OFFICE OF PEOPLE ANALYTICS
Data-Driven Solutions for Decision Makers

March 2024 | EXECUTIVE NOTE | 2025-192

Validity Evidence Supporting the Uses of ASVAB Scores

Executive Summary

This paper describes the validity evidence that supports the intended uses of the Armed Services Vocational Aptitude Battery (ASVAB) scores. It includes the definition of *validity* as well as a description of the five major sources of validity evidence. This paper then details the valid uses of ASVAB scores and provides examples of validity evidence to support those uses. Lastly, non-validated, inappropriate uses of ASVAB scores are also discussed.

What is validity?

Tests are a near-ubiquitous aspect of human society. The development and use of tests span millennia and much of the globe. Tests have been used for a myriad of uses, including to evaluate student and teacher performance, to determine graduation eligibility from secondary schools, for admittance into postsecondary schools, as a screener for various psychological qualities, and for qualification into military service and assignment to military jobs. When developed and used appropriately, a test can be an extremely useful tool for making better and more informed decisions. Alternatively, tests can do significant harm when they are poorly created or when their scores are used for purposes for which they were not intended. Thus, while it is critical to follow and maintain established testing standards for the development, administration, and use of test scores, the single most important aspect of any test is validity.

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), validity "refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." Crucial to this definition is that it is the *interpretations* of test *scores* that are validated—a test itself is neither valid nor invalid. For example, it would be inappropriate to state that "the ASVAB has been shown to be a valid test." Instead, the focus should be on the intended *uses* of ASVAB *scores,* and the evidence gathered to support those uses.

For decades, validity was compartmentalized into several types, such as construct, content, and criterion validity. While this was useful for descriptive purposes, it obscured the fact that there is no one distinct type of validity. Validity is now seen as a unitary concept: "the degree to which all accumulated evidence supports the intended interpretation of test scores for the proposed use" (AERA, APA, & NCME, 2014). The proposed uses of test scores determine the types of validity evidence that need to be gathered. There are five major sources of validity evidence: consequences, relations to other variables, test content, response processes, and internal structure. Each is briefly described here.

1. **Consequences**—Validity evidence based on *consequences* refers to the intended and unintended consequences of test score uses. Some consequences of test score use flow directly from the proposed use and are thus intended. For example, if a test (its scores) is to be used for selecting applicants into a college, then an intended consequence is that some students who apply will not be accepted. However, an *unintended* consequence might be that certain types of applicants may be disproportionately rejected from this college. While this would not necessarily mean the test is biased for or against certain people, it would nonetheless

constitute a serious consequence. The positive and negative consequences of test score use should always be assessed, and steps should be taken to mitigate the negative consequences, both intended and unintended.

2. **Relations to other variables**—Validity evidence based on *relations to other variables*, as the name implies, focuses on the statistical relations between test scores and the variables the scores are hypothesized to relate to. For example, if a test is to be used for hiring purposes, it is vital to gather evidence that demonstrates the test scores are useful predictors of job success. Similarly, if a new test is developed to measure anxiety, validity evidence for this test could be gathered by showing a relation between the new test and other pre-existing tests that also measure anxiety. This type of validity evidence is typically attained using Pearson correlations or linear regression analyses when working with continuous data (data that can have any number of possible outcomes, like age), or using logistic regression when working with binary data (data that can have only two outcomes, like true or false).

3. **Test content**—Validity evidence based on *test content*, centers on the relationship between the content of the test and the knowledge, skills, or attitudes that the test is intended to measure. This relationship often begins with the creation of a test blueprint, which should clearly define the attribute the test is intended to measure, as well as the various aspects of that attribute, and the number and types of items that will measure each aspect of the attribute. Subject matter experts may be used to judge the adequacy (breadth and depth) of the items' mapping onto the attribute of interest. Gathering validity evidence based on test content can help guard against *construct underrepresentation* (when the test does not fully capture the construct, or attribute, of interest) and *construct irrelevant variance* (when something other than the construct of interest affects scores on the test).

