ASVAB Technical Bulletin No. 3

CAT-ASVAB Forms 5–9

Personnel Testing Division
Defense Manpower Data Center

December 2008

Table of Contents

List of Tables

List of Figures

Executive Summary

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 a paper and pencil (P&P) test in the Student Testing Program. It is administered as both a P&P test and a computerized adaptive test (CAT) in the Enlistment Testing Program. Although the use of CAT-ASVAB significantly improves test security over P&P administration, repeated exposure of CAT pools over time can lead to item or test compromise. CAT-ASVAB Forms 5–9 were developed to replace existing operational pools and to support projected new uses of CAT-ASVAB. This report describes the procedures that were used to develop and administer the new pools.

Approximately 1,000 new items were developed and pretested for each of the ASVAB tests. The tryout items were calibrated along with operational items, and parameter scalings were conducted to place the parameters for the tryout items onto the scale of the operational items. The tryout items were evaluated in a variety of ways, including statistical reviews of item characteristics, differential item functioning (DIF) reviews, internal and external content reviews, and external sensitivity reviews. Items that met the criteria for operational use across all of the evaluations were retained for use in assembling new item pools. A total of five new pools (labeled CAT-ASVAB Forms 5–9) were assembled and evaluated for each ASVAB test. In the pool assembly, items with similar information functions were identified and assigned to separate pools in an attempt to minimize the differences among pool information functions.

Scores on the new pools were equated to scores on the existing pools to ensure that scores could be treated interchangeably across the new and existing CAT-ASVAB pools. Linear equating methods were used to derive constants to transform scores from the new pools to the scale of the existing pools. The linear equating procedures ensured that scores had the same mean and variance across the different pools. Data collection for the equating was conducted in three phases of operational administration to military applicants. During each phase of the data collection, it was necessary to use provisional equating transformations to provide operational scores for the applicants. Final equating transformations were then developed and applied to all subsequent examinees. Following the final equating, the accuracy of the provisional transformations was evaluated by using the final equating transformations to rescore all records of applicants taking the CAT-ASVAB during the data collection and comparing the scores to those based on the provisional transformations.

As part of the equating study data collection, one of the new pools was administered under extended time conditions, enabling comparisons of CAT-ASVAB performance across normal and extended time conditions. This was done in anticipation of the possibility of future internet administration of the CAT-ASVAB under greatly relaxed or eliminated time constraints. The results of the analyses suggested that comparable score distributions would be obtained regardless of the time limits used.

1. Introduction

The Armed Services Vocational Aptitude Battery (ASVAB) was first introduced in 1968 as part of the Student Testing Program. Since 1976, the ASVAB has been administered to all military applicants as part of the Enlistment Testing Program. The battery 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.

1.1. Overview of the ASVAB

The ASVAB is administered as a paper and pencil (P&P) test in the Student Testing Program. It is administered as both a P&P test and a computerized adaptive test (CAT) in the Enlistment Testing Program. Approximately two-thirds of military applicants take the CAT version of the ASVAB (CAT-ASVAB). The ASVAB tests are designed to measure aptitudes in four domains: Verbal (V), Math (M), Science and Technical (T), and Spatial (S). Table 1.1 describes the content of the ASVAB tests across the testing programs and administration platforms. The tests are presented in the order in which they are administered.

Table 1.1. CAT-ASVAB Content Summary

		Domain			
Test	Description	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 version.
[b]AO is not administered in the Student Testing Program.

Scores on the ASVAB tests are reported as standard scores with a mean of 50 and a standard deviation of 10. The standard scores are used to compute a variety of classification composites that are used to qualify applicants for specific military occupations. A Verbal score (VE), which is computed as a weighted composite of WK and PC scores, is also reported. 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 is computed as 2(VE) + AR + MK. AFQT scores are reported on a percentile metric. An AFQT percentile score indicates the percentage of examinees in a reference group that scored at or below that particular score. For AFQT percentile and ASVAB standard scores, the reference group is a sample of 18–23 year old youth who took the ASVAB as part of a national norming study conducted in 1997 (Segall, 2004).

1.2. History of CAT-ASVAB

The CAT-ASVAB is administered in all Military Entrance Processing Stations (MEPS) and in a few Mobile Examining Team (MET) sites. The first operational implementation of CAT-ASVAB took place at selected test sites in 1990. CAT-ASVAB was implemented operationally at all MEPS in 1996–1997 and at a few MET sites in 2000.

The full-scale implementation of CAT-ASVAB in the MEPS was preceded by 20 years of extensive research and evaluation. The research and development of the CAT-ASVAB is summarized in detail in Sands, Waters, and McBride (1997), and in ASVAB Technical Bulletin #1 (DMDC, 2006). The decision to operationally implement the CAT-ASVAB was based on the administrative and psychometric advantages of CAT-ASVAB over P&P administration. These advantages included reduced testing times, more flexible scheduling, greater standardization of administration procedures, immediate scoring, increased measurement precision, and increased test security (Sands & Waters, 1997).

Prior to the effort described here, there were four CAT pools, referred to as CAT-ASVAB Forms 1–4. Table 1.2 summarizes the test lengths and pool sizes for CAT-ASVAB Forms 1–4. Forms 1 and 2 were introduced when CAT-ASVAB was first implemented. Forms 3 and 4 were introduced in 1999. Forms 1–3 were used for regular administrations, while Form 4 was used for special administrations only (i.e., equating and linking studies). The procedures used to develop Forms 1 and 2 are discussed in Sands, et al. (1997) and in ASVAB Technical Bulletin #1 (DMDC, 2006). The procedures used to develop Forms 3 and 4 are discussed in ASVAB Technical Bulletin #2 (DMDC, 2009).

Table 1.2. Test Lengths and Pool Sizes for CAT-ASVAB Forms 1–4

Test	Test Length	Pool Size Form 1	Form 2	Form 3	Form 4
GS	15	72	67	135	133
AR	15	94	94	137	136
WK	15	95	99	137	137
PC	10	50	52	68	70
MK	15	84	85	126	132
EI	15	61	61	92	92
AI	10	53	53	77	73
SI	10	51	49	73	72
MC	15	64	64	106	104
AO[a]	15	89	89	89	89

[a]CAT-ASVAB Forms 1–3 share a common item pool for AO.

CAT-ASVAB administration and scoring is based on an Item Response Theory (IRT) model. IRT is a theory that enables test questions and examinee abilities to be placed on the same scale, thereby allowing tests to be tailored to the specific ability level of each examinee and scores to be expressed on the same scale, regardless of the combination of items that are taken. The IRT model underlying the CAT-ASVAB is the three-parameter logistic (3PL) model:

$$P(\theta) = c + \frac{(1-c)}{1+e^{-1.7a(\theta-b)}}. \tag{1.1}$$

The 3PL model represents the probability that an examinee at a given level of ability (θ) will respond correctly to an individual item with given characteristics. Specifically, the item characteristics represented in the 3PL model are discrimination (a), i.e., how well the item discriminates among examinees of differing levels of ability; difficulty (b); and guessing (c), i.e., the likelihood that a very low-ability examinee would respond correctly simply by guessing.

2. Development of CAT-ASVAB Forms 5–9

Although the use of CAT-ASVAB significantly improves test security over P&P administration, repeated exposure of CAT pools over time can lead to item or test compromise. CAT-ASVAB Forms 5–9 were developed to replace CAT-ASVAB Forms 1–3 and to support projected new uses of CAT-ASVAB. The procedures used to develop CAT-ASVAB Forms 5–9 are outlined below.

On August 25, 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 form for future equating and linking studies.

2.1. Item Development and Pretesting

Approximately 1,000 new items were developed for each of the 10 ASVAB tests in accordance with ASVAB content specifications. 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 seeded into the operational CAT-ASVAB administration of the corresponding test. Once each item in the block had been administered to at least 1,200 examinees (the minimal sample size required for item calibrations), the item block was replaced with another block of 100 items.

Table 2.1 summarizes the seeded item design used in the data collection. The tryout items were administered in either the second, third, or fourth position in the test for all tests but AO. For AO, the tryout items were administered in either the second, third, or fourth position if they were Connection items, or in either the ninth, tenth, or eleventh position if they were Puzzle items. The seeded item position was randomly assigned for each test/content area. The use of multiple positions early in the test/content sequence (i.e., before ability estimation stabilized) was intended to make it difficult for examinees to identify when they were receiving a tryout item, ensuring that they would give the same level of effort on the tryout items as the operational items.

Table 2.1. Seeded Item Design for Pretesting New Items

Test	Item Type	Seeded Position	# Seeded Items	# Operational Items	Total Items
GS	All	2, 3, or 4	1	15	16
AR	All	2, 3, or 4	1	15	16
WK	All	2, 3, or 4	1	15	16
PC	All	2, 3, or 4	1	10	11
MK	All	2, 3, or 4	1	15	16
EI	All	2, 3, or 4	1	15	16
AI	All	2, 3, or 4	1	10	11
SI	All	2, 3, or 4	1	10	11
MC	All	2, 3, or 4	1	15	16
AO	Connection Puzzle	2, 3, or 4 9, 10, or 11	1	15	16

The effect of the seeded item position on calibrated item parameters was subsequently examined via a special study (Krass & Nicewander, 2004). For each test, 10 previously seeded items were seeded in each of the 11 or 16 possible positions, according to the length of the test. The 10 items for each test were selected so that their distribution of item difficulty parameters would approximate the distribution of examinee ability. Thus, most items were of medium difficulty, with a few challenging and easy items. Approximately 2,000 responses were collected per seeded item per position. The items were then calibrated under the 3PL model using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003). For each test, the resulting item discrimination, difficulty, and guessing parameters differed minimally across all of the seeded positions, which supported the practice of seeding tryout items in different positions during operational administration.

2.2. Item Calibration and Parameter Scaling

Tryout items were calibrated along with operational items (i.e., CAT-ASVAB Forms 1–4), and parameter scalings were conducted to place the parameters for the tryout items onto the scale of the operational items. Placing item parameters on a common scale ensures that CAT-ASVAB performance can be compared across all examinees, regardless of the pool or combination of items they received. 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; this created a sparse matrix of operational responses. Second, sample sizes varied wildly 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; this is potentially problematic because item calibration is most effective when the item responses reflect a broad range of examinee ability. Fourth, each examinee took only one tryout item; because the items varied across examinees, this created 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.

2.2.1. Comparison of Calibration Methods

Because of the complexity of the calibration problem, a large-scale simulation study was conducted to evaluate and compare the performance of different calibration methods. The goal of the research was to select a calibration method 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), 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). The simulation study was conducted over six rounds (labeled Rounds 0–5). Round 0 established initial CAT pools, while Rounds 1–5 simulated successive cycles of operational CAT + seeded tryout administrations, followed by item calibrations and assembly of new pools.

2.2.1.1. Round 0

Round 0 was designed to develop four CAT pools that mimicked CAT-ASVAB Forms 1–4 for the AR test. The 3PL item parameters for 1,200 items were generated from a trivariate normal distribution. The trivariate normal distribution was derived from the means, variances, and covariances of the log(a), b, and logit(c) parameters associated with CAT-ASVAB Forms 1 and 2 for AR. For each of the 1,200 items, log(a), b, and logit(c) parameters were simultaneously sampled from the trivariate normal distribution so as to be correlated to the degree observed in AR Forms 1 and 2. The log(a) and logit(c) parameters were then transformed back to their original scale. Item responses for the 1,200 items were generated under the 3PL model using the generated parameters (treated as "true" parameters). The item responses for the 1,200 items were then calibrated using BILOG-MG, and the parameter estimates (treated as "operational" parameters) were used to build four equal-sized CAT pools (labeled Forms S01–S04). Exposure control parameters were then computed for the pools. In the pool assembly, items with similar information functions were identified and assigned to separate pools in an attempt to minimize the differences among pool information functions. The method used to compute the exposure control parameters is discussed in Hetter and Sympson (1997) and in ASVAB Technical Bulletin #1 (DMDC, 2006).

2.2.1.2. Rounds 1–5

Table 2.2 shows the administration design used across Rounds 1–5. Form S04 was treated as the reference form and was administered across all rounds. Within each round, nine waves of item response data (operational CAT + seeded tryout administrations) were generated using nine different distributions of ability. In Rounds 1–4, 100 new tryout items were administered within each wave, resulting in 900 new items per round. The 3PL item parameters for the 900 new items were generated in the manner discussed above and were treated as true parameters. In Round 5, the true item parameters for the 900 tryout items were set to be the true parameters from Forms S01–S03, enabling the evaluation of the effect of multiple cycles of pool development on the degree of parameter drift. Over all rounds, responses to all items (operational + tryout) were generated under the 3PL model using the true item parameters, while item selection and scoring during the operational CAT administration were done using estimated parameters that were treated as operational parameters.

After each round, item calibrations and parameter scalings were conducted to place the parameters for new items onto the scale of Form S04. (The procedure used to place the new items onto the reference-form scale is discussed in detail in Section 2.2.1.3.) Separate calibrations were conducted using the three methods of study: BILOG-MG, ForScore, and IFACT. Using each set of parameter estimates, three new CAT pools were then developed for use in the next round, replacing the previous three CAT pools. For example, in the BILOG-MG track,

- BILOG-MG parameter estimates for the tryout items from Round 1 were used to create Forms S05–S07, used in the operational CAT administration in Round 2.
- BILOG-MG parameter estimates for the tryout items from Round 2 were used to create Forms S08–S10, used in the operational CAT administration in Round 3.
- BILOG-MG parameter estimates for the tryout items from Round 3 were used to create Forms S11–S13, used in the operational CAT administration in Round 4.
- BILOG-MG parameter estimates for the tryout items from Round 4 were used to create Forms S14–S16, used in the operational CAT administration in Round 5.

The ForScore and IFACT tracks followed the same form development and administration cycle as the BILOG-MG track, except that ForScore parameter estimates were used in the ForScore track, and IFACT parameter estimates were used in the IFACT track.

Table 2.2. Administration Design for the Calibration Simulation Study

		Form															
Round	Wave	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
1	1 . . . 9																
2	1 . . . 9																
3	1 . . . 9																
4	1 . . . 9																
5	1 . . . 9																

Across all rounds, examinee true abilities were sampled from a normal distribution. Within each round, the mean and standard deviation of the normal distribution varied across waves, simulating shifts in the population distribution of ability. Table 2.3 shows the mean and standard deviation of the ability distributions by wave. Note that the values shown for a wave were used across all rounds. For example, a mean of –0.75 and a standard deviation of 1.00 was used to generate responses for Wave 1 in each of Rounds 1–5.

Table 2.3. Ability Distribution Means and Standard Deviations, by Wave

Wave	Mean	Standard Deviation
1	–0.75	1.00
2	–0.50	1.20
3	–0.40	1.00
4	–0.30	0.80
5	0.00	1.00
6	0.30	0.80
7	0.40	1.00
8	0.50	1.20
9	0.75	1.00

Within each wave within each round, three operational forms were administered to 98% of the examinees, while Form S04 was administered to 2% of all examinees. Form S04 was administered infrequently, as would be done in practice with the reference form to ensure minimum exposure. Table 2.4 shows the tryout item sample sizes by CAT pool for each wave within each round. The three non-reference operational forms were administered to a total of 40,000 examinees each (400 per tryout item per pool). The reference pool was administered to a total of 2,400 examinees (24 per tryout item). Over all pools, each tryout item was administered to a total of 1,224 examinees, resulting in a total of 122,400 examinees for each set of 100 tryout items. The item sample sizes were designed to mimic sample sizes that would be used operationally.

Table 2.4. Sample Sizes for Tryout Items and CAT Pools for each Wave within each Round

Operational CAT Pool	Number of Examinees Taking Each Tryout Item	Number of Examinees Over All Tryout Items
S01	400	40,000
S02	400	40,000
S03	400	40,000
S04	24	2,400
Total	1,224	122,400

2.2.1.3. Parameter Scaling

The means by which the item parameters for the tryout items were placed onto the scale of the reference form (Form S04) differed across the calibration methods studied. When the items were calibrated using ForScore, Form S04 was used to estimate the distribution of ability for the group, thereby fixing the scale of the estimated parameters to that of Form S04.

When the items were calibrated using BILOG-MG, the parameters for the operational items on the three non-reference forms and the tryout items were estimated separately by wave within each round. The estimated *a* and *b* parameters for the tryout items were then placed onto the scale of Form S04 using the following transformations[1]:

$$a^* = \frac{a}{A} \qquad b^* = (A \times b) + B \,. \tag{2.1}$$

The transformation constants *A* and *B* were computed using the means and standard deviations of the group taking Form S04 (μ_1 and σ_1) and the calibration group (μ_2 and σ_2):

$$A = \frac{\sigma_1}{\sigma_2} \qquad B = \mu_1 - (A \times \mu_2) \,. \tag{2.2}$$

In all of the transformations, the mean and standard deviation for the calibration group were fixed at 0.0 and 1.0, respectively, which is the default scale for BILOG-MG calibrations. The mean and variance of the group taking Form S04 were estimated using a maximum likelihood procedure that maximized the likelihood of the observed responses given the population distribution:

$$\begin{aligned} L(\mu_1, \sigma_1^2 \mid U) &= \prod_{j=1}^{N} P(u_j \mid \mu_1, \sigma_1^2) \\ &= \prod_{j=1}^{N} \int P(u_j \mid \theta) f(\theta \mid \mu_1, \sigma_1^2) d\theta, \end{aligned} \tag{2.3}$$

where $u_j = \{u_{j1}, u_{j2}, \ldots, u_{jn}\}$ for $j = 1, \ldots, N$ examinees, $f(\theta|\mu_1, \sigma_1^2)$ denotes a normal density, and

$$P(u_j \mid \theta) = \prod_{i=1}^{n} \left[P(\theta \mid a_i, b_i, c_i)^{u_{ji}} \, Q(\theta \mid a_i, b_i, c_i)^{1-u_{ji}} \right] \tag{2.4}$$

for $i = 1, \ldots, n$ items with 3PL parameters a_i, b_i, and c_i.

When the items were calibrated using IFACT, the parameters for the operational items on the three non-reference forms and the tryout items were estimated separately by wave within each round. The *a* and *b* parameters for the tryout items were transformed to be on the scale of the original operational parameters for Forms S01–S04 using the

---

[1] BILOG-MG has the capability of allowing selected item parameters to be fixed at pre-defined starting values, and the rescaling of the latent distribution that is typically done may be suppressed. In theory, these features may be used to place parameter estimates for non-fixed items onto the scale of the fixed items. However, Pommerich and Segall (2003) demonstrated that estimated parameters for the non-fixed items can be biased if the underlying ability distribution for the calibration sample is shifted from N(0,1). For this reason, the parameters were estimated for all items (operational + tryout) and then transformed to be on the scale of Form S04.

transformations given in Equation 2.1. The transformation constants *A* and *B* were computed as given in Equation 2.2, where μ_1 and σ_1 were the mean and standard deviation of $\hat{\theta}$ estimates in the calibration sample computed using the original operational parameters, and μ_2 and σ_2 were the mean and standard deviation of the $\hat{\theta}$ estimates in the calibration sample computed using the re-estimated parameters for the operational items. All $\hat{\theta}$ estimates were computed via Bayes mode using a fixed N(0,1) prior.

An iterative process was used to obtain final transformed IFACT parameters. Initial estimates of μ_2 and σ_2 were computed as described above. Transformation constants were then computed and used to transform the re-estimated operational and tryout parameters to the scale of the original operational parameters. Updated estimates of μ_2 and σ_2 were then computed using the transformed operational parameters, new transformation constants were computed, and the operational and tryout parameters were transformed again. The iterative process of updating estimates of μ_2 and σ_2 and transforming the operational and tryout parameters was repeated until the transformed operational parameters provided the same mean and standard deviation for the $\hat{\theta}$ estimates as the original operational parameters, implying that the transformed and original operational parameters were on the same scale. The transformed and original operational parameters typically yielded the same mean and standard deviation for the $\hat{\theta}$ estimates after 3–4 iterations.

2.2.1.4. Results

Parameter drift was evaluated after each round by comparing the true and estimated parameters. Table 2.5 summarizes the true parameters and the estimated *a*, *b*, and *c* parameters across the three calibration methods for the 900 tryout items after the completion of Round 1.

Table 2.5. True and Estimated *a, b,* and *c* Parameters for Tryout Items After Round 1

Parameter	Method	Mean	Standard Deviation	Correlation	Min	Max
a	True	1.276	0.438	–	0.427	3.461
	BILOG-MG	1.269	0.423	0.747	0.068	3.559
	IFACT	1.211	0.328	0.771	0.265	2.423
	ForScore	1.360	0.328	0.721	0.398	2.500
b	True	–0.397	1.197	–	–4.837	3.661
	BILOG-MG	–0.411	1.614	0.821	–19.965	4.404
	IFACT	–0.250	1.088	0.927	–4.431	5.758
	ForScore	–0.424	1.191	0.969	–4.508	3.000
c	True	0.197	0.094	–	0.017	0.621
	BILOG-MG	0.220	0.079	0.617	0.037	0.500
	IFACT	0.245	0.164	0.359	0.030	0.979
	ForScore	0.181	0.109	0.490	0.050	0.400

All three calibration methods showed similar recovery of the average and standard deviation of the true *a* parameters. Results for the *b* parameters suggest that the IFACT estimates were slightly biased and that the BILOG-MG estimates were less highly correlated with the true parameters than the IFACT or ForScore estimates. The lower correlation for BILOG-MG can be explained by the fact that there were no apparent distributional constraints on the parameter estimates. Hence, several very easy items where more than 99.5% of the sample responded correctly had extreme parameter values (i.e., $a < .07$ and $b < -19.0$), which lowered the value of the correlation between the true and estimated parameters.[2] Because IFACT and ForScore place constraints on their parameter distributions, the parameter estimates for these items were not as extreme as those from BILOG-MG, and the correlations between true and estimated parameters were higher. Results for the *c* parameter suggest that BILOG-MG showed the best recovery of the true *c* parameters. In general, *c* parameters are less well estimated than *a* and *b* parameters under the 3PL model.

Table 2.6 summarizes the true parameters and the estimated *a*, *b*, and *c* parameters across the three calibration methods for the 900 tryout items after the completion of Round 5. BILOG-MG and IFACT showed similar recovery of the average and standard deviation of the *a* parameters. ForScore showed a lower correlation between estimated and true *a* parameters than did BILOG-MG and IFACT. Results for the *b* parameter suggest that the IFACT and ForScore estimates were slightly biased. BILOG-MG showed a lower correlation between the true and estimated *b* parameters than the other two calibration methods; however, this is again attributable to the occurrence of very easy items that were answered correctly by almost all examinees. Results for the *c* parameter suggest that BILOG-MG showed the best recovery of the true *c* parameters.

Table 2.6. True and Estimated *a, b,* and *c* Parameters for Tryout Items After Round 5

Parameter	Method	Mean	Standard Deviation	Correlation	Min	Max
a	True	1.274	0.400	–	0.451	3.147
	BILOG-MG	1.292	0.401	0.805	0.103	3.114
	IFACT	1.233	0.313	0.784	0.424	2.414
	ForScore	1.345	0.647	0.383	0.058	2.500
b	True	–0.420	1.155	–	–4.013	3.120
	BILOG-MG	–0.384	1.296	0.904	–12.967	3.806
	IFACT	–0.264	1.052	0.933	–3.522	4.285
	ForScore	–0.154	1.138	0.922	–5.000	3.000
c	True	0.196	0.090	–	0.037	0.578
	BILOG-MG	0.216	0.070	0.634	0.056	0.500
	IFACT	0.239	0.157	0.379	0.020	0.977
	ForScore	0.154	0.111	0.285	0.050	0.400

[2] In practice, very easy items such as these would not be considered for operational use.

Table 2.7 summarizes the root mean squared deviation (RMSD) between the true and estimated abilities for the examinees across the three calibration methods and rounds, where ability was computed using a Bayes modal estimator. BILOG-MG showed the smallest RMSD across all rounds, while ForScore showed the largest. Table 2.8 summarizes the average RMSD between the true and estimated item response functions (IRFs) across the three calibration methods and rounds, where the RMSD values were averaged over theta points. Again, BILOG-MG showed the smallest average RMSD across all rounds, while ForScore showed the largest.

Table 2.7. RMSD Between True and Estimated Abilities

	Round				
Method	1	2	3	4	5
BILOG-MG	0.310	0.294	0.319	0.306	0.315
IFACT	0.314	0.301	0.319	0.311	0.320
ForScore	0.332	0.348	0.398	0.400	0.394

Table 2.8. Average RMSD Between True and Estimated IRFs

	Round				
Method	1	2	3	4	5
BILOG-MG	0.022	0.025	0.023	0.024	0.022
IFACT	0.024	0.028	0.025	0.029	0.028
ForScore	0.028	0.059	0.055	0.070	0.109

In all, the results suggest a slight edge to BILOG-MG in terms of the recovery of true parameters and true abilities. The IFACT results suggested a small degree of bias in the parameter estimates. The ForScore results suggested some degree of parameter drift may have occurred over the multiple rounds. Thus, for items that are truly 3PL in practice, BILOG-MG appears to be the preferable calibration method. In the case of items that do not conform to a 3PL model (not studied here), ForScore may show a better relative performance because it employs a non-parametric approach to calibration.

2.2.2. Operational Calibration Procedures

Based on the outcome from the simulation study, BILOG-MG was chosen as the calibration method for the operational data. For each test, separate calibrations were conducted for each block of 100 tryout items (discussed previously in Section 2.1). In all, a total of 100 calibrations were conducted (10 calibrations for each of the 10 ASVAB tests). The calibration datasets consisted of large samples of examinees that took either CAT-ASVAB Form 1, 2, or 3, plus one tryout item from the tryout block. For each block of tryout items that was administered, the CAT-ASVAB pools were administered to randomly equivalent groups of examinees. Likewise, all items within a tryout block were administered to randomly equivalent groups of examinees.

2.2.2.1. Calibration Design

Figure 2.1 summarizes the design that was used for the individual calibrations. A large block matrix of items by examinees was created. The grey-highlighted areas represent sub-matrices of sparse item responses for examinees that were administered the particular CAT-ASVAB pool (Form 1, 2, or 3) indicated in the column and row headings, plus the tryout block. These matrices are sparse because examinees took only a subset of all of the items contained in the pool or tryout block. For example, if CAT-ASVAB 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. (Note that the n_1, n_2, and n_3 possible operational items for CAT-ASVAB Forms 1, 2, and 3, respectively, are reported in Table 1.2.) For the CAT-ASVAB form administered, responses to items that were taken were coded as 0 or 1, according to whether or not the examinee answered the items correctly. Responses to items within the pool that were not administered were treated as "not presented" in the calibrations. For the remaining CAT-ASVAB pools that were not administered, all items in those pools were treated as not presented in the calibrations. The white areas in Figure 2.1 indicate the CAT-ASVAB pools that were not presented. Note that the only common items administered across the three groups defined by pool were the single tryout items administered to each examinee. All other operational items administered across the three groups defined by pool were unique to the pool. However, the random equivalent groups feature of the data collection design helped to ensure that the final parameter estimates were all on a common metric.

Figure 2.1. Operational Calibration Design for the Tryout Items

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

Note: CAT-ASVAB Form 3 was put into operational use after the first four blocks of pretest administration had been completed. The design for the calibration of the tryout items that were administered in the first four blocks of data collection was similar to the design shown here, with the exception that the rows and columns representing Form 3 were excluded.

2.2.2.2. Calibration Data Set

Table 2.9 summarizes the average number of responses per item for each of the 10 blocks of tryout items, rounded to the nearest whole number. The average was computed across the 100 tryout items included in each block. The total number of examinees included in the calibrations (N_1+ N_2+ N_3) was equal to the unrounded average multiplied by 100. The N-counts for AO were smaller than the other tests in Blocks 3 and 4 because an alternate-purpose experimental test was administered in place of AO for some examinees.

Table 2.9. Average Calibration Sample Sizes for Tryout Items Across Blocks

	Block									
Test	1	2	3	4	5	6	7	8	9	10
All but AO	1,786	1,826	1,924	1,704	1,250	1,545	1,521	1,877	2,417	1,503
AO	1,786	1,826	1,894	1,252	1,250	1,545	1,521	1,877	2,417	1,503

2.2.2.3. Calibration Process

BILOG-MG was used for each test and block to simultaneously calibrate parameters for the operational pools (Forms 1–2 in Blocks 1–4, and Forms 1–3 in Blocks 5–9) and the tryout items. For the reasons identified in Section 2.2, the calibration process was difficult and a number of steps had to be taken to obtain converged solutions. The steps discussed here were also applied in the simulation study to obtain convergence.

Starting values of 1.0 and 0.0 were used for the *a* and *b* parameters, respectively, for all items on Forms 1–3. These start values were used to override the default starting values (based on classical statistics) to avoid items being automatically omitted from the calibrations because of low negative biserial correlations. Negative biserial correlations occur often for CAT administered items because the items are administered to a group with a restricted range of ability. The default starting values were appropriate for the tryout items because those items were administered at random to examinees and were not targeted toward a particular ability. Note that the simulation study indicated that in some cases where the starting values for the operational CAT items were far from the true parameter values, the true parameter values were not well recovered by BILOG-MG. Parameter estimation of the tryout items, however, appeared unaffected by poor estimation of some operational items.

BILOG-MG uses two methods of solving the marginal likelihood equations during parameter estimation: the EM algorithm is implemented first and, upon convergence, is followed by Newton-Gauss (Fisher scoring) iterations. In some calibrations, it was not possible to get convergence in the Newton-Gauss step. In those cases, the Newton-Gauss step was suppressed, and the converged results from the EM step were used as the final parameters. In some calibrations, it was not possible to obtain convergence in the EM step. In those cases, the item(s) that were not converging were identified and excluded from the calibration so that a converged solution could be obtained. It was also necessary to invoke a prior distribution for the *b* parameters to obtain a converged solution. The default prior for the *b* parameters was used, in addition to the default priors for the *a* and *c* parameters that are automatically invoked by BILOG-MG (Zimowski, et al., 2003).

