Tools to Increase Diversity and Validity in Hiring Police Officers - Part II

Joel P. Wiesen, Ph.D.

Previously, I described the imperative to revisit methods of hiring police officers and to find selection tools or approaches to maintain or improve police officer job performance while allowing the hiring of a higher proportion of minority police officers. Six tools were described and supported, the sixth being the use of a traditional test of g on a pass-fail basis or not at all. This second column (of three) presents seven additional tools, with supporting logic or illustrative citations, some of which can be used to address the gap left by the use of Tool 6.

Tool 7:
Rank Applicants Based on High School Rank or Grades

New police officers must learn many laws, rules, and procedures, so measuring learning ability is important. The validity of high school grades may be evaluated in comparison to the validity of SAT scores for predicting college grades. High school grades and SAT scores are approximately equally good predictors of college grades (e.g., r = .37 vs .33 for cumulative college GPA; Berry & Sackett, 2009, Table 1). The SAT is a well-funded, thoroughly researched test. A locally developed employment test would be unlikely to measure the readiness to learn as accurately as the SAT. Accordingly, high school grades could be used to help select police officers (but see below concerning differential validity). High school grades might reflect conscientiousness and emotional stability as well as g, so we might select applicants with high learning ability who are also conscientious and stable. Given the possibility of differences between grading standards in different schools, we might use high school rank (in percentile terms) rather than grades, when possible.

This tool (both high school rank and grades) might work to minimize adverse impact in jurisdictions with multiple high schools, some with predominantly minority student bodies. A federal study (US GAO, 2016) reports that 16% of high schools are “segregated” (i.e., at least 75% minority student body), with the proportion of such schools rising over the past 12 years. These segregated schools are concentrated in poorer neighborhoods (approximately 60% of schools with high-poverty enrollment are segregated). Segregated schools tend to have fewer academic resources, supporting the use of high school rank to level the playing field in identifying higher g applicants (compared to tests of academic accomplishment). Jurisdictions should consider recruiting at highly segregated schools.

Tool 8:
Rank Based on a Structured Oral Exam

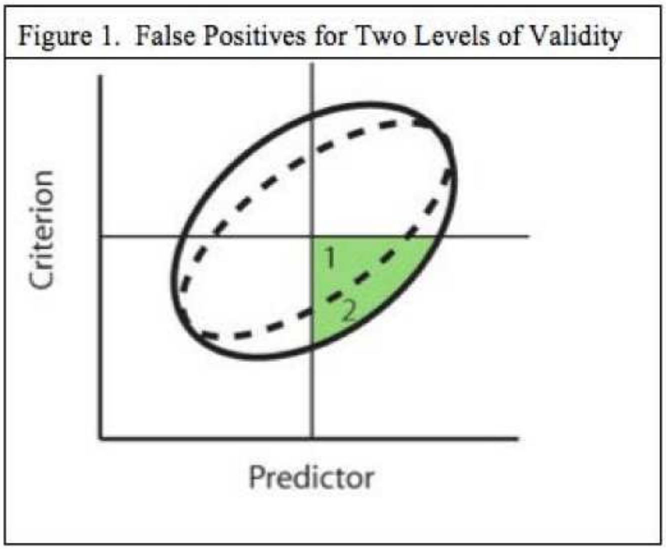
The structured oral exam (i.e., structured interview), as commonly implemented, is a highly valid selection technique (the most valid, r = .57, according to Aamodt, 2016, Table 5.2, page 194) and can assess some of the same abilities as a traditional multiple choice (M/C) test of g as well as additional abilities, with much less adverse impact but at a higher per person cost. Research has not determined the reason for this lower adverse impact, but it might simply be due to the oral nature of this type of test versus the written nature of a M/C test, or due to the more complex and realistic nature of the questions and the less structured format of the answers, tapping problem solving in a more realistic fashion, or tapping other abilities or attributes, and giving credit for innovative answers.

