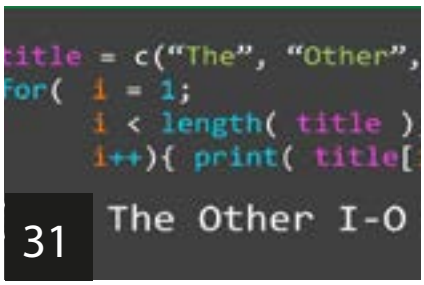
ptcmw

NEWSLETTER

CONNECT.
SHARE.
GROW.

Featured Mini-Series: Adverse Impact

Dr. Joel Wiesen provides advice about how to increase diversity and the validity in hiring for police officers.



New Column: The Other I-O

Dr. Garett Howardson introduces his new column focused on advancing the use of computer science in I-O Psychology.

NEXT ISSUE: February 2017

ptcmw
CONNECT.SHARE.GROW.

NEWSLETTER

www.ptcmw.org

newsletter.ptcmw@gmail.com

Twitter: @PTCMW

Reprinted with permission from the Personnel Testing Council of Metropolitan Washington.

How to cite this column

This column may be cited in APA format as follows:

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

Note: This is the first column in a three part series.