The problems noted during the operational calibrations demonstrate that there are some atypical calibration issues that arise when applying BILOG-MG to sparse CAT data. Calibrations of CAT data may be affected not only by the sparseness of the data but by the relationship between the item and the ability of the calibration sample. For example, in the case of a very difficult item that is administered to a below-average ability group, it may be difficult to calibrate that item no matter how many people take it. Problems in calibrating CAT data are discussed further in Pommerich and Segall (2003). Based on the findings from the simulation study, it was expected that the parameter estimates for the tryout items would be largely unaffected by the occasional poor calibration of some operational items. Since optimal calibration of the tryout items was the primary concern, it was deemed possible to overlook the noise caused by a few poorly calibrated items from CAT-ASVAB Forms 1–3.

2.2.2.4. Parameter Scaling

The *a* and *b* parameters for the tryout items were placed onto the scale of the operational parameters for CAT-ASVAB Forms 1–4 using the transformations given in Equation 2.1 in Section 2.2.1.3. The transformation constants *A* and *B* were computed as given in Equation 2.2, where μ_1 and σ_1 were the mean and standard deviation of $\hat{\theta}$ estimates in the calibration sample computed using the operational parameters, and μ_2 and σ_2 were the mean and standard deviation of the $\hat{\theta}$ estimates in the calibration sample computed using the re-estimated parameters for the operational items. All $\hat{\theta}$ estimates were computed via Bayes mode using a fixed N(0,1) prior. Final transformed parameters were obtained using the iterative process that was used to transform IFACT parameters in the calibration simulation study, as described in section 2.2.1.3.

2.2.2.5. Check of AO Parameter Calibration and Scaling

An alternate calibration and scaling was conducted for AO to check the calibration and scaling procedures that were to be implemented operationally. The alternate calibration/scaling was conducted for AO because the final transformed parameters for the tryout items demonstrated a noticeable shift in difficulty from CAT-ASVAB Forms 1–4, and it was necessary to ensure that the difficulty shift was not caused by the calibration/scaling method.

In the alternate calibration/scaling, BILOG-MG was again used to calibrate the data set described in Section 2.2.2.2, and the item parameters were transformed using the transformations given in Equation 2.1 in Section 2.2.1.3. However, the calibration differed in that the FIX command was used to fix the parameters for all of the operational CAT-ASVAB items to their existing operational values, and only the tryout items were calibrated. To avoid the possibility of inducing bias in the parameters for the tryout items (see Footnote 1 in Section 2.2.1.3), an iterative calibration/transformation process was used to control for the underlying ability distribution in the calibration sample. Final transformations were conducted following the iterations.

Specifically, the operational parameters were fixed, and the tryout items were calibrated. Transformation constants *A* and *B* were then computed as given in Equation 2.2, where μ_2 and σ_2 were equal to the estimated mean and standard deviation of the posterior distribution, as output from the calibration. μ_1 and σ_1 were fixed to 0.0 and 1.0, respectively, which was the approximate posterior distribution for the calibrations of the parameters used in operational administrations of AO. The transformation constants were then used to transform the operational parameters, as given in Equation 2.1. The transformed operational parameters were then fixed, and the tryout items were recalibrated. The process of transforming the operational parameters, fixing the transformed operational parameters, and recalibrating the tryout items was then repeated until the mean and standard deviation of the output posterior distribution was approximately 0.0 and 1.0, respectively, implying that the tryout parameters were on a N(0,1) scale.

Final transformation constants *A* and *B* were then computed as the slope (*A*) and intercept (*B*) of the line connecting the coordinates for two operational items, where the coordinates were defined by the original untransformed *b*-parameter values and the most recent transformed *b*-parameter values for the items. The final transformation constants were then applied to the operational and tryout item parameters output from the most recent calibration. Following the final transformation, the parameters for the operational items matched their original untransformed values (to within four decimal places), implying that the original scale was maintained and that the tryout item parameters were on the same scale as the original untransformed operational parameters.

Figures 2.2–2.4 compare the *a*, *b*, and *c* parameters across the two different methods of calibration/scaling for 646 AO tryout items that survived the initial item screening process (described in more detail in Section 2.3.1 below). Figure 2.2 shows that the *a* parameters were similar across the two methods of calibration/scaling for most items. The correlation between the two sets of *a* parameters was 0.92. Figure 2.2 suggests a small degree of bias in the parameter estimates, in that more items appeared to fall below the identity line than above. Figure 2.3 shows that the *b* parameters were very similar across the two methods of calibration/scaling. The correlation between the two sets of *b* parameters was 0.99, which implies that the items were rank ordered almost identically across the two calibration/scaling methods. The finding that the two sets of *b* parameters were more similar than the two sets of *a* parameters across the calibration/scaling methods was not unexpected, given that *a* parameters are typically less well estimated than *b* parameters. The *c* parameters were not transformed in either scaling; thus, the parameter differences observed across the original and alternate calibration/scalings reflect the differences in the calibration procedures. The two sets of *c* parameters were highly correlated (0.94); however, Figure 2.4 shows evidence of bias in the parameter estimates in that most items exhibited lower values in the alternative calibration/scaling (operational items fixed, tryout items estimated) than in the original calibration/scaling (operational and tryout items estimated).

Figure 2.2. Comparison of *a* Parameters for AO Across the Two Different Methods of Calibration/Scaling

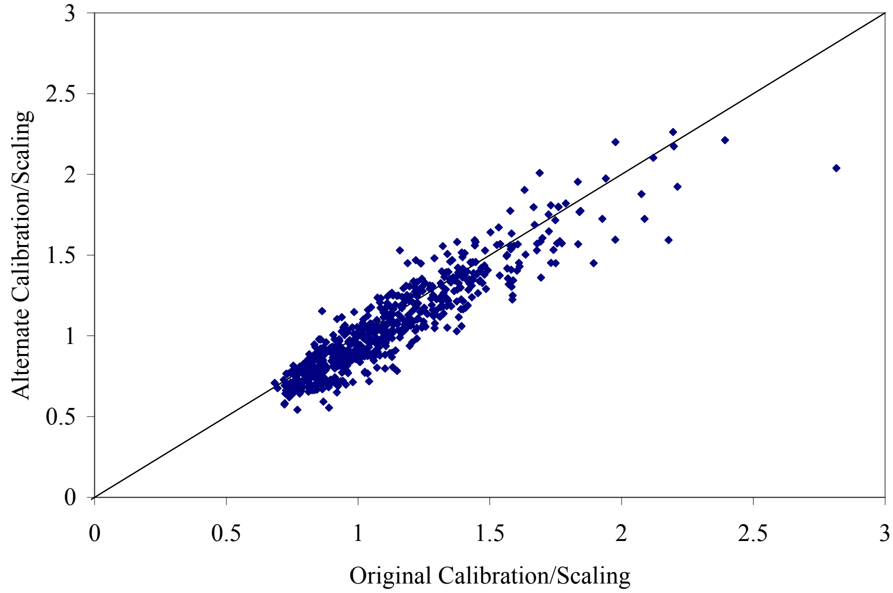


Figure 2.3. Comparison of *b* Parameters for AO Across the Two Different Methods of Calibration/Scaling

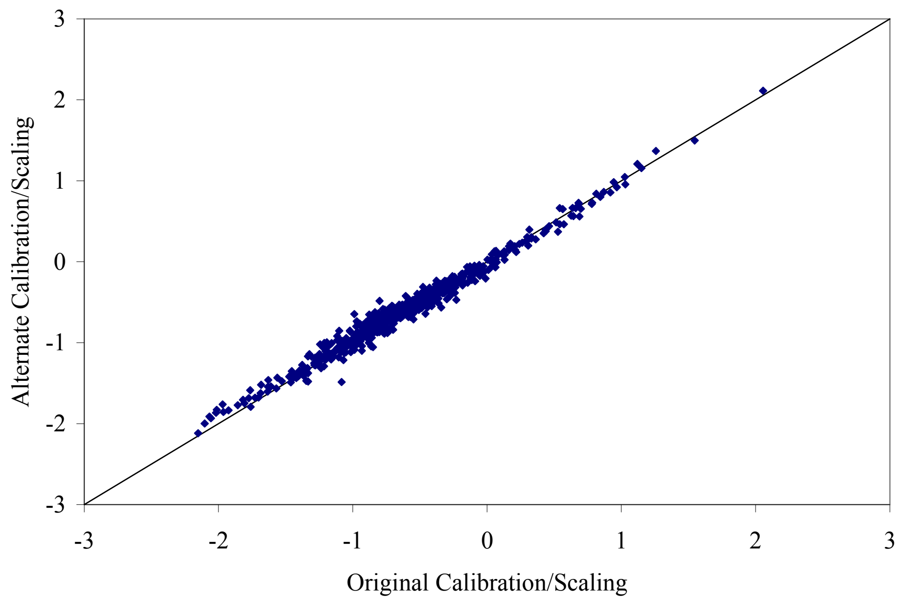
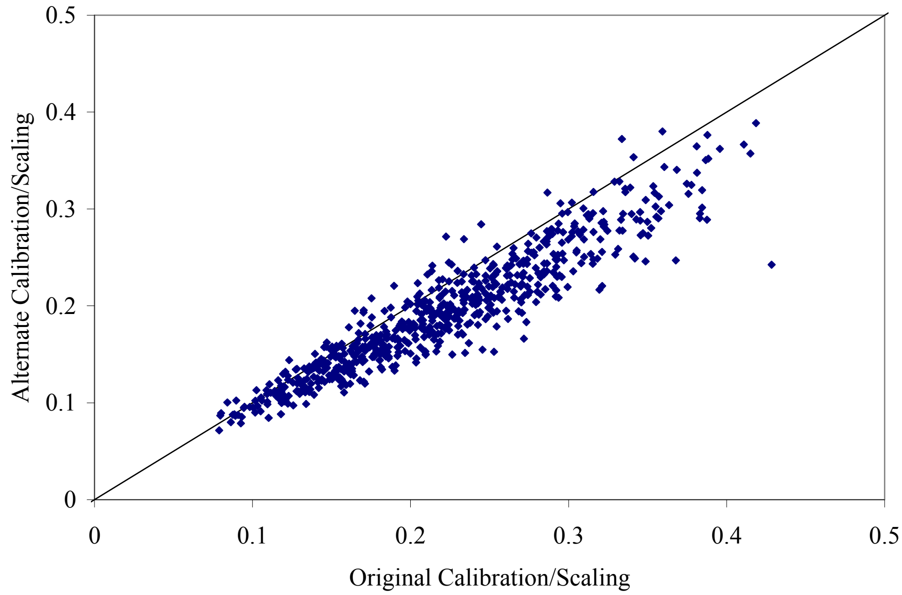


Figure 2.4. Comparison of *c* Parameters for AO Across the Two Different Methods of Calibration/Scaling

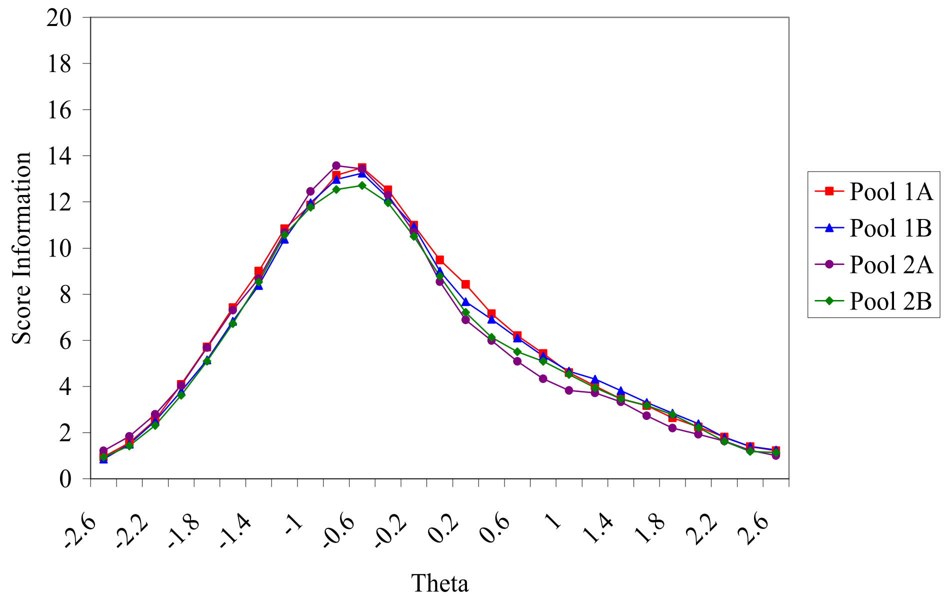


The results for the *c* parameters suggest that the parameter estimation may have been constrained by the range of the fixed parameters in the case of the alternate calibration. The results for the *a* parameters likewise suggest that the parameter estimates may have been similarly constrained by the range of the fixed parameters in the alternate calibration, although to a lesser degree than occurred with the *c* parameters.

In an additional check of the two sets of AO item parameters, two item pools of equal size (323 items in each pool) were built using each set of parameters. For each set of parameters, the pools were built so that total item information was nearly equal across the two pools. Using each pool, 15-item CATs were simulated for 2,000 examinees at each of 31 equally spaced points between ± 3.0 (62,000 examinees total). The simulated 15-item CAT pools were built using the same procedures used to build operational CAT pools, and they were administered using the same procedures used to operationally administer CAT-ASVAB for AO. Score information functions were computed for each simulated CAT pool based on the responses of the 62,000 examinees using a smoothed approximation of Lord's (1980) Equation 10-7, as presented in Segall, Moreno, and Hetter (1997) and ASVAB Technical Bulletin #1 (DMDC, 2006). (Score information functions are discussed in more detail in Section 2.5.1 below.)

Figure 2.5 shows the score information functions for the four AO CAT pools. Pools 1A and 1B were built using the parameters from the original calibration/scaling. Pools 2A and 2B were built using the parameters from the alternate calibration/scaling. The score information functions were similar across all four pools, which suggests that the precision of AO CAT scores would be similar regardless of which calibration/scaling method is used. The fact that the alternate calibration/scaling (where the operational parameters were fixed and not recalibrated) showed similar results to the original calibration/scaling (where the operational parameters were recalibrated) suggests that the occasional poor calibration of some operational items, noted earlier, is not likely to negatively affect tryout item parameters transformed via the original scaling method. The results also suggest that the shift in difficulty noted in the AO tryout items was attributable to factors other than the calibration/scaling method.

Figure 2.5. Comparison of Score Information Functions Across AO CAT Pools Built Using Parameters from Different Calibrations/Scalings



2.3. Evaluation of Tryout Items

The 10,000 tryout items were evaluated in a variety of ways, including statistical reviews of item characteristics, differential item functioning (DIF) reviews, internal and external content reviews, and external sensitivity reviews. The reviews are discussed in more detail in the following sections.

2.3.1. Initial Screening and Selection

An initial screening of the tryout items was conducted to eliminate items that were clearly poor performers from a statistical perspective. All of the tryout items were evaluated based on (a) comparisons of biserial correlations across keyed responses and distractors; (b) comparisons of percentages of low, medium, and high ability examinees responding to keyed responses and distractors; (c) comparisons of item information functions; and (d) comparisons of parameters and item characteristic curves (ICCs) across BILOG-MG, IFACT, and ForScore.

Items with the lowest information within each test were automatically dropped (approximately 25% of each tryout block). Items that were not dropped for low information were flagged for review if one or more of the following occurred: (a) a negative or low positive biserial correlation for the keyed response; (b) a very low percentage of correct responses; (c) an overly attractive distractor; (d) highly discrepant parameters across the BILOG-MG, IFACT, and ForScore calibration programs; (e) a non-parametric ICC (computed by ForScore) appeared to be non-monotonic; or (f) a BILOG-MG or IFACT ICC appeared to be non-3PL (indicated by large discrepancies between the 3PL and non-parametric ICCs). All flagged items were independently reviewed with respect to both content and statistics. If the review suggested that an item had acceptable content and statistics, it was retained for the next round of evaluations. Table 2.10 summarizes the total number of items that were dropped for each test after the initial screenings and reviews were conducted.

Table 2.10. Number of Items Dropped after Initial Screening

Test	# Dropped	Test	# Dropped
GS	344	EI	400
AR	281	AI	390
WK	312	SI	449
PC	298	MC	374
MK	324	AO	267

2.3.2. DIF Analyses

Analyses of DIF were conducted for all items that survived the initial screening. To evaluate DIF, item performance for a *focal* group is compared to item performance for a *reference* group, where the focal group is a potentially disadvantaged group, and the reference group is a potentially advantaged group. If an item performs differently for focal and reference groups that are matched on ability, the item is said to display DIF. DIF is a necessary (but not sufficient) condition for the occurrence of *bias*. Bias occurs when an item or test unfairly favors one group over another. The occurrence of bias is problematic because it can negatively affect test validity.

Table 2.11 shows the pairs of subgroups for which DIF analyses were conducted. These subgroups all had sufficient sample sizes from which to conduct the analyses. An Empirical Bayes Mantel-Haenszel statistic (Zwick, Thayer, & Lewis, 1997, 1999) was computed for each item and subgroup pair. The magnitude of the Mantel-Haenszel value (labeled as EB_{MH}) was summarized across all items using the A, B+, B−, C+, C− classification system as outlined in Zwick, et al. (1999) and shown in Table 2.12.

Table 2.11. Pairs of Subgroups Used in the DIF Analyses

Label	Reference Group	Focal Group
M/F	Males	Females
W/B	Whites	Blacks
N/H	Non-Hispanic Whites	Hispanics

Table 2.12. Classification Scheme Used to Summarize the DIF Results

Notation	Description	Mantel-Haenszel Value
A	Negligible DIF	$\lvert EB_{MH} \rvert < 1.0$
B	Slight to moderate DIF	$1.0 \le \lvert EB_{MH} \rvert < 1.5$
C	Moderate to severe DIF	$\lvert EB_{MH} \rvert \ge 1.5$
+	Direction favors Focal group	$EB_{MH} > 0.0$
−	Direction favors Reference group	$EB_{MH} < 0.0$

Tables 2.13–2.22 show the summary of the DIF results for each test. Each table summarizes the number of items falling in each classification category for each subgroup pair (M/F, W/B, and N/H). The tables show that most items were classified in category A across all the tests. Items that were flagged for categories B or C DIF were reviewed with respect to both content and statistics. Items with content that suggested a possible unfair advantage for either the focal or reference group were dropped. Evaluations of the statistics for the flagged items showed that many items classified in category C had floor or ceiling effects (i.e., the items were answered correctly or incorrectly by almost all examinees). Flagged items displaying floor or ceiling effects were retained for further evaluations if there was no apparent content reason for the occurrence of DIF.

Table 2.13. DIF Results for GS

Subgroup Pair	# A Items	# B+ Items	# C+ Items	# B− Items	# C− Items
M/F	631	18	7	0	0
W/B	654	1	1	0	0
N/H	648	6	2	0	0
Total	1933	25	10	0	0

Table 2.14. DIF Results for AR

Subgroup Pair	# A Items	# B+ Items	# C+ Items	# B− Items	# C− Items
M/F	695	20	4	0	0
W/B	719	0	0	0	0
N/H	713	5	1	0	0
Total	2127	25	5	0	0

Table 2.15. DIF Results for WK

Subgroup Pair	# A Items	# B+ Items	# C+ Items	# B− Items	# C− Items
M/F	632	39	16	1	0
W/B	658	17	13	0	0
N/H	628	34	26	0	0
Total	1918	90	55	1	0

Table 2.16. DIF Results for PC

Subgroup Pair	# A Items	# B+ Items	# C+ Items	# B− Items	# C− Items
M/F	668	16	9	0	0
W/B	693	0	0	0	0
N/H	690	2	1	0	0
Total	2051	18	10	0	0

Table 2.17. DIF Results for MK

Subgroup Pair	# A Items	# B+ Items	# C+ Items	# B− Items	# C− Items
M/F	669	7	0	0	0
W/B	667	9	0	0	0
N/H	675	1	0	0	0
Total	2011	17	0	0	0

Table 2.18. DIF Results for EI

Subgroup Pair	# A Items	# B+ Items	# C+ Items	# B− Items	# C− Items
M/F	596	4	0	0	0
W/B	599	1	0	0	0
N/H	595	5	0	0	0
Total	1790	10	0	0	0

Table 2.19. DIF Results for AI

Subgroup Pair	# A Items	# B+ Items	# C+ Items	# B− Items	# C− Items
M/F	608	2	0	0	0
W/B	609	1	0	0	0
N/H	602	6	2	0	0
Total	1819	9	2	0	0

Table 2.20. DIF Results for SI

Subgroup Pair	# A Items	# B+ Items	# C+ Items	# B− Items	# C− Items
M/F	531	3	4	13	0
W/B	551	0	0	0	0
N/H	542	8	1	0	0
Total	1642	11	5	13	0

Table 2.21. DIF Results for MC

Subgroup Pair	# A Items	# B+ Items	# C+ Items	# B− Items	# C− Items
M/F	622	3	0	0	0
W/B	625	0	0	0	0
N/H	625	0	0	0	0
Total	1872	3	0	0	0

Table 2.22. DIF Results for AO

Subgroup Pair	# A Items	# B+ Items	# C+ Items	# B− Items	# C− Items
M/F	728	5	0	0	0
W/B	733	0	0	0	0
N/H	732	1	0	0	0
Total	2193	6	0	0	0

Because of a lack of in-house foreign language expertise, all B and C category items flagged in the Non-Hispanic White/Hispanic (N/H) analyses were reviewed by two linguists from the Defense Language Institute Foreign Language Center who were fluent in both English and Spanish. (Note that all the flagged items in the N/H analyses favored Hispanics.) The linguists reviewed the items with respect to the following: (a) Spanish equivalents/translations for key words in the stem and response options, (b) the relationship between the stem and correct response in English and the Spanish translations, (c) whether Spanish speakers could rule out any options that English speakers could not, (d) the register associated with the stem and options in each language (i.e., the level of ability and/or education at which the terms and comparable translations are used), (e) the frequency of use for terms and comparable translations in each language, and (f) the cultural context in which the terms and comparable translations are used in each language.

Table 2.23 summarizes the total number of items dropped after conducting content and statistical reviews of the B and C category DIF items. Most of the dropped items were viewed as being potentially advantageous to Spanish speaking examinees.

Table 2.23. Number of Items Dropped after DIF Analyses

Test	M/F	W/B	N/H	Total
GS	1	0	2	3
AR	0	0	0	0
WK	0	0	31	31
PC	0	0	0	0
MK	0	0	0	0
EI	0	0	0	0
AI	0	0	0	0
SI	1	0	1	2
MC	0	0	0	0
AO	0	0	0	0
Total	2	0	34	36

2.3.3. External Content Reviews

Items that made it through the initial screenings and DIF evaluations were submitted to a content review by content specialists external to the ASVAB testing program. All content areas were externally reviewed, with the exception of AO. Reviewers were solicited through a national advertising campaign for content specialists with relevant expertise relevant. All participating external reviewers had verifiable credentials.

The reviewers were instructed to verify the following for all items: (a) the item content was accurate; (b) the item was keyed correctly; (c) there were no misstatements of fact or confusion of principles involved in the item stem or answer choices; and (d) any artwork associated with the items was accurate and understandable. Items that were flagged by the content experts as questionable on any of the factors noted above underwent further review. Flagged items for the AI, EI, SI, and MC tests were independently reviewed by a second external content expert. Flagged items for GS and MK were reviewed by an in-house editor with content expertise. Flagged items for AR, PC, and WK were reviewed by two in-house editors. Recommendations to drop or keep the flagged items were made based on the external content expert input, in-house item development expertise, and ASVAB style guides. Items were always dropped in cases where a second reviewer agreed with the initial content expert's recommendation. Table 2.24 shows the percentage of externally reviewed items that were dropped after all of the content reviews were completed.

Table 2.24. Percentage of Items Dropped after Content Reviews

Test	% Dropped	Test	% Dropped
GS	7.3	EI	10.3
AR	1.9	AI	9.8
WK	2.8	SI	5.4
PC	11.1	MC	1.4
MK	4.1		

2.3.4. External Sensitivity Reviews

Items that survived the external content review also underwent an external sensitivity review. All content areas except AO and MK were reviewed for potential bias and insensitivity. AO and MK were not reviewed because they were deemed to be unsusceptible to insensitivity issues due to their language-free content. The sensitivity reviews were conducted by the American Institutes for Research (AIR), which assembled a team of ethnically diverse reviewers. All reviewers were trained using guidelines developed by both AIR and the Defense Manpower Data Center (DMDC). Teams of two-to-three people reviewed all items, and flagged items were then reviewed by a senior editor at AIR who either agreed or disagreed that the item should be flagged. An in-house review of all flagged items was then conducted at DMDC where test developers considered the item content and the rationale for flagging in making a final decision whether to drop or keep flagged items.

Some general findings from the external sensitivity review were that flagged items typically (a) used gender-specific pronouns and names; (b) used Caucasian-sounding names; (c) could be deemed as ageist, elitist, or regional; or (d) referred to "America" rather than the "United States." The majority of flagged items were for use of gender-specific pronouns. Table 2.25 shows the number of items that were dropped for various bias and sensitivity concerns after all the reviews were completed. PC displayed the most items dropped for possible insensitivity, which is to be expected because of the predominance of human-interest content in the reading passages.

Table 2.25. Number of Items Dropped after Sensitivity Reviews

Sensitivity Issue	GS	AR	WK	PC	EI	AI	SI	MC	Total
Culture	1	0	0	0	0	0	0	0	1
Disabilities	0	0	0	1	0	0	0	0	1
Distracting Material	0	0	0	3	0	0	0	0	3
Elderly	0	0	0	2	0	0	0	0	2
Ethnicity	0	0	0	5	1	0	0	0	6
Gender	0	1	0	27	0	0	0	0	28
Geographical Area	0	0	0	0	0	1	0	0	1
Medical Conditions	0	0	0	1	0	0	0	0	1
Religion	0	0	0	6	0	0	0	0	6
Sexual Orientation	0	0	1	0	0	0	0	0	1
Vocabulary	0	1	0	0	0	0	0	0	1
Weight Issues	0	0	1	0	0	0	0	0	1
Other	0	0	0	2	0	0	0	0	2
Total	1	2	2	47	1	1	0	0	54

2.3.5. Summary

All tryout items underwent extensive reviews. Items that did not meet the criteria for operational use across all of the evaluations were dropped from consideration. Items that met the criteria for operational use across all of the evaluations were retained for use in assembling new item pools (i.e., new CAT-ASVAB forms). Table 2.26 summarizes the final numbers of retained items, following all evaluations.

Table 2.26. Number of Items Retained after All Evaluations

Test	# Retained
GS	581
AR	665
WK	616
PC	545
MK	606
EI	503
AI	473
SI	490
MC	584
AO	646

2.4. Evaluation of Local Dependence

Preliminary work to assemble new item pools from the tryout items showed some unexpected results for the MK test. Namely, score information functions for the new pools did not always meet the targeted score information functions (defined by CAT-ASVAB Form 1 operational parameters). This phenomenon was initially attributed to methodological differences between the software programs used to calibrate the parameters (ASCAL [Vale & Gialluca, 1985] for Form 1 parameters versus BILOG-MG for the tryout parameters). Comparisons of the Form 1 parameters as originally calibrated using the ASCAL computer program and the Form 1 parameters recalibrated using BILOG-MG showed similar results for the *b* parameters across all items but substantially different *a* parameters for some items. Examination of item content for the items in question showed that each item shared very similar content with another item in the calibration data, which suggested that local dependence was another possible explanation for the parameter differences.

Local independence is one of the primary assumptions behind Item Response Theory. It assumes that conditional on ability, an examinee's responses to different items on a test are statistically independent:

$$P(\underline{U} = \underline{u} \mid \theta) = \prod_{i=1}^{n} P_i(U_i = u_i \mid \theta), \qquad (2.5)$$

such that the probability of an observed response pattern ($\underline{U}$) for an examinee is equal to the product of the probability of the observed response on each individual item *i*, multiplied over all *n* items taken. Namely, after taking into account an examinee's underlying ability level, knowledge of the examinee's performance on one item must not provide additional information about his or her responses to any other item. When an examinee is administered items that share highly similar content or format, it is possible that knowledge of his or her performance on one item could provide additional information about his or her response to another item, even after taking into account his or her ability level.

Local dependence (LD) is said to exist when local independence does not hold. In the presence of LD, the estimation of item parameters may be affected due to the violation of Equation 2.5. Research studies have shown that measurement characteristics such as test information, reliability, and item discrimination are inflated when a 3PL model is used to score passage-based items where LD is inherent (e.g., Zenisky, Hambleton, & Sireci, 2002; Yen, 1993; Sirici, Thissen, & Wainer, 1991; and Thissen, Steinberg, & Mooney, 1989). Oshima (1994) showed that speededness (identified by Yen [1993] as a possible cause of local dependence) resulted in overestimation of the 3PL item discrimination and difficulty parameters, with a more serious effect for the discrimination (*a*) parameter. Inflated *a* parameters lead to inflated estimates of score information, which can result in examinees being measured with less precision than is apparent.

2.4.1. Effect of LD on MK Score Precision

Because the occurrence of LD can have negative consequences for the accuracy of examinee scores, a study was conducted to thoroughly evaluate the existence and effect of LD in the MK test. The existence of LD is typically detected by applying diagnostic tools to responses between item pairs. However, the seeded item design used for pretest administrations of the tryout items (summarized in Table 2.1) prohibited application of the diagnostic tools since each examinee took only one tryout item. Thus, CAT-ASVAB Forms 1 and 2 were used as the basis for evaluating the existence of LD in the MK test, as their items were administered conventionally in a tryout study before they were incorporated into CAT pools. A simulation study was then used to evaluate the effects of LD on the precision of examinee CAT scores. Inferences from the diagnostic evaluations of CAT-ASVAB Forms 1–2 tryout data and the simulation study were used to make procedural decisions regarding LD in the development of CAT-ASVAB Forms 5–9. Complete details about the item tryout for CAT-ASVAB Forms 1–2 are available in Prestwood, Vale, Massey, & Welsh (1985), while complete details about the evaluation of LD in the item tryout data and the simulation study are presented in Pommerich and Segall (2008).

