Requirements and the Demographic Profile of the Eligible Population

The Use of Standardized Aptitude Tests in Determining Eligibility

Abstract

The U.S. military uses minimum scores on aptitude tests, such as the Armed Services Vocational Aptitude Battery (ASVAB), the SAT, and the ACT, to determine eligibility for both enlistment and admission into pre-commissioning officer programs. However, it is well documented that average scores on these tests tend to differ by race/ethnicity and gender: Racial/ethnic minorities usually score lower than whites, while women often score slightly higher than men on tests of verbal ability, but lower on tests of quantitative ability. These differences in test scores affect the demographic mix of those who are eligible for service. They also raise questions of test bias and discrimination in admission decisions. However, differences in average test scores alone do not indicate that bias or discrimination is present. In the case of aptitude tests, extensive research shows that well-developed aptitude tests are valid predictors of future performance and are typically not found to be discriminatory according to testing standards and the law. And, although other selection tools and methods can be used that tend to show less average demographic differences, they tend to be poorer predictors of future performance, and they present their own issues and obstacles to implementation. Therefore, although replacing aptitude tests with alternative tools may eliminate some of the demographic differences in meeting eligibility requirements, doing so could also result in a less-efficient selection system, potentially producing lower-quality accessions.

In the issue paper "How Requirements Shape the Demographic Profile of the Eligible Population," we showed that many of the Services' eligibility requirements cause the demographic mix of the eligible population to differ from that of the U.S. population as a whole. One of those eligibility requirements is scores on standardized aptitude tests. For example, all individuals who want to enlist in the military must take and meet a minimum score on the Armed Forces Qualification Test (AFQT).[1] Similarly, scores on the SAT and the ACT are used to determine admission into the Service academies and other colleges and universities that have Reserve Officer Training Corps (ROTC) programs. The Air Force in particular uses an additional aptitude test—the Air Force Officer Qualifying Test (AFOQT)—for selection into both the ROTC and Officer Training School (OTS) programs. However, average scores on these tests tend to differ by race/ethnicity and gender, and these differences in test scores contribute to a key disconnect between the U.S. population as a whole and the population that is eligible for service in the enlisted and officer ranks.

This issue paper defines standardized aptitude tests, provides an overview of how they are used and how different demographic groups tend to perform on them, and examines how issues of test bias and discrimination in their use are addressed in the scientific and legal communities.

Standardized Aptitude Tests

According to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), a *standardized test* is any test in which individuals' responses are scored and evaluated in a consistent manner. Consistent with the use of

MLDC Research Areas

Definition of Diversity
Legal Implications
Outreach & Recruiting
Leadership & Training
Branching & Assignments
Promotion
Retention
Implementation & Accountability
Metrics
National Guard & Reserve

This issue paper aims to aid in the deliberations of the MLDC. It does not contain the recommendations of the MLDC.

Military Leadership Diversity Commission
1851 South Bell Street
Arlington, VA 22202
(703) 602-0818

the term, standardized tests can include measures of ability, aptitude, or achievement. However, the term standardized test is also used by selection and assessment experts to refer to measures of "attitudes, interests, personality, cognitive functioning, and mental health," among others (American Educational Research Association et al., 1999, p. 3).

Therefore, aptitude tests (i.e., cognitive-ability tests) are one type of standardized test that assesses such areas as verbal, mathematical, and spatial ability (Roth, Bevier, Bobko, Switzer, & Tyler, 2001). Standardized aptitude tests (e.g., the SAT, the ACT, the GRE, the ASVAB) are often used in organizational hiring decisions and in educational admission decisions because they have been consistently found to be among the best predictors of future job performance and academic achievement (Sackett, Schmitt, Ellingson, & Kabin, 2001). Several studies have also shown that aptitude tests are highly predictive of training success, including several studies linking scores on the ASVAB to training and job performance in military specialties (e.g., McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Ree & Earles, 1992). Furthermore, the ability of aptitude tests to predict performance has been found to persist over time (Sackett, Borneman, & Connelly, 2008). For example, SAT scores have been found to predict academic achievement throughout college (Bolt, 1986; Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008; Wilson, 1983), and, in a military context, Armor and Roll (1994) found that AFQT scores predicted performance in the military across at least a four-year period.

