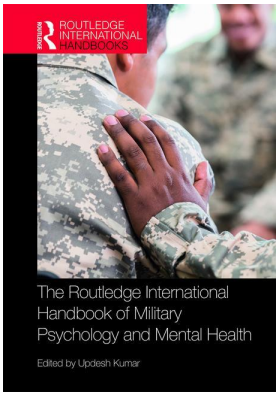

The Routledge International Handbook of Military Psychology and Mental Health

Updesh Kumar

Four stages in the evolution of military Enlistment testing

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9780429281266-3
Michael G. Rumsey
Published online on: 19 Dec 2019

3

FOUR STAGES IN THE EVOLUTION OF MILITARY ENLISTMENT TESTING

Michael G. Rumsey

Military personnel decision-making has become more scientific, expansive, and effective as the environment in which the military operates has become more complex and demanding. This progression is particularly evident in the history of psychological testing for enlisted selection and classification in the United States military. This chapter focuses on how this history has gone through three progressive stages: 1) screening for basic cognitive ability in a single service, 2) matching individual cognitive capabilities to job requirements separately for each service, and 3) developing a single cognitive battery to enable person-job matching for all services, and how it is currently embarking on a fourth stage: developing the capacity to select and classify applicants using a combination of cognitive and noncognitive measures. This last phase, while continuing the evolution, is revolutionary in terms of both character and pace (Rumsey, in press). These advancements have been facilitated by a number of technical advances, including the development and administration of group ability tests, the development of computer adaptive tests, the development of a model of job performance, and the development of fake-resistant personality tests.

While focusing on these four stages, it is important to acknowledge related developments that have unavoidably been left out in this chapter, including those with respect to officer selection, selection and classification outside the United States, and personnel allocation systems to augment person-job classification. While these developments are outside the scope of the current paper, they are nevertheless important. While one may search in vain for an inclusive, integrative treatment of the vast number of military selection programs throughout the world, proceedings from a North Atlantic Treaty Organization meeting (Research and Technology Organization, 2000) did provide a snapshot of various officer selection programs at the time.

Basic ability tests: Army Alpha and Beta

While both the Revolutionary War and the Civil War required the mobilization of large armies, the process of enlisting troops during these periods might be characterized more as desperation than any type of systematic selection and classification. When the United States entered World War I in 1917, the situation was similarly desperate, with the need to rapidly mobilize large numbers of Army troops to avoid a disastrous outcome at the European front, but this time the reaction was more thoughtful. Robert Yerkes, president of the American Psychological

Association, initiated a project to develop a test to ensure that new enlistees had sufficient ability to perform basic tasks. Such luminaries as Henry Goddard, Walter Bingham, Lewis Terman, and Edward Thorndike participated in this effort. The concept for this test was influenced by the Binet Simon scales developed to measure intelligence. Ultimately, two test batteries were developed. These contained a variety of tasks, but the common element of all was a requirement that the applicant demonstrate general cognitive ability. One of the new batteries, the Army Alpha, was composed of verbal tasks for those who could read and write. Representative tests in this battery included general information, arithmetic, and analogies. The other, the Army Beta, was composed of non-verbal items, such as mazes, picture completion, and number checking (Ramsberger, 2012).

Ideally, selection occurs before one is actually placed in a job. However, perhaps partly because an infrastructure for pre-enlistment testing had not yet been created, this is not the way it worked in World War I. Soldiers were tested after enlistment by Commanders in the field. Based on the results of these tests, a limited number were discharged or transferred.

World War I was a time of transition, during which the Army gave tentative approval to limited use of testing for decision-making, but with clear ambivalence over whether such testing actually provided better information than a Commander's judgment. After August 14, 1918, Commanders were given the option of not using the results at all. Nevertheless, the Army did allow 1,750,000 men to be tested on either the Army Alpha or the Army Beta during this unprecedented World War (Ramsberger, 2012).

Even in these early days of standardized testing, the Army recognized the need to determine how well test scores related to other measures. To the extent that could be measured, the results were quite impressive. For example, it was reported that rankings by officers who had known their subordinates 6 to 12 months correlated an average of .536 with ranked scores on the Army Alpha for an enlisted sample of 965 men (Yoakum & Yerkes, 1920). This extraordinarily high correlation would appear to be an anomaly, as more recent correlations between performance on cognitive tests and performance ratings have tended to range between –.02 and.20 (Borman, White, Pulakos, & Oppler, 1991; Vineberg & Taylor, 1972). However, the data reported by Yoakum and Yerkes was not, strictly speaking, based on ratings of performance. Two of the three rating categories specifically mentioned "ability," while only the third referenced "efficiency" (Yoakum & Yerkes, 1920, p. 31).

