Graduate
Management
Admission
Council®

The Nine Lives of CAT-ASVAB: Innovations and Revelations

Mary Pommerich
Daniel O. Segall
Kathleen E. Moreno

Defense Manpower Data Center

*Presented at the U.S. Government Supported
CAT Programs and Projects Paper Session, June 2, 2009*

Abstract

The Armed Services Vocational Aptitude Battery (ASVAB) is administered annually to more than one million military applicants and high school students. ASVAB scores are used to determine enlistment eligibility, assign applicants to military occupational specialties, and aid students in career exploration. The ASVAB is administered as both a paper-and-pencil (P&P) test and a computerized adaptive test (CAT). CAT-ASVAB holds the distinction of being the first large-scale adaptive test battery to be administered in a high-stakes setting. Approximately two-thirds of military applicants currently take CAT-ASVAB; long-term plans are to replace P&P-ASVAB with CAT-ASVAB at all test sites. Given CAT-ASVAB's pedigree—approximately 20 years in development and 20 years in operational administration—much can be learned from revisiting some of the major highlights of CAT-ASVAB history. This paper traces the progression of CAT-ASVAB through nine major phases of development including: research and development of the CAT-ASVAB prototype, the initial development of psychometric procedures and item pools, initial and full-scale operational implementation, the introduction of new item pools, the introduction of Windows administration, the introduction of Internet administration, and research and development of the next generation CAT-ASVAB. A background and history is provided for each phase, including discussions of major research and operational issues, innovative approaches and practices, and lessons learned.

Acknowledgment

The views expressed in this paper are those of the authors and not necessarily those of the Department of Defense or the United States Government. Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2009 by the Authors

Citation

Pommerich, M., Segall, D.O., & Moreno, K.E. (2009). The nine lives of CAT-ASVAB: Innovations and revelations. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

**Mary Pommerich, DMDC, DoD Center Monterey Bay, 400 Gigling Rd.,
Seaside, CA 93955, U.S.A. Email: mary.pommerich@osd.pentagon.mil**

The Nine Lives of CAT-ASVAB: Innovations and Revelations

The Armed Services Vocational Aptitude Battery (ASVAB) is administered annually to more than one million military applicants and high school students. Military applicants take the ASVAB through the Enlistment Testing Program (ETP), while high school students take the ASVAB through the Career Exploration Program (CEP). ASVAB scores from ETP administrations are used to determine enlistment eligibility and to assign qualified applicants to military occupational specialties. ASVAB scores from CEP administrations are used to aid students in career exploration (scores can also be used to enlist). The ASVAB is administered as both a paper-and-pencil (P&P) test and a computerized adaptive test (CAT) in the ETP. Approximately two-thirds of military applicants currently take CAT-ASVAB; the rest take P&P-ASVAB. Students testing via the CEP take only P&P-ASVAB.

CAT-ASVAB holds the distinction of being the first large-scale adaptive test battery to be administered in a high-stakes setting. Given CAT-ASVAB's pedigree—approximately 20 years in development and 20 years in operational administration—much can be learned from revisiting some of the major phases of CAT-ASVAB history. This paper traces the progression of CAT-ASVAB through nine major phases of development including: research and development of the CAT-ASVAB prototype, the initial development of psychometric procedures and item pools, initial and full-scale operational implementation, the introduction of new item pools, the introduction of Windows administration, the introduction of Internet administration, and research and development of the next generation CAT-ASVAB. Relevant details are provided for select phases, including background and history, major research and operational issues, innovative approaches and practices employed, and lessons learned.

ASVAB Overview

The first ASVAB was introduced in 1968 as part of the Student Testing Program (now called the CEP). In 1976, all branches of the Services began to use the ASVAB as a selection and classification test battery. Since its inception, the individual tests comprising the ASVAB have undergone several changes. Table 1 summarizes the history of ASVAB content across P&P-ASVAB forms and CAT-ASVAB pools since 1968. Current ASVAB tests are designed to measure aptitudes in four domains: Verbal (V), Math (M), Science and Technical (T), and Spatial (S). Table 2 describes the content and identifies the domain measured by each current ASVAB test. The tests are presented in the order in which they are administered.

Table 1. History of ASVAB Content

	Forms/Pools				
	P&P 1-3	P&P 5-7	P&P 8-22	P&P 23-26	CAT 1-9
Test	(1968-75)	(1976-80)	(1980-2002)	(2002 →)	(1990 →)
Word Knowledge	×	×	×	×	×
Arithmetic Reasoning	×	×	×	×	×
Tool Knowledge	×				
Space Perception	×	×			
Mechanical Comprehension	×	×	×	×	×
Shop Information	×	×	×	×	×
Automotive Information	×	×	×	×	×
Electronics Information	×	×	×	×	×
Coding Speed[a]	×		×		
Mathematics Knowledge		×	×	×	×
Numerical Operations[a]		×	×		
Attention to Detail		×			
General Science		×	×	×	×
General Information		×			
Paragraph Comprehension			×	×	×
Assembling Objects				×	×

[a] Originally included in CAT-ASVAB, but subsequently dropped.

Current test lengths and administration times for both CAT-ASVAB and P&P-ASVAB are summarized in Table 3. The individual test lengths reported for CAT-ASVAB include one tryout item that is seeded into the operational administration. Although CAT-ASVAB allows more total testing time than P&P-ASVAB, on average a CAT-ASVAB session is completed in about half the time it takes to complete a P&P-ASVAB session (approximately 1.5 hours versus 3 hours). For both CAT-ASVAB and P&P-ASVAB, scores on the individual tests are reported as standard scores with a mean of 50 and a standard deviation of 10. A Verbal score (VE) is also reported, computed as a weighted composite of PC and WK scores. Standard scores for VE, AR, and MK are used to compute Armed Forces Qualification Test (AFQT) scores; AFQT scores are used to determine enlistment eligibility. Specifically, the AFQT score is computed as 2(VE) + AR + MK. AFQT scores are reported on a percentile metric. Various composite scores are also computed for each Service from ASVAB test scores and used for placement into military occupations.