Race/Ethnicity and Gender Differences on Aptitude Tests

Aptitude tests do not come without criticism, however, as it has been shown that average test scores differ significantly across demographic groups. Although individual members of all demographic groups receive scores that range from low to high, studies of aptitude tests generally find that, on average, blacks and Hispanics tend to score significantly lower than whites. In particular, on average, Hispanics tend to score significantly below whites, and blacks tend to have average scores below those of both Hispanics and whites. Finally, Asians tend to score somewhat higher than whites, especially on measures of mathematical ability, but often score lower on measures of verbal ability (Sackett et al., 2001). In terms of gender, average score differences between men and women are much smaller, with women often scoring slightly higher than men on verbal ability and men scoring slightly higher than women on quantitative ability. However, there is some variability in these gender differences across studies and subtests (Sackett et al., 2008). Similar average group differences have been found in scores on the aptitude tests used by the military as part of its selection process (e.g., Asch, Buck, Klerman, Kleykamp, & Loughran, 2009; Carretta,1997; Roberts & Skinner, 1996).

How Selection-Test Experts Address Issues of Bias in Using Aptitude Tests

Given the existence of demographic differences in average aptitude-test scores, questions of test bias often arise. However, the existence of these group differences does not necessarily mean that bias is present. According to the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999), *test bias* occurs when a selection test does not accurately predict the future performance of a particular group. Specifically, a test is considered to be biased *against* a particular group (or to underestimate the true ability of individuals in that group) if it underpredicts how its members will perform in the future. A test can also be biased *in favor* of a particular group if it overpredicts how its members will perform. In the case of aptitude tests, considerable research has shown that such tests are generally not biased against racial/ethnic minorities. Instead, these tests actually tend to overpredict performance for (or are biased in favor of) blacks and Hispanics (e.g., Bartlett, Bobko, Mosier, & Hannan, 1978; Hartigan & Wigdor, 1989; Young, 2001). In the case of gender, the story is more mixed: Aptitude tests sometimes predict that women will do better than they do in reality and sometimes that they will do worse than they do in reality (e.g., Carretta, 1997; Leonard & Jiang, 1999; Roberts & Skinner, 1996).[2]

How Civil Rights Law Defines Discrimination in Using Aptitude Tests for Selection

Unlawful discrimination is assessed differently than test bias. Public employers are prohibited from discriminating on the basis of race, color, religion, sex, and national origin by Title VII of the Civil Rights Act of 1964 (as amended) and the Fourteenth Amendment of the U.S. Constitution's guarantee of the equal protection under the laws. Several courts have determined that Title VII does not apply in the military context as a matter of law.[3] However, the military generally does apply the substantive rules of Title VII to its servicemembers (Naval Inspector General, 1995, chap. 11).

Title VII forbids several types of employment discrimination. They are generally divided into two types of discrimination: disparate treatment and disparate (adverse) impact.[4] *Disparate treatment* occurs when employers intentionally treat protected groups (e.g., any race or gender) differently. *Disparate impact* occurs when an employment policy or practice has an adverse effect on members of a particular race, color, religion, sex, or national origin, regardless of whether differential treatment was intended. For example, in the case of hiring or selection, this could involve showing that a particular selection test results in women being hired or accepted at significantly lower rates than men. The most commonly used rule of thumb for determining when disparate impact has occurred in the selection or hiring context is when the proportion of applicants that are hired or accepted from a protected group is less than four-fifths (80 percent) of the proportion of applicants that are hired or accepted from the group with the

highest selection rate (usually white or male applicants).[5] This is often referred to as the *four-fifths rule*, and it is the approach used by testing experts.[6]

However, just because a test has a disparate impact does not make it unlawful. Title VII shields employers from liability under disparate impact if they can demonstrate that "the challenged practice is job related for the position in question and consistent with business necessity" (Civil Rights Act of 1991, sec. 105). With regard to selection tests, *disparate impact as a result of average test score differences is not considered unlawful discrimination if the selection test that caused it is a valid predictor of an important job-related outcome*. Thus, if evidence shows that a test is a significant predictor of an important job-related outcome, such as performance, and that there is no equally effective but less-discriminatory test available, then the disparate impact caused by the test does not violate Title VII.[7]