Army Alpha scores were also found to correlate with another cognitive measure, the Stanford-Binet, in the range of .80 to .90, and the Army Alpha was found to correlate with the Army Beta at about .80. The Army Alpha scores were also related to success in training, rank, and later success in field operations (Yoakum & Yerkes, 1920).

Classification batteries

Although the Army Alpha and Beta test batteries were not completely irrelevant to person-job matching, as those scoring in the upper ranges might be considered best qualified to perform particularly cognitively challenging jobs, this was not their primary function. Later tests, such as the Army General Classification Test (AGCT; Staff, Personnel Research Section, 1945), were only marginally more suited to matching individuals to jobs, a function known as personnel classification. The next major development was to create test batteries designed specifically to support such matching. Since each service had somewhat different needs, separate test classification batteries were developed for each one (Rumsey, 2012).

The problem of matching individuals to jobs is a complicated one when the numbers of jobs and individuals are as great as they are in each military service. First, consider that there might be

a number of jobs that are very similar in terms of the attributes needed to perform successfully. For example, there might be a number of electronics jobs that are distinguished primarily by the type of equipment used in each. Does there need to be a separate combination of test scores to determine an individual's qualifications for each? Is it even possible to identify how each job might be different in terms of individual attributes needed? Is it feasible to develop an instrument that would reliably make the distinctions necessary to make differential assignments across such jobs?

Second, consider how each person-job assignment affects every other potential person-job assignment. If Person A is assigned to Job X, because that is the best fit for Person A, that has implications for other people who also might have been assigned to Job X, as well as for other jobs to which Person A might have been assigned. Person B might not be as well qualified as Person A for Job X, but because Person A is so much better qualified for Job Y than Person B, the Army might be better served by having Person B in Job X and Person A in Job Y. These are just some examples of complications associated with large-scale person-job classification.

Faced with such conundrums, and lacking the high-speed computational power available today, the services opted for a partial solution. Rather than developing a test composite for every single job, the services developed composites, or combinations of tests, for groups of jobs. A 1972 version of the Army Classification Battery (ACB), for example, included composites for such job categories, or Aptitude Areas, as Combat, Electronics Repair, General Maintenance, and Clerical. Of the 16 scores yielded by the ACB, a subset was selected for each of nine Aptitude Areas. The composite for Electronics Repair, for example, consisted of the sum of scores on Arithmetic Reasoning, Electronics Information, Trade Information, Mechanical Comprehension, and a self-reported Electronics inventory measure, each equally weighted (Maier & Fuchs, 1972).

Maier (1993) noted the great care taken by the services in developing their classification tests at each stage in the process, including defining content, constructing test items, collecting data to refine the items, and testing the correspondence between test scores and course grades. Validities for predicting course grades were consistently high, such as .63 for the Air Force's 1960 version of the Airman Qualification Examination (Weeks, Mullins, & Vitola, 1975) and .57 for the Navy's Basic Test Battery (Thomas, 1970).

Joint-service battery

The story of the development of a joint-service selection and classification instrument, the Armed Services Vocational Aptitude Battery (ASVAB), is in some respects a cautionary one. Ultimately, the outcome was a battery that has persevered in various forms for roughly 50 years. Along the way, however, the developmental process encountered severe problems that proved disruptive and embarrassing for each of the military services.

The development of the very first iterations of the ASVAB proceeded without notable controversy. In 1966, the Assistant Secretary of Defense for Manpower and Reserve Affairs directed that a common aptitude battery be developed that all services could use for high school testing. Content areas of the existing Army, Navy, and Air Force classification tests that were determined to be interchangeable (e.g., word knowledge) across the services were identified, and items were drawn from each of the service tests in that content area to form a new subtest. New content was added as needed. Appropriate analyses were conducted to ensure the new battery met scientific and practical standards (Bayroff & Fuchs, 1970). The first version of the ASVAB was administered from 1968 to 1973 to high school students for vocational counseling and recruiting purposes. Two parallel forms of a second version were administered from 1973 to 1975 (Seeley, Fischl, & Hicks, 1978). One form was used for high school testing and the other for Air Force and Marine Corps operational screening (ASVAB Working Group, 1980).