Structured oral exams generally have considerably lower adverse impact on black applicants than g tests (e.g., McCarthy, van Iddekinge & Campion, 2010, Table 2). 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 skills, e.g., Huffcutt, Roth, Conway & Stone, 2001, Table 3; Levashina et al, 2014, page 262, par 5). One study involving 1,334 police officer applicants found d slightly below 0 (McFarland, Ryan, Sacco & Kriska, 2004, Table 2).

Use a Test of g on a Pass-Fail Basis

My previous arguments for using a test of g on a pass-fail basis, if such a test is used, are germane to the next tool. Additionally, differential validity and differential prediction provide further reasons not to use a test of g for ranking, or at all. Differential validity results in hiring decisions that can be considered unfair. (Differential validity was found in two large N studies: Mattern, Patterson, Shaw, Kobrin & Barbuti, 2008, Table 2, SAT used for college admission; and by Berry, Clark and McClure, 2011, Table 1, cognitive ability tests used for employee selection, although the pattern is not the same for all employment types.) Differential validity will result in somewhat lower mean job performance for the applicant group for which the test has the lower validity, even if the same selection standard is applied and even if the people in the two groups have identical means and s.d. on both the predictor and the criterion. The reason can be understood as follows. Lower validity results in more false positives and false negatives, but the false negatives will not be hired (or admitted to college). In Figure 1, the false positives are indicated in green. The false positives for the higher validity (dashed) oval are depicted in area “1”. The false positives for the lower validity (solid) oval are depicted in areas 1 and 2. If the test validity is lower for black applicants than for white applicants, then there will be relatively more false positives among newly hired black applicants, compared to white. Therefore, the average job performance for newly hired black applicants will be somewhat lower than for white. Further, the higher false negative rate for black applicants could itself be considered unfair to black applicants as a group.

Figure 1. False Positives for Two Levels of Validity



The literature reports that g tests overpredict job and college performance for black but not white applicants. For example, the SAT overpredicts college GPA by .2 s.d. for black applicants (Mattern, Patterson, Shaw, Kobrin & Barbuti, 2008, Table 3). As a result, the black applicants who are hired or admitted to college will do somewhat less well, on average, on the job or in college (d = -.2 s.d. for college GPA) than white applicants with equal scores. Arguably, such somewhat lower mean job performance for black applicants is not unfair to any individual black applicant, yet any lower group job performance might be damaging to the self-esteem of newly hired or admitted black applicants, and could lead to the formation of negative stereotypes. If a jurisdiction chooses to employ a traditional test of g on a pass-fail basis, it might consider the next tool.

Tool 9:
Allow Alternatives for a Pass-Fail Test of g

If a traditional M/C test of g is used on a pass-fail basis, consider offering alternatives to qualifying by passing the g test. I will suggest two possible alternatives. First, an honorable, three-year military discharge might be allowed in lieu of passing a traditional M/C test. The military uses cognitive ability tests as a selection tool, ensuring a certain level of cognitive ability. This type of honorable discharge indicates the ability to learn the procedures of a paramilitary organization and a military work specialty, and a willingness to follow orders. Based on national data on the racial and ethnic composition of veterans, this tool is not expected to have an adverse impact on blacks, but may have some adverse impact on Hispanics (e.g., Lee & Beckhusen, 2012). Regional data allows more community-specific impact projections. This tool would give veterans with modest education levels an easy route of entry into the selection process, providing an incentive for applying.

Second, a college degree, indicating a certain level of learning ability, is another possible alternative to passing a traditional test of g. Parenthetically, if a college degree is required for hire, that is another reason to use a test of g on a pass-fail basis, if such a test is used. Only 36% of the US population aged 25-34 has attained a Bachelor's degree or higher (Ryan & Bauman, 2016). If the validity of g is .27 for the general population, then the validity of g can be expected to be .15 among this top 36%. Allowing a g test with such modest validity to drive severe adverse impact in the applicant ranking should only be done for compelling reasons. I see none.