Diagnostic assessments of LD in the tryout study data were made using Pearson's X^2 statistic (e.g., Chen & Thissen, 1997), the Q_3 statistic (Yen, 1984), and density plots of residual tetrachoric correlations. Preliminary content evaluations of the test booklets for the tryout study showed many instances of items with similar, finely specified content administered together in the same booklet. The diagnostic evaluations confirmed that many item pairs exhibited LD, some to a very large degree. The item pairs exhibiting the biggest degree of LD were item clones; namely, items testing the same concept using the same notation but containing different numerical values in the stem of the question and the response options. A large number of item pairs were flagged by all three diagnostic measures; content evaluations of those items showed content justification for the flagging.

The diagnostic evaluations of the data from the tryout study displayed strong evidence that LD occurred between certain types of item pairs that were administered to the same examinee. Prior to the diagnostic evaluations of LD discussed here, results from the tryout study had been used to conduct calibrations, select items, and assemble CAT-ASVAB Forms 1–2. Unfortunately, it was not feasible to conduct diagnostic assessments of LD in response data from operational administrations of CAT-ASVAB Forms 1–2. Although Pommerich and Ito (2008) demonstrated that the Q_3 statistic might be usable with adaptive test data, it was not possible to apply the Q_3 procedure systematically to CAT-ASVAB data because of the widely varying sample sizes observed across the item pairs (resulting from varying administration rates across items). Because of the difficulties inherent with applying the LD statistics to adaptive CAT data, the probability of the occurrence of LD in operational administrations of CAT-ASVAB Forms 1 and 2 was instead inferred by extrapolating the findings from the tryout study data to the CAT pools. These evaluations suggested that LD likely occurred to some degree in operational CAT administrations of the MK test.

The effect of LD on CAT score precision was then evaluated using a simulation study designed to evaluate the effect of two sources of LD on the precision of examinee scores: (a) LD in CAT item parameters, and (b) LD in examinees' CAT item responses. The former was evaluated by comparing results when LD was induced in the item parameters with results when no LD was induced in the parameters, while the latter was evaluated by comparing results when LD was induced in examinee responses with results when no LD was induced in the responses. The analyses consistently showed that under the simulated conditions, (a) LD in CAT item parameters had a very minimal effect on the precision of examinee CAT scores, while (b) LD in examinees' CAT item responses had a fairly substantial effect on score precision, depending on the degree of LD present.

Although the simulation findings suggest that examinees that are administered CAT-ASVAB Forms 1 and 2 could be measured at a lower precision than intended, empirical reliability and validity studies of the CAT scores show sufficiently high levels of reliability and validity when compared to P&P testing. If local dependence had caused a large or significant decrement in precision, results of the empirical reliability and validity studies would be expected to reflect it. Nevertheless, the results from the diagnostic assessments of LD and the simulation study suggest that it would be prudent to guard against the occurrence of LD in future administrations of the MK test.

2.4.2. Evaluation of LD in Other ASVAB Tests

Additional diagnostic evaluations were conducted on data from the tryout study (Prestwood, et al., 1985) to evaluate the likelihood of LD occurring during CAT-ASVAB administration of the other ASVAB tests. All tests but AO were evaluated using Pearson's X^2 statistic (AO was not administered in the tryout study). Results for the PC, WK, and AR tests showed little evidence of LD, with the exception of some items pairs with similar levels of difficulty (i.e., both very easy or very hard) that tended to be flagged for LD even when there was no content justification for the flagging. Results for GS, AI, SI, and EI showed evidence of LD in only a handful of items with highly similar content. However, results for MC showed consistent evidence of LD among items testing similar concepts.

2.4.3. Evaluation of LD in the Tryout Items

The diagnostic measures used to assess LD on the ASVAB tests operate on item pairs that have been taken by the same examinee. Because the data collection design prohibited application of the LD statistics to the new tryout items, the probability of the occurrence of LD was inferred by extrapolating the findings from evaluations of the item tryout data for CAT-ASVAB Forms 1 and 2 to the new tryout items. Content evaluations of the new tryout items suggested the necessity of taking steps to control for LD in future CAT-ASVAB administrations of the items, particularly in the MK and MC tests.

2.4.4. Determination of Item Enemies

Various models have been proposed to control for the occurrence of LD in item responses. These include grouping LD items into testlets and using a polytomous model to score the testlets (Wainer & Kiely, 1987), using a model that accounts for LD between items (Tuerlinckx & De Boeck, 2004), and using a model that does not assume local independence (Jannarone, 1997). Another solution was proposed by Pommerich and Segall (2008) that would allow the use of a 3PL model for item selection and scoring. Their approach was to identify groupings of item enemies and allow only one item per enemy grouping to be administered to the same examinee. The term "item enemies" is used here to refer to items that are likely to trigger LD in responses if administered to the same examinee. All of the above approaches would require the identification of groupings of LD items.

To allow the continued use of a 3PL model with CAT-ASVAB administration, the item enemy solution was selected for implementation with the new item pools. In the case of the MK and MC tests, extensive enemy lists were developed, where enemies were defined as items likely to display LD because they addressed highly similar concepts. For MK, a total of 155 enemy groupings were specified, and all tryout items were classified into an enemy group. Some sample groupings included reciprocals, prime numbers, least common dominators, and order of operations. For MC, a total of 95 enemy groupings were specified, and all tryout items were classified into an enemy group. Some sample groupings included the expansion/contraction of materials, buoyancy in fluids and gases, pulley effort, and acceleration. Note that an enemy group could contain only one item if that item was not deemed to be an enemy of any other items. In the case of the AI, AR, EI, GS, SI, and WK tests, a handful of enemy groupings were defined for the purpose of controlling content that might be viewed as redundant, such as two questions addressing different aspects of photosynthesis. For all content areas, if an item was administered from an enemy group, all other items in that group were subsequently blocked from administration during the test session.

2.5. Pool Assembly

CAT administrations utilize pools of items from which a (potentially) unique subset of items are selected for administration to an individual. Because CAT administrations are targeted toward an individual examinee's level of ability, the item pools must contain items that range in difficulty from very easy to very difficult. Also, there should be sufficient numbers of items across the range of difficulty to precisely measure each examinee's level of ability. Formal procedures are typically used to assign items to pools (i.e., assemble pools).

Prior to developing final pools from the retained tryout items, preliminary research was conducted to evaluate the quality of item pools built using (a) parameters obtained from the three calibration methods discussed earlier (BILOG-MG, IFACT, and ForScore), and (b) different form-assembly methods. Findings were used to determine procedures for assembling the final pools.

2.5.1. Comparison of Pools across Calibration Methods

Using each set of parameters, two-to-four pools were constructed for the AR, MK, PC, and WK tests (i.e., the tests used to compute AFQT scores) with the goal of maximizing conditional precision levels under the constraint that the pools be similar to one another. Specifically, items with similar information functions were identified and assigned to separate pools in an attempt to minimize the differences among the pool information functions. For each number of pools created (two, three, or four), the precision of the new pools was compared across the three calibration methods (BILOG-MG, IFACT, and ForScore). The number of assembled pools was varied to help evaluate the number of pools that could be built while maintaining adequate precision.

Pool precision was evaluated using score information functions. Equation 2.6 presents the information function for a test score.

$$I(\theta) = \frac{\left[\frac{\partial}{\partial\theta}\mu\left(\hat{\theta} \mid \theta\right)\right]^2}{\sigma^2\left(\hat{\theta} \mid \theta\right)} . \qquad (2.6)$$

Higher information at a given ability level corresponds to a smaller asymptotic confidence interval for ability estimation. Score information functions were computed based on the responses of 2,000 examinees at each of 31 equally spaced points between ± 3.0 (62,000 examinees total), using a smoothed approximation of Lord's (1980) Equation 10-7, as presented in Segall, Moreno, and Hetter (1997). Item responses were simulated using CAT-ASVAB item selection and scoring algorithms. More details about operational administration conditions for CAT-ASVAB are given in Segall, Moreno, Bloxom, and Hetter (1997) and ASVAB Technical Bulletin #1 (DMDC, 2006).

Comparisons of the score information functions showed a high level of agreement across pools assembled using the BILOG-MG and IFACT parameters, with IFACT displaying slightly lower estimated precision levels. Pools assembled using the ForScore parameters provided the most optimistic precision projections (i.e., precision levels were consistently higher using the ForScore parameters than the BILOG-MG or IFACT parameters). This finding was consistent with the high degree of sampling variance that was observed among the ForScore discrimination parameters relative to the BILOG-MG and IFACT parameters (see Table 2.6) and may also be attributable to methodological differences in linking the tryout parameters to the scale of the operational parameters. In all, the results suggested the suitability of using BILOG-MG parameters for operational implementation.

2.5.2. Comparison of Pool Assembly Methods

Using the BILOG-MG parameters, three different methods were used to assemble two pools for each ASVAB test (AI, AO, AR, EI, GS, MC, MK, PC, SI, and WK): (a) a simple matching of information across pools utilizing no enemy lists (discussed in Section 2.4.4.), (b) a non-linear optimization utilizing no enemy lists, and (c) a non-linear optimization utilizing enemy lists (where relevant). The first method had previously been used in the comparison of pools across calibration methods. Form assembly using the non-linear optimization approach was conducted using Markov Chain Monte Carlo methods, as described in Gilks, Richardson, and Spiegelhalter (1996). All pool assembly methods utilized the basic constraint that every item would appear in only one pool. Where enemy lists were utilized, items from an enemy group were constrained to be distributed evenly across the pools.

For GS and AO, additional constraints were incorporated to ensure appropriate representation of content taxonomies in the pools as the CAT-ASVAB controls for content taxonomy in administrations of the GS and AO tests. For pool assembly of GS and AO using the matching method, the assignment to pools was conducted separately by content taxonomy. For pool assembly of GS and AO using the non-linear optimization approach, additional constraints were incorporated to set upper and lower bounds on the number of items per content taxonomy per pool.

The basic steps for the non-linear optimization approach are as follows. The method began by roughly dividing all available items into the desired number of pools. Simulation methods were then used to compute score information curves for each pool, and items with usage rates lower than a given bound were released into a “free” pool. For each pool in areas where score information did not meet a target information function, items were swapped between the different pools and the free pool in an attempt to improve the goal function. This process was repeated until a maximum number of iterations were reached or no improvement in the goal function was noted.

For all tests except MC and MK (both of which had extensive enemy lists), comparisons of score information functions showed minimal differences across the different pool assembly methods. Under all of the pool assembly methods, estimates of score information and test-retest reliability sufficiently exceeded that of a P&P-ASVAB form for these tests. Results for these tests suggested that final pools could be assembled using the simpler method of matching information across pools, rather than the more computationally intense non-linear optimization approach.

For MC and MK, there was a noticeable loss in precision when enemy lists were used in pool assembly and CAT administration. A loss in precision is to be expected if items that would normally be selected under maximum information criteria are blocked from administration because of enemy constraints. Failing to control for local dependence in responses, however, can result in inflated estimates of actual score precision (Pommerich & Segall, 2008). The loss in precision demonstrated here is believed to be more indicative of the true precision in cases where local dependence is a concern. Further, both the MC and MK tests showed estimates of score information and test-retest reliability that well exceeded that of a P&P form of the ASVAB, even when enemy constraints were utilized in the pool assembly.

2.5.3. Final Pools

Final pools were assembled using BILOG-MG parameters and matching information across pools, incorporating enemy lists into the process where relevant. Use of these practices was supported by the findings from the evaluations of the pool assembly methods. A total of five new pools were assembled for each ASVAB test. The new pools were labeled CAT-ASVAB Forms 5, 6, 7, 8, and 9. The final pools were reviewed by content editors to ensure adequate representation of content taxonomies and non-redundancy of content within pools.

For each content area, an item information function (Lord, 1980) was computed for each retained tryout item (summarized in Table 2.26). A weighted information value was then computed for each item, where the item information functions were weighted across θ using a normal distribution. All items that were not contained in an enemy group were sorted by their weighted information value, and the five items with the highest weighted information values were randomly distributed across the five pools. The five items with the next highest weighted information values were then randomly distributed across the five pools, and so on until all of the items were distributed into the pools. A similar process was then followed within each enemy group (where relevant). Namely, all items in an enemy group were sorted by their weighted information value, with the top five items randomly distributed across the five pools, followed by the next five, and so on until all items in the enemy group were distributed.

For GS and AO, the above steps were conducted separately for items classified into each content taxonomy, as the CAT-ASVAB controls for content taxonomy in administrations of the GS and AO tests.

2.5.3.1. Exposure Control

Controlling the exposure of test questions is an important issue in CAT administration. A CAT achieves maximum precision when each item administered is the most informative for the current estimate of the examinee's ability level (i.e., a maximum-information rule is used to select items for administration). Under maximum-information item selection, items with more information at a given ability level will always be selected for administration over items with less information at the same ability level. If the differential selection rate is not controlled for, the higher information items are likely to be over-exposed, particularly in regions of the ability distribution with greater frequencies of examinees. Over-exposure is a particular concern at the beginning of the test, where all examinees are assumed to have equal abilities and could feasibly be administered the same initial sequence of items. As such, exposure control procedures are implemented to ensure that individual items are not over-exposed.

To reduce exposure of items in the new pools, exposure control parameters were computed for each item using the Sympson-Hetter probabilistic algorithm (Sympson & Hetter, 1985). The Sympson-Hetter algorithm was designed to reduce the predictability of item administration sequences and limit over-exposure of the most informative items. The algorithm controls for overall item use in such a way that the probability of an item being administered (and thereby exposed) to any examinee can be approximated to a pre-specified maximum value. Exposure control parameters are computed for each item within a pool using this algorithm. The resulting exposure control parameters are subsequently applied during CAT administration to help control item selection.

In computing exposure control parameters for the items in the new pools, a target maximum exposure rate of 2/3 was set across all the ASVAB tests. The 2/3 rate was selected based on the assumption that four of the new pools will be used simultaneously at proctored test sites, with the fifth reserved for internet administration. This yields an overall maximum exposure rate of 1/6 across pools:

$$\frac{2}{3}(rate) \times \frac{1}{4}(pools) = \frac{1}{6}, \tag{2.7}$$

which matches the rate of item exposure that occurs in administrations of the P&P-ASVAB.

Simulation methods were used to compute the item-level exposure control parameters. For each new pool, CAT-ASVAB administration was simulated using N = 2,000 examinees randomly sampled from a N(0,1) ability distribution. Exposure control parameters were computed based on the observed usage rates of the items. The CAT-ASVAB administrations were repeated using updated exposure control parameters until the maximum probability that an item was administered, given that it was selected, converged to the pre-specified 2/3 exposure rate. A complete description of the steps followed in the computation of the exposure control parameters is presented in Hetter and Sympson (1997) and ASVAB Technical Bulletin #1 (DMDC, 2006). The same procedures were followed here, with the exception of using a different maximum exposure rate.

Final exposure control parameters ranged in value between 0.67 and 1.0 across items. A value of 0.67 ensures that items projected to be administered to more than 2/3 of the population if no exposure control were used would ultimately be seen by no more than 2/3 of the population. A value of 1.0 ensures that items projected to be administered to less than 2/3 of the population if no exposure control were used would always be administered upon selection. Items assigned exposure control parameters values of less than 1.0 are likely to be highly discriminating items of moderate difficulty that would be selected in the early stages of CAT administration. The use of multiple pools provides an additional means of reducing item exposure above and beyond the 2/3 maximum exposure rate associated with items within each individual pool.

2.5.3.2. Item Parameters

Tables 2.27–2.36 summarize the item parameters for the five new pools for each ASVAB test. The tables present the average, standard deviation, minimum, and maximum values observed over all items in a pool.

Table 2.27. Summary of Item Parameters for New GS Pools

Parameter	Statistic	Form 5	Form 6	Form 7	Form 8	Form 9
	Ave	1.05	1.05	1.07	1.07	1.09
a	SD	0.28	0.29	0.30	0.32	0.34
	Min	0.52	0.54	0.52	0.55	0.56
	Max	1.76	2.01	1.83	2.27	2.34
	Ave	0.10	0.12	0.16	0.15	0.19
b	SD	1.29	1.27	1.14	1.28	1.22
	Min	–3.66	–2.51	–2.53	–3.25	–2.28
	Max	2.38	2.46	2.20	2.38	2.20
	Ave	0.20	0.18	0.20	0.20	0.20
c	SD	0.07	0.07	0.09	0.10	0.08
	Min	0.06	0.04	0.07	0.04	0.05
	Max	0.45	0.37	0.42	0.46	0.46

Table 2.28. Summary of Item Parameters for New AR Pools

Parameter	Statistic	Form 5	Form 6	Form 7	Form 8	Form 9
	Ave	1.33	1.33	1.31	1.33	1.29
a	SD	0.28	0.30	0.27	0.33	0.25
	Min	0.81	0.79	0.78	0.82	0.85
	Max	2.10	2.22	1.93	2.64	1.85
	Ave	0.01	–0.09	–0.08	–0.02	–0.06
b	SD	1.12	1.14	1.10	1.10	1.11
	Min	–2.59	–2.93	–2.98	–2.67	–2.77
	Max	1.91	1.86	2.14	2.13	1.83
	Ave	0.17	0.18	0.18	0.18	0.18
c	SD	0.08	0.08	0.07	0.08	0.08
	Min	0.03	0.04	0.05	0.04	0.05
	Max	0.48	0.47	0.37	0.49	0.42

Table 2.29. Summary of Item Parameters for New WK Pools

Parameter	Statistic	Form 5	Form 6	Form 7	Form 8	Form 9
	Ave	1.54	1.52	1.54	1.47	1.53
a	SD	0.36	0.40	0.43	0.37	0.39
	Min	0.92	0.87	0.85	0.79	0.89
	Max	2.44	2.71	2.94	2.48	2.41
	Ave	–0.15	–0.12	–0.15	–0.14	–0.20
b	SD	1.24	1.25	1.22	1.24	1.22
	Min	–2.61	–2.74	–2.51	–2.58	–2.61
	Max	1.91	2.23	1.93	2.14	1.98
	Ave	0.23	0.22	0.23	0.20	0.22
c	SD	0.08	0.08	0.08	0.07	0.09
	Min	0.03	0.06	0.07	0.08	0.05
	Max	0.43	0.44	0.48	0.37	0.50

Table 2.30. Summary of Item Parameters for New PC Pools

Parameter	Statistic	Form 5	Form 6	Form 7	Form 8	Form 9
	Ave	1.27	1.25	1.22	1.30	1.23
a	SD	0.26	0.26	0.24	0.26	0.23
	Min	0.83	0.79	0.73	0.85	0.73
	Max	1.84	2.09	1.77	1.92	1.65
	Ave	–0.38	–0.37	–0.36	–0.34	–0.38
b	SD	1.06	1.08	1.13	1.09	1.10
	Min	–2.20	–2.07	–2.47	–2.10	–2.56
	Max	1.83	1.93	1.51	1.59	1.50
	Ave	0.20	0.19	0.16	0.20	0.19
c	SD	0.05	0.06	0.05	0.06	0.06
	Min	0.09	0.08	0.08	0.08	0.09
	Max	0.36	0.35	0.31	0.39	0.44

Table 2.31. Summary of Item Parameters for New MK Pools

Parameter	Statistic	Form 5	Form 6	Form 7	Form 8	Form 9
	Ave	1.41	1.43	1.50	1.47	1.41
a	SD	0.46	0.46	0.58	0.55	0.45
	Min	0.72	0.74	0.74	0.76	0.77
	Max	3.37	2.93	3.75	4.48	2.96
	Ave	0.46	0.45	0.46	0.44	0.41
b	SD	0.98	0.90	0.96	0.93	0.95
	Min	–1.84	–1.81	–1.55	–1.85	–1.88
	Max	2.39	2.01	2.20	2.06	2.30
	Ave	0.16	0.16	0.18	0.17	0.17
c	SD	0.08	0.07	0.08	0.08	0.08
	Min	0.02	0.05	0.04	0.05	0.03
	Max	0.36	0.33	0.47	0.39	0.39

Table 2.32. Summary of Item Parameters for New EI Pools

Parameter	Statistic	Form 5	Form 6	Form 7	Form 8	Form 9
	Ave	1.17	1.11	1.15	1.25	1.16
a	SD	0.41	0.38	0.48	0.54	0.39
	Min	0.55	0.57	0.55	0.55	0.59
	Max	2.41	2.16	3.34	3.35	2.36
	Ave	0.01	–0.04	–0.03	–0.07	–0.03
b	SD	1.32	1.36	1.23	1.31	1.29
	Min	–3.03	–3.15	–2.68	–3.54	–2.87
	Max	2.07	2.35	2.08	2.24	2.13
	Ave	0.22	0.19	0.20	0.23	0.19
c	SD	0.07	0.07	0.08	0.09	0.07
	Min	0.06	0.06	0.06	0.04	0.04
	Max	0.40	0.35	0.39	0.50	0.41

Table 2.33. Summary of Item Parameters for New AI Pools

Parameter	Statistic	Form 5	Form 6	Form 7	Form 8	Form 9
	Ave	1.48	1.40	1.50	1.55	1.43
a	SD	0.55	0.44	0.54	0.54	0.43
	Min	0.72	0.64	0.67	0.84	0.66
	Max	3.61	2.49	3.03	3.70	2.39
	Ave	–0.06	–0.13	–0.05	–0.01	–0.13
b	SD	1.16	1.13	–0.05	1.09	1.17
	Min	–2.82	–2.39	–2.14	–2.58	–2.52
	Max	1.77	1.82	1.72	1.86	1.99
	Ave	0.19	0.19	0.20	0.18	0.17
c	SD	0.10	0.09	0.10	0.10	0.07
	Min	0.03	0.03	0.04	0.04	0.03
	Max	0.50	0.50	0.43	0.43	0.34

Table 2.34. Summary of Item Parameters for New SI Pools

Parameter	Statistic	Form 5	Form 6	Form 7	Form 8	Form 9
	Ave	1.22	1.24	1.18	1.18	1.19
a	SD	0.33	0.32	0.24	0.29	0.27
	Min	0.71	0.67	0.72	0.71	0.74
	Max	2.34	1.99	1.73	2.12	1.85
	Ave	0.05	–0.06	0.11	0.01	–0.09
b	SD	1.16	1.26	1.18	1.26	1.18
	Min	–2.00	–2.79	–2.38	–2.35	–2.92
	Max	2.32	2.14	2.32	2.48	1.96
	Ave	0.20	0.20	0.17	0.19	0.19
c	SD	0.08	0.08	0.09	0.09	0.08
	Min	0.05	0.08	0.03	0.03	0.08
	Max	0.41	0.50	0.45	0.40	0.50

Table 2.35. Summary of Item Parameters for New MC Pools

Parameter	Statistic	Form 5	Form 6	Form 7	Form 8	Form 9
	Ave	0.96	0.96	0.96	0.96	0.93
a	SD	0.23	0.23	0.24	0.21	0.23
	Min	0.62	0.61	0.54	0.58	0.53
	Max	1.90	1.64	1.77	1.56	1.75
	Ave	0.03	–0.03	–0.04	0.06	0.04
b	SD	1.15	1.29	1.31	1.22	1.22
	Min	–2.00	–2.62	–2.47	–2.52	–2.84
	Max	2.19	2.08	2.53	2.29	2.19
	Ave	0.19	0.20	0.20	0.20	0.18
c	SD	0.07	0.07	0.07	0.08	0.07
	Min	0.04	0.03	0.08	0.07	0.03
	Max	0.33	0.37	0.43	0.40	0.41

Table 2.36. Summary of Item Parameters for New AO Pools

Parameter	Statistic	Form 5	Form 6	Form 7	Form 8	Form 9
	Ave	1.23	1.22	1.21	1.22	1.21
a	SD	0.38	0.34	0.34	0.37	0.31
	Min	0.74	0.70	0.73	0.70	0.73
	Max	2.39	2.20	2.20	2.82	2.08
	Ave	–0.55	–0.55	–0.58	–0.57	–0.56
b	SD	0.70	0.68	0.71	0.65	0.73
	Min	–2.15	–1.96	–2.06	–2.07	–2.02
	Max	1.55	1.39	1.26	1.03	2.06
	Ave	0.23	0.23	0.22	0.21	0.22
c	SD	0.08	0.07	0.08	0.07	0.07
	Min	0.09	0.08	0.11	0.09	0.09
	Max	0.40	0.38	0.42	0.37	0.38

2.5.3.3. Item Enemies

Table 2.37 summarizes the distribution of item enemies across the five pools. Specifically, it indicates the frequency with which enemy groups of various sizes occur across the pools. For example, across the MK pools, there are seven occurrences of an enemy group containing three items, 34 occurrences of an enemy group containing two items, and 237 occurrences of an enemy group containing 1 item (i.e., the item had no enemies within the pool). The table shows that with the exception of MC and MK, most items within a pool had no enemies. MC had the biggest distribution of enemies, showing several occurrences of four-to-five items per enemy group within a pool.

Table 2.37. Summary of Enemy Groups Across the Five Pools

Test	# Items Per Enemy Group	Frequency
GS	1	311
	2	1
AR	1	344
	2	1
WK	1	371
PC	1	206
MK	1	237
	2	34
	3	7
EI	1	296
	2	2
AI	1	215
	2	1
SI	1	204
	2	4
MC	1	164
	2	48
	3	11
	4	2
	5	2
AO	1	334

2.5.3.4. Estimated Score Information Functions

Figures 2.6–2.15 show estimated score information functions for the five new pools for each ASVAB test. The score information functions were computed based on the responses of 2,000 simulated examinees at each of 31 equally spaced points between ± 3.0 (62,000 examinees total), using a smoothed approximation of Lord's (1980) Equation 10-7 as presented in Segall, Moreno, and Hetter (1997) and ASVAB Technical Bulletin #1 (DMDC, 2006). In the simulation, CAT-ASVAB item selection and scoring algorithms were applied to the item parameters and exposure control parameters. The figures suggest that for each test, examinees throughout the ability distribution are likely to be measured at a similar level of precision across the five new pools.