Table 2. Summary of Current ASVAB Content

Test	Description	Domain			
		V	M	T	S
General Science (GS)	Knowledge of physical and biological sciences			×	
Arithmetic Reasoning (AR)	Ability to solve arithmetic word problems		×		
Word Knowledge (WK)	Ability to select the correct meaning of words presented in context and to identify best synonym for a given word	×			
Paragraph Comprehension (PC)	Ability to obtain information from written passages	×			
Math Knowledge (MK)	Knowledge of high school mathematics principles		×		
Electronics Information (EI)	Knowledge of electricity and electronics			×	
Auto Information (AI) [a]	Knowledge of automobile technology and auto shop practices			×	
Shop Information (SI) [a]	Knowledge of tools and shop terminology and practices			×	
Mechanical Comprehension (MC)	Knowledge of mechanical and physical principles			×	
Assembling Objects (AO) [b]	Ability to figure out how an object will look when its parts are put together				×

Note. Domains measured are Verbal (V), Math (M), Science and Technical (T), and Spatial (S).

[a] AI and SI are administered as separate tests in the computer administration but combined into one single score (labeled AS). AI and SI are combined into one test (AS) in the P&P administration.

[b] AO is not administered in the CEP.

Table 3. ASVAB Test Lengths and Time Limits

	CAT-ASVAB		P&P-ASVAB	
Test	No. Items[a]	Minutes	No. Items	Minutes
GS	16	8	25	11
AR	16	39	30	36
WK	16	8	35	11
PC	11	22	15	13
MK	16	20	25	24
EI	16	8	20	9
AI	11	7	--	--
SI	11	6	--	--
AS	--	--	25	11
MC	16	20	25	19
AO	16	16	25	15
Total	145	154	225	149

[a] Includes one tryout item that is seeded into the operational administration.

Phase 1: Research and Development of the CAT-ASVAB Prototype

From approximately 1970-1985, the concept of CAT was evaluated and an experimental CAT-ASVAB system was developed. In the early to mid 1970s, preliminary research on CAT procedures was conducted via theoretical analyses and simulation studies. In the late 1970s and early 1980s, empirical studies were conducted to evaluate the feasibility of using CAT for personnel selection. Given demonstration of the viability of CAT for this use, an experimental CAT-ASVAB battery was developed along with a delivery system for administering it. Large-scale studies were then conducted in the mid 1980s to validate the experimental CAT-ASVAB system.

This paper cannot possibly do justice in summarizing the momentous work that was undertaken and the groundbreaking achievements that were obtained in the initial phase of CAT-ASVAB history. Work conducted at this stage overcame major hurdles to prove that CAT-ASVAB could be used for large-scale, high-stakes testing and that it demonstrated economic and psychometric advantages over the traditional P&P-ASVAB administration. Because a comprehensive summary of the history of CAT-ASVAB research and development is provided in Sands, Waters, and McBride (1997), additional details of this stage will not be presented here.

Phase 2: Development of a Full-Scale CAT-ASVAB System

Following the success of the validation studies of the experimental CAT-ASVAB system, a full-scale CAT-ASVAB system was developed starting in about 1985. As part of the development of the full-scale CAT-ASVAB system, two unique item pools were developed for each ASVAB test (referred to as CAT-ASVAB Forms 1–2) and psychometric procedures were selected for operational use. The full-scale CAT-ASVAB system employed procedures and practices that are still in use today, including:

1. 3–parameter logistic model used for item selection and scoring.
2. All multiple-choice item format.
3. Passages and items constrained to fit on one screen.
4. Maximum information item selection with Sympson-Hetter exposure control (Sympson & Hetter, 1985; Hetter & Sympson, 1997).
5. Provisional ability estimates computed using Owen's Bayesian procedure (1969, 1975); final ability estimates computed using a Bayesian modal estimator.
6. Forms assembled to minimize the weighted sum-of-squared differences between form information functions.
7. Content balancing not used during administration of most tests.

Content Balancing Practices

Implementation choices made at this stage created some notable issues that are worthy of further discussion. Perhaps the most striking of these issues was the choice not to content balance during administration of most ASVAB tests. That practice still holds today. Due to concerns about multidimensionality, the CAT-ASVAB does currently control for content taxonomy in administrations of the AO and GS tests, balancing the number of administered items from targeted content areas in a test session. For all other ASVAB tests, no constraints are placed on item content for each examinee, relying instead on a natural content balancing created by the proportional representation of content areas within item pools.

Too many constraints on item selection can lead to what Martha Stocking called "barely adaptive tests" or "BATs" ("Computerized adaptive testing," 2009). The use of fewer constraints leads to a truer CAT administration. With the CAT-ASVAB, it is possible to obtain a desired level of precision at fairly short test lengths (i.e., 10 and 15 items), but for most of the tests this precision is obtained at the expense of overtly controlling content representation for individuals. The CAT-ASVAB is somewhat unique among testing programs in its choice to remain more of a CAT than a BAT. Research in support of the content-balancing/no content balancing practices for the individual CAT-ASVAB tests is described in detail in Segall, Moreno, and Hetter (1997) and ASVAB Technical Bulletins #1 (DMDC, 2006) and #2 (DMDC, 2009).

The choice not to content balance during administration of most CAT-ASVAB tests is further supported by empirical evidence. In general, correlations between CAT-ASVAB scores and P&P-ASVAB scores are higher than correlations between scores on two alternate P&P-ASVAB forms (see Moreno, Wetzel, McBride, & Weiss, 1984). This suggests that relying on natural content balancing in CAT-ASVAB rather than imposing explicit content constraints (as occurs with P&P-ASVAB forms) does not significantly degrade the reliability or validity of CAT-ASVAB when compared to P&P-ASVAB. However, further improvements in validity might be possible by using more proactive content-related constraints in pool assembly and item selection. Future CAT-ASVAB research will look carefully into issues of multidimensionality and content balancing during administration.

Presentation Constraints

Another implementation choice of note in the development of the full-scale CAT-ASVAB system was that all content for individual items (i.e., passage, graphics, stem, response options) was constrained to fit on one screen. This constraint was implemented because of usability concerns; namely, that the employment of scrolling or paging mechanisms would be too complicated for users. Most of the tests were unaffected by this constraint, as their P&P-ASVAB items could easily be displayed in one computer screen. In the PC test, however, the P&P-ASVAB format was not conducive to the CAT-ASVAB presentation constraint, as most passages were too long to display in one screen and also contained multiple items. Thus, the PC test was modified for the CAT-ASVAB to contain shortened passages with only one item per passage, so that each passage and item could be presented all on one screen.