4. **Response processes**—Validity evidence based on *response processes* relates to the cognitive processes that an examinee exhibits while engaging with a test. For example, imagine a series of test items intended to measure reading comprehension. Examinees must first read a passage and then answer items based on that passage. For these items, validity evidence based on response processes is vital, because it is important to determine whether examinees are actually reading the passage before answering the items. Otherwise, the construct of reading comprehension is not being measured. This type of validity evidence can be gathered in a number of ways—if the test is administered via computer, data regarding eye movement or time spent on each item can provide useful evidence on response processes. Another way would be to conduct focus groups with examinees regarding their interactions with the test. There are also other statistical models, such as IRT modeling (Böckenholt, 2012; Thissen-Roe & Thissen, 2013), which can provide nuanced statistical evidence that examinees are displaying appropriate cognitive processes while engaging with test items.

5. **Internal structure**—Validity evidence based on *internal structure* pertains to the intended relationships among items on a test. For example, imagine a test that is intended to measure a single mathematical ability, such as two-digit addition. One would expect that the items on this test would all be closely related to each other, or "similar," given they are all measuring the same ability. This means that, in terms of statistical relationships, inter-item correlations (the correlation between two items) would be similar across all the item pairings on a test.

Conversely, imagine a test that is intended to measure two distinct domains, like vocabulary and reading comprehension. One would expect that the vocabulary items would be more closely related to each other, or "similar," than they are to the reading comprehension items, and vice versa. This means that the vocabulary items would correlate more highly with each other than they do with reading comprehension items, and vice versa. Validity evidence based on internal structure is typically gathered using statistical techniques, such as reliability analysis (score consistency and precision) (Traub & Rowley, 1991) and factor analysis (a commonly used technique for analyzing patterns of correlations) (Bandalos & Finney, 2018). These techniques can provide statistical evidence of the intended patterns of relationships among a set of test items.

Evidence to Support Valid Uses of ASVAB Scores

The ASVAB is a well-researched and well-established test that, in its various forms and iterations,[1] has been used by U.S. Armed Services for decades. Researchers have conducted numerous studies specifically on validity evidence that support ASVAB score uses. ASVAB scores have been validated for three primary uses: 1) for selection of recruits into the U.S. Armed Forces, wherein potential recruits must meet a minimally accepted score on the Armed Forces Qualification Test (AFQT), a composite of the Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Math Knowledge subtests of the ASVAB; 2) for classification of new recruits into specific jobs within the various branches of the U.S. Armed Forces, by using composites of various ASVAB subtests that are known to be effective predictors of success in specific jobs; and 3) for career exploration for high school and early postsecondary students (Office of People Analytics, 2020).

Of course, these uses of ASVAB scores must continue to be validated, especially given the high-stakes nature of these uses. Validity evidence based on *consequences* has been and is routinely considered by DTAC and the various branches of the U.S. Armed Forces. Responsibility for gathering validity evidence related to consequences is shared by the Defense Testing and Assessment Center (DTAC), a division of the Office of People Analytics (OPA) and the research divisions of each of the U.S. Armed Forces. Generally, these research divisions have focused on gathering validity evidence in *relations to other variables*, while DTAC has focused on evidence regarding *test content*, *response processes*, and *internal structure*. Even though certain types of validity evidence are typically gathered by different stakeholders, the task of validating ASVAB score uses has been quite collaborative. These various stakeholders meet monthly via the Manpower Accession Policy Working Group (MAPWG), which consists of policy and technical representatives across the U.S. Armed Services. The MAPWG provides a forum for joint-service review, discussion, and collaboration of ongoing work pertaining to the research, development, implementation, and maintenance of the ASVAB.

The uses of ASVAB scores can certainly be considered high stakes, both for the individual applicants who want careers in the U.S. Armed Forces, and for the health and success of the United States military. This is the primary reason for ensuring the extensive *construct* validity evidence that exists for the ASVAB. For example, reliability analyses are conducted for all examinees and groups to determine adequate precision

[1] The ASVAB has been continually updated since its first use as a paper-and-pencil (P&P) test in the late 1960s. Since then, specific subtests have been dropped from or added to the battery as the needs of the U.S. Armed Forces have changed. New test items with updated content are regularly tried out and then used operationally. A computer-adaptive version of the ASVAB (CAT-ASVAB) was introduced in 1990 (Defense Testing and Assessment Center, 2021).

of scores for all examinees. All ASVAB test items are screened for Differential Item Functioning (DIF), which occurs when examinees of the same ability across groups have differential performance on a test item. Items exhibiting DIF are reviewed by experts and discarded if deemed to be biased. Adverse impact analyses are also regularly conducted to investigate differences in test performance across various groups. These practices are well-established and considered the industry standard in the measurement and testing arena, and a crucial step in ensuring test score use validity.