Figure 2.6. Estimated Score Information Functions for the Five New GS Pools

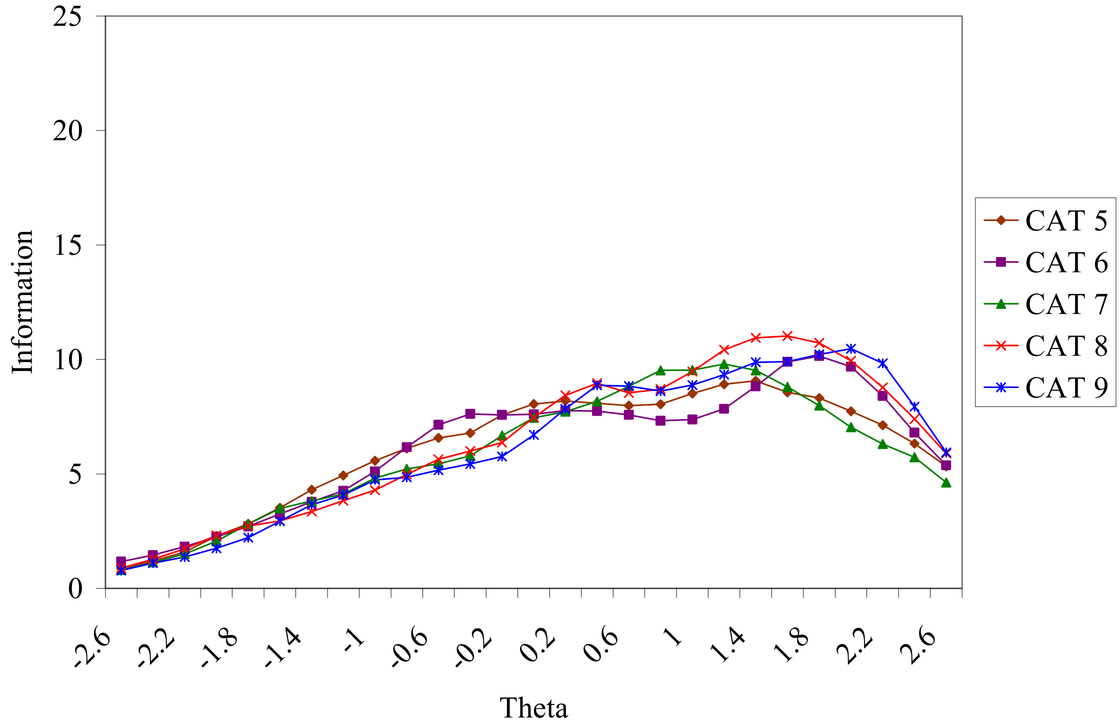


Figure 2.7. Estimated Score Information Functions for the Five New AR Pools

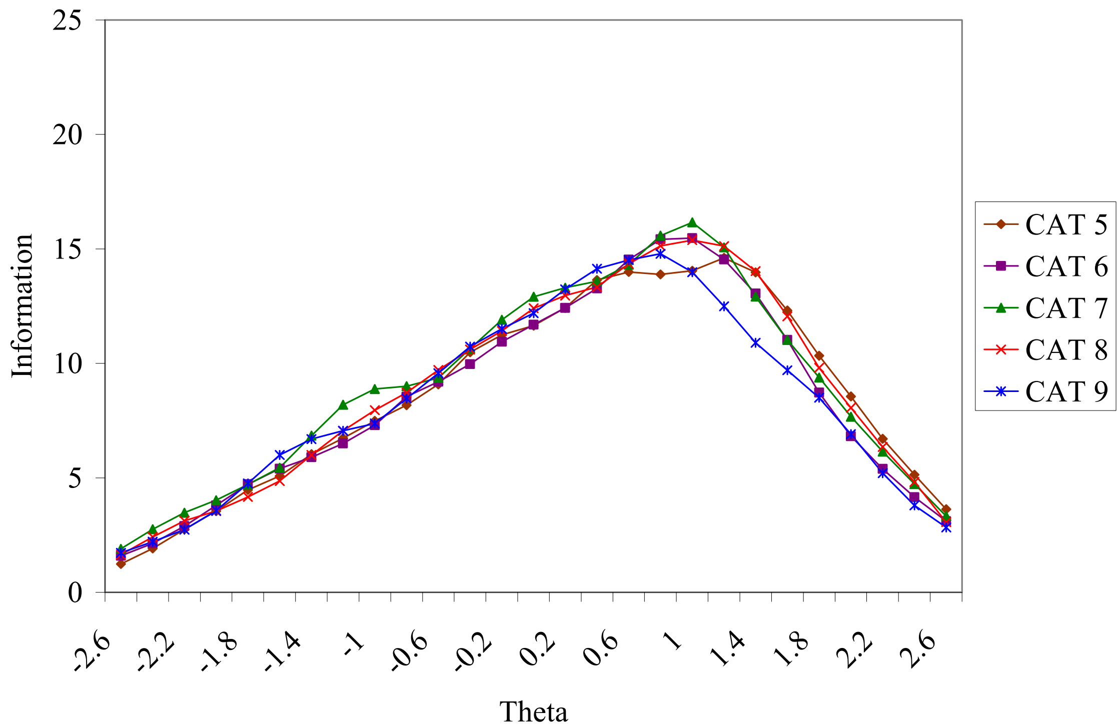


Figure 2.8. Estimated Score Information Functions for the Five New WK Pools

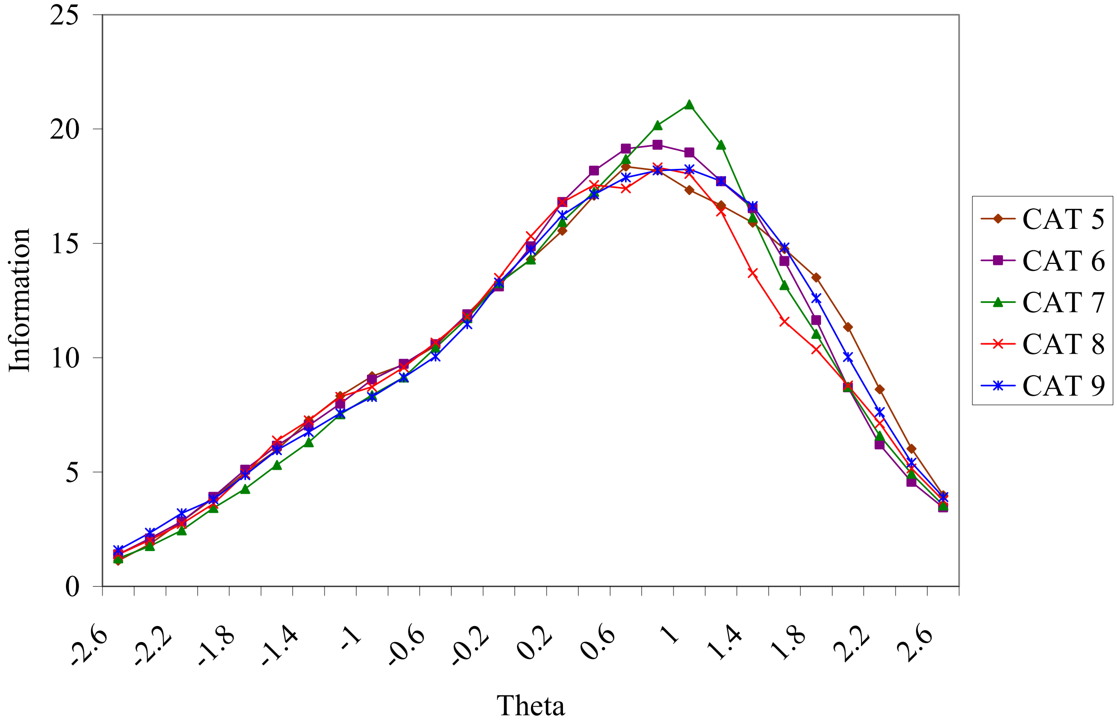


Figure 2.9. Estimated Score Information Functions for the Five New PC Pools

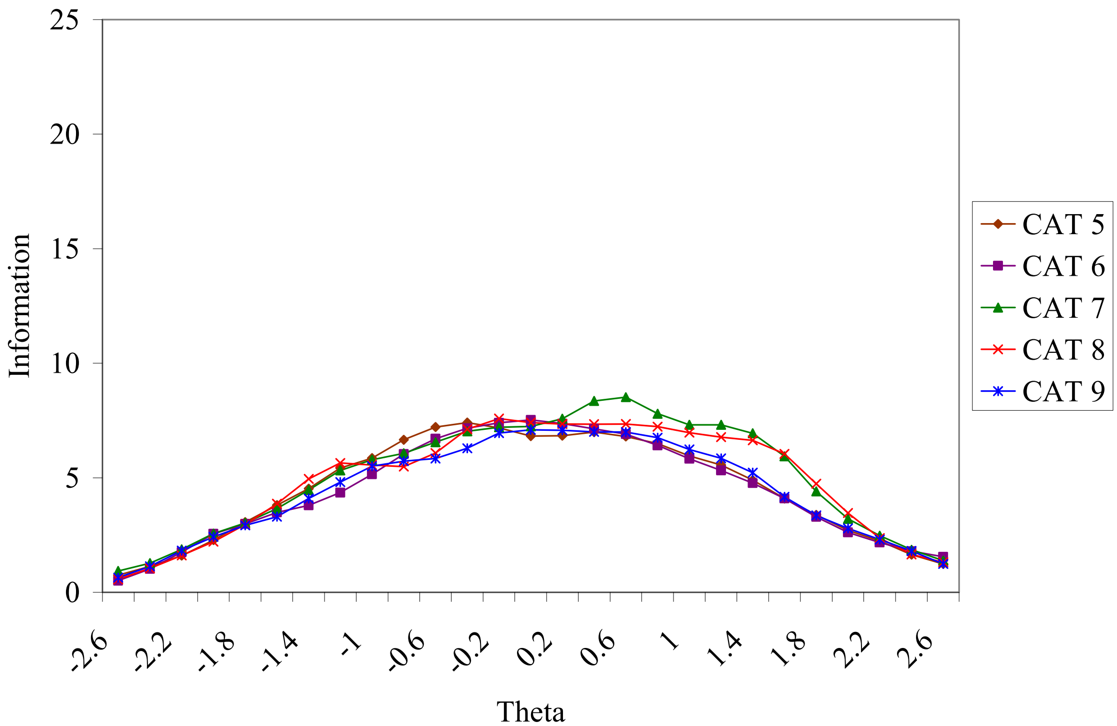


Figure 2.10. Estimated Score Information Functions for the Five New MK Pools

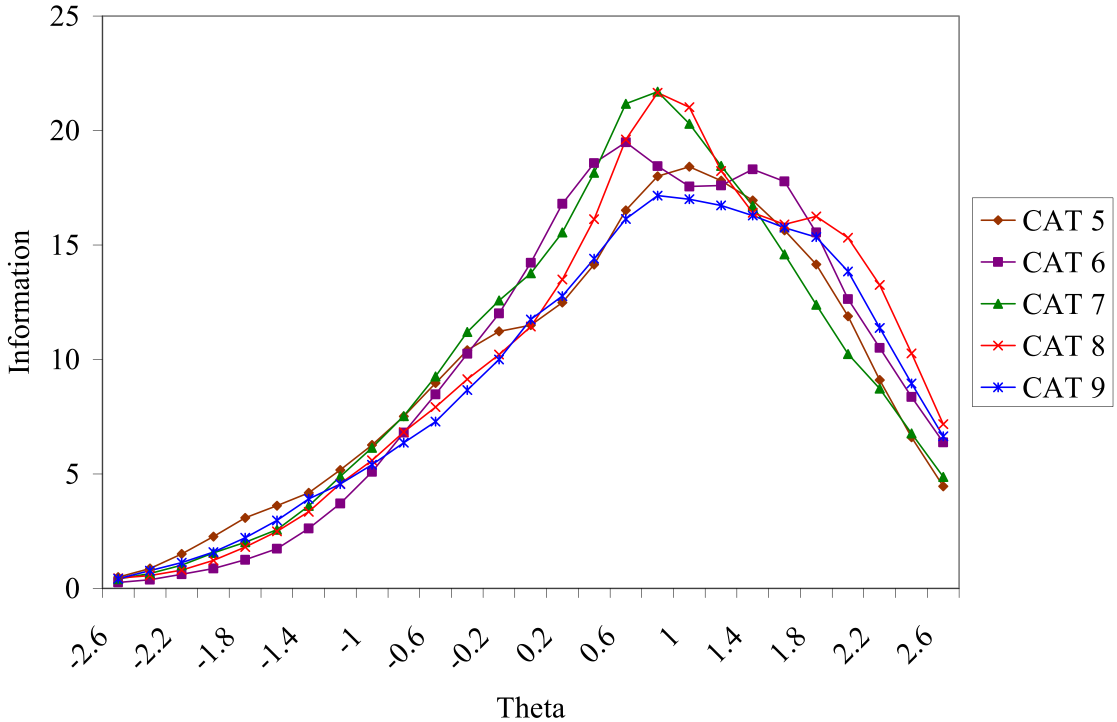


Figure 2.11. Estimated Score Information Functions for the Five New EI Pools

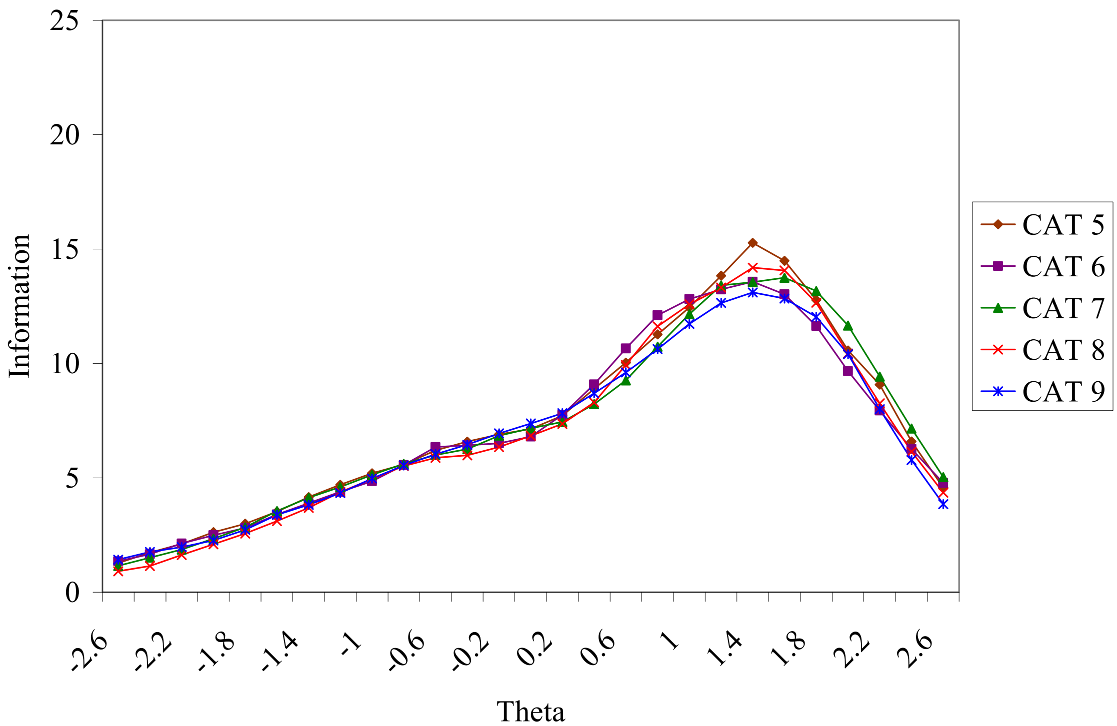


Figure 2.12. Estimated Score Information Functions for the Five New AI Pools

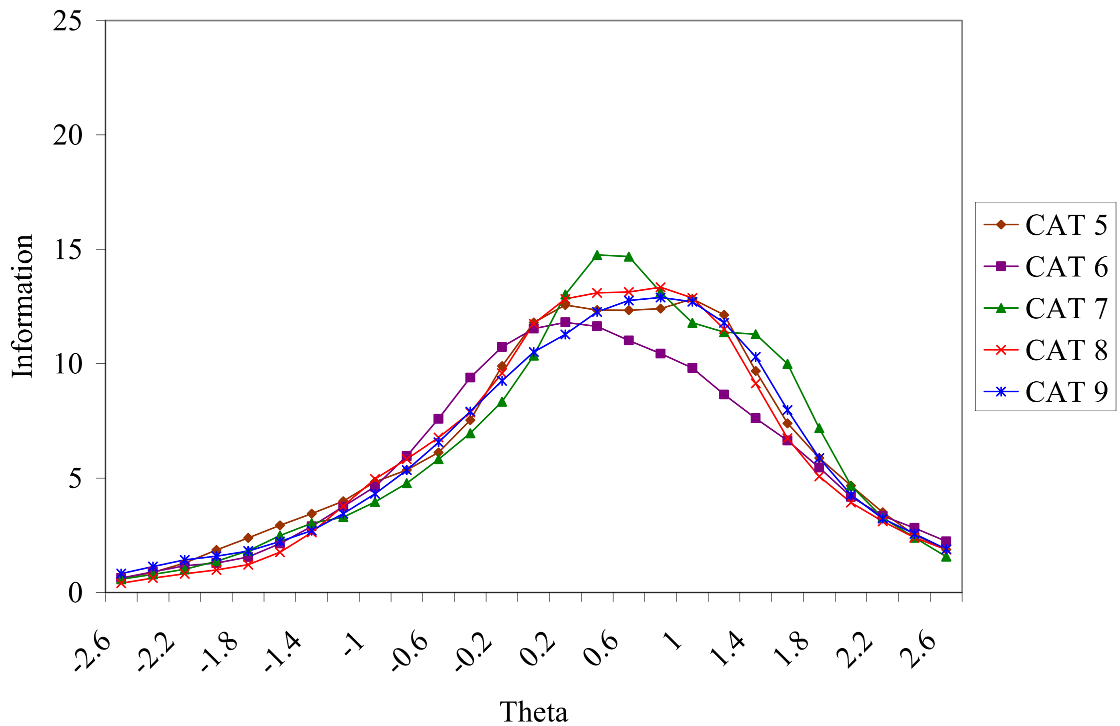


Figure 2.13. Estimated Score Information Functions for the Five New SI Pools

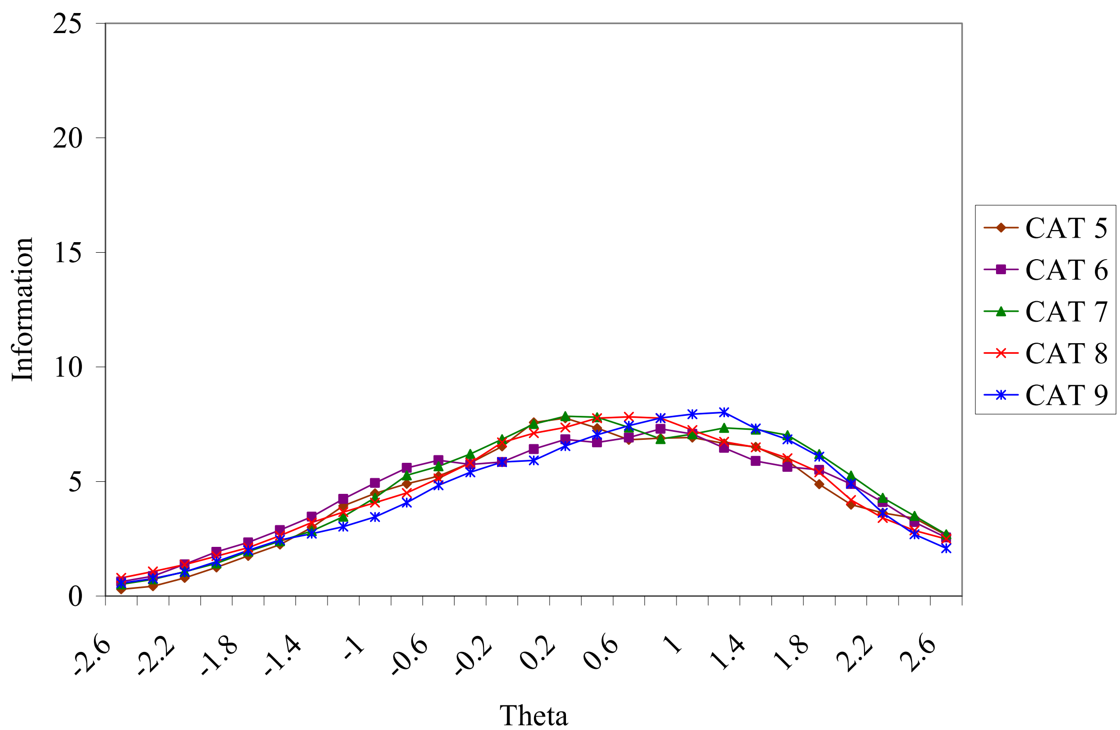


Figure 2.14. Estimated Score Information Functions for the Five New MC Pools

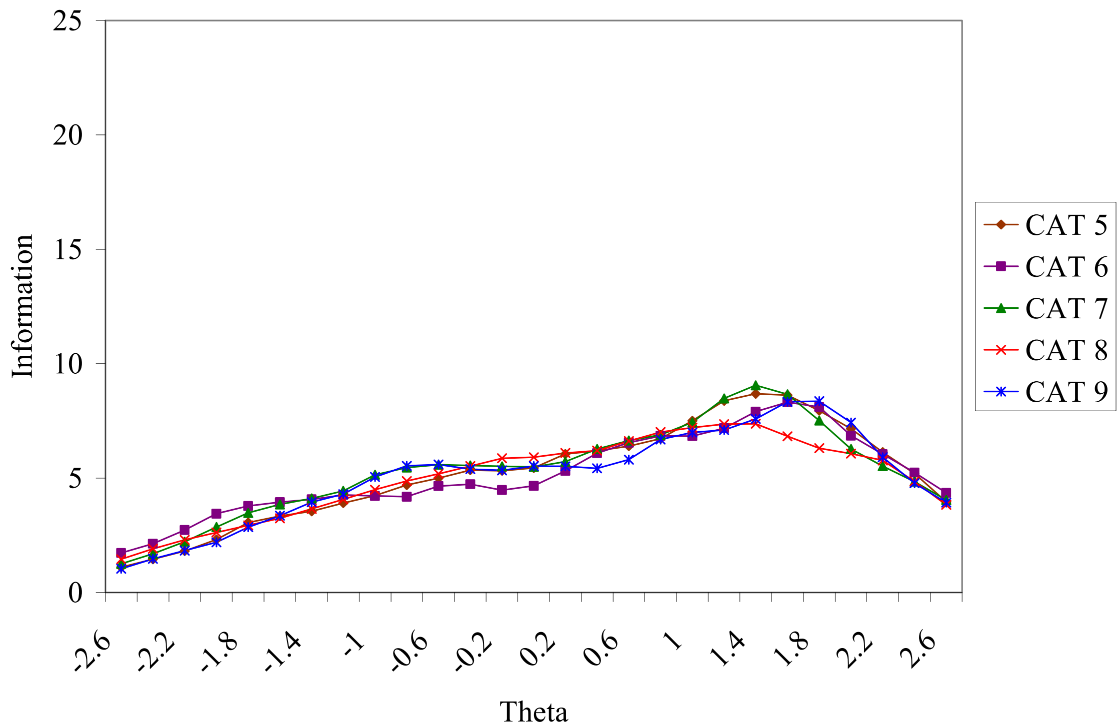
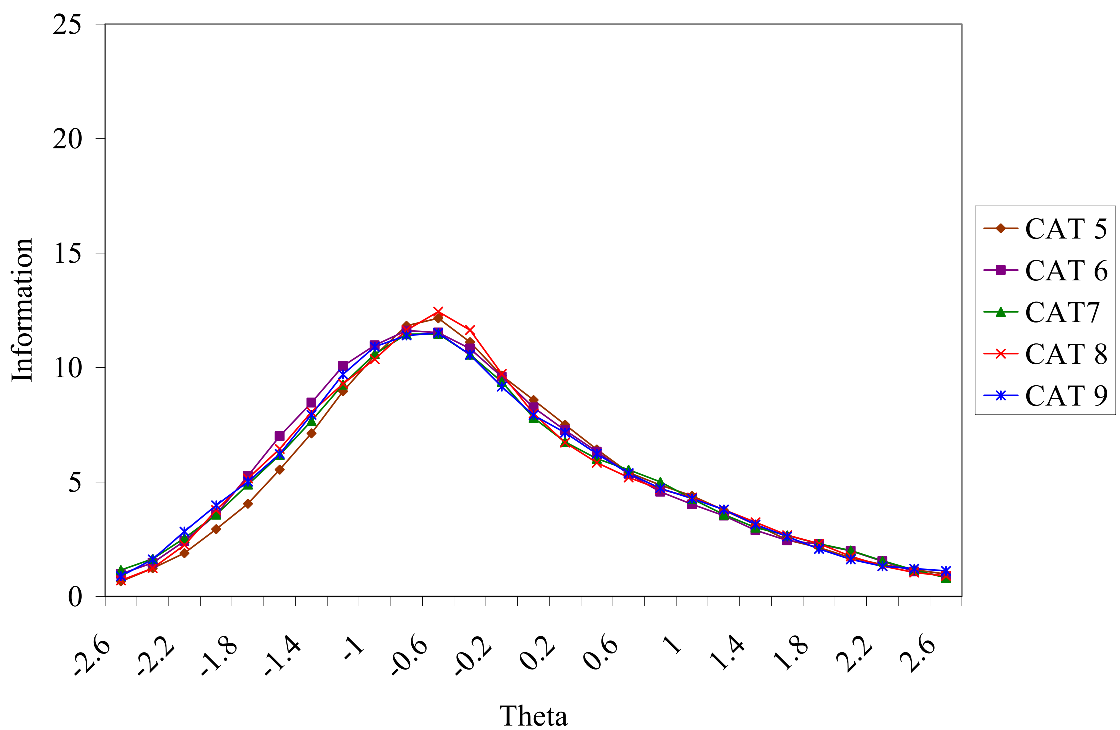


Figure 2.15. Estimated Score Information Functions for the Five New AO Pools



Figures 2.16–2.25 summarize for each ASVAB test the estimated score information functions for the five new pools relative to estimated score information functions for CAT-ASVAB Forms 1–4. Score information for CAT-ASVAB Forms 1–4 was computed using the same simulation methods as discussed for the new pools. In the figures, the estimated score information functions are averaged over the five new pools (labeled CAT 5–9) and averaged over Forms 1–4 (labeled CAT 1–4). Figures 2.16–2.24 also show score information functions for P&P-ASVAB Form 09A, computed from number right scores. Figure 2.25 excludes results for Form 09A because the AO test is not contained in Form 09A.

Figure 2.16. Comparison of GS Score Information Functions Across CAT Pools 5–9 (Averaged), CAT Pools 1–4 (Averaged), and P&P Form 09A

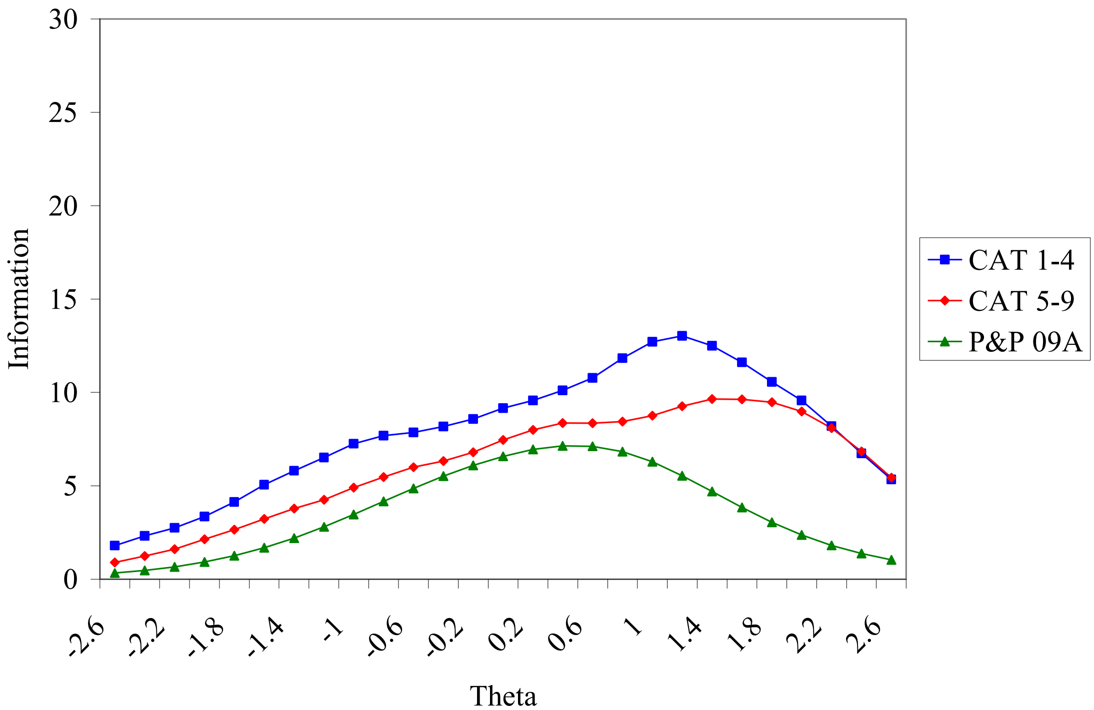


Figure 2.17. Comparison of AR Score Information Functions Across CAT Pools 5–9 (Averaged), CAT Pools 1–4 (Averaged), and P&P Form 09A

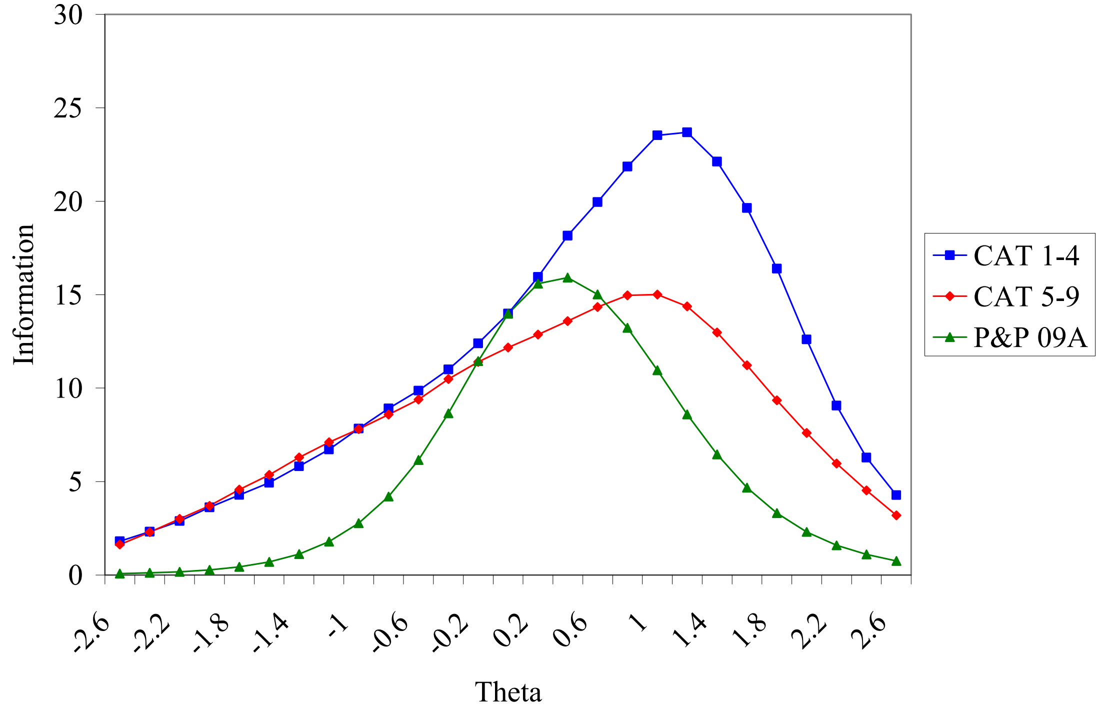


Figure 2.18. Comparison of WK Score Information Functions Across CAT Pools 5–9 (Averaged), CAT Pools 1–4 (Averaged), and P&P Form 09A