One possible side effect of the presentation constraint was the lack of mode effects observed in comparisons of scores across P&P-ASVAB and CAT-ASVAB administrations for the AR, GS, SI, and WK tests (reported in Hetter, Segall, & Bloxom, 1997). The fact that no navigation was required to view all of the information for an item in CAT-ASVAB likely played a direct role in this finding, as comparability research suggests that tests that require navigation might be more subject to mode effects than tests that do not require navigation (Pommerich, 2004). The finding of no mode effects was also particularly notable for the graphics-rich SI test, as the screen resolution associated with the MS-DOS processing system was of much poorer quality than screen resolutions in today's computing environment.

Because the presentation constraint for CAT-ASVAB resulted in an incompatibility of the PC test across administration modes, it was not possible to study mode effects directly for PC. Since scores are equated across P&P-ASVAB and CAT-ASVAB, any indirect effects due to mode of administration are adjusted for in the scoring. The use of different formats across administration modes also raises the possibility that the two versions of the PC test could measure different constructs. Fortunately, validation research has not provided any evidence to suggest that different constructs are being measured for PC across P&P-ASVAB and CAT-ASVAB. Interestingly, recent research on PC test scores has identified a ceiling effect, where scores are curtailed in the upper tail of the ability distribution. It might be logical to attribute that finding to the shorter passage lengths that are necessary to meet the CAT-ASVAB presentation constraint; however, the ceiling effect is evident in both CAT-ASVAB and P&P-ASVAB, suggesting that other factors are contributing to that finding.

Inclusion of Speeded Tests

Another implementation choice of note in the development of the full-scale CAT-ASVAB system was the inclusion of two speeded tests in the battery. The two tests, Coding Speed (CS) and Numerical Operations (NO), had been part of the P&P-ASVAB battery since 1980. As shown in Table 1, CS had also been a part of the battery from 1968-1975. In contrast to the power tests in the CAT-ASVAB battery, CS and NO were purposefully speeded tests that were administered non-adaptively. The two tests were subsequently dropped from both P&P-ASVAB and CAT-ASVAB because characteristics of the input media (i.e., hardware in the CAT-ASVAB and answer sheets in the P&P-ASVAB) were found to have a strong effect on test scores. Namely, the tests were very sensitive to changes in presentation or to the input media (i.e., changing the shape of the response bubbles on answer sheets affected scores), creating the need to equate scores whenever any changes were made. The expense of conducting equating studies could have been avoided by the use of highly standardized equipment, but in the case of CAT-ASVAB, such an approach was viewed as prohibitive to long-term goals of shifting to Internet administration, where standardization of equipment could not be easily controlled.

The problems noted with the (purposefully) speeded tests are also of interest because they have implications for the power tests in CAT-ASVAB. Ideally, pure power tests should be administered without time limits; realistically, however, time limits are an administrative necessity. The challenge, therefore, is to set time limits that are reasonable from an administrative perspective, yet liberal enough to ensure that no unintended speededness occurs during administration. Original time limits for the CAT-ASVAB were set so that over 95% of the examinees taking the test would complete all items without having to rush (Segall, Moreno, Bloxom, & Hetter, 1997).

Use of Modified Keyboards

A final implementation choice of note in the development of the full-scale CAT-ASVAB system was the use of a special keyboard during operational administration. Unlike today's computing environment, where 99.6% of public schools report having Internet access (Provasnik, KewalRamani, Coleman, Gilbertson, Herring, & Xie, 2007), lack of computer experience was a significant concern with the applicant population at the time CAT-ASVAB was developed. Hence, in the interest of ensuring that examinees without computer experience would not be disadvantaged by taking the test on computer, a modified keyboard was used for responding and no mouse was provided. Only usable keys were labeled on the modified keyboard; all remaining keys were covered with blank keycaps. The usable keys were the A-E response keys (relocated to be equally spaced in the center row), the space bar (relabeled as "ENTER"), and the F1 key (relabeled as "HELP"). All examinees were trained on using the keyboard prior to testing.

Phase 3: Initial Operational Implementation of CAT-ASVAB

The first operational implementation of CAT-ASVAB took place in 1990 at six selected test sites. Although CAT-ASVAB had been shown to have psychometric advantages over P&P-ASVAB, there was substantial resistance on the part of the Services to changing a system that was viewed as effective, especially when no other testing programs were using CAT for large-scale, high-stakes testing. Because of this resistance, the operational implementation was treated as a multi-year pilot study or pseudo beta test of the system. Administering CAT-ASVAB in this fashion allowed all affected parties to adjust to the paradigm shift and confirm that CAT-ASVAB was equally effective as P&P-ASVAB (if not more so). Fears about CAT-ASVAB were further eased when the financial benefits associated with decreased testing time were demonstrated. With the passage of sufficient time, all parties became comfortable enough with the concept of CAT-ASVAB to implement it nation-wide at all Military Entrance Processing Sites (MEPS).

Phase 4: Full-Scale Operational Implementation of CAT-ASVAB at All MEPS

In 1996-1997, CAT-ASVAB was implemented operationally at the MEPS (at 65 total locations throughout the United States and Puerto Rico). Consistent with P&P-ASVAB practices, CAT-ASVAB was offered on a continuous basis to applicants testing at the MEPS. In spite of the proven advantages of CAT-ASVAB, it was not possible to eliminate P&P-ASVAB

completely, as P&P-ASVAB continued to be used at remote Military Entrance Testing (MET) sites. Because P&P-ASVAB was offered on a limited, fixed basis at MET sites, it was not viewed as economically feasible to implement CAT-ASVAB at the MET sites. The inability to implement CAT-ASVAB at MET sites created psychometric challenges associated with maintaining two different administration platforms for ASVAB, such as ensuring score interchangeability across administration modes.

In 2000, CAT-ASVAB was installed at a few MET sites with larger testing volumes. However, as many as one-third of all applicants today still take P&P-ASVAB at the remaining MET sites. The low volume of testing at most MET sites makes the full-scale implementation of CAT-ASVAB (and the corresponding elimination of P&P-ASVAB) one of the greatest current challenges to the ASVAB program.