Many studies have been conducted to demonstrate the predictive ability of the ASVAB in terms of individual subtests, the AFQT composite, and the various composites used by the individual branches of the U.S. Armed Forces. The typical outcomes in these studies are variables like final grades in training courses, attrition rates, job knowledge tests, job performance reviews, and level of training success. Predictive validity, or *relations to other variables*, is generally assessed by calculating correlations between ASVAB subtests/composites and the outcome of interest. Overall, the studies have demonstrated that ASVAB subtests and composites are valid predictors of various training outcomes used by the Army (Maier & Grafton, 1981), Navy (Held & Monzon, 1991), Air Force (Welsh et al., 1990), and Marines (Mayberry, 1990). The predictive validity of the ASVAB is similar regardless of whether it is administered via paper and pencil (P&P) or the computer (CAT-ASVAB) (Defense Manpower Data Center [DMDC], 2006).

When new test forms of the ASVAB are developed, validity evidence related to *test content* is always gathered (Adams et al., 2022; Bejar et al., 2020; DMDC, 2006, 2008, 2009, 2012; Oppler et al., 1997; Ramsberger, 2024; Waters et al., 2009; Waugh et al., 2015). Test items are reviewed by content experts to ensure item quality, that the content maps to the appropriate test blueprint, and that items are free from bias (screened for DIF). There are also ongoing analyses being conducted to help ensure the ASVAB content remains current. For example, DTAC and the Human Resources Research Organization (HumRRO) conducted one extensive validity study on the alignment of ASVAB items with modern conceptualizations (HumRRO, 2015), which resulted in improved item writer guidelines that could help increase the breadth of the subject matter covered (DPAC, 2014, 2015, 2019, 2020a, 2020b, 2020c, 2020d, 2020e, 2020f, 2024; Pommerich, 2016; Reeder, 2023).

Validity evidence related to *response processes*, while less commonly gathered than some of the other types of evidence, is still an important part of the ASVAB validation process. One such example involved a statistical model that was developed to test the hypothesis that response processes for answering Paragraph Comprehension test items can be represented by two main steps: 1) reading and deciphering the text, and 2) comparing the question being asked (the stem) and response options for accuracy to the text. This model was found to be a good fit of a set of ASVAB data (i.e., the sample data "fit" what was expected according to the model), indicating support for this response process theory for the Paragraph Comprehension subtest (Embretson & Wetzel, 1987). Analyses have also been conducted to study response processes for the Assembling Objects subtest (Embretson & Gorin, 2001) and the Arithmetic Reasoning and General Science subtests (Luecht, 2014).

Validity evidence related to *internal structure* is periodically gathered to ensure that the ASVAB subtests are functioning appropriately, i.e., each subtest is unidimensional (i.e., measuring a single attribute). Factor analyses have shown that the ASVAB subtests can, indeed, be treated as unidimensional (DMDC, 2006). The reliability of scores from each subtest has also been extensively studied. Reliability, simply put, is a measure of the precision of test scores. Reliability typically ranges from 0 to 1, with higher values indicating more precise (more trustworthy) scores. Estimated reliabilities are calculated for all ASVAB

subtests and the AFQT composite for both the P&P and CAT-ASVAB (available at www.officialasvab.com), which show adequate to excellent reliability for each subtest.

Non-Validated and Inappropriate Uses of ASVAB Scores

Unfortunately, ASVAB scores are occasionally used in a manner for which they were never designed or validated. For example, there are some who would like to use the ASVAB as a predictor of success in postsecondary education, as a proxy for SAT or ACT scores. Others have proposed using ASVAB scores to evaluate educational outcomes for secondary education. While some of the ASVAB subtests may appear to have similar content to the SAT and ACT or to various high-stakes tests used by states for measuring educational outcomes, ASVAB scores have *not* been validated for these uses. Extensive validity evidence would need to be gathered to support these types of uses, and as this evidence has not yet been gathered or analyzed, making such uses of ASVAB scores would be inappropriate. More information about appropriate uses of ASVAB scores is available in Office of People Analytics' (2020) executive note/information paper, "Appropriate Use of ASVAB Scores."