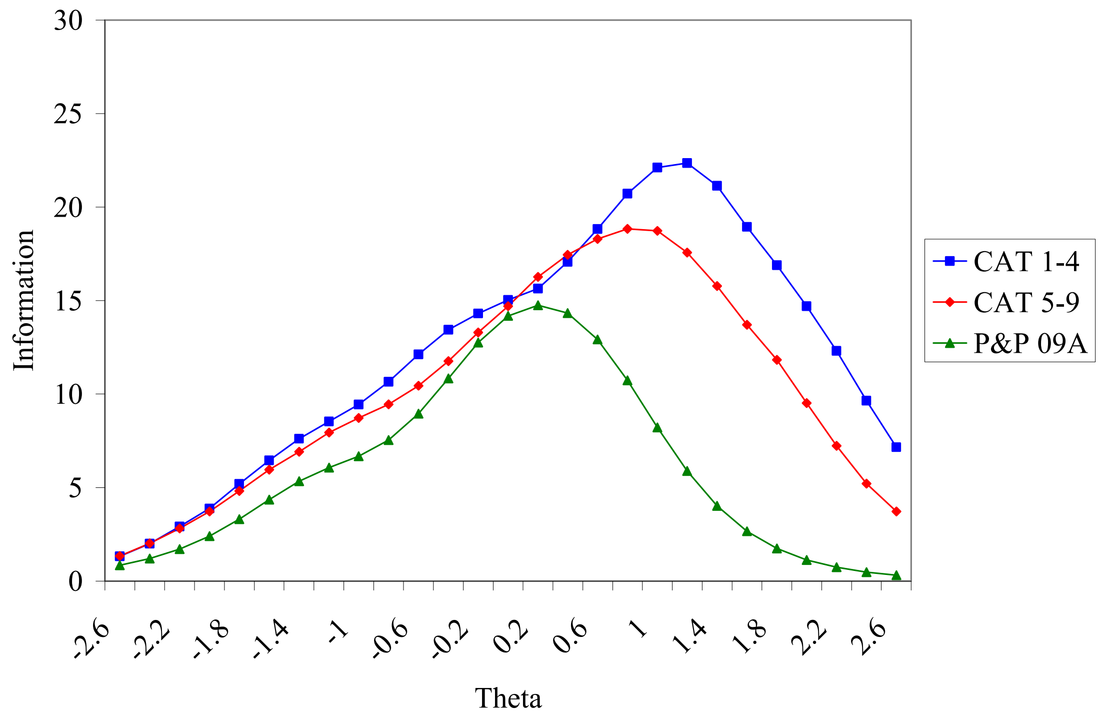


Figure 2.19. Comparison of PC Score Information Functions Across CAT Pools 5–9 (Averaged), CAT Pools 1–4 (Averaged), and P&P Form 09A

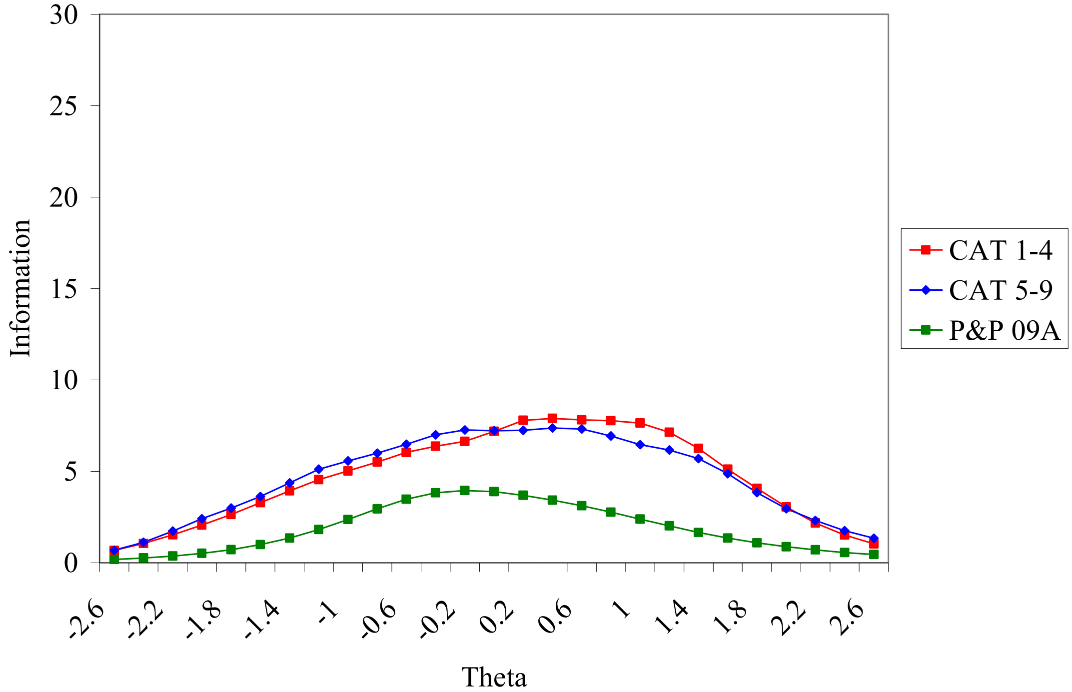


Figure 2.20. Comparison of MK Score Information Functions Across CAT Pools 5–9 (Averaged), CAT Pools 1–4 (Averaged), and P&P Form 09A

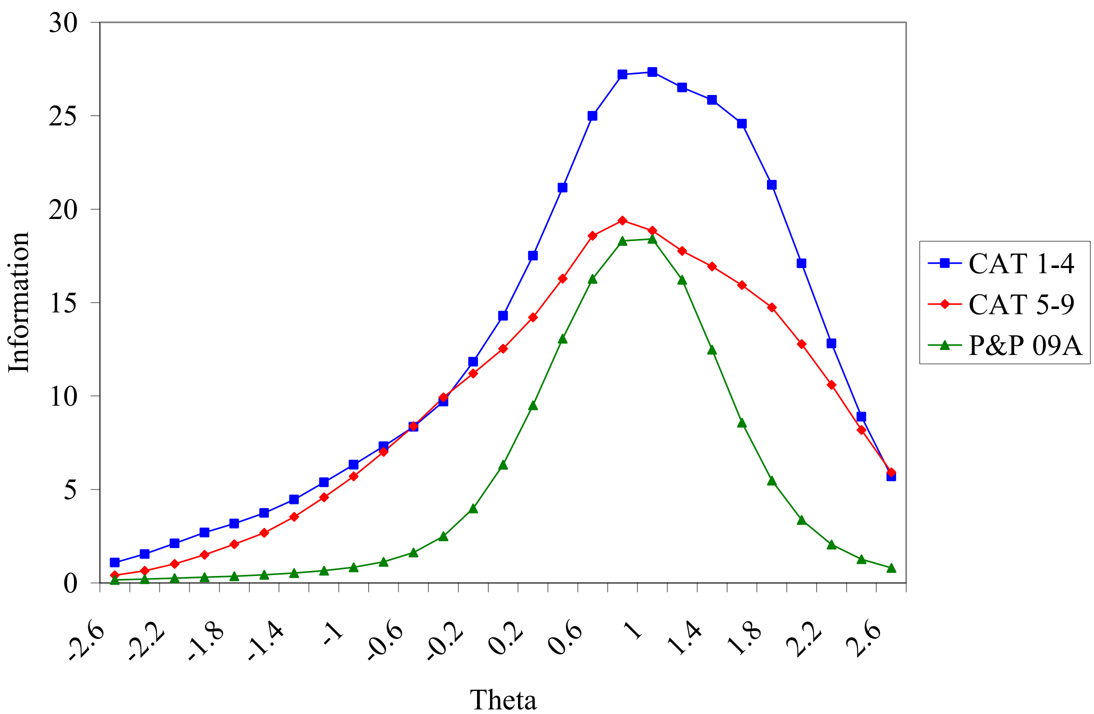


Figure 2.21. Comparison of EI Score Information Functions Across CAT Pools 5–9 (Averaged), CAT Pools 1–4 (Averaged), and P&P Form 09A

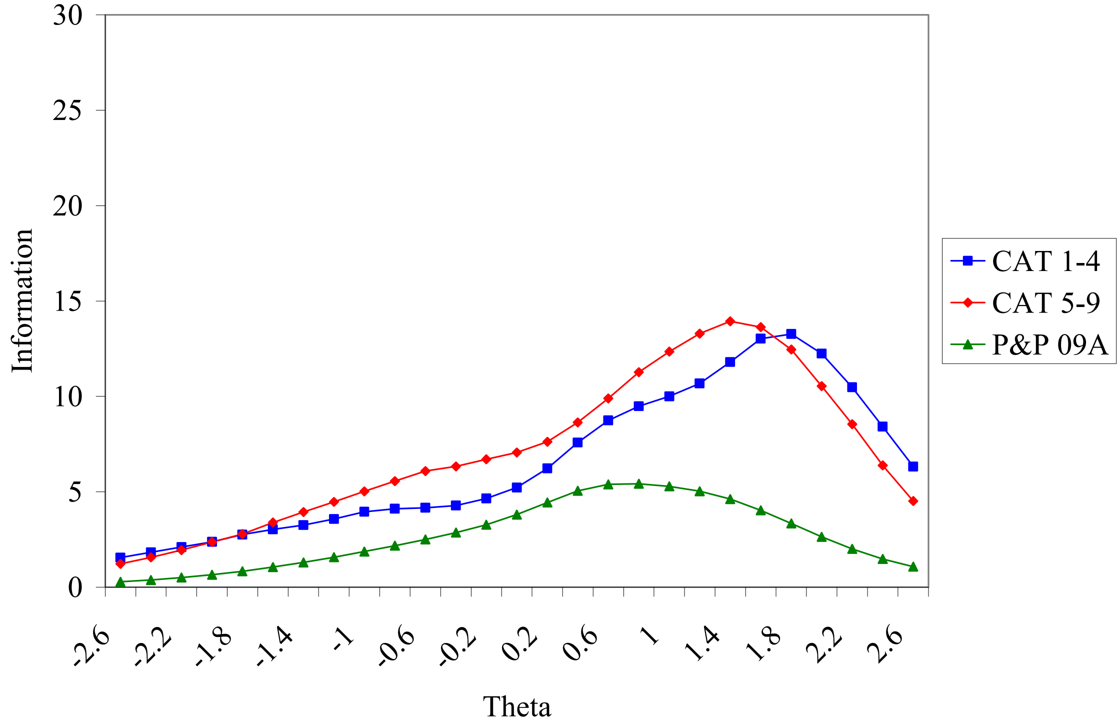


Figure 2.22. Comparison of AI Score Information Functions Across CAT Pools 5–9 (Averaged), CAT Pools 1–4 (Averaged), and P&P Form 09A

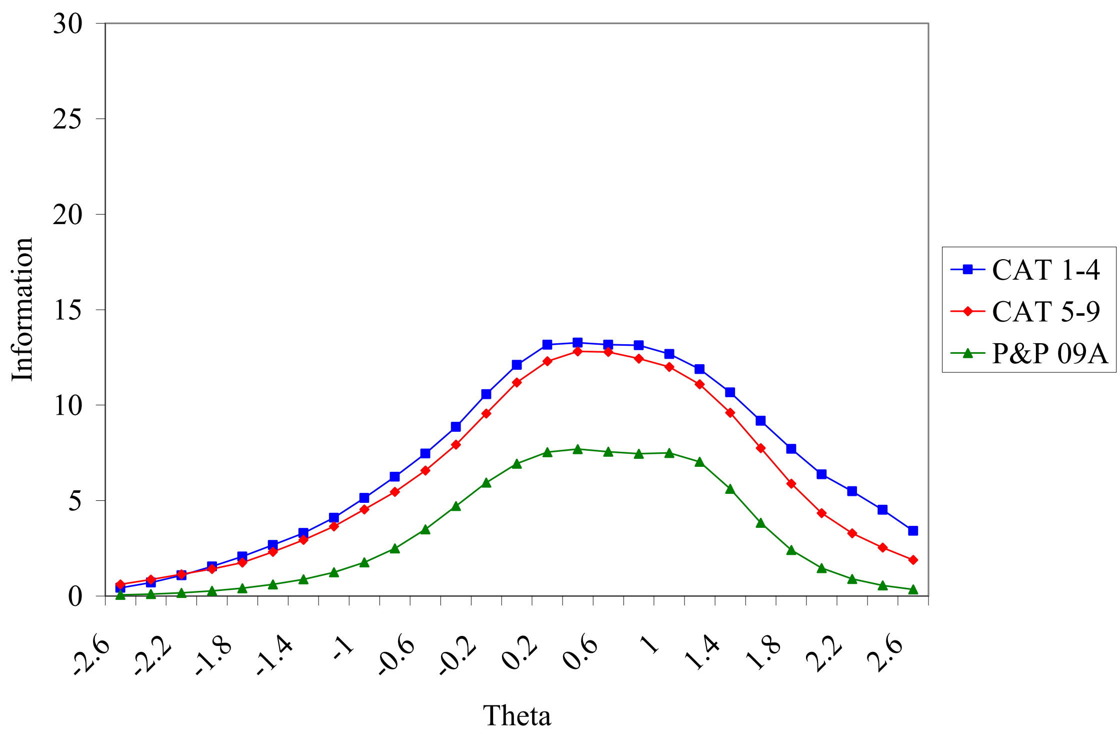


Figure 2.23. Comparison of SI Score Information Functions Across CAT Pools 5–9 (Averaged), CAT Pools 1–4 (Averaged), and P&P Form 09A

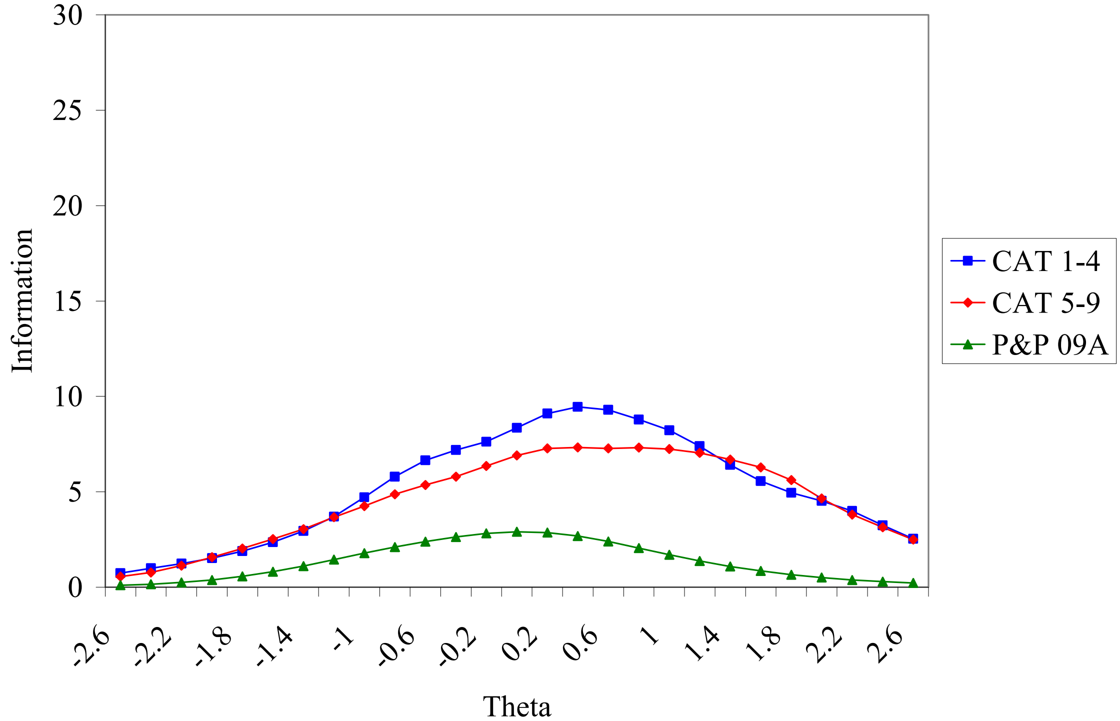


Figure 2.24. Comparison of MC Score Information Functions Across CAT Pools 5–9 (Averaged), CAT Pools 1–4 (Averaged), and P&P Form 09A

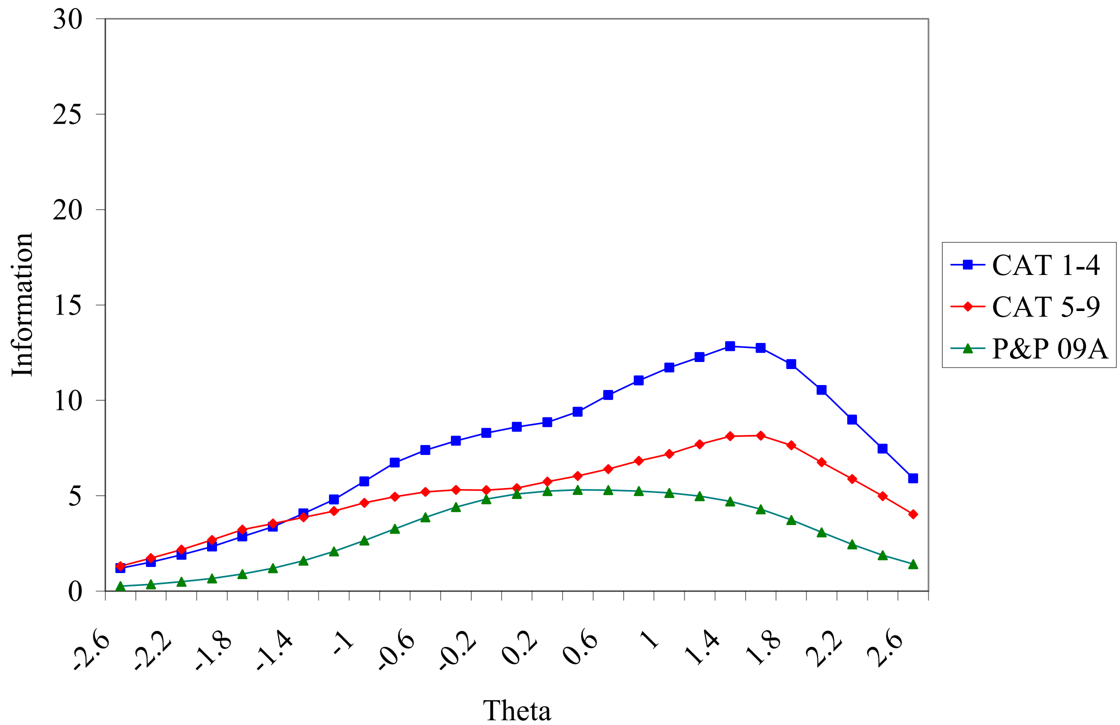
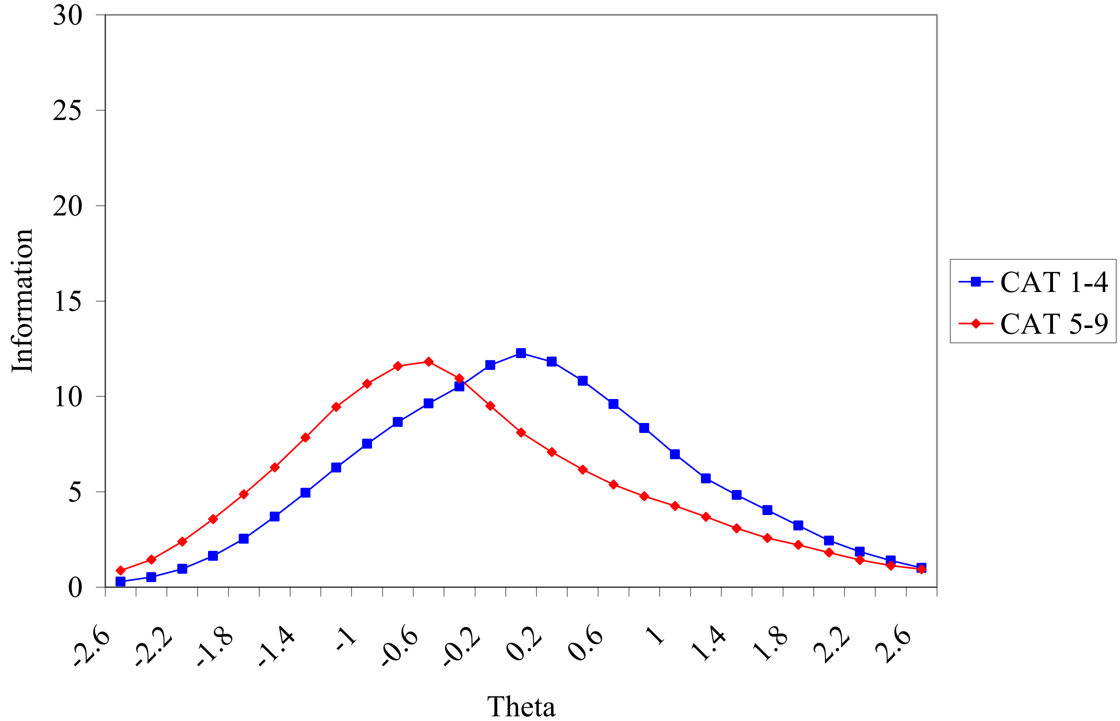


Figure 2.25. Comparison of AO Score Information Functions Across CAT Pools 5–9 (Averaged) and CAT Pools 1–4 (Averaged)



Construction of fewer than five new pools would have resulted in more score precision than observed with CAT-ASVAB Forms 5–9. However, the creation of five new pools enables the implementation of CAT-ASVAB concepts that have emerged in recent years (such as an internet administered practice CAT-ASVAB), while simultaneously reducing compromise threats through the use of more pools. Figures 2.16–2.24 also suggest that the precision of scores from the five new pools is likely to equal or exceed the precision of scores from P&P-ASVAB forms, as the average information across the five new pools exceeded P&P Form 09A information throughout the ability distribution for most ASVAB tests.

2.5.3.5. Test-Retest Reliabilities

Test-retest reliabilities for the new pools were estimated by correlating theta scores across two simulated CAT sessions for 2,000 examinees sampled from a N(0,1) distribution. Table 2.38 summarizes the test-retest reliability estimates for each of the five new CAT pools. Average test-retest reliability estimates over the five new pools are also provided (labeled CAT 5–9), along with average test-retest reliability estimates for CAT-ASVAB Forms 1–4 (labeled CAT 1–4) and test-retest reliability estimates for an operational P&P form (labeled P&P). Table 2.38 shows that the reliabilities of the new CAT pools meet or exceed the reliability of the P&P form for all ASVAB tests, while falling only slightly below the reliability of CAT-ASVAB Forms 1–4.

Table 2.38. Test-Retest Reliability Estimates

Test	CAT 5	CAT 6	CAT 7	CAT 8	CAT 9	CAT 5–9[a]	CAT 1–4[b]	P&P
GS	0.87	0.86	0.86	0.87	0.87	0.87	0.89	0.82
AR	0.91	0.91	0.91	0.91	0.91	0.91	0.92	0.89
WK	0.92	0.93	0.92	0.93	0.93	0.92	0.93	0.90
PC	0.86	0.86	0.87	0.86	0.84	0.86	0.86	0.74
MK	0.91	0.91	0.91	0.91	0.90	0.91	0.92	0.87
AI	0.88	0.88	0.88	0.88	0.89	0.88	0.89	0.81
SI	0.85	0.84	0.85	0.85	0.84	0.85	0.86	0.67
MC	0.85	0.83	0.85	0.85	0.85	0.85	0.88	0.79
EI	0.88	0.88	0.88	0.88	0.88	0.88	0.85	0.75
AO	0.87	0.88	0.87	0.88	0.86	0.87	0.89	0.87

[a]Averaged across CAT-ASVAB Forms 5–9
[b]Averaged across CAT-ASVAB Forms 1–4

3. Equating of CAT-ASVAB Forms 5–9

Although the calibration and scaling procedures were designed to ensure that theta scores based on CAT-ASVAB Forms 5–9 were on the same scale as CAT-ASVAB Forms 1–4, theta scores based on the new pools were equated to theta scores based on the reference pool (Form 4) as an extra precaution. Equating ensures that reported scores can be treated interchangeably across the different CAT-ASVAB pools; namely, as if they had come from the same CAT pool and had the same meaning regardless of which pool was actually used during administration. Linear equating methods were used to derive constants to transform theta scores from the new pools to the scale of the reference pool. The linear equating procedures ensure that theta scores have the same mean and variance across the reference pool and the new pools. Unique transformations were estimated for each ASVAB test for each new pool.

Data collection for the equating was conducted in three phases of operational administration of Forms 5–9 to military applicants. Table 3.1 describes the testing dates and resulting sample sizes for each phase. The equating was conducted in phases to maximize the accuracy of the reported operational scores (i.e., equated Form 5–9 scores). Sample sizes increased with each phase. In the first phase of the data collection, provisional score transformations were provided based on the IRT invariance assumption that theta scores were on the same scale across the different pools. In the second and third phases of the data collection, the reported operational scores were based on provisional score transformations computed from an equating conducted in the previous phase. Upon completion of the data collection, final equating transformations were developed and applied to all subsequent examinees.

Table 3.1. Testing Dates and Final Sample Sizes for the Equating Study

Phase	Start Date	End Date	Sample Size
I	June 11, 2007	June 26, 2007	2,091
II	July 23, 2007	August 17, 2007	5, 853
III	September 17, 2007	January 31, 2008	103,438

A random groups design was used in all three phases of the equating study. Each examinee was randomly assigned to one of the following CAT-ASVAB pools: Form 1, Form 4, Form 5, Form 6, Form 7, Form 8, Form 9 with regular time limits (9R), or Form 9 with lengthened time limits (9L). The lengthened time administration of Form 9 was intended to more closely resemble the condition of administration with no time limits such as would occur with internet administration of CAT-ASVAB. It was included so that the effect of time limits on test scores could be evaluated, as longer time limits might result in increased scores on some ASVAB tests. Only scores from applicants testing for the first time were included in the equating and time limits analyses. Each step in the study is described in more detail below.

3.1. Data Collection and Provisional Score Transformations

3.1.1. Phase I

Table 3.2 shows the assignment probabilities used in Phase I of the data collection. (Note that the same assignment probabilities were also used in Phase II of the data collection.) All scores from new pools administered under the regular time limits were equated to the scale of Form 4, the reference pool. Form 1 was included as a double check to evaluate whether Form 4 was an accurate representative of operational performance on CAT-ASVAB, since Form 4 had not been used operationally for some time. A 5/16 assignment probability was selected for the operational and reference pools so that their usage rates would equal that of the five new pools with regular time limits. This enabled a pooled equating to be conducted over the new pools (i.e., simultaneously equating all of the new pools to the reference pool) for use with Phase II applicants. A 1/16 assignment probability was also selected for Form 9L so that it had the same assignment probability as Form 9R. Assignment to a pool was made by computer, using a pseudo random number generator.

Table 3.2. Assignment Probabilities for Phases I and II

Form #	Description	Assignment Probability
1	Operational	5/16
4	Reference	5/16
5	New	1/16
6	New	1/16
7	New	1/16
8	New	1/16
9R	New, regular time	1/16
9L	New, lengthened time	1/16

Table 3.3 shows the targeted sample sizes for Phase I, along with the actual sample sizes that were obtained. The actual sample sizes exceeded the target sample sizes for all pools.

Table 3.3. Targeted and Actual Sample Sizes for Phase I

Form	Target	Actual
1	300	645
4	300	678
5, 6, 7, 8, 9R, 9L	360	768
Total	960	2,091

Provisional score transformations were provided for examinees testing during the Phase I data collection. They were based on the IRT assumption that ability estimates are invariant across items and pools; namely, if the item parameters are on the same scale, an examinee should get the same ability estimate regardless of the set of items or pool administered. Because the parameters for the new pools were calibrated to be on the same scale as the parameters for the operational pools, theta scores computed from the new pools were assumed to be on the same scale as theta scores computed from Form 4. Thus, theta scores based on the new pools were converted to a standard score using existing theta to standard score transformations for Form 4. The accuracy of the Phase I provisional equating transformations was evaluated after the final score transformations were computed and is discussed in Section 3.4 below.

3.1.2. Phase II

Table 3.2 above shows the assignment probabilities used in Phase II of the data collection. (Note that the same assignment probabilities were used in Phase I of the data collection.) Table 3.4 shows the targeted sample sizes for Phase II, along with the actual sample sizes that were obtained. The actual sample sizes exceeded the target sample sizes for all pools.

Table 3.4. Targeted and Actual Sample Sizes for Phase II

Form	Target	Actual
1	1,000	1,828
4	1,000	1,849
5, 6, 7, 8, 9R, 9L	1,200	2,176
Total	3,200	5,853

Provisional score transformations were applied to applicants testing during the Phase II data collection. They were based on an equating conducted using Phase I data, where linear methods were used to derive constants to transform theta scores from the new pools to the scale of the reference pool. Scores from Forms 5, 6, 7, 8, and 9R were pooled together to conduct the equating since within-pool sample sizes were insufficient to equate separately by pool. Unique transformations were estimated for each ASVAB test and applied to all Phase II applicants.

The procedures used to equate the pools and compute the provisional score transformations are outlined in more detail in Section 3.2 below. (Note that although Section 3.2 discusses final score transformations, the procedures described there were also used to compute the provisional score transformations for scores pooled across Forms 5–9R.) The provisional score transformation functions showed little difference from the transformation functions used operationally with CAT-ASVAB Forms 1–3. The accuracy of the Phase II provisional equating transformations was evaluated after the final score transformations were computed and is discussed in Section 3.4 below.

3.1.3. Phase III

Table 3.5 shows the assignment probabilities used in Phase III of the equating study. A 1/8 assignment probability was selected for all pools, which enabled (a) separate equatings to be conducted linking scores on each new pool to the reference form, (b) scores to be compared across the regular and lengthened time conditions for Form 9, and (c) scores to be compared across the reference and operational form.

Table 3.5. Assignment Probabilities for Phase III

Form #	Description	Assignment Probability
1	Operational	1/8
4	Reference	1/8
5	New	1/8
6	New	1/8
7	New	1/8
8	New	1/8
9R	New, regular time	1/8
9L	New, lengthened time	1/8

Table 3.6 shows the targeted and actual sample sizes for Phase III. For each pool, the target sample size was exceeded by about N = 3,000. Assignment to a pool was made by computer, using a pseudo random number generator. To check the hypothesis that the sample was uniformly distributed across the different pools, a χ^2 goodness of fit test was conducted. The results indicated that the within-pool sample sizes were consistent with expectation: $\chi^2(7) = 8.60$, $p = 0.28$.