Tool 10:
Rank Based on KSAPs and Test Modes with Little or No Adverse Impact

There are KSAPs (a.k.a. KSAOs), tests, and composite scores (more on this below) that have very little or no adverse impact, and some have demonstrated at least a modest level of criterion-related validity for the selection of police officers (e.g., Barrett, Miguel, Hurd, Lueke & Tan, 2003) and many other jobs. For example, there are personality factors and facets with r's in the .15 to .20 range and with small or zero d's (e.g., conscientiousness; e.g., Hough & Johnson, 2013).

There is a substantive body of literature on achievement-oriented personality measures. Briley, Domiteaux & Tucker-Drob (2014) report zero order r's with a college GPA of .20 for conscientiousness and .33 for effort. I do not know of studies of B-W d's for all the various achievement-oriented personality measures, but to the extent that they are measures of personality, the d's are likely to be relatively small.

Concerning test modes, video based tests, for example, have been shown to have promise for increasing validity and decreasing B-W d (e.g., Chan & Schmitt, 1997; Lievens & Sackett, 2006). Additionally, some new measures of intelligence show smaller d's than traditional g tests (e.g, Agnello, Ryan & Yusko, 2015, page 52, par 5).

One critique of such measures is that they have lower validity than g. Yet, even if true, in some situations simply decreasing the selection ratio will maintain the expected level of job performance, making up for any lower validity.

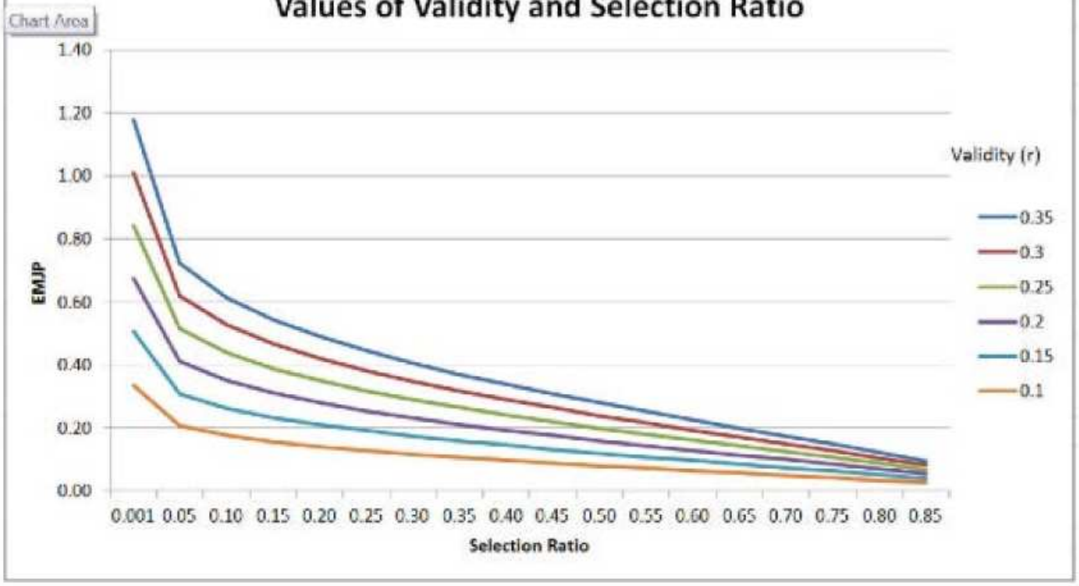
Tool 11:
Maintain or Improve Job Performance by Using Smaller Selection-Ratios

There are three ways to try to achieve a desired level of predicted job performance based on employee selection: improve the test validity, select a smaller proportion of the applicants, and recruit a more qualified applicant pool. This tool focuses on the selection ratio. The utility literature shows that the predicted average level of job performance is a function of both test validity and selection ratio, with validity in the numerator and selection ratio in the denominator. (See Cascio & Aguinis, 2011, page 333 for the basic Naylor-Shine formula for mean criterion score.) This formula demonstrates that we can increase the expected level of job performance by increasing the validity or decreasing the selection ratio (i.e., being more selective). This can be turned into more specific guidance, as follows.