Armed Services Vocational Aptitude Battery misnorming

The problems arose with the third iteration of the ASVAB. In 1974, the Assistant Secretary of Defense for Manpower and Reserve Affairs directed that the ASVAB be used as a common instrument for all the services—not just for high school students, as before, but now for operational enlisted selection and classification as well (ASVAB Working Group, 1980). Along with this decision came three immediate implications for the next iteration of the ASVAB: 1) as a determinant of selection and classification decisions for all the services, it would be more consequential than previous versions; 2) there was a perceived need to implement the new forms quickly, in part because the single form then used for operational screening was viewed as inadequate by the Marine Corps and in part by a desire by the Department of Defense to make the transition to joint testing expeditiously (ASVAB Working Group, 1980); and 3) for the first time, with the draft having ended in 1972 (Rostker, 2006), the test would be developed in an all-volunteer environment.

An ambitious schedule for fielding the new ASVAB collided with numerous delays. In the end, this meant little time was available for determining how the raw scores on the ASVAB could be interpreted in making selection and classification decisions (ASVAB Working Group, 1980). This involved an estimation of how individuals in general would score on this test, not just those individuals tested at this particular time and place. To make this estimation, scores on the ASVAB were linked to another test, a reference test, already calibrated against a representative population. However, later analyses suggested that the linkage process was flawed in multiple respects (Maier & Truss, 1983). The test used for linkage in the Army was the one then being used for operational selection and classification decisions. Thus, examinees were motivated to perform higher on the reference test and recruiters were more likely to coach examinees how to perform higher on this test than would have been the case for the earlier representative population. Also, the reference test used for the Navy and Air Force was found to be incorrectly scored (Maier & Truss, 1983).

The new ASVAB was fielded in all services in 1976. Almost immediately, there were suspicions that the standards being applied with respect to the new forms were incorrect, but pinpointing the problem proved to be difficult (ASVAB Working Group, 1980). Finally, research was conducted that generated standards that the services could agree on, and tests based on these were implemented in 1980 (Maier & Grafton, 1981). By then, it had been determined that the scores used for selection and classification had been seriously inflated, leading to the enlistment of many more with marginal scores on the ASVAB than had been intended. To Army General Maxwell Thurman, who was "arguably the single most important person in the history of the All-Volunteer Force" (Rostker, 2006, p 387), the result was a "calamity" (Rostker, 2006, p. 398). The entire episode, known as the ASVAB misnorming, was described as a "tragedy of errors. A travesty of sound psychometric practices and common sense" (Laurence & Ramsberger, 1991, p. 72, quoted in Rostker, 2006, p. 481).

However, the ultimate impact of the misnorming was that sound psychometric practices and common sense came back to the fore as never before. A Defense Advisory Committee (DAC) on Military Personnel Testing was established in 1981 composed of outside experts to oversee the process of updating military test practices. The DAC and technical and policy military groups met multiple times each year to provide oversight to the enlistment testing program (Maier, 1993).

Computerized Armed Services Vocational Aptitude Battery

As enlisted selection and classification was becoming a joint-service enterprise, a research program was initiated to generate a computerized version of the ASVAB (Wiskoff, 1997).

Computerization offered some benefits in terms of efficiency, but more importantly it could permit tailoring test items to an applicant's ability (Sellman & Arabian, 1997). An item answered correctly could trigger the administration of a more difficult item; an item answered incorrectly could trigger the administration of an easier item. Each item would provide a more precise estimate of the applicant's true ability level. Time would not be wasted on the administration of items far above or below that level (Sands & Waters, 1997).

Interest in a computer adaptive test heightened as a result of the misnorming. The computer adaptive technology was viewed as a partial antidote to the coaching and scoring issues that emerged during this crisis (Wiskoff, 1997). It is difficult to coach a candidate on the correct answers to a test that has no set sequence of items and presents different items to different applicants. Also, computer scoring would reduce the likelihood of human error.

One lesson that apparently had not been learned from the ASVAB misnorming was that pressure to meet an ambitious schedule may have unforeseen consequences. After preliminary research had been completed in 1982, the Navy was identified as the lead service, and completion of the project was scheduled for 1985 (Wiskoff, 1997). However, it soon became apparent that this schedule could not be met. Available computer hardware and technical capabilities were not adequate for the requirements of full-scale enlisted testing in multiple sites across the nation (McBride, 1997). Funding and management issues (Martin & Hoshaw, 1997) also emerged. Finally, in 1993, implementation of CAT-ASVAB was approved (Martin & Hoshaw, 1997). While the development time was much longer than anticipated, the services, under the Navy's lead, persevered to overcome enormous technical and management problems.

Job Performance Measurement Project

As the misnorming revealed that the services were unintentionally not meeting the standards they had set for themselves, the question arose of what importance these standards actually had. What did it mean for someone to receive a low score on the ASVAB? This was a question Congress was very much interested in, and it laid the groundwork for what was to become known as the Job Performance Measurement Project (JPM). Each of the services was directed to conduct research to determine what the link between ASVAB enlistment standards and job performance was (Knapp, 2006).