Table 3.6. Targeted and Actual Sample Sizes for Phase III

Form	Target	Actual
1	10,000	12,932
4	10,000	12,838
5	10,000	13,104
6	10,000	12,811
7	10,000	12,966
8	10,000	13,123
9R	10,000	12,848
9L	10,000	12,816
Total	80,000	103,438

To check the assumption of the random equivalence of the groups, χ^2 tests of independence were conducted to evaluate the relationship between analysis group (i.e., pool) and available demographic variables (gender, education, and race/ethnicity). Results for gender and education were non-significant, which suggested that the distributions of gender and education were similar across the analysis groups. Results for race/ethnicity were significant at $\alpha = .05$: $\chi^2(24, N = 88{,}808) = 38.016$, $p = 0.035$. Table 3.7 shows the frequency distribution table for race/ethnicity by pool. The distribution of race/ethnicity appears to be very similar across the different pools. Because the χ^2 statistic is known to be sensitive to sample size, the significant finding for race/ethnicity may be attributable to the very large sample size used in the analysis. Overall, results of the χ^2 tests and distributional evaluation suggest that the groups can be considered to be randomly equivalent.

Table 3.7. Frequency Distribution of Race/Ethnicity by Pool for Phase III Data

	Pool							
Race	1	4	5	6	7	8	9[a]	Total
Asian	2.9	2.6	2.4	2.7	2.5	3.0	2.8	2.7
Black	12.8	13.1	13.2	13.0	14.4	13.5	13.7	13.4
White	58.1	58.7	58.2	57.8	57.9	58.0	57.8	58.0
Hispanic	16.6	16.6	16.8	16.5	16.3	16.6	16.7	16.6
Other	9.6	9.1	9.3	10.0	9.0	9.0	9.1	9.3
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

[a]Includes Form 9R and Form 9L data combined.

Provisional score transformations were applied to applicants testing during the Phase III data collection. They were based on an equating conducted using Phase II data, where linear equating methods were used to derive constants to transform theta scores from the new pools to the scale of the reference pool. Scores from Forms 5, 6, 7, 8, and 9R were pooled together to conduct the equating, since within-pool sample sizes were insufficient to equate separately by pool. The procedures described in Section 3.2 below were used to compute the provisional score transformations for scores pooled across Forms 5–9R. Unique transformations were estimated for each ASVAB test and applied to all applicants. The provisional score transformation functions showed little difference from the transformation functions used operationally with CAT-ASVAB Forms 1–3. The accuracy of the Phase III provisional equatings was evaluated after the final score transformations were computed and is discussed in Section 3.4 below.

3.2. Final Score Transformations

Data from Phase III were used to compute final score transformations for the new pools. In linear equating, scores on one test form (Form X) are converted to the scale of another test form (Form Y) such that the Form X converted scores have the same mean and variance as the Form Y scores. The linear equation for converting CAT scores on Form X to the scale of Form Y operates on the means and standard deviations of the ability estimates on Form X [$\mu(\hat{\theta}_x), \sigma(\hat{\theta}_x)$] and Form Y [$\mu(\hat{\theta}_y), \sigma(\hat{\theta}_y)$] and is given as

$$\hat{\theta}_{x(y)} = B + A\hat{\theta}_x, \quad (3.1)$$

where

$$A = \frac{\sigma(\hat{\theta}_y)}{\sigma(\hat{\theta}_x)} \quad \text{and} \quad B = \mu(\hat{\theta}_y) - A\mu(\hat{\theta}_x). \quad (3.2)$$

A more detailed discussion of linear equating is available in Kolen & Brennan (2004).

For each ASVAB test within each new pool, Equations 3.1 and 3.2 were used to transform ability estimates to the scale of Form 4 ability estimates. For example, GS theta scores for Form 5 were converted to the scale of GS theta scores for Form 4 (denoted as $\hat{\theta}_{GS5(4)}$) by first computing the *A* and *B* transformation constants:

$$A_{GS5} = \frac{\sigma(\hat{\theta}_{GS4})}{\sigma(\hat{\theta}_{GS5})} \quad \text{and} \quad B_{GS5} = \mu(\hat{\theta}_{GS4}) - A_{GS5}\mu(\hat{\theta}_{GS5}) \quad (3.3)$$

and then applying the transformation constants to the linear equation:

$$\hat{\theta}_{GS5(4)} = B_{GS5} + A_{GS5}\hat{\theta}_{GS5}. \quad (3.4)$$

In a likewise manner, separate conversions were obtained for each of the other ASVAB tests (AR, WK, PC, MK, EI, AI, SI, MC, and AO), converting Form 5 ability estimates to the scale of Form 4. Form 6 ability estimates for each ASVAB test were converted to the scale of Form 4 using a similar process, as were Forms 7, 8, and 9R ability estimates. Ability estimates for applicants taking Form 9L were transformed using the Form 9R transformation constants.

All of the final transformed ability estimates were then converted to standard scores using transformation constants (α and β) derived from the reference form (Form 4).[3] For example, GS Form 5 transformed ability estimates were converted to standard scores using the following equation:

$$SS_{GS5} = \alpha_{GS} + \beta_{GS}\hat{\theta}_{GS5(4)}. \quad (3.5)$$

For ease of computation, the two linear transformations of ability estimates (i.e., from the equating and standard score conversions) were combined into a single linear

[3] Scores on all ASVAB tests are reported as standard scores, i.e., scores that have a fixed mean and standard deviation in the population of examinees. A standard score indicates how many units of the standard deviation a particular score is above or below the mean. In the case of the ASVAB tests, the mean is set to 50 and the standard deviation is set to 10.

transformation during the actual implementation. For example, for GS Form 5, the final reported score was computed by combining Equations 3.4 and 3.5 into Equation 3.6:

$$SS_{GS5} = \alpha^*_{GS5} + \beta^*_{GS5}\hat{\theta}_{GS5}, \tag{3.6}$$

where α^*_{GS5} and β^*_{GS5} were functions of A_{GS5}, B_{GS5}, α_{GS}, and β_{GS}. In a likewise manner, standard score conversions were obtained for all new pools within each of the ASVAB tests. A single linear transformation, as described here, was also used operationally with CAT-ASVAB Forms 1–3 to directly convert ability estimates that were on the scale of the administered pool to standard scores that were on the scale of the reference pool.

3.3. Pool Equivalence

Following the computation of the final score transformations, the transformations were applied to Phase III examinees and used to evaluate whether the transformed scores could be treated interchangeably across the different pools. A number of analyses were conducted to evaluate pool equivalence across the new pools and the reference pool. They included comparisons of (a) correlations across the different ASVAB tests, (b) composite score distributions, and (c) subgroup performance. In addition, equated scores from the new pools were compared to scores on an operational pool. Each of these comparisons is discussed in more detail below.

3.3.1. Score Correlations

Correlations were computed between all reported standard scores (GS, AR, WK, PC, MK, MC, EI, AO, AS, and VE) across pools 5, 6, 7, 8, and 9R. Of interest was whether the score correlations were similar across the new pools and whether the new pools displayed the same score correlations as the reference and operational pools. Tables A.1–A.10 in Appendix A display the results for each ASVAB score.

The results suggested that the correlations between the different ASVAB scores were similar across the new pools. The variation among correlations was small. Specifically, the variation among score correlations for the new pools was smaller than the variations in score correlations between the reference pool (Form 4) and the operational pool (Form 1). The correlations averaged across the new pools were also similar to the reference pool correlations. The size of the differences between the new pool correlations and the reference pool was generally about the same, or smaller, compared to differences between the reference pool and the operational pool.

3.3.2. Composite Distributions

Composite scores computed from ASVAB standard scores are used to help classify new recruits into military occupations. Each Service uses its own set of composites based on the combination of tests that are most highly correlated with on-the-job performance for clusters of occupations. Table B.1 in Appendix B gives the composite scores that are computed for each Service and shows how they are computed. Because the equating was conducted at the test level only (and not at the composite level), the similarity of composite score distributions across the new pools is not guaranteed. Since most of the Service composites are computed as linear transformations of the standard scores (with the exception that Air Force and AFQT composites are transformed to percentiles), the composites will have the same means across pools. However, to the extent that pools display different patterns of test correlations, the composites can have different variances (and possibly higher order moments) across pools.

Several analyses were conducted to evaluate the similarity of composite distributions across the new pools (Forms 5, 6, 7, 8, and 9R) and the reference pool. The analyses evaluated differences in the first two moments and in cumulative distribution functions. Table B.2 in Appendix B displays the first and second moments for each composite score across each pool. The means were identical within the limits of rounding error, while the standard deviations showed very small differences across the pools. The nearly identical means are a result of the fact that most composites are linear transformations of the standard scores. The similarity of standard deviations across the pools is due to their similar test score correlations. Table B.3 in Appendix B displays the results of Kolmogorov-Smirnov (K-S) tests comparing the score distribution for each new pool to the score distribution for the reference pool. Most composites displayed similar distributions across the new and reference pools. Five composites consistently showed statistically significant differences ($p < 0.01$) across the five pools (i.e. on average, the score distributions for the new pools were significantly different from the score distribution for the reference pool at the $\alpha = 0.01$ level).

Figures 3.1–3.5 display the magnitude of the score differences at each point along the score scale for the five composites that showed statistically significant differences on average between the new and reference pools. The score differences are plotted in terms of qualification rate differences (qualification rate using the reference pool minus the qualification rate using the new pool). Positive values for the difference indicate that the percentage of applicants qualifying on the new pool (at a given cut-score) is lower than the percentage qualifying on the reference pool. Negative values indicate lower qualification rates using the reference pool. Figures 3.1–3.5 demonstrate that the qualification rate differences are mainly confined to ± 2.0 percentage points (which is very close to the range of sampling error) with a maximum difference of about 3.0 percentage points. The qualification rate differences observed here are smaller than qualification rate differences observed across CAT-ASVAB Forms 1–3 and P&P-ASVAB administrations. Overall, the qualification rate differences were viewed as being within a tolerable range.

Figure 3.1. Qualification Rate Differences for Army Mechanical Maintenance Composite

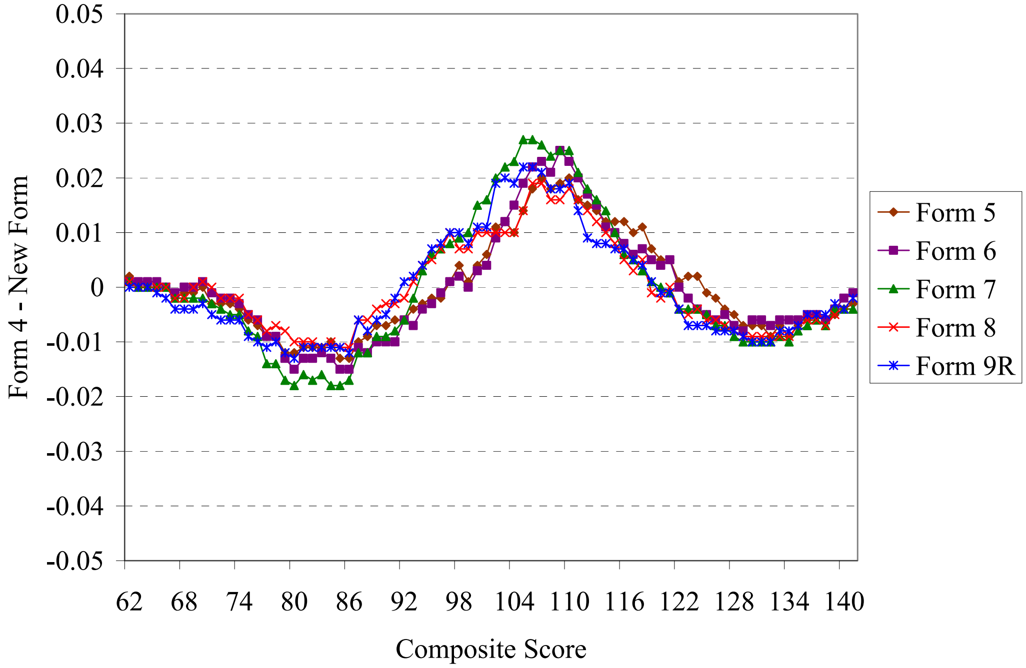


Figure 3.2. Qualification Rate Differences for Navy Mechanical 1 Composite

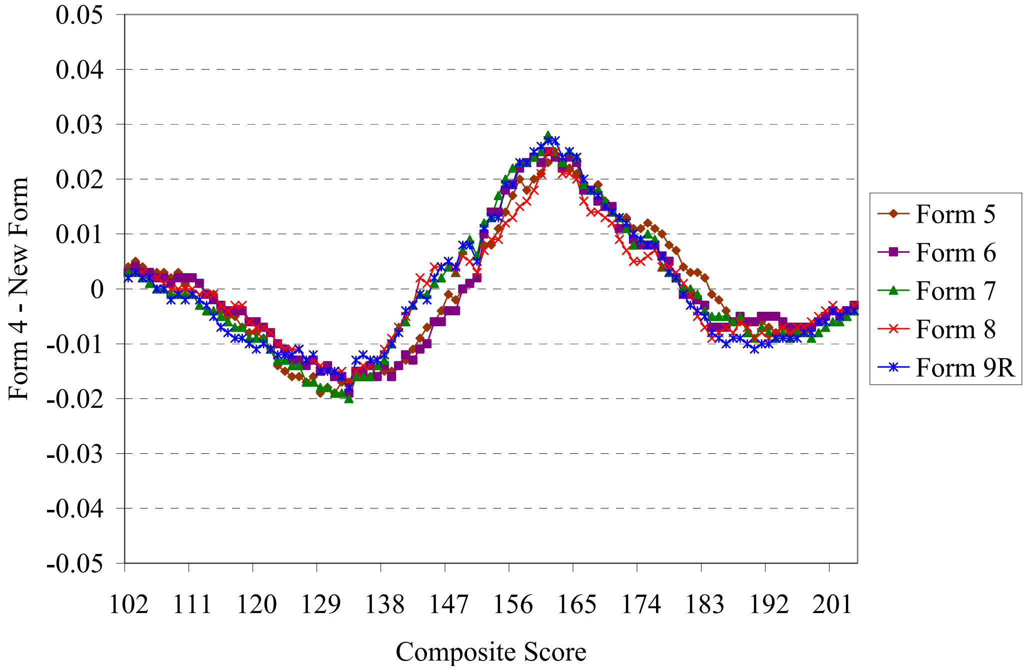


Figure 3.3. Qualification Rate Differences for Navy Mechanical 2 Composite

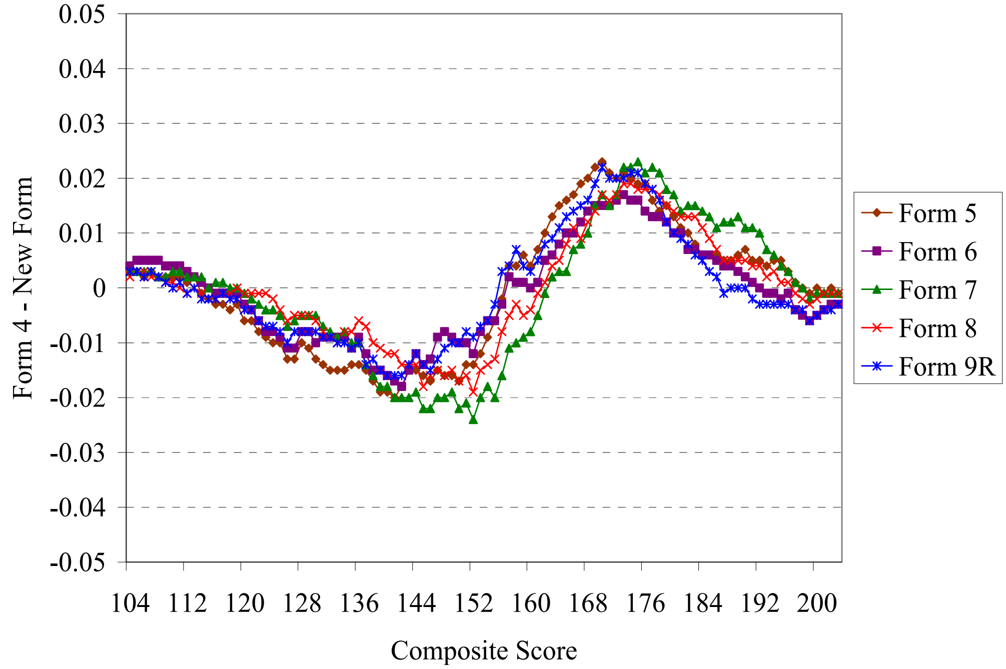


Figure 3.4. Qualification Rate Differences for Marine Corp Mechanical Composite

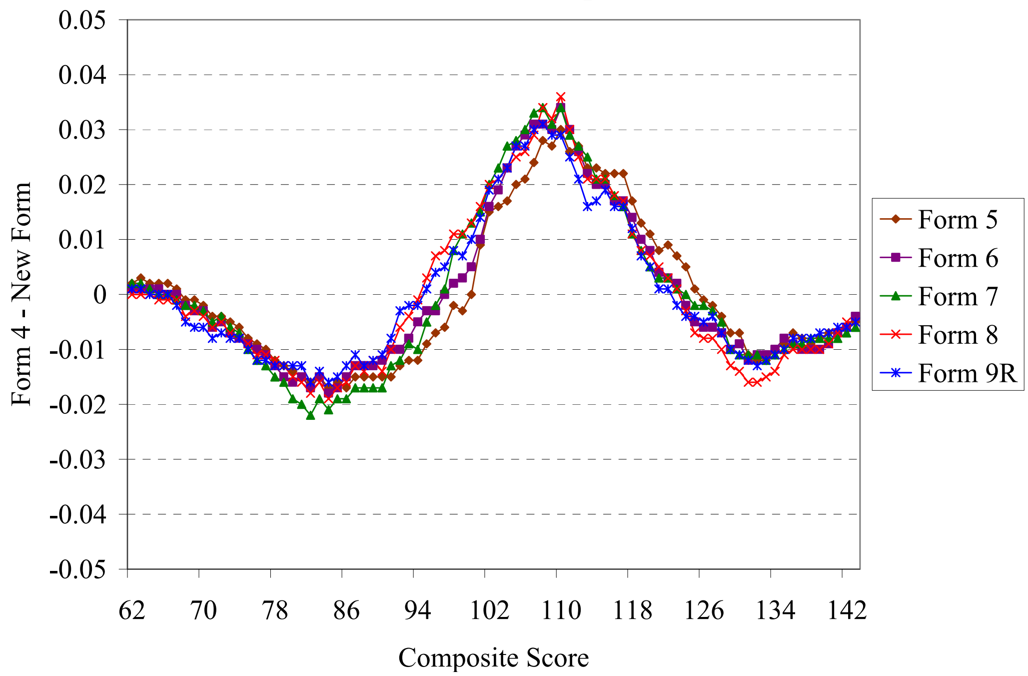
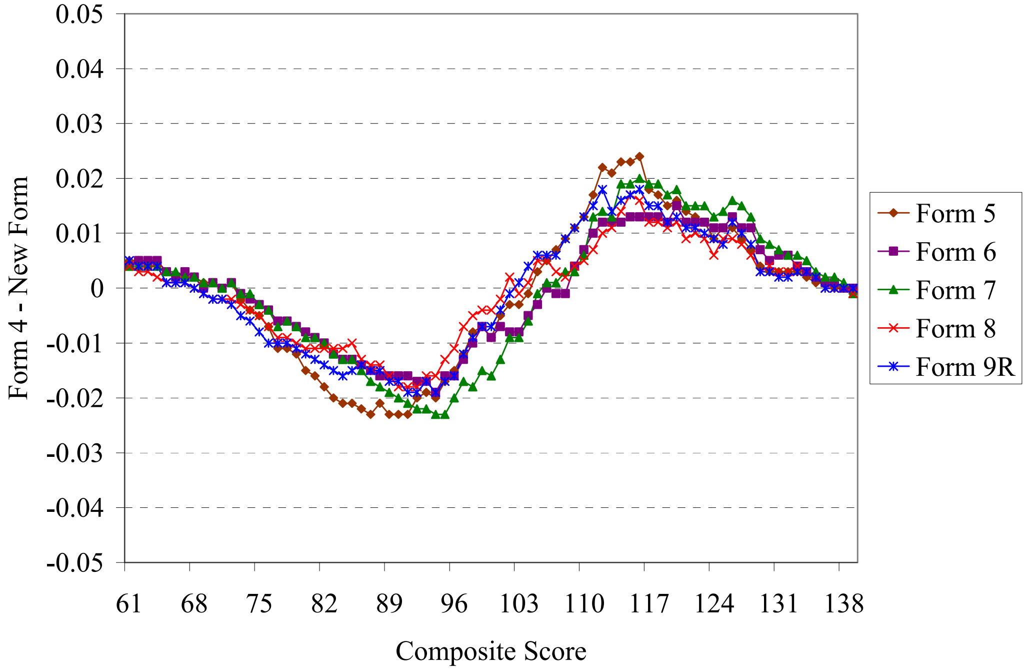


Figure 3.5. Qualification Rate Differences for Marine Corp General Technician Composite



3.3.3. Subgroup Performance

Population invariance is widely viewed as a necessary requirement for a successful equating; namely, the choice of (sub)population used to estimate the linking function between scores should not matter (Dorans & Holland, 2000). One means of assessing population invariance is to compare the performance of important subgroups across the different pools. If population invariance holds, subgroups would be expected to perform similarly across the pools.

To evaluate the question of whether subgroups perform at the same level across the new and reference pools, the equating results were applied to subgroup members taking each pool, and subgroup performance was compared across the new and reference pools. The subgroups evaluated were Females, Blacks, and Hispanics. Male and White subgroups were excluded from the analyses because they were predominantly represented in the total sample from which the equating was conducted and, as such, were not likely to display any performance differences across pools.

For each subgroup of interest, a one-way analysis of variance (ANOVA) with six groups defined by pool taken (4, 5, 6, 7, 8, or 9R) was computed for the following ASVAB scores: GS, AR, MK, MC, EI, AO, AS, VE, and AFQT. PC and WK were excluded from the analyses since they are represented in the VE score. Simultaneous 99% confidence intervals were also computed for all pairwise differences of means using the Dunn-Sidak Method. Tables C.1–C.9 in Appendix C give the results for each score for Females, Tables C.10–C.18 give the results for Blacks, while Tables C.19–27 give the results for Hispanics.

The ANOVA results for Females showed significant differences ($p < 0.01$) for the GS, EI, AS, and VE scores. For those scores, there were some confidence intervals where the pairwise mean differences did not span zero, suggesting significant mean differences across some pools. However, even the biggest mean differences between pool pairs were small, with effect sizes of 0.12 standard deviation units or less. The ANOVA results for Hispanics showed significant differences ($p < 0.01$) for the AS score. The confidence intervals for this score showed one significant mean difference across pools (Form 4 versus Form 8). Again, even the biggest mean differences between pool pairs were small, with effect sizes of 0.11 standard deviation units or less. The ANOVA results for Blacks showed no significant differences for any of the ASVAB scores, with maximum effect sizes of 0.11 standard deviation units or less. Thus, although subgroup mean differences were statistically significant across some pools, the mean differences did not appear to be practically significant.

3.3.4. Comparison with Operational Pool

An operational pool (Form 1) was included in the random assignment during the data collection, enabling a comparison of equated scores on the new pools with scores from an operational pool. The score comparison was done indirectly, comparing mean score differences across the operational and reference pools. This approach was taken because the means of the new pools are based on transformed scores, and hence, direct significance tests between new and operational pool means would have provided inaccurate sampling distributions.

Table 3.8 shows the results of a one-tailed t-test comparing the score means across Form 1 and Form 4, for GS, AR, WK, PC, MK, MC, EI, AO, AS, VE, and AFQT scores. The Form 1 and Form 4 mean scores and mean differences are reported along with the mean scores across Forms 5, 6, 7, 8, and 9R. Upper and lower bounds for a 99% confidence interval for the difference between means are also reported, in addition to the results of the t-test and associated p-values. The score means were significantly different ($p < 0.01$) for MK, MC, and AO, with confidence intervals that did not span zero. However, the largest mean score difference was 0.11 standard deviation units, which suggests that the mean differences were not practically significant. The close agreement between the reference and operational pool means suggests that the equated new pools would also display means that were very similar to the operational pool. Assuming the score differences between Forms 1 and 4 are representative of the differences among other operational pools, the new pool implementation would not be expected to have a negative (or systematic) impact on qualification rates.

Table 3.8. Comparison of Form 1 and Form 4 Means

Score	Form 1 Mean	Form 4 Mean	Forms 5–9R Mean	Form 1– Form 4	Lower Limit 99% CI	Upper Limit 99% CI	t	p
GS	50.9	50.9	50.9	0.0	–0.3	0.3	–0.1	0.934
AR	51.4	51.5	51.5	–0.1	–0.1	0.4	1.1	0.252
WK	50.2	50.3	50.3	–0.1	–0.2	0.4	1.1	0.269
PC	51.9	51.7	51.7	0.2	–0.4	0.1	–1.7	0.088
MK	52.6	51.8	51.8	0.8	–1.1	–0.6	–8.4	0.000
MC	53.4	53.0	53.0	0.3	–0.6	–0.1	–3.1	0.002
EI	51.8	51.6	51.6	0.2	–0.5	0.1	–1.6	0.101
AO	55.0	53.9	53.9	1.1	–1.4	–0.9	–10.9	0.000
AS	50.9	50.6	50.6	0.3	–0.6	0.1	–2.1	0.034
VE	50.9	50.9	50.9	0.0	–0.2	0.3	0.2	0.834
AFQT	54.5	54.0	53.8	0.5	–1.3	0.3	–1.6	0.102
N	12,932	12,838	64,852					

3.4. Accuracy of Provisional Equating Transformations

Since the final equating transformations were not computed until after completion of the Phase III data collection, it was necessary to use provisional equating transformations to compute operational scores for applicants testing during Phases I–III. This raises the question of how different scores would have been had the applicants been scored using the final transformations rather than the provisional transformations.

The accuracy of the provisional transformations was evaluated by using the final transformations to rescore all records of applicants taking CAT-ASVAB Forms 5, 6, 7, 8, or 9R during Phase I–III data collection and comparing the scores to those based on the provisional transformations. For each examinee and each test, the difference was computed between scores calculated using the provisional and final transformations. Tables 3.9–3.11 summarize the accuracy of the provisional scores for Phase I, Phase II, and Phase III examinees, respectively. The tables present the square root of the mean squared difference (RMSD) between the two scores, bias (i.e., the mean of the difference), and the standard deviation of the difference. The standard deviation may be viewed as the error of equating and is labeled as SD(E). The standard error of measurement (SEM) is also presented to provide a baseline against which SD(E) can be judged. Total error is also computed as a function of the sum of the errors of equating:

$$E(Total) = \sqrt{SD(E)^2 + SEM^2}\ . \tag{3.7}$$

Table 3.9 shows that the theory-based IRT provisional transformations used for applicants testing during Phase I displayed moderate bias and moderate contributions to total measurement error. This suggests that their performance would have been more accurately represented had the final score transformations been applied rather than the provisional score transformations. Overall, very few applicants were affected by the use of the provisional score transformation in Phase I, as the applicants represented less than 1% of the total sample across Phases I–III.

Table 3.9. Accuracy of Provisional Scores for Phase I Examinees (N = 768)

Test	RMSD	Bias	SD(E)	SEM	E(Total)
GS	2.26	1.8	1.40	3.61	3.87
AR	1.97	1.0	1.72	3.00	3.46
WK	1.02	–0.9	0.51	2.83	2.87
PC	2.79	–2.1	1.84	3.74	4.17
MK	0.64	0.3	0.58	3.00	3.06
MC	3.57	–2.6	2.47	3.87	4.59
EI	1.47	–1.1	0.95	3.46	3.59
AO	0.88	–0.3	0.82	3.61	3.70
AS	1.14	0.3	1.10	3.16	3.35

Tables 3.10–3.11 show that the empirically based provisional transformations used for applicants testing during Phases II and III displayed small bias and small contributions to total measurement error, with Phase III generally displaying smaller errors than Phase II. This suggests that the provisional scores provided a fair representation of their performance.

Table 3.10. Accuracy of Provisional Scores for Phase II Examinees (N = 2,176)

Test	RMSD	Bias	SD(E)	SEM	E(Total)
GS	0.62	0.4	0.49	3.61	3.64
AR	0.45	0.1	0.43	3.00	3.03
WK	0.49	–0.2	0.46	2.83	2.87
PC	0.49	0.0	0.49	3.74	3.77
MK	0.49	0.1	0.47	3.00	3.04
MC	1.04	0.9	0.52	3.87	3.91
EI	0.78	0.4	0.69	3.46	3.53
AO	1.09	1.0	0.52	3.61	3.64
AS	0.53	–0.1	0.51	3.16	3.20

Table 3.11. Accuracy of Provisional Scores for Phase III Examinees (N = 77,668)

Test	RMSD	Bias	SD(E)	SEM	E(Total)
GS	0.60	0.3	0.53	3.61	3.64
AR	0.60	0.3	0.50	3.00	3.04
WK	0.60	0.3	0.52	2.83	2.88
PC	0.66	0.4	0.51	3.74	3.78
MK	0.58	0.3	0.47	3.00	3.04
MC	0.53	0.2	0.49	3.87	3.90
EI	0.72	0.1	0.70	3.46	3.54
AO	0.91	0.5	0.73	3.61	3.68
AS	0.58	0.1	0.58	3.16	3.21

3.5. Time Limit Impact Analysis

The CAT-ASVAB is administered at MEPS or MET sites under fixed time constraints. It is possible that in the future, the CAT-ASVAB will be administered via the internet. If internet administration occurs, constraints on testing time will likely be greatly relaxed or eliminated completely. Anticipating such a possibility, Form 9 was administered in the equating study under regular time limits (Form 9R) and lengthened time limits (Form 9L) so that performance could be compared across normal and extended time conditions. Comparisons of performance across Form 9R and Form 9L were conducted during Phases II and III of the equating study using Phase I and Phase II data, respectively.