The Naylor-Shine formula sets forth the relationship between test validity, selection ratio, and expected mean job performance (EMJP). Using this formula, we can see that there is only modest improvement in EMJP for commonly observed levels of test validity and selection ratio, and that using stringent selection ratios can sometimes maintain EMJP in the face of decreases in validity. The EMJP for select values of test validity and a wide range of selection ratios are shown in Figure 2 in terms of standard deviation units for a job performance distribution with a mean of zero and a standard deviation of 1.0. (EMJP for any validity and selection ratio can be calculated at: http://jpwphd.com/emjp.) For the selection ratios shown in Figure 2, the EMJP difference is about 0.1 s.d. between tests of abilities with validities that differ by .05. To put 0.1 s.d. in perspective, it corresponds to a difference of 10 points on the SAT.

Figure 2 shows an EMJP decrease of about 0.10 for a difference in validity of 0.05 at a selection ratio of .05, independent of the size of the initial validity coefficient, and the decrease is smaller for larger selection ratios. Assuming that the status quo is a test with r = .25 used at a selection ratio of .15, if we switched to a test with r = .20, we could maintain the EMJP by decreasing the selection ratio to .05. This means recruiting 20 applicants, rather than 3.3, per opening. Almost all the EMJP values in Fig. 2 are less than 0.5 s.d. above the mean for police officers hired randomly.

Fig 2. Expected Mean Job Performance (EMJP) for Select Values of Validity and Selection Ratio



Reducing the selection ratio will result in more adverse impact unless the selection measure has a d of zero or less (see my first column, Tool 4). This tool should be used with caution.

Even if we decide not to recruit additional applicants, the EMJP only shows minor changes resulting from a decrease in validity from .25 to .20. At a selection ratio of .10 and validity of .25, the EMJP is .44. (If selection was by chance, the EMJP would be zero.) With the same selection ratio but a validity of .20, the EMJP is .35, a difference of .09 s.d. units, corresponding in size to 9 SAT points. This difference is a small loss to achieve diversity in hiring. With a validity of .15 and a selection ratio of .10, the EMJP is .26, only .18 s.d. units less than this example's status quo. Even this larger difference, comparable to 18 SAT points, is a small loss to achieve diversity in hiring.

Tool 12:
Use Measures with Reverse Impact: Measures of Implicit Bias

A little used type of test, now some 30 years old, known as the Implicit Association Test (IAT) purports to measure implicit bias in an unobtrusive fashion. (While I was preparing this column, the moderator of the second presidential debate asked Mrs. Clinton whether she "believed that police are implicitly biased against black people." Mrs. Clinton responded, "Implicit bias is a problem for everyone, not just police." So this is a topic with which the general public has some familiarity.) Tests of freedom from bias tend to favor minority applicants because minority group members tend to be less biased than white people against other minority group members (Axt, Ebersole & Nosek, 2014).

One meta-analysis (by a SIOP Fellow) argues against using IATs, saying IATs have shown only low levels of validity for predicting practically useful (non-EEG) criteria (Oswald, Mitchell, Blanton & Tetlock, 2013). However, even that study reported meta-analyses showing IAT validities of .19 for person perception and .14 for microbehavior (the two categories of studies with the largest N's in their Table 3, excluding EEG studies). These validities are comparable to those of other non-cognitive factors often used for employee selection.