The Department of Defense determined that the gold standard for performance assessment was hands-on performance. Each service was to include a measure of hands-on performance in its validation strategy. However, beyond this, the services diverged in terms of other criteria they employed (Knapp, 2006). The most ambitious strategy was the one used by the Army. It was driven by the philosophy that no single type of measure truly captured the complexity of performance. Hands-on performance reflected the maximum capability of an individual on certain well-defined tasks. Job knowledge was another measure of capability. However, ratings reflected performance over a longer period of time, and were more indicative of an individual's motivation to perform, even when not being tested. Similarly, administrative measures, such as awards or disciplinary history, provided a view of performance that could not be completely represented by a task-oriented proficiency measure (each of these types of measures was described in Knapp, Campbell, Borman, Pulakos, & Hanson, 2001).

The Army's research effort, known as Project A, was conducted over a 13-year period, from 1982 to 1995, and completely revamped conceptions of job performance and individual differences related to such performance. These individual differences were measured not just by ASVAB, a cognitive measure, but also by measures of personality, vocational interests, spatial ability, and psychomotor ability (Russell & Peterson, 2001). For the first time since the early

1970s, there was active service interest in using a broader perspective than cognitive ability alone to guide selection and classification decisions.

This multidimensional approach to both performance and the predictors of performance proved highly illuminating. It generated results that revealed that one could not simply ask whether a measure predicted performance, but rather one needed to ask the question in terms of what type of performance was being predicted. The combination of hands-on and job knowledge yielded one type of performance, sometimes referred to as can-do or task performance. Ratings that consisted of dimensions that reflected behavior that essentially transcended specific task performance, combined with administrative measures of behavior, yielded another category, which could be labeled will-do performance (Rumsey, Peterson, Oppler, & Campbell, 1996).

To those vested in the credibility of ASVAB, the results were a relief. ASVAB was a strong predictor of can-do performance. To those who believed that a broader combination of predictors was needed for selection, there was also support. ASVAB did not predict will-do performance as well as it predicted can-do performance, but personality and interests showed promise in being able to add to the ASVAB in predicting this type of behavior (McHenry, Hough, Toquam, Hanson, & Ashworth, 1990). In combination with the results from the other services participating in the JPM (Sellman, Russell, & Strickland, 2017), the prevailing conclusion was that ASVAB was doing the job intended for it, predicting technical performance, but that the selection system could likely be improved by the addition of noncognitive individual difference measures.

Armed Services Vocational Aptitude Battery today

Since 1976, the ASVAB has been a dual-use instrument. Its core is a combination of verbal and quantitative tests used for selection. These tests and more specialized technical tests are combined in various ways by the different services for job placement decisions. The ASVAB's utility for selection has been well demonstrated. Recently, its utility for classification has become somewhat controversial. Because it consists of cognitive tests, and such tests have been found to be highly correlated with one another, some have concluded (e.g., Murphy & Davidshofer, 1991) that the different ASVAB tests measure essentially the same thing: general cognitive ability. However, other scientists have conducted simulations showing that, even with high correlations between different ASVAB components, substantial classification utility is still possible (e.g., Zeidner & Johnson, 1994).

The rise of the noncognitives

Personality measures

Project A stimulated interest in measurement of individual differences that could predict will-do performance, particularly personality variables. Despite this interest, progress toward including noncognitive variables in selection moved slowly at first. Early efforts to introduce a joint-service personality measure were unsuccessful due to concerns about faking (Rumsey, 2012). Faking was indeed a serious issue. If an individual responded to a personality item in terms of what he or she thought would impress the examiner, then the test was not measuring what it was intended to measure.

For a number of years, researchers labored somewhat inconspicuously to overcome this obstacle. Two researchers in particular, Leonard White and Mark Young, at the Army Research Institute for Behavioral and Social Sciences, explored building an instrument using a forced choice methodology. The idea was to disguise each item so it would not be clear what response would put the examinee in the most favorable light. The result of their endeavors was a measure

known as the Assessment of Individual Motivation (AIM) (White & Young, 1998). Their work came to sudden prominence when the Army needed a new approach to the screening of those without high school diplomas, whose proclivity towards attrition made them poor risks for the Army. Research indicated that the AIM could be useful in identifying those non–high school diploma graduates who did not present the same level of risk (White, Rumsey, Mullins, Nye, & LaPort, 2014). The AIM was implemented, but early results were discouraging (White, Young, Hunter, & Rumsey, 2008). The researchers persevered, identified and replaced problematic items, and a revised version of the AIM proved to be a successful component of a new screening system directed specifically at nongraduates (White et al., 2014).