Prior to Phase I data collection, the operational time limits for two ASVAB tests were changed for administration of Forms 5, 6, 7, 8, and 9R. The time for MK was increased from 18 to 20 minutes, while the time for AO was increased 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. The time limits for MK and AO were increased in administrations of the new pools so that speed would play a smaller role in test performance. Additional changes in time limits were implemented in the Phase III data collection for AI, SI, and AO for administrations of Forms 5, 6, 7, 8, 9R, and 9L. The changes were made in response to findings from analyses conducted during Phase II. Administrations of CAT-ASVAB Forms 1–4 utilized the original operational time limits during the Phase I, II, and III data collections. Table 3.12 summarizes the time limits used across the different pools at different phases of the equating study.

Table 3.12. Time Limits (in Minutes) Used During the Equating Study

	Forms 1–4	Forms 5, 6, 7, 8, 9R		Form 9L	
Test	All Phases	Phases I–II	Phase III	Phases I–II	Phase III
GS	8	8	8	10	10
AR	39	39	39	47	47
WK	8	8	8	10	10
PC	22	22	22	26	26
MK	18	20	20	24	24
EI	8	8	8	10	10
AI	6	6	7	7	8
SI	5	5	6	6	7
MC	20	20	20	24	24
AO	12	13	16	16	19

3.5.1. Evaluation of Phase I Time Limits

Analyses were conducted during Phase II of the data collection comparing the performance of Phase I examinees across Forms 9R and 9L. One-way ANOVAs were conducted for ASVAB scores GS, AR, WK, PC, MK, MC, EI, AO, AS, and VE, where the pool served as the independent measure (9R or 9L), and the standard score served as the dependent measure. Test time distributions for each test were also computed, comparing results for examinees that completed the test against results for examinees that did not complete the test, for all administered pools. There were no significant differences in score means at the $\alpha = 0.01$ level for any of the reported scores, although results were significant at the $\alpha = 0.05$ level for MC, AO, and AS, with higher scores on Form 9L. The evaluation of test time distributions displayed evidence of a ceiling effect on test times for examinees testing under regular time limits for MK, AO, AI, and SI.

As a result of the findings, the Phase III regular time limits were increased by one minute for AI (from 6 minutes to 7 minutes), by one minute for SI (from 5 minutes to 6 minutes), and by three minutes for AO (from 13 minutes to 16 minutes). The Phase III lengthened time limits were also increased by one minute for AI (from 7 minutes to 8 minutes), by one minute for SI (from 6 minutes to 7 minutes), and by three minutes for AO (from 16 minutes to 19 minutes). No adjustments were made to the regular time or lengthened limits for any other tests in Phase III. Specifically, the MK time limit was not adjusted because the analyses showed that the extra time allocated to Form 9L had no significant effect on test scores. Likewise, although there was a significant difference in MC scores across Forms 9R and 9L, the MC time limit was not adjusted because there did not appear to be a ceiling effect on test time imposed by the regular time limit.

3.5.2. Evaluation of Phase II Time Limits

Analyses were conducted during Phase III of the data collection comparing the performance of Phase II examinees across Forms 9R and 9L. A one-tailed t-test was conducted for ASVAB scores GS, AR, WK, PC, MK, MC, EI, AO, AS, VE, and AFQT, where the pool served as the independent measure (9R or 9L), and the standard score served as the dependent measure. Table 3.13 presents the t-test results and associated p-values for the mean differences, along with completion rates across Forms 9R and 9L. In addition, upper and lower bounds for a 99% confidence interval for the difference between means are also reported. There were no significant differences in score means at the $\alpha = 0.01$ level for any of the reported scores, although results were significant at the $\alpha = 0.05$ level for AR, WK, and AO. However, the large sample sizes used in the analyses produced narrow confidence intervals for the mean differences, suggesting that the significant mean differences could have arisen from small effect sizes.

Table 3.13. Comparison of Form 9R and Form 9L Means

	Completion Rates		Means		Mean Differences: 9L – 9R				
Test	9R	9L	9R	9L	Diff	Lower Limit 99% CI	Upper Limit 99% CI	t	p
GS	97.3	99.3	50.9	50.9	0.0	–0.3	0.3	0.1	0.905
AR	99.0	99.7	51.5	51.8	0.2	–0.5	0.0	–2.3	0.023
WK	99.5	99.8	50.3	50.1	–0.2	–0.1	0.5	2.0	0.047
PC	99.0	99.6	51.7	51.8	0.1	–0.4	0.1	–1.3	0.202
MK	98.2	99.4	51.8	51.9	0.1	–0.3	0.2	–0.9	0.393
MC	99.9	100.0	53.1	53.0	–0.1	–0.2	0.3	0.6	0.525
EI	98.4	99.6	51.6	51.7	0.0	–0.3	0.3	–0.2	0.804
AO	99.1	99.7	53.9	54.2	0.3	–0.5	0.0	–2.5	0.011
AS	99.3	99.7	50.6	50.6	0.0	–0.3	0.3	0.0	0.990
VE[a]	98.7	99.5	50.9	50.8	–0.1	–0.2	0.3	0.9	0.383
AFQT[a]	96.2	98.7	53.9	54.0	0.2	–0.1	0.6	–0.6	0.578
N	12,848	12,816							

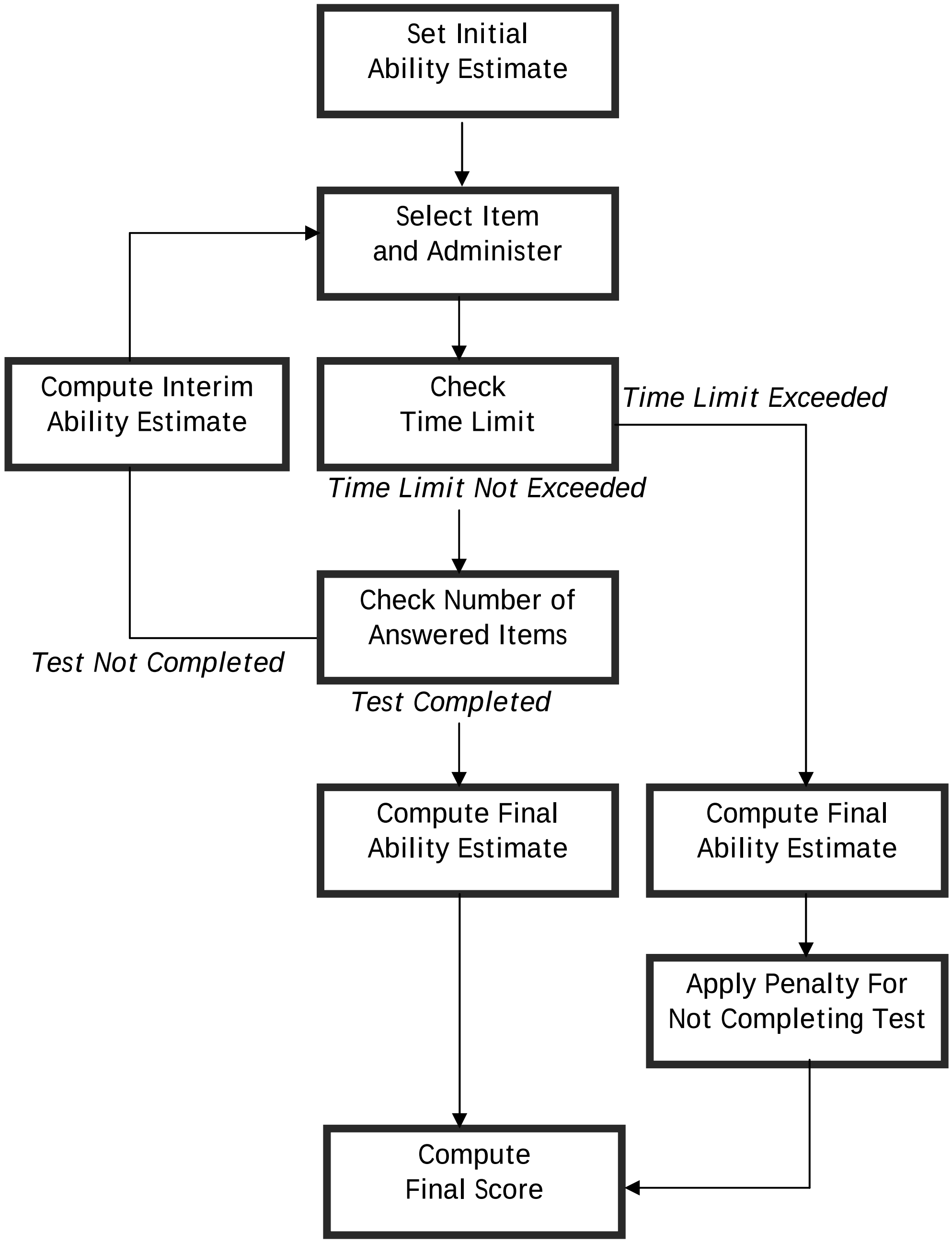
[a]VE and AFQT completion is defined as completing all tests that are part of the composite score.

The completion rates showed that nearly all examinees finished all tests under the regular time limits, implying that additional (or unlimited) time would not impact scores to a substantial degree. The results of the Phase II time limit analyses suggest that the regular time limits would be likely to produce score distributions that are comparable to those produced from untimed tests (such as those proposed for use with an internet administration).

4. Administration Procedures for CAT-ASVAB Forms 5–8

CAT-ASVAB Forms 5–8 were implemented on August 25, 2008 for administration at MEPS and MET sites under standard operational procedures. Form 9 has been reserved for internet administration of a practice or operational CAT-ASVAB. The basic steps used in item selection and scoring for CAT-ASVAB Forms 5–8 are displayed in Figure 4.1 and discussed briefly below. The standard CAT-ASVAB administration procedures are discussed in more detail in Segall, Moreno, Bloxom, and Hetter (1997) and in ASVAB Technical Bulletin #1 (DMDC, 2006), along with rationales for selecting the procedures.

Figure 4.1. Steps in CAT-ASVAB Item Selection and Scoring



4.1. Administration Steps

For each examinee, the initial ability estimate is set to $\theta = 0$. The item with maximum information at that ability level is then selected for administration. A random number between 0 and 1 is generated and compared to the exposure control parameter for the selected item. If the value of the exposure control parameter is greater or equal to the random number, the item is administered. If the value of the exposure control parameter is less than the random number, the item is not administered and is blocked from administration at any other point in the test for that examinee. (The computation of the exposure control parameters is discussed in more detail in Section 2.5.3.1.) If an item is blocked from administration, an alternate item is selected using the same procedures until an item is administered. After an item has been administered, the time limit is checked. If the time limit has been exceeded before the test is completed, then final scoring is conducted applying a penalty for non-completion of the test. If the time limit has not been exceeded, then the number of answered items is checked. If the test is completed (i.e., the fixed test length is met), then final scoring is conducted with no penalty. If the test is not completed, then an interim ability estimate is computed, the item with maximum information from among the available items is selected for administration, and exposure control is checked. This cycle is repeated until either the time limit is exceeded or the test is completed.

Interim ability estimates are computed using Owen's Bayesian procedure (1969, 1975). Final ability estimates are computed using a Bayes modal estimator. For examinees that complete the test before the time limit is exceeded, the final ability estimate is then transformed to a standard score as described in Section 3.2. For examinees that do not complete the test before the time limit is exceeded, a penalty function is applied to their final ability estimate prior to transforming to the standard score. The penalty function has the following properties:

- The size of the penalty is related to the number of unfinished items.
- Examinees who answer the same number of items and have the same ability estimate receive the same penalty.
- The penalty eliminates the possibility of using "coachable" test-taking strategies to artificially increase test scores.

The final ability estimate computed using the penalty procedure is equivalent to the score that would be obtained if the examinee guessed at random on the unfinished items.

The time limits and test lengths used in operational administration of CAT-ASVAB Forms 5–8 are given in Table 4.1. The operational time limits match the regular time limits used in Phase III of the equating study (see Table 3.12).

Table 4.1. Time Limits (in Minutes) and Test Lengths for Operational Administration of CAT-ASVAB Forms 5–8

Test	Time Limit	Test Length
GS	8	16
AR	39	16
WK	8	16
PC	22	11
MK	20	16
EI	8	16
AI	7	11
SI	6	11
MC	20	16
AO	16	16

4.2. Evaluation of Prior Distributions

A characteristic of the Bayesian scoring procedures is the use of prior distributions in the estimation of ability. For ease of implementation, a N(0,1) prior is often assumed. Several analyses were conducted to evaluate the suitability of using a N(0,1) prior in operational CAT-ASVAB administrations. The analyses assessed whether estimates of prior distributions were stable across years and whether a N(0,1) prior could be used in place of the estimated prior distributions without a loss in score precision.

For the analyses, three random samples of N = 10,000 were selected from the population of examinees taking CAT-ASVAB Forms 1–2 during Fiscal Years (FY) 2001, 2002, and 2004. (Note that due to a special data collection, FY 2003 data was not available in a sufficient quantity to include in these analyses.) The distribution means (μ) and variances (σ^2) were estimated for each sample using a maximum likelihood procedure that maximized the likelihood of the observed responses given the population distribution. (The procedure is summarized in Equations 2.3 and 2.4 in Section 2.2.1.3.)

Table D.1 in Appendix D summarizes the μ and σ^2 estimates across the ASVAB tests. Figures D.1–D.10 in Appendix D plot the μ and σ^2 estimates as a N(μ,σ^2) distribution (i.e., as an estimate of the prior distribution), along with a N(0,1) distribution, for each ASVAB test. The results showed similar μ and σ^2 estimates across the FY data and similar μ and σ^2 estimates across CAT-ASVAB Forms 1–2, which suggested that estimates of the prior distributions were stable across years. The stability of the estimated prior distributions implied that a fixed prior distribution could be used in administrations of CAT-ASVAB, provided the distribution was representative of the underlying population ability distribution.

When plotted as a N(μ,σ^2) distribution, some of the estimated distributions appeared different from the N(0,1) distribution, which raised the question of whether it would be suitable to use a N(0,1) prior in CAT-ASVAB computations of ability estimates. To answer this question, administrations of CAT-ASVAB Form 1 were simulated using both a N(0,1) prior distribution and the estimated N(μ,σ^2) prior distribution during ability estimation. Figures 4.2–4.11 plot estimated score information functions from the simulated administrations for each ASVAB test. Estimated test-retest reliabilities based on the use of each prior distribution are also given in the legend. The plots show that, with a few exceptions in the tails of the ability distribution (where there are very few examinees), the score information functions were very similar across the two prior distributions. The reliabilities were also similar across the two prior distributions. This suggests that the CAT-ASVAB tests were long enough that use of a different prior did not have a substantial effect on score precision. As such, a N(0,1) prior was implemented in administration of CAT-ASVAB Forms 5–9.

Figure 4.2. Estimated Score Information Functions by Prior Distribution for GS Form 1

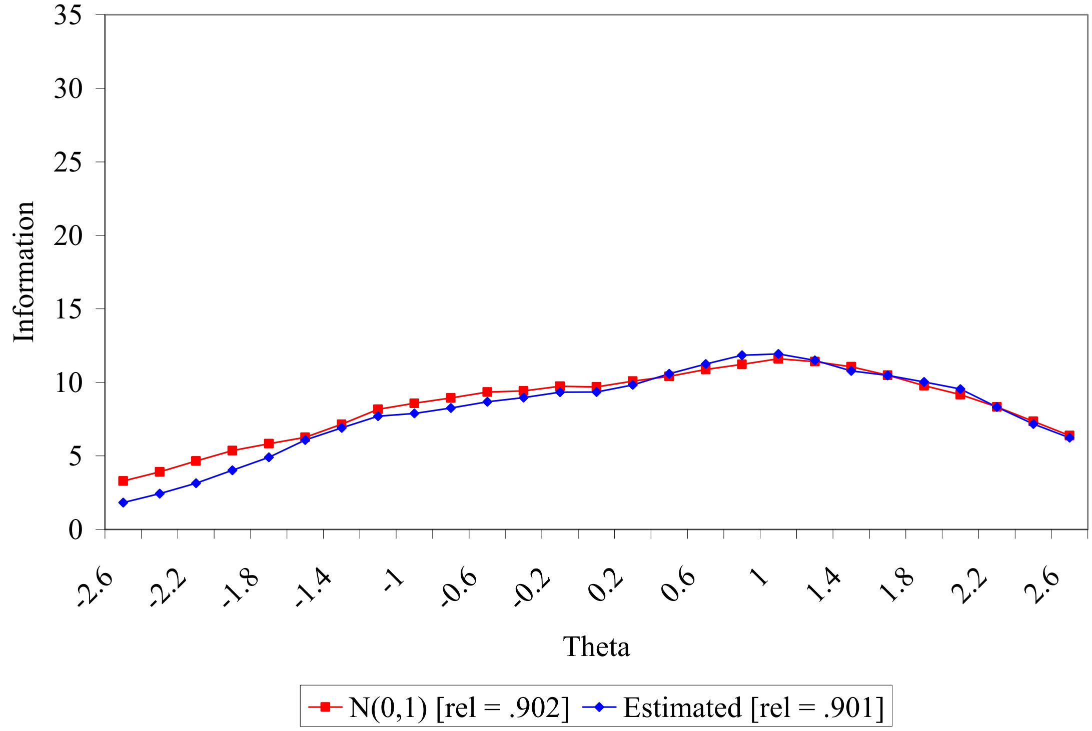


Figure 4.3. Estimated Score Information Functions by Prior Distribution for AR Form 1

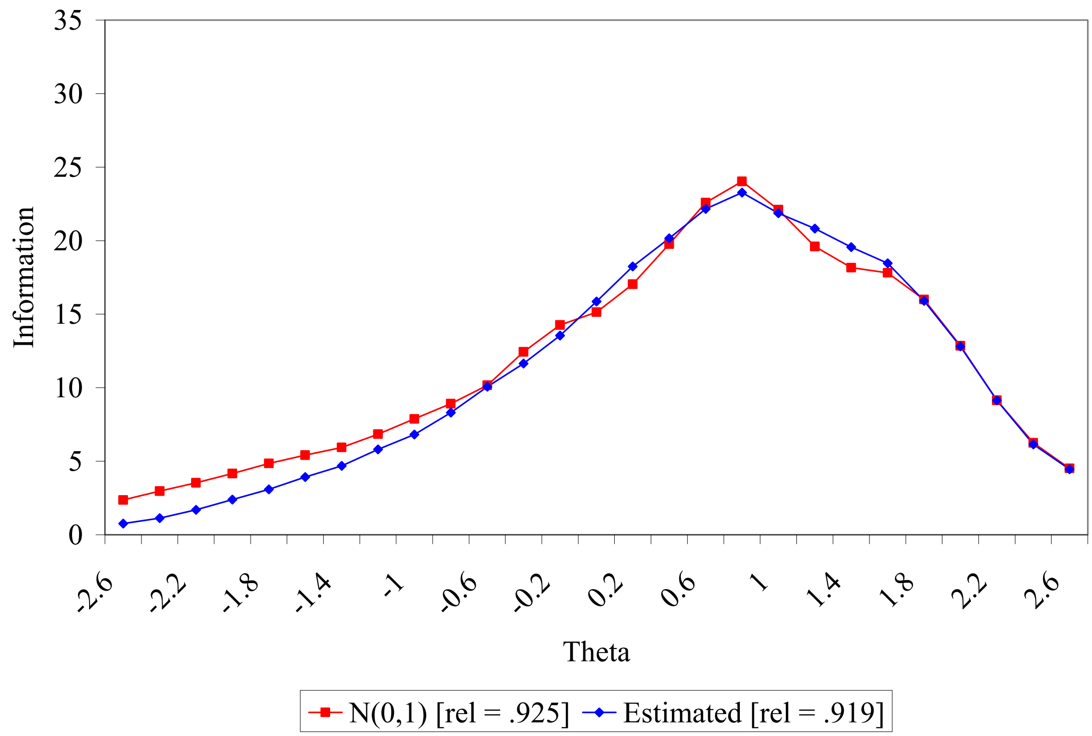


Figure 4.4. Estimated Score Information Functions by Prior Distribution for WK Form 1

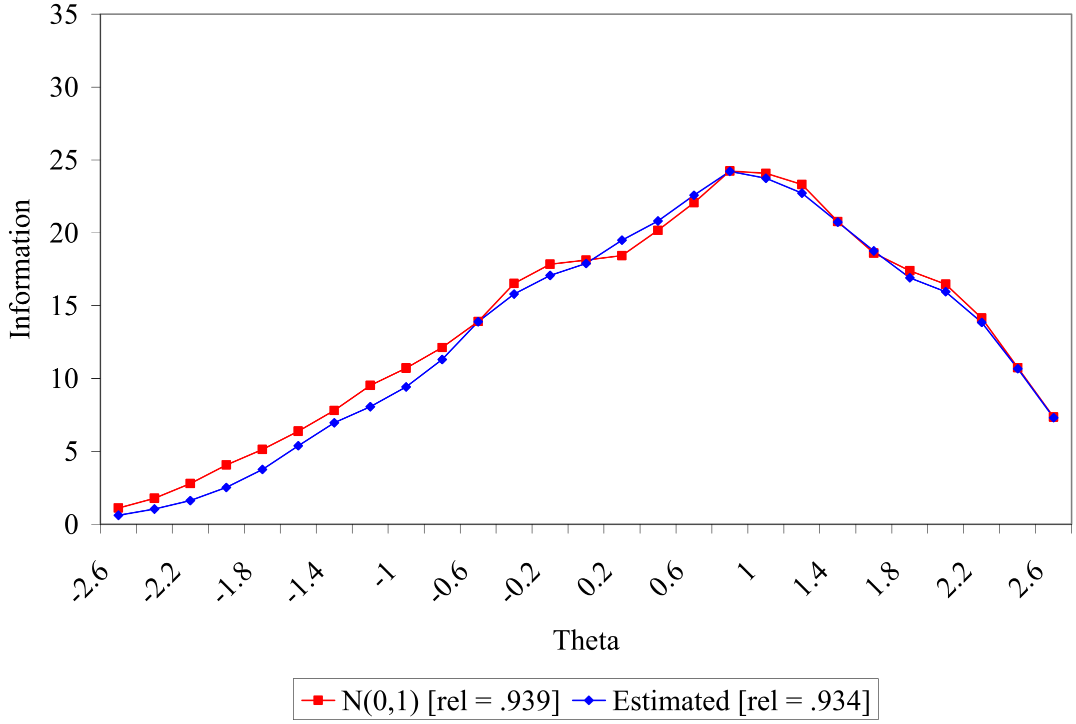


Figure 4.5. Estimated Score Information Functions by Prior Distribution for PC Form 1

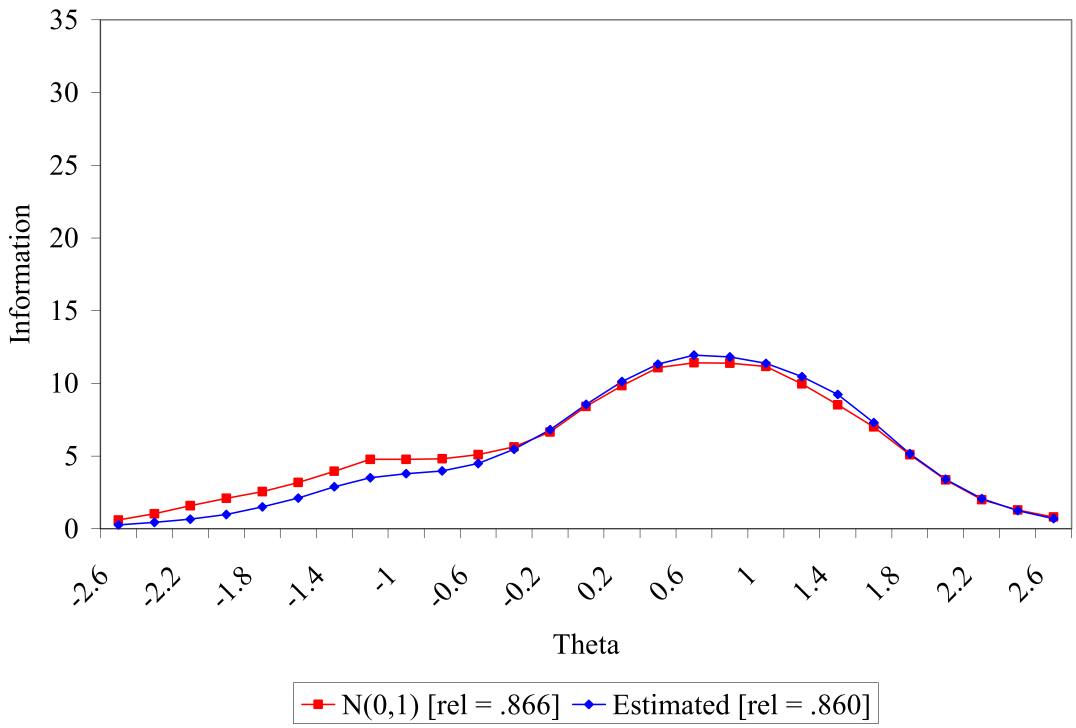


Figure 4.6. Estimated Score Information Functions by Prior Distribution for MK Form 1

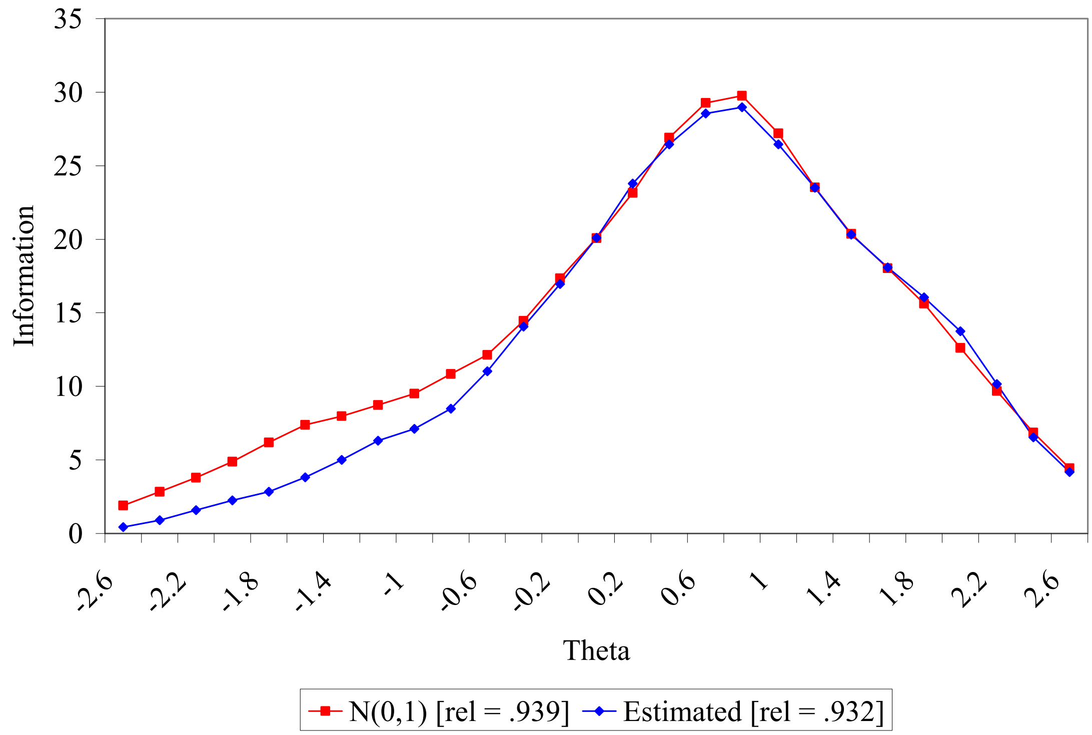


Figure 4.7. Estimated Score Information Functions by Prior Distribution for EI Form 1

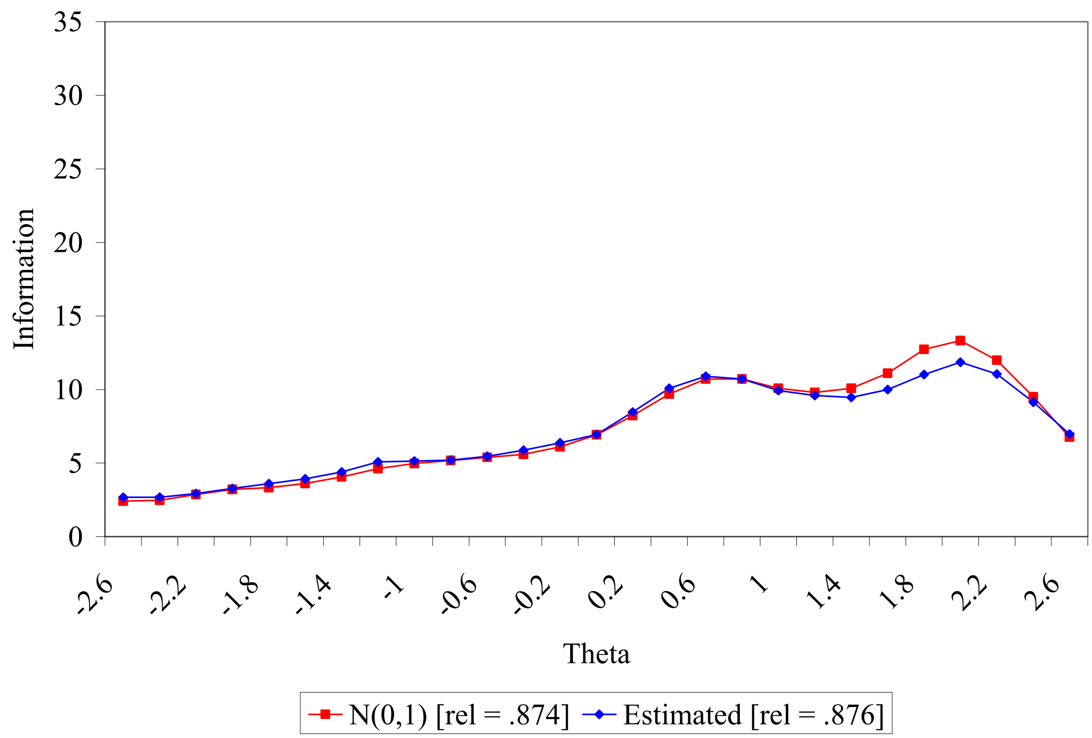


Figure 4.8. Estimated Score Information Functions by Prior Distribution for AI Form 1