Tool 13:
Use a Composite with No Mean Score Difference

A composite including g that has no mean score difference can be formed by adding one or more valid measures with sufficient reverse impact (negative d; e.g., face recognition or IAT). This may be criticized as adding measures of low or unknown validity to a measure of g, thereby lowering validity and diluting standards. Let's examine this critique quantitatively.

What happens to a validity of .27 if we add a measure with reverse impact and then rank based on the composite? The validity of the composite depends on the validity of the measure with reverse impact. Let's consider two cases, adding a measure with reverse impact with a validity of .10 or zero. If the measure with reverse impact has a validity of .10, the equally weighted composite will

have a validity of .262 (based on the formula for a composite of unweighted measures, Guilford, 1965, page 427). That is almost indistinguishable from the (overly corrected) r = .27 of g alone for police officers (Aamodt, 2004; working the correction formula backward, r = .24 when uncorrected for unreliability of the predictor). We would be giving up under ½ of 1 percent of explained variance in exchange for the ability to hire a diverse group of police officers. This outcome, diversity in hiring with essentially equal validity, is just what we have spent several decades searching for.

If the measure with reverse impact has zero validity (e.g., a promising measure that turned out to lack validity), the equally weighted composite will have a validity of .19. We would be giving up under 3.7% of explained variance in exchange for the ability to hire a diverse group of police officers. That is the worst case scenario. We would surely only include a measure with unknown criterion-related validity based on content validity support for so doing. Therefore, it is unlikely this worst case scenario would occur.

We should note that g has a modest level of validity for police job performance. Aamodt reports a meta-analysis that found r = .27, corrected for unreliability in the predictor. Therefore, the value of .27 is an overestimate of the operational validity, since we cannot correct for unreliability in the predictor when we are using that predictor to make selection decisions. A test with an r of .27 to select police officers and a selection ratio of .10, would make only modest contributions to the EMJP of those hired. Figure 2 shows that the EMJP would be about .45 s.d. above the mean. To interpret this in practical terms, a mini Monte Carlo study indicates that over 31% of those hired would be below the mean in job performance. (I refer here to the mean job performance if 100% of applicants were hired.) So, with a selection ratio of .10, g supports only a modest improvement over random hiring of police officers. (Even with a selection ratio of .001, one in a thousand, over 17% of those selected would be below the mean in job performance.) This has at least two implications: (a) maintaining the level of EMJP is not a high hurdle for an alternative selection tool, and (b) our profession should redouble efforts to find previously unmeasured job-related abilities to help select police officers.

Some may be incredulous of Aamodt's meta-analysis finding of r of only .27 for g, but a recent study supports the low validity of g for police selection. This study looked at predictors used by police managers making real-life hiring decisions (hiring 56 of 245 applicants) in a jurisdiction in The Netherlands apparently not bound to select in order of test score. The police managers considered applicants based on multiple measures including g, language proficiency, and personality (neuroticism, extroversion, openness, agreeableness, and conscientiousness), along with a structured interview and a role play. The police managers gave virtually no weight to the cognitive ability scores (r = .03 between cognitive ability and selection decision). Instead, the selection decisions were driven by the structured interview (r = .36), the role play (r = .25), and a measure of agreeableness (r = .15; De Soete, Lievens, Oostrom & Westerveld, 2013, last line of Table 2).

The third column of this series will provide additional tools, describe several real-life and hypothetical examples of these tools in application, comment on legal issues, and provide closing comments to put the series in perspective.

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.

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.

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. Psychological Science, 25, 1804-1815.

Barrett, G. V., Miguel, R. F., Hurd, J. M., Lueke, S. B. & Tan, J. A. (2003) Practical Issues in the Use of Personality Tests in Police Selection. Public Personnel Management, 32, 497-517.

Berry, C. M., Clark, M. A. & McClure, T. K. (2011) Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. Journal of Applied Psychology, 96, 881-906.