Phase 5: Implementation of CAT-ASVAB Forms 3-4

Two new CAT pools (referred to as CAT-ASVAB Forms 3-4) were operationally implemented in 1999. Form 3 was used with Forms 1-2 for regular CAT-ASVAB administrations, while Form 4 was used for special administrations only (i.e., equating and linking studies). Generally, the development of Forms 3-4 paralleled that of Forms 1-2 with a few exceptions: greater consideration of content taxonomy during form assembly (but not during test administration), and the use of different maximum exposure control values in exposure control simulations (DMDC, 2009). The biggest change that occurred with the implementation of the new pools was the addition of an adaptive version of the Assembling Objects test (AO had previously been administered in CAT-ASVAB as a fixed form).

Revisions to the Battery

The change to AO administration highlighted a challenge associated with maintaining two different administration platforms for ASVAB, namely that a dual platform restricts the types of tests that can be added to the battery. For example, several promising new computerized tests that measured domains not represented on the ASVAB had been identified in a large-scale validation study (Wolfe, 1997). However, most tests could not be added to the ASVAB battery because they could only be administered by computer. AO had been added to the battery because of its desirable psychometric qualities (i.e., it added incremental validity to the ASVAB and did not show practice effects or adverse impact for gender), but other promising tests measuring psychomotor skills and working memory could not be added because they could not be administered in P&P-ASVAB.

Equating Practices

Development of CAT-ASVAB Forms 3-4 highlighted another challenge associated with the co-existence of both P&P-ASVAB and CAT-ASVAB—namely, that a dual administration platform complicates equating. In implementing the new pools, it was necessary to ensure score interchangeability not only across administration mode, but also across the new and existing CAT pools. As part of the development of Forms 3-4, calibration and parameter scaling procedures were designed to place parameters for the new items onto the existing ASVAB scale; scores on the new CAT pools were then equated to P&P scores as an extra precaution.

Table 4 summarizes two different linking designs that were evaluated during a provisional equating. An indirect (chained) linkage was conducted first, linking CAT Forms 3-4 to CAT

Form 1 and then linking CAT Form 1 to P&P Form 8A. Non-operational data were used to link CAT Forms 3-4 to CAT Form 1, while operational data were used to link CAT Form 1 to P&P Form 8A. A direct linkage was also conducted, linking CAT Forms 3-4 directly to P&P Form 8A. Non-operational data were used to conduct the direct linkage. A preferable approach would have been to administer all forms operationally, but it was not feasible to provide operational scores for examinees taking CAT Forms 3-4 during the equating study.

Table 4. Linking Designs Compared in Equating New CAT Pool Scores to P&P Scores

	Form				
Type	P&P 8A		CAT 1		CAT 3 & 4
Indirect			Non-Operational	⇐	Non-Operational
	Operational	⇐	Operational		
Direct	Non-Operational		⇐		Non-Operational

Because the cumulative effects of chained equating typically result in more equating error when multiple links are used, a direct linkage would normally be expected to provide a cleaner equating than the indirect linkage. However, lack of motivation on the part of examinees taking P&P Form 8A non-operationally was a serious concern. Although the CAT forms were also administered non-operationally for some of the links, a novelty effect was expected to improve examinee motivation since computerized test administration was still a relatively new concept at the time the study was conducted. Because of motivation concerns for examinees taking P&P Form 8A non-operationally, the indirect linkage was expected to result in less equating bias than the direct linkage. Based on the equating results, it was recommended that the indirect approach be used to provide provisional scores and conduct the final equating. More details on the equating study are given in DMDC (2009).

Pretesting Practices

Items developed for use in CAT-ASVAB Forms 3-4 were pretested in standard-length tryout booklets using P&P administration. The tryout items were administered in a P&P setting rather than the CAT-ASVAB environment because the pretesting took place prior to full-scale operational implementation of CAT-ASVAB. As Pommerich and Harris (2003) warned, context effects can occur when item characteristics are created from an item administered in one context, and those characteristics are then used to represent the same item when it is administered in a different context. Using P&P pretested items in operational CAT administrations opens the possibility that examinees could be administered inappropriate items and scored at a lower level of precision than desired.

Although Hetter, Segall, and Bloxom (1997) concluded that paper calibrated parameters could be used in a CAT administration without changing the construct being measured or reducing reliability, a more ideal approach would have been to administer the tryout items in the same mode as their intended operational use. Likewise, it would have been preferable to administer the tryout items in an operational setting, since administering tryout items in a non-operational environment can lead to problems with motivation, as discussed above. However, pretesting items in a CAT context creates new psychometric hurdles that must be overcome, such as how to conduct accurate calibrations from sparse data matrices.

Phase 6: Implementation of Windows Administration

The CAT-ASVAB operating system was converted from MS-DOS to Windows in 2003. A platform effects study showed no differences in performance across MS-DOS or Windows administration for any of the ASVAB tests. This finding was notable, particularly for tests with a high graphics load, given that screen resolution was drastically improved in shifting from MS-DOS to the Windows operating system. The lack of platform effects provided indirect evidence supporting some of the implementation choices made during the development of the full-scale CAT-ASVAB system. Specifically, the lack of platform effects might have resulted, in part, from the following implementation choices: (1) constraining all content for individual items (i.e., passage, graphics, stem, response options) to one screen, and (2) setting liberal time limits to ensure that no unintended speededness occurs.

Also of note at this phase was the addition of a mouse as an input device. Although computer usage had become increasingly common in schools (i.e., the U. S. Department of Education reported that about 97% of 9th–12th graders used computers in 2003; Debell & Chapman, 2006), lack of computer experience was still a concern with the applicant population at the time the conversion to Windows administration was planned. Thus, the modified keyboard continued to be used in place of a regular keyboard. Under Windows administration, examinees could respond using either the mouse or the modified keyboard. Prior to administering the ASVAB tests, examinees were explicitly instructed that they could use either the mouse or the keyboard to respond, and were given a tutorial demonstrating how to use the keyboard and mouse to answer test questions.