Another common misuse of ASVAB scores is related to *face validity*: a subjective judgment of whether a measurement instrument seems to measure what it is supposed to measure. A test is considered to have face validity when it has the appearance (i.e., at face value) of measuring what it purports to measure. Face validity is not identified as a component of validity in the *Standards for Educational and Psychological Testing.* Nevertheless, it remains a widely regarded type of "validity" by some testing stakeholders, although it only *appears* to measure something without the evidence that it *actually* measures something.

One such example specific to the ASVAB pertains to the use of calculators. Currently, calculators are prohibited when taking the ASVAB. This policy has garnered criticism given the ubiquity of calculator usage on other assessments and in the classroom. The assertion that calculators should be permitted when taking the ASVAB is an example of face validity. It *appears* that allowing calculators would allow for better measurement of the math knowledge required for service in the U.S. Armed Forces, when in fact, there is no *actual* validity evidence supporting such a claim. More importantly, there are compelling arguments to be made that calculators should not be allowed, given the nature of mathematical requirements for numerous jobs within the U.S. Armed Forces. If there were a policy change allowing calculators on the ASVAB, scores for the math subtests would need to be re-validated, as the validity studies conducted thus far have been under the policy of no calculator use. New validity studies that build on evidence based on *consequences*, *relations to other variables*, and *test content* would need to be conducted. For example, validity studies would need to show that use of calculators on the ASVAB would not diminish the predictive relationship between ASVAB scores and job performance. Validity studies would also need to show that calculator use would not exacerbate differences in performance across groups, instead of helping to reduce differences in performance. Evidence would also need to indicate that calculator usage does not fundamentally change the content being assessed in a test item. More details regarding the issue of calculator use and face validity can be found in Office of People Analytics' (2018) executive note/information paper, "The Use of Calculators on the ASVAB."

Conclusion

Validity is the single most important aspect of any testing program. It ensures that test scores are meaningful, accurate, and used appropriately. Without validity, test scores would not help the U.S. Armed

Forces select the best recruits or accurately predict their training success, making the scores unusable. Over the decades, researchers have put forth a significant amount of effort toward providing extensive validity evidence for the three primary uses of ASVAB scores. In the future, new validiation studies will be conducted to ensure that the ASVAB reflects the latest requirements and changing needs of the U.S. Armed Forces and that the intended uses of ASVAB scores continue to be valid.

References

Adams, K., Oppler, S. H., Prendez, J. Y., & Robertson, S. (2022). *Training relevance survey for the Armed Services Vocational Aptitude Battery (ASVAB)*. (Technical Report 2022-059). Human Resources Research Organization.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), & Joint Committee on Standards for Educational and Psychological Testing (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bandalos, D. L., & Finney, S. J. (2018). Factor analysis: Exploratory and confirmatory. In *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (pp. 98–122). Routledge.

Bejar, I. I., Morley, M., Weeks, J., Holtzman, S., Graf, A. E., & Fife, J. (2020). *Development of the ASVAB AR and MK automated generation workflow: Final report*. Defense Personnel Assessment Center.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods, 17*(4), 665–678.

Defense Manpower Data Center. (2006). *CAT-ASVAB forms 1 & 2* (Technical Bulletin No. 1). Defense Manpower Data Center.

Defense Manpower Data Center. (2008). *CAT-ASVAB forms 5–9* (Technical Bulletin No. 3). Defense Manpower Data Center.

Defense Manpower Data Center. (2009). *CAT-ASVAB forms 3 & 4* (Technical Bulletin No. 2). Defense Manpower Data Center.

Defense Manpower Data Center. (2012). *P&P-ASVAB forms 23–27* (Technical Bulletin No. 4). Defense Manpower Data Center.

Defense Personnel Assessment Center (DPAC). (2014). Guidelines for writing Word Knowledge (WK) test items. Unpublished manuscript for official use only.

Defense Personnel Assessment Center (DPAC). (2015). Guidelines for writing General Science (GS) test items. Unpublished manuscript for official use only.