Berry, C. M. & Sackett, P. R. (2009) Individual differences in course choice result in underestimation of the validity of college admissions systems. Psychological Science, 20, 822-830.
Briley, D. A., Domiteaux, M., & Tucker-Drob, E. M. (2014). Achievement-relevant personality: Relations with the Big Five and validation of an efficient instrument. Learning and Individual Differences, 32, 26-39.

Cascio, W. F. & Aguinis, H. (2011) Applied Psychology in Human Resource Management (7th ed.) Saddle River, NJ: Prentice Hall.

Chan, D. & Schmitt, N. (1997) Video-based versus paper-and-pencil method of assessment in situational judgment tests: subgroup differences in test performance and face validity perceptions. Journal of Applied Psychology, 82, 143-159.

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

Guilford, J. P. (1965) Fundamental Statistics in Psychology and Education (4th ed.) New York: McGraw Hill.

Hough, L. M., & Johnson, J. W. (2013). Use and importance of personality variables in work settings. In I. B. Weiner (Ed.-in-Chief) & N. Schmitt & S. Highhouse (Vol. Eds.), Handbook of Psychology: Vol. 12. Industrial and Organizational Psychology (pp. 211-243). New York: Wiley.

Huffcutt, A. I., Roth, P. L., Conway, J. M. & Stone, N. J. (2001) Identification and Meta-Analytic Assessment of Psychological Constructs Measured in Employment Interviews. Journal of Applied Psychology, 86, 897-913.

Lee, J. H., & Beckhusen, J. B. (2012). Veterans' racial and ethnic composition and place of birth: 2011 (Issue brief No. ACSBR/11-22). U.S. Department of Commerce Economics and Statistics Administration: U.S. Census Bureau.

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.

Lievens, F. & Sackett, P. R. (2006) Video-based versus written situational judgment tests: a comparison in terms of predictive validity. Journal of Applied Psychology, 91, 1181-1188.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L. & Barbuti, S. M. (2008) Differential validity and prediction of the SAT. College Board Research Report No. 2008-4. New York: The College Board. Downloaded 1/3/2017 from https://collegereadiness.collegeboard.org/pdf/redesigned-sat-pilot-predictive-validity-study-first-look.pdf

McCarthy, J. M., Iddekinge, C. H., & Campion, M. A. (2010). Are highly structured job interviews resistant to demographic similarity effects?. Personnel Psychology, 63, 325-359.

McFarland, L. A., Ryan, A. M., Sacco, J. M., & Kriska, S. D. (2004). Examination of structured interview ratings across time: the effects of applicant race, rater race, and panel composition. Journal of Management, 30, 435-452.

Oswald, F. L., Mitchell, G., Blanton, H. & Tetlock, P. E. (2013) Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. Journal of Personality and Social Psychology, 105, 171-192.

Ryan, C. L. & Bauman, K. (March 2016) Educational attainment in the United States: 2015 Population Characteristics. Current Population Reports, P20-578. Downloaded 1/7/2017 from http://www.census.gov/content/dam/Census/library/publications/2016/demo/p20-578.pdf

US Government Accountability Office (GAO) (2016) K-12 Education; Better Use of Information Could Help Agencies Identify Disparities and Address Racial Discrimination. GAO-14-345, a report to congressional requesters.

Volume XII Issue 4

ptcmw NEWSLETTER

CONNECT.

SHARE.

GROW.

6

Featured Mini-Series: Adverse Impact

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

24

Fall Event Coverage

Did you miss the fall event? No worries! We have covered the highlights for you in this issue.

All statements expressed in this publication are those of the authors and do not necessarily reflect the official opinions of policies of PTCMW.

The views, opinions, and/or findings expressed in this newsletter are those of the authors and do not necessarily represent the official position or policy of PTCMW.

NEXT ISSUE: April/May 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. (2017, March). Tools to Increase Diversity and Validity in Hiring Police Officers - Part II. *The Personnel Testing Council of Metropolitan Washington. Newsletter, XII*, 6-15.

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