Phase 7: Introduction of CAT-ASVAB Forms 5-9

Five new CAT pools (CAT-ASVAB Forms 5-9) were introduced in 2008. Forms 5–8 were implemented operationally, and Forms 1–3 were retired. Form 9 has been reserved for Internet administration of a practice or operational CAT-ASVAB. Form 4 will continue to be used for special administrations and will serve as the reference pool for future equating and linking studies. Some notable procedural changes were implemented at this phase, including conducting item pretesting and score equating in the operational CAT environment, and using item enemies in test administration.

Pretesting Practices

In a significant departure from previous pretesting practices, items developed for use in CAT-ASVAB Forms 5-9 were pretested as part of the operational CAT-ASVAB administration of Forms 1-3. The items were pretested in blocks of 100 items per test. In all, 10 blocks of 100 tryout items were administered for each test. During administration of a block, one item from the set of 100 was randomly selected and embedded (or seeded) into the operational CAT-ASVAB administration of the corresponding test (see Table 3). The tryout item was randomly assigned to one of three possible positions in the administration sequence for each test.

The new pretesting approach had several advantages over the previous practice of pretesting a large set of tryout items via a special P&P administration. Pretesting during operational CAT administration ensured that the tryout items would be administered in the same mode for both pretesting and operational use, minimizing the possibility of context effects. The embedding likely made it difficult to detect tryout items, minimizing concerns about lack of motivation (i.e., examinees were expected to give the same level of effort across tryout and operational items). The administration of a single tryout item ensured there would be no local dependence between responses to tryout items. It also had the advantage of being minimally invasive to examinees and therefore less likely to affect their performance on operational items. Likewise, the approach resulted in minimal increases in applicant processing time.

While demonstrating numerous advantages over previous pretesting practices, the new pretesting approach had one key disadvantage. Namely, the process was much slower and more inefficient than a special pretesting study would have been. A lengthy amount of time was required to collect sufficient item response data for conducting calibrations and building new CAT pools.

Calibration Practices

Tryout items were calibrated alongside operational items (i.e., CAT-ASVAB Forms 1-3), and parameter scalings were conducted to place the parameters for the tryout items onto the scale of the operational items. The calibration of the tryout and operational items was a difficult task for several reasons. First, each examinee took only a subset of the operational items, creating a sparse matrix of operational responses. Second, sample sizes varied considerably across the operational items; some items had very small numbers of responses, while other items had very large numbers of responses. Third, because the item selection was tailored to each individual examinee's level of ability, the operational items were administered to examinees within a restricted range of ability. Fourth, each examinee took only one of 100 possible tryout items, creating a sparse matrix of tryout responses. Thus, the resulting calibration design was contrary to the typical calibration design where a fixed number of examinees with varying abilities take a fixed set of items.

Figure 1 summarizes the design that was used for the calibrations. The grey-highlighted areas represent sub-matrices of sparse item responses for examinees that were administered an operational test using the particular CAT-ASVAB pool (Form 1, 2, or 3) indicated in the column and row headings, plus the tryout block. The white areas in Figure 1 indicate the CAT-ASVAB pools that were not presented. For example, if Form 1 was administered, then each of the N_1 total examinees taking Form 1 took 15 of n_1 possible operational items and 1 of 100 possible tryout items. For those N_1 examinees, responses to items contained in CAT-ASVAB Forms 2 and 3 were treated as not presented. In total, there were ($n_1 + n_2 + n_3 + 100$) columns of items and ($N_1 + N_2 + N_3$) rows of examinees represented in the calibration data matrix.

Figure 1. Operational Calibration Design for the Tryout Items

		Items			
		Form 1 (1 × n_1)	Form 2 (1 × n_2)	Form 3 (1 × n_3)	Tryout Block (1 × 100)
Examinees	Form 1 (N_1 × 1)	Sparse	Not Presented	Not Presented	Sparse
	Form 2 (N_2 × 1)	Not Presented	Sparse	Not Presented	
	Form 3 (N_3 × 1)	Not Presented	Not Presented	Sparse	

Because of the complexity of the calibration problem, a large-scale simulation study was conducted to evaluate and compare the performance of different calibration and scaling methods. The goal of the research was to select a calibration/scaling method for operational use that would best represent the tryout data and maintain a consistent scale over time. The calibration methods studied included marginal maximum likelihood (MML) methods, applied using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003); nonparametric and adjusted MML methods, applied using multilinear formula score theory (Levine, 2003) and a suite of model-fitting programs collectively called ForScore; and MCMC methods, applied using the computer program IFACT (Segall, 2002). The calibration methods are discussed in more detail in Pommerich and Segall (2003), Krass and Williams (2003), and Segall (2003), respectively. The simulation study is described in detail in DMDC (2008).

The simulation study was conducted over multiple rounds that simulated successive cycles of operational administrations of CAT-ASVAB plus tryout items, followed by item calibrations/ scalings and assembly of new pools. Parameter drift was evaluated after each round by comparing the true and estimated parameters. The simulation results showed a slight edge for BILOG-MG in terms of the recovery of true parameters and true abilities, and demonstrated that calibrations/scalings based on BILOG-MG could be expected to adequately represent the tryout data and maintain the CAT-ASVAB scale over time, when applied to items that fit a three-parameter logistic model.

Based on the simulation findings, BILOG-MG was used operationally to simultaneously calibrate the tryout items with the operational items. The operational calibrations were not without problems— for example, a number of steps had to be taken to obtain convergence (discussed in DMDC, 2008) and some operational items were poorly calibrated. Similar problems were noted in the simulation study. Because the simulation study showed that parameter estimates for the tryout items were largely unaffected by these types of problems, it was deemed possible to overlook the noise caused by a few poorly calibrated operational items.

Equating Practices

In a significant departure from previous practices, score equating for Forms 5-9 was conducted entirely in the operational CAT environment. Although the calibration and scaling procedures were designed to ensure that the ability estimates $\left(\hat{\theta}s\right)$ based on administrations of CAT-ASVAB Forms 5–9 would be on the same scale as those based on administrations of CAT-ASVAB Forms 1–4, a score equating was conducted as an extra precaution. In the equating, $\hat{\theta}s$ based on the new pools were equated to $\hat{\theta}s$ based on the reference pool (Form 4). Three factors made it possible to conduct the equating entirely in the operational CAT environment. First, the P&P reference form that was used to set the ASVAB scale was replaced with a CAT reference pool (Segall, 2004a). This meant that it was no longer necessary to equate new CAT pools to a P&P reference form. This greatly simplified the equating design and data collection. Second, a new procedure was implemented using an IRT-based theoretical equating that made it possible to specify an initial equating to provide operational scores for examinees taking CAT-ASVAB Forms 5-9, with no additional data collection. Third, by using a phased approach to the data collection for the final equating, it was possible to provide operational scores while minimizing error in the reported (i.e., equated) scores.