Following this success, further developments in personality testing proceeded with new momentum. Another group of researchers working in the Drasgow Consulting Group (DCG) developed a new personality test technology that could be applied to an entire applicant pool. A paper-and-pencil instrument with limited items could, with repeated and frequent use, be particularly vulnerable to compromise. The DCG technology streamlined the item development process to facilitate production of a massive number of items. Further, it employed a computer adaptive process (Stark et al., 2014). The ASVAB had already been converted to a computer adaptive format, but applying this approach to a personality measure was something new. The basic idea is the same, however: approach an individual's true position on a scale by iteratively selecting the items administered to that individual based on his or her responses to previous items until the individual's position is identified with sufficient accuracy.

The combination of an enormous item pool and the lack of a predetermined ordered list of items made the new adaptive tests highly fake resistant (Stark et al., 2014). Also, as noted above, with a computer adaptive approach, items could be limited to those that provided the most information about the individual on the attribute in question. Thus, a computer adaptive test could be administered in much less time than a traditional paper-and-pencil test. For these reasons, this approach was highly desirable for screening large numbers of applicants in a limited period of time.

Both the Navy and the Army have had success implementing computer adaptive personality measures. The Navy's version, the Navy Computer Adaptive Personality Scales (NCAPS), was introduced as a tool used in selection for Special Operations training (Stark et al., 2014). The Army version, the Tailored Adaptive Personality Assessment System (TAPAS), was provisionally implemented as a supplement to the ASVAB in enlisted selection. It replaced the AIM as a tool for identifying non–high school diploma graduates who were acceptable attrition risks (White et al., 2014) and was also used as a screen for high school diploma graduates whose ASVAB scores put them in jeopardy of missing the cut (Knapp, Heffner, & White, 2011).

TAPAS has been found to predict performance in training (White et al., 2014) and on the job (Kirkendall, Bynum, Kennedy, & Hughes, in press), as well as attrition (Hughes, O'Brien, Reeder, & Purl, in press; Kirkendall, Nye et al., in press). Both the Army (Wolters, Heffner, & Sams, 2015) and Air Force (Trent, Barron, Rose, & Carreta, in press) have found it useful for predicting counterproductive work behaviors. The Air Force has conducted research demonstrating its resistance to faking and its stability over time. Navy recruits have also participated in TAPAS research (Stark et al., 2014).

Vocational interest measures

The new noncognitive advances in military testing have also extended to vocational interest measurement. An early version of the ASVAB included interest items carried over from the Army Classification Battery, but the interest measures were subsequently dropped "due to low predictive validity" (Greenston, 2012). Each of the services has been conducting sporadic interest

measurement research since 1949 (Ingerick & Rumsey, 2014) but with little to show for it in terms of operational success up to the turn of the century. Since then, however, such research has gained momentum. The Navy has, having identified key job characteristics as a basis for determining applicant interest in particular Navy jobs (Watson, in press), developed an interest measure that of this writing is scheduled for implementation for all Navy enlisted accessions in October 2018. The Air Force has developed a tool based on the Navy's work called the Air Force Work Interest Navigator (AF-WIN; Johnson, Romay, & Barron, in press). The Army has built upon past work to identify a new instrument covering 20 basic interest categories (Kirkendall, Nye et al., in press).

The controversy over the ASVAB's classification potential suggests the need for further examination of how much can be gained through optimal combinations of cognitive measures for job placement. Since, as shown in Project A (Peterson et al., 1992), cognitive measures and noncognitive measures are not highly correlated, it also suggests that incorporating noncognitive measures into the classification process may be particularly helpful. Preliminary research provides some support for the potential value of TAPAS as a classification tool (Nye et al., in press).

Conclusion

Military enlistment testing has gone through a number of progressive changes. First, with the Army Alpha and Beta, there were service-specific tests of general cognitive ability. Such tests were primarily useful for selection but not well designed for person-job matches. Then each service developed its own classification test battery, facilitating classification decisions.

The development of a joint-service battery was another step in the maturation of military testing. Initial stumbles of disastrous proportions threatened the continuing existence of joint-service testing, but the services pulled together and took a number of steps that made their testing program stronger than ever. From the misnorming disaster came increased oversight, improved testing technology, and a demonstration of the efficacy of the ASVAB for predicting job performance, as well as a better understanding of such performance and the relationship between different types of individual differences and different types of performance.