As already noted, a considerable amount of research has found that well-developed aptitude tests are the single strongest predictor of job performance across a wide range of jobs (Schmidt & Hunter, 1998). Therefore, because scores on aptitude tests have been found to be highly related to future performance, and because there are few equally predictive substitutes, using these aptitude tests would not typically violate Title VII, even if it did apply to uniformed servicemembers.

The "Diversity-Validity Dilemma" and Alternatives to Standardized Aptitude Tests

Overall, well-developed aptitude tests are considered to be the best existing selection tool for several reasons. First, evidence from numerous studies shows that they are the best single predictor of both performance and training. Second, they can be used to help select candidates for entry-level jobs, unlike other similarly valid tools which require previous work experience. Third, they are easier and less costly to administer than other selection tools (Schmidt & Hunter, 1998). However, the average differences in test scores among demographic groups means that using aptitude tests in selection leads to what is known as a *diversity-validity dilemma*: Such tests may have high validity in that they are the best single predictor of performance, but they can result in reduced organizational diversity (Pyburn, Ployhart, & Kravitz, 2008).

Searching for a resolution to this dilemma, researchers and practitioners have explored many different ways to improve diversity and reduce disparate impact. Some of these means include the following:[8]

- using both cognitive- and noncognitive-based standardized tests
- using alternative selection methods, such as validated structured interviews, that show less disparate impact but still have high predictive validity (i.e., are good predictors of important outcomes, such as performance)
- supplementing aptitude tests with other measures, such as personality tests, that produce less adverse impact (Ployhart & Holtz, 2008).

None of these options is, however, a perfect solution. Other standardized tests, such as personality tests (particularly conscientiousness)[9], integrity tests,[10] and other standardized methods of selection (such as structured interviews) are good predictors of job performance and tend to show smaller group differences (Ployhart & Holtz, 2008). However, they typically do not come close to the predicative validity of aptitude tests (Schmidt & Hunter, 1998).[11] In addition, there are potential obstacles to implementing these other tests and methods. For example, many of the alternative tests, such as personality tests, are easily coached, and applicants may lie to be selected. Additionally, some alternatives are time-consuming and labor-intensive to develop and administer, so using them can increase costs.

Conclusion

The U.S. military relies on aptitude tests to determine service eligibility as a way to help select the highest-quality applicants. Such tests have often been criticized, however, because average scores tend to differ by demographic group, with racial/ethnic minorities usually scoring lower than their white counterparts. Nonetheless, research shows that well-developed aptitude tests are strong predictors of future performance. Although there are other selection tools and methods that can be used to reduce disparate impact, they present their own problems and obstacles to implementation. Specifically, although replacing aptitude tests with alternative tools may eliminate some of the disconnect between the eligible population and the U.S. population as a whole, doing so could also result in a less-efficient selection system, potentially producing lower-quality accessions.

Notes

[1]The AFQT is used to help determine eligibility requirements for enlistment. It consists of a combination of subtests (of vocabulary, mathematics, arithmetic reasoning, and paragraph comprehension) from the ASVAB, which all applicants interested in enlisting are required to take.

[2]These are results for aptitude tests in general and do not necessarily represent the results for each of the various aptitude tests used in the military.

[3]See, e.g., Roper v. Dep't of Army, 832 F.2d 247, 248 (2d Cir. 1987); Gonzalez v. Dep't of Army, 718 F.2d 926, 928-29 (9th Cir. 1983); Taylor v. Jones, 653 F.2d 1193, 1200 (8th Cir. 1981). Title VII does apply to civilian employees of the military.

[4]*Adverse impact* is another name for disparate impact. The equal-protection clause covers disparate treatment but not disparate impact.

[5]For example, if a test results in 60 percent of white applicants being selected but only 30 percent of Hispanic applicants being selected, the selection rate for Hispanics would only be 50 percent of the selection rate for whites and is considerably below the 80-percent rule of thumb.