Data collection for the equating was conducted in three phases of operational administration. Within each phase, a random groups design was used to assign examinees (military applicants) to CAT-ASVAB Forms 1, 4, 5, 6, 7, 8, or 9. (For purposes of evaluating time limits, Form 9 was administered under regular and lengthened time conditions in each phase.) Sample sizes increased across each phase, with total N = 2,091 in Phase I, total N = 5,853 in Phase II, and total N = 103,438 in Phase III. For applicants taking Forms 5-9 during Phase I of the data collection, the IRT assumption that $\hat{\theta}s$ are invariant across pools and items was used to determine provisional score transformations to convert the $\hat{\theta}s$ to reported operational scores. Because the invariance assumption implies that examinees should get the same $\hat{\theta}$ regardless of the pool administered, $\hat{\theta}s$ based on Forms 5-9 were converted to a standard score (i.e., the reported operational score) using existing θ to standard score transformations for Form 4. For applicants taking Forms 5-9 during Phases II-III of the data collection, $\hat{\theta}s$ were converted to standard scores using provisional score transformations computed from a linear equating conducted in the previous phase. Upon completion of the data collection, data from Phase III was used to conduct a final equating and develop final score transformations to use in subsequent computations of operational scores.

The use of provisional score transformations during the three phases of data collection invokes questions about how different the reported operational scores would have been had the final score transformations been used instead. Likewise, reliance on the IRT invariance assumption to determine provisional score transformations in Phase I invokes questions about the reasonableness of such an approach. The accuracy of the provisional score transformations was evaluated by using the final score transformations to rescore the records for applicants taking CAT-ASVAB during each phase of the data collection and comparing the new scores to the reported scores (i.e., those based on the provisional score transformations). For each examinee, the difference was computed between scores calculated using the provisional and final transformations. The differences were summarized in several ways, including bias (i.e., the mean of the difference) and contribution to total error. Results showed that the provisional transformations used for applicants testing during Phase I displayed moderate bias and moderate contributions to total measurement error. In comparison, the linear equating-based provisional transformations reported during Phases II and III of the equating study displayed small bias and small contributions to total measurement error. The smaller sample sizes used in Phase I could have contributed to the finding of greater bias and greater contribution to total measurement error than in Phases II-III, but the results serve to remind us that relying solely on the IRT invariance assumption to achieve score interchangeability might not be the safest practice. More details about the equating study and the accuracy of the provisional transformations are provided in DMDC (2008).

Administration Practices

During development of CAT-ASVAB Forms 5-9, we became alerted to the possibility that local dependence (LD) could occur between certain types of item pairs (i.e., items sharing a similar, finely specified level of content) if administered to the same examinee. Because the occurrence of LD can have negative consequences for score accuracy, a study was conducted to first evaluate the possibility of LD occurring in the MK test and then assess its effect on score precision. The pretesting practices used with tryout items for CAT-ASVAB Forms 5-9 (i.e., administration of a single tryout item per examinee) prohibited directly evaluating the existence of LD in the tryout items, as LD is diagnosed between item pairs taken by the same examinees. Instead, the possibility of LD in the Forms 5-9 tryout items was inferred based on diagnostic evaluations of tryout items for CAT-ASVAB Forms 1-2 that had been pretested in standard-length booklets using P&P administration. Following the diagnosis of LD, a simulation study was then used to evaluate the effect of two sources of LD on the precision of examinee CAT scores: LD in CAT item parameters and LD in examinees' CAT item responses.

The diagnostic evaluations of the data from the P&P tryout study confirmed the hypothesis that LD could occur between item pairs sharing a similar, finely specified level of content when administered to the same examinee. The simulation results showed that under the conditions studied, LD in CAT item parameters had a very minimal effect on the precision of examinee CAT scores, while LD in examinees' CAT item responses had a fairly substantial effect on score precision, depending on the degree of LD present. Complete details about the evaluation of LD in the item tryout data and the simulation study are presented in Pommerich and Segall (2008).

Although there was no empirical evidence to suggest that LD was a problem on CAT-ASVAB (empirical reliability and validity studies of CAT-ASVAB scores showed sufficiently high levels of reliability and validity when compared to P&P-ASVAB scores), the results from the diagnostic assessments of LD and the simulation study suggested that it would be prudent to guard against the occurrence of LD in item responses in future administrations of CAT-ASVAB. As a result, groupings of item enemies were identified, where the term "item enemies" is used to refer to items that are likely to trigger LD in responses if administered to the same examinee. CAT-ASVAB administration procedures were then modified to allow only one item per enemy grouping to be administered to the same examinee. Such an approach allowed the continued use of a three-parameter logistic model for item selection and scoring during CAT-ASVAB.

Phase 8: Introduction of Internet Administration

In 2008, Internet administration of CAT-ASVAB (referred to as *i*CAT) was operationally implemented at one MET site. In theory, *i*CAT can be used to administer CAT-ASVAB at any location that has a computer, browser, and Internet connection, eliminating the need for a set of networked computers (or specialized equipment, such as a modified keyboard). As a result, the introduction of *i*CAT makes administration of CAT-ASVAB for applicant testing more economically feasible at remote locations where it previously was viewed as too costly. Plans are currently underway to implement *i*CAT at all MET sites that still use P&P-ASVAB, a move that could ultimately enable the complete elimination of P&P-ASVAB testing in the ETP (Enlistment Testing Program).

Plans are also underway to evaluate the feasibility of using *i*CAT in the CEP (Career Exploration Program). Implementing *i*CAT in the CEP is a much more daunting task than implementing *i*CAT in the ETP, as issues such as where testing will take place and who will administer the tests need to be resolved in order to do so. Although *i*CAT could facilitate the administration of CAT-ASVAB in participating schools, the logistics of administering the test would likely be more difficult than observed with the current use of P&P-ASVAB administration only. For example, although 93.6% of public schools in 2005 reported having instructional classrooms with Internet access (Provasnik, et al, 2007), it might not be possible to test the same quantities of students in one sitting as can be done with P&P-ASVAB, or to control against the occurrence of technology glitches in administering CAT-ASVAB.