New technological advances and accumulation of evidence of the validity of noncognitive tests have led to renewed attention to personality and vocational interest measures, not only as enlisted selection tools, but also as tools for Reserve Officers' Training Corps scholarship screening (Legree, Kilcullen, Putka, & Wasko, 2014) and in-service selection (Muhammad, Wolters, & Jane, in press; Nye et al., in press) tools. They may also have utility in matching individuals to jobs. Despite this, with their limited operational history, their acceptance remains tenuous.

The history of change in the military with respect to selection and classification is consistent with Gersick's (1991, p. 12) paradigm of organizational evolution: "relatively long periods of stability (equilibrium), punctuated by compact periods of qualitative, metamorphic change (revolution)." Poor organizational performance or major environmental changes "often precipitate discontinuous periods of change" (Colarelli, 1998, p. 1049).

Major environmental changes have been associated with each of the four evolutionary stages of military enlistment testing. The first scientifically developed group selection tests, the Army Alpha and Beta, came as a direct response to the personnel needs triggered by World War I. The first true service-specific classification batteries, the Navy Basic Test Battery (completed in 1943; Odell, 1947), the Air Force's Airman Classification Battery (completed in 1948; Weeks, Mullins, & Vitola, 1975), and the Army Classification Battery (completed in 1949; Zeidner, Harper, & Karcher, 1956), were all developed during or after World War II, again in response to critical personnel needs.

The move toward an all-volunteer force was a significant incentive for the services to develop a joint-service selection and classification battery. The prospect of discouraging recruits by requiring separate enlistment tests for each service was not an enticing one. The shift from the Cold War to a world of multiple, diverse challenges could be seen as a factor in the rise of noncognitive testing. The new environment required adaptable service members who could handle uncertainty and ambiguity. Accordingly, personnel managers' perspective shifted from a focus on cognitive ability alone to a broader, "whole person" orientation.

The progressive evolution of military testing has never been, and will never be, inevitable. Not all organizations adapt to change. "Once a social technology becomes routine, a tacit consensus emerges that this is the way things are done" (Colarelli, 1998, p. 1048), and thus resistance to change, especially discontinuous change, is natural. Continued exploration of alternatives and a willingness to scientifically evaluate these alternatives and to commit to the alternative determined to be most functional is necessary to ensure the military will continue to adapt.

Military enlistment testing has thus far adapted well. The ASVAB misnorming was a great setback but one it overcame. Questions about the value of ASVAB for classification and about the value of noncognitive testing for selection or classification will limit the acceptability of these tools until these questions can be definitely answered. Yet even these questions are catalysts for positive change. They should stimulate further research to sharpen our classification tools and improve our noncognitive instruments.

A key element of the success of the military testing program has been its scientific rigor. Colarelli (1998, p. 1048) noted the "low use of scientifically valid technologies in organizations." Not only does the military take validation seriously, but it has a history of developing highly valid tests. The selection portion of the ASVAB was found to correlate with hands-on job performance at a level of .40 across all the services in the JPM project (Sellman, Russell, & Strickland, 2017).

The military has benefited from research in civilian testing technology, but it has also led the way in terms of its own research advancements, whether in terms of large-scale cognitive testing, computerized testing, performance measurement, personality measurement, or vocational interest measurement. It has developed a state-of-the-art selection and classification system but will constantly need to refine and advance its procedures to meet the challenges of a complex and evolving threat environment and to maintain acceptability from both its own leadership and the general public.

References

ASVAB Working Group. (1980). *History of the Armed Services Vocational Aptitude Battery, 1974–1980* (AD-E750743). Washington, DC: ASVAB Working Group.

Bayroff, A. G., & Fuchs, E. F. (1970). *The Armed Services Vocational Aptitude Battery* (Tech. Research Rep. No. 1161). Arlington, VA: U. S. Army Behavior and Systems Research Laboratory.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology*, *76*(6), 863–872.

Colarelli, S. M. (1998). Psychological interventions in organizations: An evolutionary perspective. *American Psychologist*, *53*(9), 1044–1056.

Gersick, C. J. G. (1991). Revolutionary change theories: A multilevel exploration of the punctuated equilibrium paradigm. *Academy of Management Review*, *16*(1), 10–36.

Greenston, P. (2012). Classification research. In P. F. Ramsberger, N. R. Wooten, & M. G. Rumsey (Eds.), *A History of the Research into Methods for Selecting and Classifying U. S. Army Personnel* (pp. 275–338). New York: Mellen Press.

Hughes, M., O'Brien, E., Reeder, M., C., & Purl, J. (in press). Attrition and reenlistment in the Army: Using the Tailored Adaptive Personality Assessment System to improve retention. *Military Psychology*.

Ingerick, M., & Rumsey, M.G. (2014). Taking the measure of work interests: Past, present, and future. *Military Psychology*, *26*(3), 165–181.