Defense Personnel Assessment Center (DPAC). (2019). Guidelines for writing Auto Information (AI) test items. Unpublished manuscript for official use only.

Defense Personnel Assessment Center (DPAC). (2020a). Guidelines for writing Arithmetic Reasoning (AR) test items. Unpublished manuscript for official use only.

Defense Personnel Assessment Center (DPAC). (2020b). Guidelines for writing Mathematics Knowledge (MK) test items. Unpublished manuscript for official use only.

Defense Personnel Assessment Center (DPAC). (2020c). Guidelines for writing Paragraph Comprehension (PC) test items. Unpublished manuscript for official use only.

Defense Personnel Assessment Center (DPAC). (2020d). Guidelines for writing Mechanical Comprehension (MC) test items. Unpublished manuscript for official use only.

Defense Personnel Assessment Center (DPAC). (2020e). Guidelines for writing Shop Information (SI) test items. Unpublished manuscript for official use only.

Defense Personnel Assessment Center (DPAC). (2020f). Guidelines for writing Electronics Information (EI) test items. Unpublished manuscript for official use only.

Defense Personnel Assessment Center (DPAC). (2024). *CAT-ASVAB forms 11–15* (Technical Bulletin No. 7). Defense Personnel Assessment Center.

Defense Testing and Assessment Center. (2021). *History of military testing, Armed Services Vocational Aptitude Battery.* https://www.officialasvab.com/researchers/history-of-military-testing/

Embretson S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343–368.

Embretson, S. E, & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11*(2), 175–193.

Held, J. D., & Monzon, R. I. (1991). *Validation study of Armed Services Vocational Aptitude Battery (ASVAB) selector composites: Operations control (OA) occupational group*. Navy Personnel Research and Development Center.

Human Resources Research Organization. (2015). *Refining ASVAB item and test development procedures*. Human Resources Research Organization.

Leucht, R. (2014). *Task 5: Assessment engineering task modeling for the ASVAB AR and GS subtests* [Unpublished manuscript]. Human Resources Research Organization.

Maier, M. H., & Grafton, F. C. (1981). *Aptitude composites for ASVAB 8, 9, and 10.* U.S. Army Research Institute for the Behavioral and Social Sciences.

Mayberry, P. W. (1990). *Validation of ASVAB against infantry job performance*. Center for Naval Analyses.

Office of People Analytics. (2020). *Appropriate use of Armed Services Vocational Aptitude Battery (ASVAB) scores* [Executive note/information paper]. Office of People Analytics.

Office of People Analytics. (2022). *The use of calculators on the ASVAB* [Executive note/information paper]. Office of People Analytics.

Oppler, S. H., Felker, D. B., & Rossmeissl, P. G. (1997). *Item evaluation for ASVAB science and technical test specifications: Conduct training course analysis* (DMDC Technical Report 97-021). Defense Manpower Data Center.

Pommerich, M. (2016, January). *Seeding and form replacement strategy* [PowerPoint briefing]. Defense Manpower Data Center DAC-MPT Meeting, Carmel, CA, United States.

Ramsberger, P. (2024). *High school curriculum study* (2024 No. 165). Human Resources Research Organization.

Reeder, M. (2023, August). *ASVAB item development process: Item analysis*. [PowerPoint briefing]. Defense Manpower Data Center DAC-MPT Meeting, Chicago IL, United States.

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics, 38*(5), 522–547.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability: An NCME instructional module *Educational Measurement: Issues and Practice*, *10*(1), 37–45.

Waters, S. D., Russell, T. L., Shaw, M. N., Allen, M. T., Sellman, W. W., & Geimer, J. L. (2009). *Development of a methodology for linking ASVAB content to military occupation information and training curricula* (FR-09-64). Human Resources Research Organization.

Waugh, C. G., Knapp, D., Ramsberger, P., & Caramagno, J. (2015). *Refining ASVAB item and test development procedures: Final report* (Report 2014 No. 082). Human Resources Research Organization.

Welsh, J. R., Kucinkas, S. K., & Curran, L. T. (1990). *Armed Services Vocational Aptitude Battery (ASVAB): Integrative review of validity studies.* (Technical Report No. 90-22). Air Force Systems Command.