[6]The four-fifths rule comes from the *Uniform Guidelines on Employee Selection Procedures*, which were jointly promulgated in 1978 by the Equal Employment Opportunity Commission, the U.S. Departments of Labor and Justice, and the Civil Service Commission (now called the Office of Personnel Management) to provide guidance to help employers comply with Title VII. Courts and federal enforcement agencies have also adopted other statistical means of identifying disparate impact.

[7]The uniform guidelines provide instructions on how to determine whether a selection test validly predicts a job-related outcome and is consistent with business necessity.

[8]For a broader discussion of alternatives see Ployhart and Holtz (2008).

[9]*Conscientiousness*, which is one of the "Big Five" factors of personality, can be defined as achievement-oriented, dependable, careful, and hardworking (Barrick & Mount, 1991).

[10]*Integrity tests* are another type of personality measure and represent a combination of conscientiousness, agreeableness, and emotional stability (Ones & Viswesvaran, 2001).

[11]Work-sample tests can be slightly stronger predictors of performance, but they are more costly and can only be used with applicants that already know the job. Validated structured interviews can have similar predictive validity for performance, but they too are more costly, and they may not be appropriate for entry-level jobs when they are intended to assess job-related knowledge (Schmidt & Hunter, 1998).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Armor, D. J., & Roll, C. R., Jr. (1994). Military manpower quality: Past, present, and future. In B. F. Green & A. S. Mavor (Eds.), *Modeling cost and performance for military enlistment: Report of a workshop* (pp. 13–34). Washington, DC: National Academies Press.

Asch, B. J., Buck, C., Klerman, J. A., Kleykamp, M., & Loughran, D. S. (2005). *What factors affect the military enlistment of Hispanic youth? A look at enlistment qualification*s. Santa Monica, CA: RAND Corporation.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, pp. 1–26.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology*, *31*, pp. 233–241.

Bolt, R. F. (1986). *Generalization of SAT validity across colleges*. New York, NY: The College Board.

Carretta, T. R. (1997). Group differences on US Air Force pilot selection tests. *International Journal of Selection and Assessment*, *5*(2), pp. 115–127.

Civil Rights Act of 1964, Pub.L. 88-352, 78 Stat. 241, July 2, 1964, codified (as amended) at 42 U.S.C. §2000e-2 (2008).

Civil Rights Act of 1991, Pub. L. 102-166, Nov. 21, 1991.

Hartigan, J., & Wigdor, A. K. (1989). *Fairness in employment testing*. Washington, DC: National Academies Press.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average*. New York, NY: The College Board.

Leonard, D. K., & Jiang, J. M. (1999). Gender bias and the college predictions of the SATs: A cry of despair. *Research in Higher Education*, *40*, pp. 375–407.

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*, pp. 335–354.

Naval Inspector General. (1995, July). *Investigations manual*. Washington, DC: Department of the Navy.

Ones, D. S., & Viswesvaran, C. (2001). Integrity tests and other criterion-focused occupational personality scales (COPS) used in personnel selection. *International Journal of Selection and Assessment*, *9*, pp. 31–39.

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*, pp. 153–172.

Pyburn, K. M., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology*, *61*, pp. 143–151.

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science*, *1*, pp. 86–89.

Roberts, H. E., & Skinner, J. (1996). Gender and racial equity of the Air Force officer qualifying test in officer training school selection decisions. *Military Psychology*, *8*(2), pp. 95–113.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*, pp. 297–330.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, *63*, pp. 215–227.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing and higher Education: Prospects in a post-affirmative-action world. *American Psychologist*, *56*, pp. 302–318.

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. *Psychology Bulletin*, *124*, pp. 262–274.

Uniform guidelines on employee selection procedures, 43 Fed. Reg. 166 (1978), 29 C.F.R. part 1607 (2006).

Wilson, K. M. (1983). *A review of research on the prediction of academic performance after the freshman year*. New York, NY: The College Board.

Young, J. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis*. New York, NY: The College Board.