Non-Standardization of Equipment

With the transition to Internet administration, non-standardization of equipment becomes an issue of concern, as *i*CAT is intended to be administered on available computers regardless of their specifications. The use of non-standardized equipment is not expected to be a show stopper for *i*CAT, however, because there are no longer any speeded tests in the battery, and liberal time limits have been established for the power tests. In theory, if examinees have sufficient time to complete a test, the equipment used to take the test should not have any effect on performance.

Unproctored Internet Administration

The implementation of *i*CAT opens up the possibility of allowing unproctored Internet administration of ASVAB tests. Unproctored Internet administration could improve recruiting by allowing potential applicants to self-screen based on their performance. All applicants obtaining qualifying scores on the unproctored test would be required to take a proctored verification test (Segall, 2001; Segall, 2004b; Segall & McCloy, 2008) at a MEPS or MET site to confirm their performance. The merits of allowing unproctored Internet administration with proctored verification testing are currently under discussion for the tests comprising the AFQT score used for selection into the military (PC, WK, AR, and MK).

With unproctored Internet administration, time constraints could be greatly relaxed or eliminated completely. To prepare for that possibility, a time limit analysis was conducted during development of CAT-ASVAB Forms 5-9, comparing scores across regular and lengthened time administrations of Form 9. Comparisons of performance across Form 9R (regular time) and Form 9L (lengthened time) were conducted using the Phase I and Phase II equating data described above. In the lengthened time administrations, testing times were increased anywhere from one to eight minutes over the regular time limits, depending upon the test. The lengthened time limits were considered sufficient for all examinees to complete all tests in the battery without any unintended speededness affecting their performance.

Prior to the Phase I data collection, the operational time limits (regular time) had been increased for the MK test (from 18 to 20 minutes) and for the AO test (from 12 to 13 minutes). This change was made in response to evidence that there were a relatively large number of incomplete tests for MK and AO in operational administrations of CAT-ASVAB Forms 1–3. After comparing the performance of Phase I and II examinees across Form 9R and 9L, additional adjustments were made to the regular (i.e., operational) time limits for AI (from 6 to 7 minutes), SI (from 5 to 6 minutes), and AO (from 13 to 16 minutes).

The adjusted time limits were implemented in Phase III of the equating study and continue to be used in operational administrations, as reported in Table 3. Completion rates for Phase III examinees taking the tests under the newly adjusted (regular) time limits showed that nearly all examinees finished all tests, implying that future changes allowing additional (or unlimited) time would not impact scores to a substantial degree. Comparisons of performance across regular and lengthened time limits further suggested that future untimed tests would be likely to produce score distributions that are comparable to those produced from the regular time limits.

Phase 9: The Next Generation CAT-ASVAB

Starting in 2005, a review panel with expertise in the areas of personnel selection, job classification, psychometrics, and cognitive psychology met six times over a 15-month period to make recommendations for improvements and enhancements to the ASVAB testing program. In their report, the panel made 21 recommendations that addressed item and test development practices, test administration practices, validation practices, content of the battery, potential new domains of measurement, aptitude measurement in non-native English speakers, and detection and reduction of faking (Drasgow, Embretson, Kyllonen, & Schmitt, 2006). Upon prioritization of the 21 recommendations, the review panel’s top five priorities were identified as: (1) discontinue P&P-ASVAB testing in the ETP, (2) increase the time allocated for seeding tryout items and administering new measures under study, (3) use non-cognitive measures for classification, (4) review content specifications on a regular basis, and (5) reevaluate the contents of the ASVAB battery.

The review panel’s recommendations were also prioritized by the Manpower Accession Policy Working Group (MAPWG), a joint service group that oversees the development and maintenance of the ASVAB testing program. The MAPWG’s top five priorities showed some overlap with the review panel’s top five: (1) discontinue P&P-ASVAB testing in the ETP, (2) consider classification accuracy and not just incremental validity when evaluating the benefits of adding a new test to the battery, (3) reevaluate the contents of the ASVAB, (4) examine external validity of current tests and new measures on a regular basis, and (5) increase the time allocated for seeding tryout items and administering new measures under study. The greater emphasis on validity issues in the MAPWG’s top five priorities reflects the vital role the Services play in validating ASVAB tests.

In response to the review panel recommendations and the MAPWG priorities, a number of major research initiatives are currently underway. Several activities have been planned to evaluate cognitive tests measuring domains not currently represented on the ASVAB. They include a study comparing the performance of several nonverbal reasoning measures, the development and validation of an information and communication technology test, and the evaluation of a working memory test. In addition, several non-cognitive measures are being implemented on the CAT-ASVAB platform (not as part of the battery) to evaluate for use in selection and classification. Work is also in progress to develop standardized procedures for validating new and existing ASVAB tests, and to develop a system to inform ASVAB content specifications using military training curricula.

Discontinuing P&P-ASVAB Testing in the ETP

It is not surprising that both the review panel and the MAPWG rated discontinuing P&P-ASVAB testing in the ETP as their highest priority. Some of the challenges associated with maintaining two different administration platforms for the ASVAB were discussed earlier. Most notably, the need to ensure continuity of the battery across P&P-ASVAB and CAT-ASVAB administrations greatly constrains the types of tests that can be added to the battery. Many promising tests have gone by the wayside because they could not be administered in P&P-ASVAB. Eliminating P&P-ASVAB in the ETP would free the ASVAB testing program to incorporate more innovative types of tests into the battery.

However, current practices in the CEP create a complicating factor. The contents of the ASVAB match across the ETP and CEP, with the exception that AO is administered in the ETP, but not the CEP. (Because not all Services use AO in their classification composites, the lack of complete continuity in the batteries across the two programs is not problematic for applicant processing.) The continuity of content across the CEP and ETP allows students taking the ASVAB through the CEP to use their scores to enlist. Under current policy, CEP scores can be used for enlistment for up to two years after the date of administration. Discontinuing P&P-ASVAB testing in both the ETP and CEP would allow the CEP to remain parallel with the ETP, facilitating continued enlistment from CEP scores. However, if our evaluations suggest that it is feasible to eliminate P&P-ASVAB in the ETP but not the CEP, then it would be prudent to reconsider current CEP practices, lest we find ourselves back at square one on the issue of maintaining continuity of the battery across two administration platforms.