Johnson, J., Romay, S., & Barron, L. G. (in press). Air Force Work Interest Navigator (AF-WIN) to improve person-job match: Development, validation, and initial implementation. *Military Psychology*.

Kirkendall, C., Bynum, B., Nesbitt, C., & Hughes, M. (in press). Validation of the TAPAS for predicting in-unit soldier outcomes. *Military Psychology*.

Kirkendall, C., Nye, C., Rounds, J., Drasgow, F., Chernyshenko, O. S., & Stark, S. (in press). Adaptive vocational interest diagnostic: Informing and improving the job selection process. *Military Psychology*.

Knapp, D. (2006). The U.S. Joint-Service Job Performance Measurement Project. In W. Bennett, C. E. Lance, & D. J. Woehr (Eds.), *Performance Measurement: Current Perspectives and Future Challenges* (pp. 113–140). Mahwah, NJ: Erlbaum.

Knapp, D. J., Campbell, C. H., Borman, W. C., Pulakos, E. D., & Hanson, M. A. (2001). Performance assessment for a population of jobs. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the Limits in Personnel Selection and Classification* (pp.181–235). Mahwah, NJ: Erlbaum.

Knapp, D. J., Heffner, T. S., & White, L. (2011). *Tier One Performance Screen Initial Operational Test and Evaluation: Early Results* (Tech. Rep. No. 1285). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Legree, P. J., Kilcullen, R. N., Putka, D. J., & Wasko, L. E. (2014). Identifying the leaders of tomorrow: Validating predictors of leader performance. *Military Psychology*, *26*(4), 292–309.

Maier, M. H. (1993). *Military Aptitude Testing: The Past 50 Years* (Technical Report No. 93–007). Defense Manpower Data Center.

Maier, M. H., & Fuchs, E. F. (1972). *An Improved Differential Army Classification System* (Tech. Rep. No. 1177). Arlington, VA: U S. Army Behavior and Systems Research Laboratory.

Maier, M. H., & Grafton, F. C. (1981). *Scaling Armed Services Vocational Aptitude Battery (ASVAB) Form 8AX* (Research Rep. No. 1301). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Maier, M. H., & Truss, A. R. (1983). *Original Scaling of ASVAB Forms 5/6/7: What Went Wrong* (CRC 457). Alexandria, VA: Center for Naval Analyses (AD-A129499).

Martin, C. J., & Hoshaw, C. R. (1997). Policy and program management perspectives. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 11–20). Washington, DC: American Psychological Association.

McBride, J. R. (1997). Technical perspective. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 11–20). Washington, DC: American Psychological Association.

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*(2), 335–354.

Muhammed, R., Wolters, H., & Jayne, B. S. (in press). Using personality testing to enhance in-service selection. *Military Psychology*.

Murphy, K. R., & Davidshofer, C. O. (1991). *Psychological Testing, Principles and Applications*. Englewood Cliffs, NJ: Prentice-Hall.

Nye, C., White, L., Drasgow, F., Prasad, J., Chernyshenko, O., & Stark, S. (in press). Examining personality for the selection and classification of soldiers: Validity and differential validity across jobs. *Military Psychology*.

Nye, C., White, L., Horgen, K., Drasgow, F., Stark, S., & Chernyshenko, O. (in press). Predictors of attitudes and performance in U. S. Army recruiters: Does personality matter? *Military Psychology*.

Odell, C. E. (1947). Selection and classification of enlisted personnel. In D. B. Stuit (Ed.), *Personnel Research and Test Development in the Bureau of Naval Personnel* (pp. 21–30). Princeton: Princeton University Press.

Peterson, N., Russell, T., Hallam, G., Hough, L., Owens-Kurtz, C., Gialluca, K., & Kerwin, K. (1992). Analysis of the experimental battery: LV sample. In J. P. Campbell & L. M. Zook (Eds.), *Building and Retaining the Career Force: New Procedures for Accessing and Assigning Army Enlisted Personnel: Annual Report, 1990 Fiscal Year* (Tech. Rep. No. 952). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Ramsberger, P. F. (2012). Selection and classification in World War I: Personnel psychologists lend a hand. In P. F. Ramsberger, N. R. Wooten, & M. G. Rumsey (Eds.), *A History of the Research into Methods for Selecting and Classifying U. S. Army Personnel* (pp. 70–94). Lewiston NY: Edwin Mellen Press.

Research and Technology Organization. (2000). *Officer Selection* (Meeting Proceedings No. 55). Neuilly-Sur-Seine, Cedex, France: Research and Technology Organization, North Atlantic Treaty Organization.