Closing Comments

CAT-ASVAB has had a long and successful history, and continues to thrive in a testing environment that has demonstrated some notable struggles against cheating and unwanted item exposure (Celis, 1994; Lavelle, 2008). The continued success of the program suggests that there are fewer compromise opportunities for CAT-ASVAB than for other high-stakes tests. Several factors likely contribute to the program's seeming immunity to compromise. First, when CAT-ASVAB was first implemented, P&P-ASVAB was already being offered on a continuous basis, giving a good idea of what sorts of compromise pressures to expect with continuous administration of CAT-ASVAB. Thus, there was no adjustment period, as occurs in programs implementing continuous administration for the first time. Second, the context for taking CAT-ASVAB might be less conducive to large scale attempts to capture item pools. Potential bootleggers would need to pose as prospective applicants and work with a recruiter to obtain access to a MEPS for testing. The prospect of undergoing further applicant processing while at the MEPS could also serve as an impediment to large-scale cheating endeavors. Third, sharing networks are likely much smaller for military-bound examinees than for college-bound examinees.

With respect to unproctored Internet administration of ASVAB tests, there is considerable uncertainty regarding likely compromise pressures. As such, the use of proctored verification testing is being considered to mitigate possible effects of compromise associated with unproctored Internet administration (Segall, 2001; Segall, 2004b, Segall & McCloy, 2008). Suffice it to say, great care should be taken as we move forward in the development of the next generation CAT-ASVAB, to ensure the continued success of the program.

References

Celis, W. (1994, Dec. 16). Computer admissions test found to be ripe for abuse. *New York Times*. Retrieved July 31, 2009, from http://www.nytimes.com/1994/12/16/us/computer-admissions-test-found-to-be-ripe-for-abuse.html?scp=1&sq=Ripe%20for%20abuse&st=cse.

Computerized adaptive testing. (2009, June 30). In *Wikipedia, the free encyclopedia*. Retrieved July 14, 2009, from http://en.wikipedia.org/wiki/Computerized_adaptive_testing.

Debell, M. & Chapman, C. (2006). *Computer and internet use by students in 2003* (NCES 2006-065). U.S. Department of Education. Washington, D.C.: National Center for Education Statistics.

DMDC (2006). *CAT-ASVAB Forms 1 & 2* (Technical Bulletin No. 1). Seaside, CA: Defense Manpower Data Center.

DMDC (2008). *CAT-ASVAB Forms 5-9* (Technical Bulletin No. 3). Seaside, CA: Defense Manpower Data Center.

DMDC (2009). *CAT-ASVAB Forms 3 & 4* (Technical Bulletin No. 2). Seaside, CA: Defense Manpower Data Center.

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB).* (FR-06-25). Alexandria, VA: Human Resources Research Organization.

Hetter, R. D., Segall, D. O., & Bloxom, B. M. (1997). Evaluating item calibration medium in computerized adaptive testing. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 161–167). Washington, DC: American Psychological Association.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 141–144). Washington, DC: American Psychological Association.

Krass, I. A., & Williams, B. (2003, April). *Calibrating CAT pools and online pretest items using nonparametric and adjusted Marginal Maximum Likelihood methods.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Lavelle, L. (2008, June 23). Shutting down a GMAT cheat sheet. *Business Week.* Retrieved July 31, 2009, from http://www.businessweek.com/bschools/content/jun2008/bs20080623_153722.htm.

Levine, M. V. (2003). Dimension in latent variable models. *Journal of Mathematical Psychology*, *47*, 450–466.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 8,* 155-163.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (RB-69-92). Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70,* 351–356.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2(6).* Available from http://www.jtla.org.

Pommerich, M., & Harris, D.J. (2003, April*). Context effects in pretesting: Impact on item statistics and examinee scores*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Pommerich, M., & Segall, D. O. (2003, April). *Calibrating CAT pools and online pretest items using Marginal Maximum Likelihood methods.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Pommerich, M., & Segall, D. O. (2008). Local dependence in an operational CAT: Diagnosis and implications. *Journal of Educational Measurement, 45,* 201-223*.*

Provasnik, S., KewalRamani, A., Coleman, M. M., Gilbertson, L., Herring, W., and Xie, Q. (2007). *Status of education in rural america* (NCES 2007-040). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Segall, D. O. (2001, April). *Detecting test compromise in high-stakes computerized adaptive testing: A verification testing approach.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Segall, D. O. (2002). *IFACT computer program Version 2.0: Full information confirmatory item factor analysis using Markov chain Monte Carlo estimation* [Computer program]. Seaside, CA: Defense Manpower Data Center.

Segall, D. O. (2003, April). *Calibrating CAT pools and online pretest items using MCMC methods.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Segall, D.O. (2004a). *Development and evaluation of the 1997 ASVAB score scale* (Technical Report No. 2004-002). Seaside, CA: Defense Manpower Data Center.

Segall, D. O. (2004b, June). *Modelling and detecting collaboration: A multidimensional item response theory approach*. Paper presented at the Annual Meeting of the Psychometric Society, Pacific Grove, CA.

Segall, D. O., & McCloy, R. A. (2008, April). *Verification testing in unproctored internet testing programs*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Segall, D. O., Moreno, K. E., Bloxom, B. M., & Hetter, R. D. (1997). Psychometric procedures for administering CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 131–140). Washington, DC: American Psychological Association.

Segall, D. O., Moreno, K. E., & Hetter, R. D. (1997). Item pool development and evaluation. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 117–130). Washington, DC: American Psychological Association.

Sympson, J. B., & Hetter, R. D. (1985). *Controlling item exposure rates in computerized adaptive tests*. Paper presented at the Annual Conference of the Military Testing Association. San Diego, CA.

Wolfe, J. H. (Ed.). (1997). Enhanced computer-administered test (ECAT) battery [Special Issue]. *Military Psychology, 9(1).*

Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. D. (2003). BILOG-MG [Computer program]. Lincolnwood, IL: Scientific Software International, Inc.