Rostker, B. (2006). *I Want You: The Evolution of the All-Volunteer Force*. Santa Monica, CA: The Rand Corporation.

Rumsey, M. G. (2012). Military selection and classification in the United States. In J. H. Laurence & M. D. Matthews (Eds). *The Oxford Handbook of Military Psychology* (pp. 129–147). New York: Oxford University Press.

Rumsey, M. G. (in press). Introduction to the special issue on noncognitive testing. *Military Psychology*.

Rumsey, M. G., Peterson, N. G., Oppler, S. H., & Campbell, J. P. (1996). What's happened since Project A: The future career force. *Journal of the Washington Academy of Sciences*, *84*, 94–110.

Russell, T. L., & Peterson, N. G. (2001). The experimental battery: Basic attribute scores for predicting performance in a population of jobs. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the Limits in Personnel Selection and Classification* (pp. 269–306). Mahwah, NJ: Erlbaum.

Sands, W. A., & Waters, B. K. (1997). Introduction to ASVAB and CAT (pp. 3–9). In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 11–20). Washington, DC: American Psychological Association.

Seeley, L. C., Fischl, M. A., & Hicks, J. M. (1978). *Development of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 2 and 3* (Tech. Paper No. 289). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Sellman, W. S., & Arabian, J. M. (1997). Foreword (pp. xv–xvii). In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 11–20). Washington, DC: American Psychological Association.

Sellman, W. S., Russell, T. L., & Strickland, W. J. (2017). Selection and classification in the U. S. military. In J. L. Farr & N. T. Tippins (Eds), *Handbook of Employee Selection* (pp. 697–721). NY: Taylor and Francis.

Staff, Personnel Research Section, Classification and Replacement Branch, the Adjutant General's Office. (1945). The Army General Classification Test. *Psychological Bulletin, 42*, 760–768.

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology*, *26*(3), 153–164.

Thomas, P. J. (1970). *A comparison between the Armed Services Vocational Aptitude Battery and the Navy Basic Test Battery in predicting Navy School performance* (Tech. Bulletin No. STB 70-4). San Diego: Navy Personnel and Training Research Laboratory.

Trent, J. D., Barron, L. G., Rose, M. R., & Carreta, T. R. (in press). Tailored Adaptive Personality Assessment System (TAPAS) as an indicator for counterproductive work behavior: Comparing validity in applicant, honest, and directed faking conditions. *Military Psychology*.

Vineberg, R., & Taylor, E. N. (1972). *Performance in Four Army Jobs by Men at Different Aptitude (AFQT) Levels: 3. The Relationship of AFQT and Job Experience to Job Performance*. Alexandria, VA: Human Resources Research Organization.

Watson, S. E. (in press). Job Opportunities in the Navy (JOIN). *Military Psychology*.

Weeks, J. L., Mullins, C. J., & Vitola, B. M. (1975). *Airman Classification Batteries from 1948 to 1975: A Review and Evaluation*. (Tech. Paper No. 75–78). Lackland Air Force Base, TX: Air Force Human Resources Laboratory.

White, L. A., Rumsey, M. G., Mullins, H. M., Nye, C. D., & LaPort, K. A. (2014). Toward a new attrition screening paradigm: Latest Army advances. *Military Psychology*, *26*(3), 138–152.

White, L. A., & Young, M. C. (1998). *Development and validation of the Assessment of Individual Motivation*. Paper presented at the meeting of the American Psychological Association, San Francisco.

White, L. A., Young, M. C., Hunter, A. E., & Rumsey, M. G. (2008). Lessons learned in transitioning personality measures from research to operational settings. *Industrial and Organizational Psychology*, 1, 291–295.

Wiskoff, M. F. (1997). R&D laboratory management perspective. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 11–20). Washington, DC: American Psychological Association.

Wolters, H., Heffner, T., & Sams, M. (2015, October). *Overview and Introduction of ARI's Non-Cognitive Selection and Assignment Research: Enlisted Personnel*. Paper presented at the meeting of the Manpower Accession Policy Working Group, Seaside, CA.

Yoakum, C. S., & Yerkes, M. (1920). *Army Mental Tests*. New York: Holt.

Zeidner, J., Harper, B. P., & Karcher, E. K. (1956). *Reconstitution of the Aptitude Areas* (PRB Tech. Research Rep. No. 1095). Washington, DC: Adjutant General's Office (Army).

Zeidner, J., & Johnson, C. D. (1994). Is personnel classification a concept whose time has passed? In M. G. Rumsey, C. B. Walker, & J. H. Harris, (Eds.), *Selection and Classification Research: New Directions* (pp. 377–410). Hillsdale, NJ: Erlbaum.