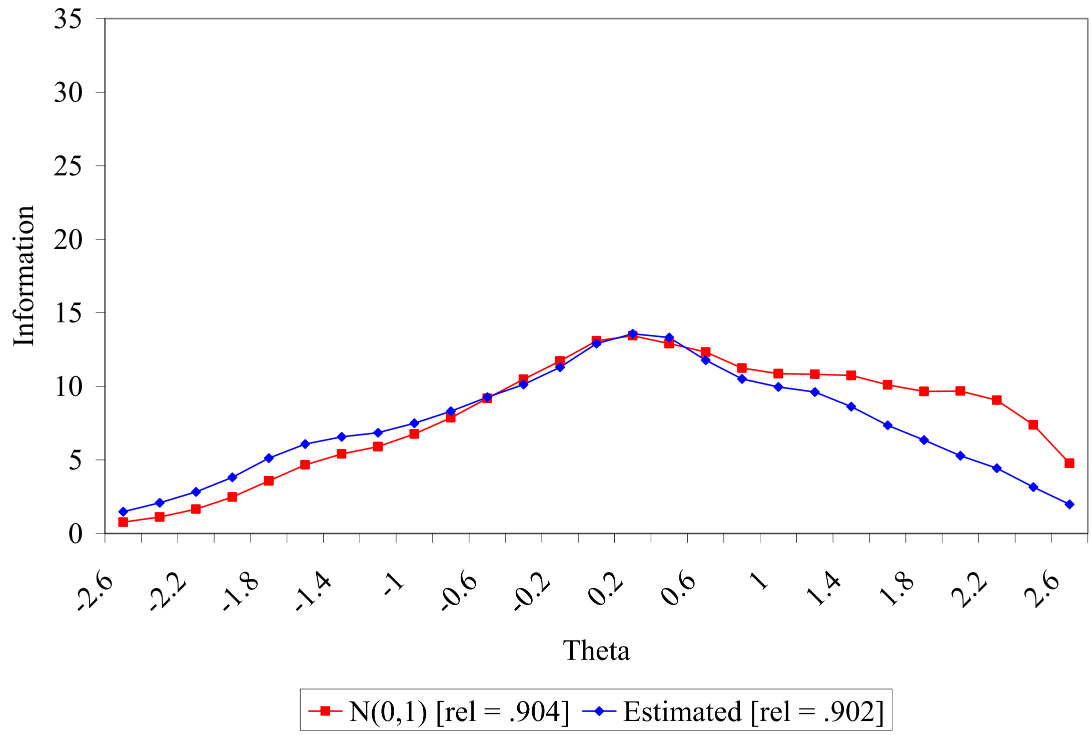


Figure 4.9. Estimated Score Information Functions by Prior Distribution for SI Form 1

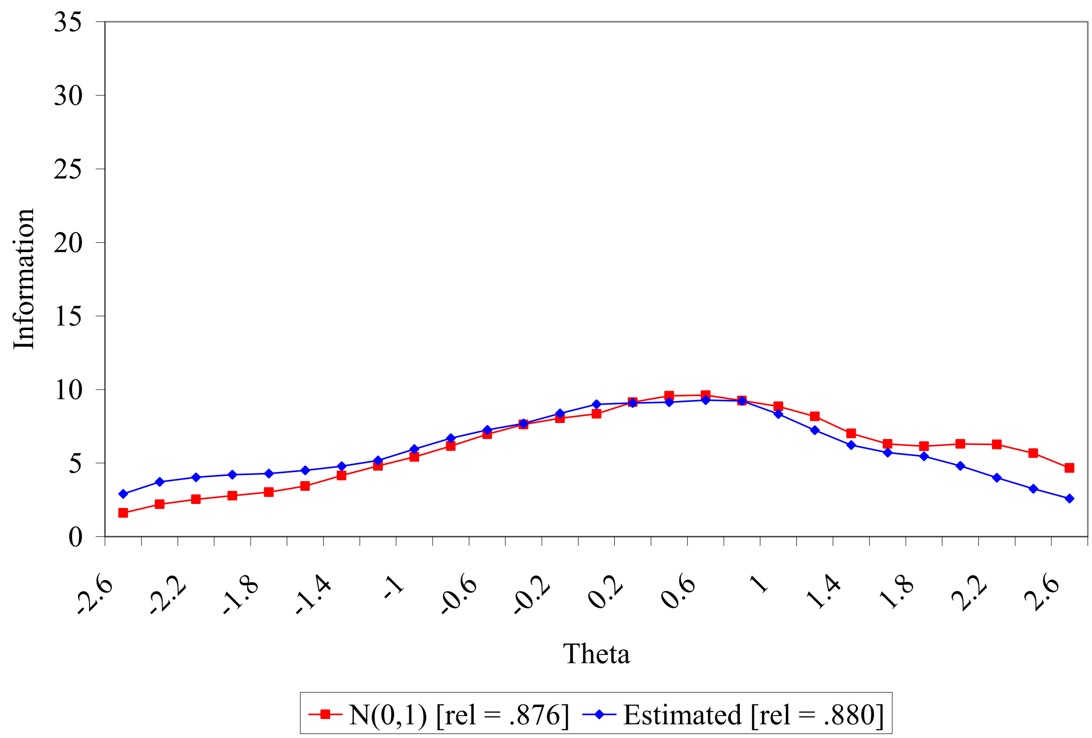


Figure 4.10. Estimated Score Information Functions by Prior Distribution for MC Form 1

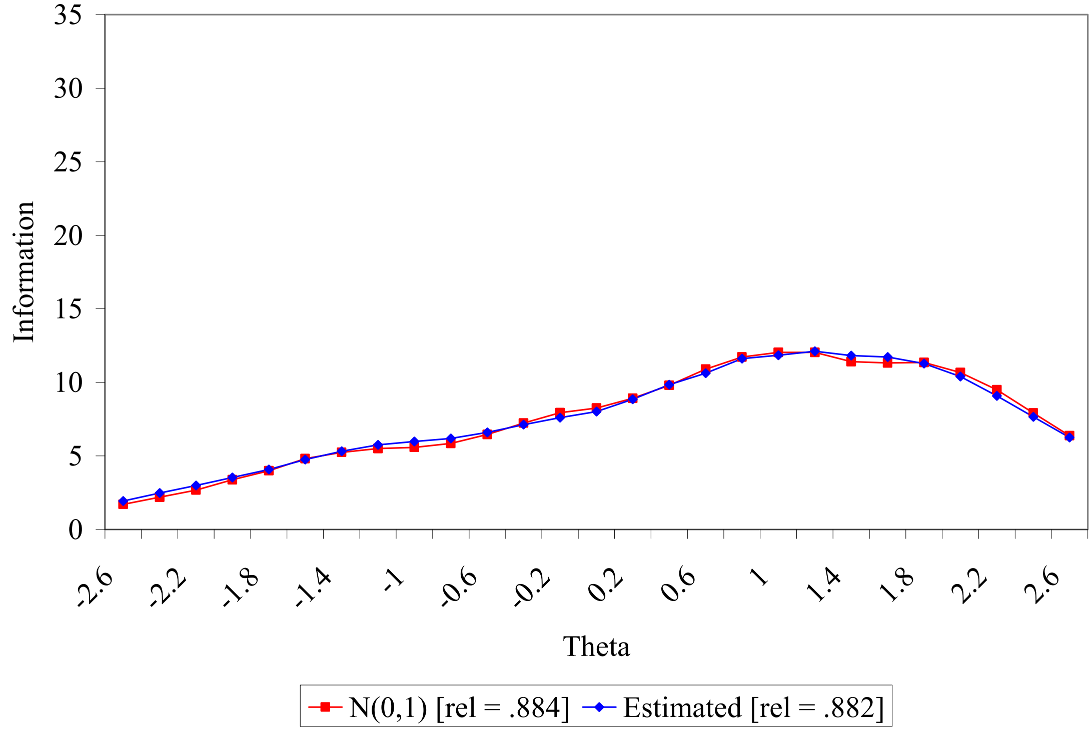
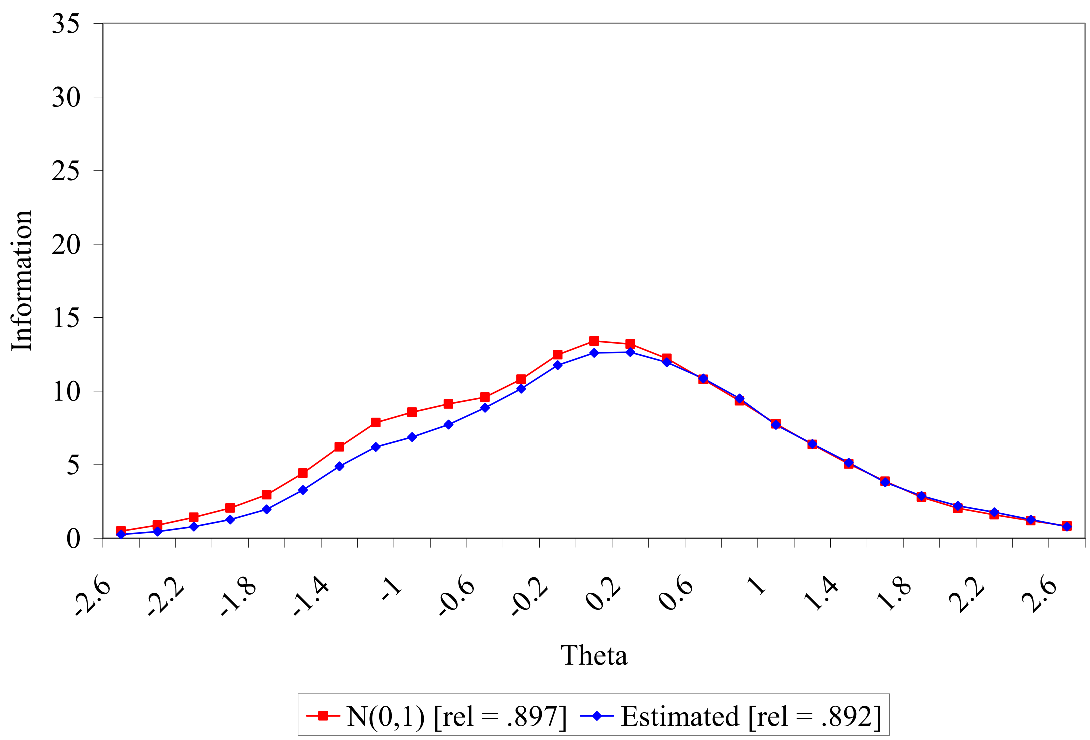


Figure 4.11. Estimated Score Information Functions by Prior Distribution for AO Form 1



4.3. Content Balancing

Due to concerns about multi-dimensionality, the CAT-ASVAB controls for content taxonomy in administrations of the AO and GS tests. This approach balances the numbers of administered items from targeted content areas in a test session. Administration of the AO test is content-balanced among the Puzzle (P) and Connection (C) content areas using the following administration sequence: CCCCCCCPPPPPPPP. Administration of the GS test is content-balanced among the Life Sciences (L) and Physical Sciences (P) content areas using the following administration sequence: PLPLPLPLPLPLPLP. Earth Science items are classified as either L or P based on the item content. Chemistry items are classified as both L and P so that they will automatically be administered if selected (given they pass the exposure control check). This approach is taken because Chemistry items are generally more difficult than all other GS item types, and blocking administration of more difficult items could restrict measurement precision.

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 pools. Research in support of the CAT-ASVAB content-balancing practices for the individual ASVAB tests is described in detail in Segall, Moreno, and Hetter (1997) and ASVAB Technical Bulletins #1 (DMDC, 2006) and #2 (DMDC, 2009).

The reliance on natural content balancing in operational administrations of most CAT-ASVAB tests does not seem to significantly degrade reliability or validity when compared to P&P-ASVAB administrations. 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. 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.

References

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

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

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

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281–306.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice.* New York: Chapman & Hall.

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.

Jannarone, R. J. (1997). Models for locally dependent responses: Conjunctive item response theory. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 465–479). New York: Springer.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking. Methods and practices* (2nd ed.). New York: Springer.

Krass, I. A., & Nicewander, W. A. (2004, June). *The effects of item position on the calibration of items seeded into adaptive tests.* Paper presented at the Annual Meeting of the Psychometric Society, Pacific Grove, CA.

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.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*, 200–219.

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., & Ito, K. (2008, March). *An examination of the properties of local dependence measures when applied to adaptive data.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

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.

Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). *Armed Services Vocational Aptitude Battery: Development of an adaptive item pool* (Technical Report no. 85-19). Brooks AFB, TX: Air Force Human Resources Laboratory.

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

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

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.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237–247.

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.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*, 247–260.

Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Eds.), *Explanatory Item Response Models* (pp. 289-316). New York: Springer.

Vale, C. D. & Gialluca, K. A. (1985). *ASCAL: A microcomputer program for estimating logistic IRT item parameters.* St. Paul, MN: Assessment Systems Corporation.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

Zenisky, A. L., Hambleton, R. K., & Sirici, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement*, *39*, 291–309.

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

Zwick, R., Thayer, D. T., & Lewis, C. (1997). *An investigation of the validity of an empirical Bayes approach to Mantel-Haenszel DIF analysis* (ETS Research Report No. 97-21). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*, 1–28.

Appendix A

Score Correlations

Table A.1. Score Correlations for GS

	-GS-	-AR-	-WK-	-PC-	-MK-	-MC-	-EI-	-AO-	-AS-	-VE-	
01E	1.000	0.601	0.772	0.681	0.451	0.628	0.720	0.422	0.537	0.790	
04E	1.000	0.607	0.747	0.636	0.527	0.659	0.664	0.455	0.523	0.765	
New	1.000	0.613	0.763	0.703	0.520	0.701	0.701	0.439	0.511	0.788	
											AveA
D14	0.000	-0.006	0.025	0.045	-0.076	-0.031	0.056	-0.033	0.013	0.025	0.035
Dn4	0.000	0.005	0.016	0.067	-0.006	0.042	0.037	-0.015	-0.012	0.023	0.025
05E	1.000	0.625	0.761	0.709	0.533	0.696	0.729	0.446	0.496	0.786	
06E	1.000	0.603	0.762	0.684	0.519	0.687	0.716	0.420	0.530	0.782	
07E	1.000	0.617	0.756	0.710	0.511	0.715	0.703	0.445	0.532	0.787	
08E	1.000	0.599	0.756	0.679	0.493	0.700	0.698	0.446	0.515	0.777	
09R	1.000	0.619	0.780	0.731	0.546	0.706	0.658	0.440	0.481	0.808	
09L	1.000	0.615	0.776	0.716	0.522	0.699	0.650	0.417	0.478	0.800	
											AVE
SDN	0.000	0.010	0.009	0.019	0.018	0.010	0.024	0.010	0.020	0.011	0.014
SD14	0.000	0.003	0.013	0.023	0.038	0.016	0.028	0.016	0.007	0.013	0.017

New: Average new (5, 6, 7, 8, 9R) form correlation
D14: Cor(Form1) – Cor(Form4)
Dn4: Cor(New) – Cor(Form4)
SDN: SD of new forms
SD14: SD of Forms 1 and 4
AveA: Average of absolute difference values
AVE: Average of SD values

Table A.2. Score Correlations for AR

	-GS-	-AR-	-WK-	-PC-	-MK-	-MC-	-EI-	-AO-	-AS-	-VE-	
01E	0.601	1.000	0.558	0.632	0.695	0.621	0.526	0.540	0.354	0.624	
04E	0.607	1.000	0.536	0.596	0.725	0.588	0.491	0.543	0.366	0.603	
New	0.613	1.000	0.560	0.637	0.735	0.656	0.525	0.554	0.354	0.624	
											AveA
D14	-0.006	0.000	0.022	0.036	-0.030	0.033	0.035	-0.003	-0.012	0.021	0.022
Dn4	0.005	0.000	0.024	0.041	0.010	0.068	0.034	0.011	-0.012	0.021	0.025
05E	0.625	1.000	0.573	0.620	0.752	0.670	0.542	0.556	0.348	0.625	
06E	0.603	1.000	0.563	0.643	0.739	0.642	0.538	0.546	0.359	0.630	
07E	0.617	1.000	0.558	0.646	0.737	0.671	0.518	0.569	0.361	0.627	
08E	0.599	1.000	0.547	0.638	0.714	0.653	0.525	0.552	0.351	0.618	
09R	0.619	1.000	0.558	0.639	0.735	0.645	0.502	0.549	0.349	0.622	
09L	0.615	1.000	0.547	0.629	0.729	0.634	0.488	0.550	0.334	0.611	
											AVE
SDN	0.010	0.000	0.008	0.009	0.012	0.012	0.014	0.008	0.006	0.004	0.009
SD14	0.003	0.000	0.011	0.018	0.015	0.016	0.017	0.002	0.006	0.011	0.011

New: Average new (5, 6, 7, 8, 9R) form correlation
D14: Cor(Form1) – Cor(Form4)
Dn4: Cor(New) – Cor(Form4)
SDN: SD of new forms
SD14: SD of Forms 1 and 4
AveA: Average of absolute difference values
AVE: Average of SD values

Table A.3. Score Correlations for WK

	-GS-	-AR-	-WK-	-PC-	-MK-	-MC-	-EI-	-AO-	-AS-	-VE-	
01E	0.772	0.558	1.000	0.727	0.419	0.521	0.633	0.362	0.414	0.964	
04E	0.747	0.536	1.000	0.682	0.425	0.553	0.588	0.375	0.465	0.958	
New	0.763	0.560	1.000	0.751	0.443	0.600	0.630	0.373	0.432	0.967	
											AveA
D14	0.025	0.022	0.000	0.045	-0.006	-0.032	0.045	-0.013	-0.051	0.007	0.028
Dn4	0.016	0.024	0.000	0.069	0.017	0.047	0.043	-0.002	-0.033	0.009	0.029
05E	0.761	0.573	1.000	0.766	0.457	0.597	0.668	0.375	0.424	0.969	
06E	0.762	0.563	1.000	0.746	0.446	0.595	0.657	0.367	0.451	0.966	
07E	0.756	0.558	1.000	0.747	0.439	0.604	0.615	0.367	0.438	0.966	
08E	0.756	0.547	1.000	0.735	0.432	0.586	0.617	0.377	0.411	0.965	
09R	0.780	0.558	1.000	0.761	0.438	0.619	0.595	0.378	0.436	0.968	
09L	0.776	0.547	1.000	0.756	0.420	0.615	0.588	0.358	0.430	0.968	
											AVE
SDN	0.009	0.008	0.000	0.011	0.008	0.011	0.027	0.005	0.014	0.001	0.011
SD14	0.013	0.011	0.000	0.023	0.003	0.016	0.023	0.007	0.026	0.003	0.014

New: Average new (5, 6, 7, 8, 9R) form correlation
D14: Cor(Form1) – Cor(Form4)
Dn4: Cor(New) – Cor(Form4)
SDN: SD of new forms
SD14: SD of Forms 1 and 4
AveA: Average of absolute difference values
AVE: Average of SD values

Table A.4. Score Correlations for PC

	-GS-	-AR-	-WK-	-PC-	-MK-	-MC-	-EI-	-AO-	-AS-	-VE-	
01E	0.681	0.632	0.727	1.000	0.500	0.552	0.566	0.437	0.353	0.880	
04E	0.636	0.596	0.682	1.000	0.516	0.520	0.493	0.433	0.361	0.861	
New	0.703	0.637	0.751	1.000	0.530	0.616	0.580	0.449	0.381	0.892	
											AveA
D14	0.045	0.036	0.045	0.000	-0.015	0.031	0.073	0.004	-0.008	0.019	0.031
Dn4	0.067	0.041	0.069	0.000	0.014	0.095	0.087	0.015	0.020	0.031	0.049
05E	0.709	0.620	0.766	1.000	0.505	0.618	0.621	0.439	0.399	0.899	
06E	0.684	0.643	0.746	1.000	0.546	0.594	0.583	0.444	0.369	0.890	
07E	0.710	0.646	0.747	1.000	0.538	0.640	0.577	0.456	0.395	0.890	
08E	0.679	0.638	0.735	1.000	0.525	0.606	0.570	0.453	0.363	0.885	
09R	0.731	0.639	0.761	1.000	0.535	0.620	0.550	0.451	0.380	0.897	
09L	0.716	0.629	0.756	1.000	0.515	0.607	0.535	0.440	0.372	0.894	
											AVE
SDN	0.019	0.009	0.011	0.000	0.014	0.015	0.023	0.006	0.014	0.005	0.013
SD14	0.023	0.018	0.023	0.000	0.008	0.016	0.037	0.002	0.004	0.010	0.015

New: Average new (5, 6, 7, 8, 9R) form correlation
D14: Cor(Form1) – Cor(Form4)
Dn4: Cor(New) – Cor(Form4)
SDN: SD of new forms
SD14: SD of Forms 1 and 4
AveA: Average of absolute difference values
AVE: Average of SD values

Table A.5. Score Correlations for MK

	-GS-	-AR-	-WK-	-PC-	-MK-	-MC-	-EI-	-AO-	-AS-	-VE-	
01E	0.451	0.695	0.419	0.500	1.000	0.414	0.349	0.441	0.073	0.478	
04E	0.527	0.725	0.425	0.516	1.000	0.426	0.331	0.479	0.132	0.495	
New	0.520	0.735	0.443	0.530	1.000	0.487	0.350	0.481	0.101	0.504	
											AveA
D14	-0.076	-0.030	-0.006	-0.015	0.000	-0.012	0.017	-0.038	-0.058	-0.017	0.030
Dn4	-0.006	0.010	0.017	0.014	0.000	0.061	0.018	0.002	-0.030	0.009	0.019
05E	0.533	0.752	0.457	0.505	1.000	0.499	0.379	0.478	0.097	0.502	
06E	0.519	0.739	0.446	0.546	1.000	0.487	0.372	0.475	0.120	0.514	
07E	0.511	0.737	0.439	0.538	1.000	0.500	0.333	0.491	0.103	0.504	
08E	0.493	0.714	0.432	0.525	1.000	0.473	0.343	0.476	0.081	0.497	
09R	0.546	0.735	0.438	0.535	1.000	0.476	0.321	0.487	0.107	0.502	
09L	0.522	0.729	0.420	0.515	1.000	0.464	0.298	0.471	0.090	0.482	
											AVE
SDN	0.018	0.012	0.008	0.014	0.000	0.011	0.022	0.006	0.013	0.006	0.012
SD14	0.038	0.015	0.003	0.008	0.000	0.006	0.009	0.019	0.029	0.008	0.015

New: Average new (5, 6, 7, 8, 9R) form correlation
D14: Cor(Form1) – Cor(Form4)
Dn4: Cor(New) – Cor(Form4)
SDN: SD of new forms
SD14: SD of Forms 1 and 4
AveA: Average of absolute difference values
AVE: Average of SD values

Table A.6. Score Correlations for MC

	-GS-	-AR-	-WK-	-PC-	-MK-	-MC-	-EI-	-AO-	-AS-	-VE-	
01E	0.628	0.621	0.521	0.552	0.414	1.000	0.649	0.594	0.605	0.569	
04E	0.659	0.588	0.553	0.520	0.426	1.000	0.695	0.555	0.670	0.586	
New	0.701	0.656	0.600	0.616	0.487	1.000	0.716	0.578	0.632	0.644	
											AveA
D14	-0.031	0.033	-0.032	0.031	-0.012	0.000	-0.046	0.039	-0.065	-0.018	0.034
Dn4	0.042	0.068	0.047	0.095	0.061	0.000	0.021	0.023	-0.038	0.058	0.051
05E	0.696	0.670	0.597	0.618	0.499	1.000	0.726	0.602	0.634	0.640	
06E	0.687	0.642	0.595	0.594	0.487	1.000	0.716	0.556	0.636	0.633	
07E	0.715	0.671	0.604	0.640	0.500	1.000	0.715	0.591	0.620	0.656	
08E	0.700	0.653	0.586	0.606	0.473	1.000	0.709	0.580	0.623	0.634	
09R	0.706	0.645	0.619	0.620	0.476	1.000	0.713	0.562	0.645	0.657	
09L	0.699	0.634	0.615	0.607	0.464	1.000	0.705	0.552	0.644	0.649	
											AVE
SDN	0.010	0.012	0.011	0.015	0.011	0.000	0.006	0.017	0.009	0.011	0.011
SD14	0.016	0.016	0.016	0.016	0.006	0.000	0.023	0.019	0.033	0.009	0.017

New: Average new (5, 6, 7, 8, 9R) form correlation
D14: Cor(Form1) – Cor(Form4)
Dn4: Cor(New) – Cor(Form4)
SDN: SD of new forms
SD14: SD of Forms 1 and 4
AveA: Average of absolute difference values
AVE: Average of SD values

Table A.7. Score Correlations for EI

	-GS-	-AR-	-WK-	-PC-	-MK-	-MC-	-EI-	-AO-	-AS-	-VE-	
01E	0.720	0.526	0.633	0.566	0.349	0.649	1.000	0.390	0.642	0.652	
04E	0.664	0.491	0.588	0.493	0.331	0.695	1.000	0.404	0.722	0.600	
New	0.701	0.525	0.630	0.580	0.350	0.716	1.000	0.413	0.710	0.651	
											AveA
D14	0.056	0.035	0.045	0.073	0.017	-0.046	0.000	-0.014	-0.080	0.052	0.046
Dn4	0.037	0.034	0.043	0.087	0.018	0.021	0.000	0.009	-0.012	0.051	0.035
05E	0.729	0.542	0.668	0.621	0.379	0.726	1.000	0.437	0.689	0.690	
06E	0.716	0.538	0.657	0.583	0.372	0.716	1.000	0.406	0.713	0.671	
07E	0.703	0.518	0.615	0.577	0.333	0.715	1.000	0.411	0.714	0.640	
08E	0.698	0.525	0.617	0.570	0.343	0.709	1.000	0.419	0.706	0.641	
09R	0.658	0.502	0.595	0.550	0.321	0.713	1.000	0.391	0.727	0.614	
09L	0.650	0.488	0.588	0.535	0.298	0.705	1.000	0.375	0.723	0.604	
											AVE
SDN	0.024	0.014	0.027	0.023	0.022	0.006	0.000	0.015	0.013	0.026	0.019
SD14	0.028	0.017	0.023	0.037	0.009	0.023	0.000	0.007	0.040	0.026	0.023

New: Average new (5, 6, 7, 8, 9R) form correlation
D14: Cor(Form1) – Cor(Form4)
Dn4: Cor(New) – Cor(Form4)
SDN: SD of new forms
SD14: SD of Forms 1 and 4
AveA: Average of absolute difference values
AVE: Average of SD values

Table A.8. Score Correlations for AO

	-GS-	-AR-	-WK-	-PC-	-MK-	-MC-	-EI-	-AO-	-AS-	-VE-	
01E	0.422	0.540	0.362	0.437	0.441	0.594	0.390	1.000	0.298	0.416	
04E	0.455	0.543	0.375	0.433	0.479	0.555	0.404	1.000	0.322	0.429	
New	0.439	0.554	0.373	0.449	0.481	0.578	0.413	1.000	0.292	0.425	
											AveA
D14	-0.033	-0.003	-0.013	0.004	-0.038	0.039	-0.014	0.000	-0.025	-0.013	0.020
Dn4	-0.015	0.011	-0.002	0.015	0.002	0.023	0.009	0.000	-0.030	-0.003	0.012
05E	0.446	0.556	0.375	0.439	0.478	0.602	0.437	1.000	0.300	0.422	
06E	0.420	0.546	0.367	0.444	0.475	0.556	0.406	1.000	0.298	0.420	
07E	0.445	0.569	0.367	0.456	0.491	0.591	0.411	1.000	0.297	0.424	
08E	0.446	0.552	0.377	0.453	0.476	0.580	0.419	1.000	0.283	0.431	
09R	0.440	0.549	0.378	0.451	0.487	0.562	0.391	1.000	0.282	0.429	
09L	0.417	0.550	0.358	0.440	0.471	0.552	0.375	1.000	0.277	0.412	
											AVE
SDN	0.010	0.008	0.005	0.006	0.006	0.017	0.015	0.000	0.008	0.004	0.009
SD14	0.016	0.002	0.007	0.002	0.019	0.019	0.007	0.000	0.012	0.006	0.010

New: Average new (5, 6, 7, 8, 9R) form correlation
D14: Cor(Form1) – Cor(Form4)
Dn4: Cor(New) – Cor(Form4)
SDN: SD of new forms
SD14: SD of Forms 1 and 4
AveA: Average of absolute difference values
AVE: Average of SD values

Table A.9. Score Correlations for AS

	-GS-	-AR-	-WK-	-PC-	-MK-	-MC-	-EI-	-AO-	-AS-	-VE-	
01E	0.537	0.354	0.414	0.353	0.073	0.605	0.642	0.298	1.000	0.419	
04E	0.523	0.366	0.465	0.361	0.132	0.670	0.722	0.322	1.000	0.464	
New	0.511	0.354	0.432	0.381	0.101	0.632	0.710	0.292	1.000	0.440	
											AveA
D14	0.013	-0.012	-0.051	-0.008	-0.058	-0.065	-0.080	-0.025	0.000	-0.045	0.040
Dn4	-0.012	-0.012	-0.033	0.020	-0.030	-0.038	-0.012	-0.030	0.000	-0.024	0.024
05E	0.496	0.348	0.424	0.399	0.097	0.634	0.689	0.300	1.000	0.440	
06E	0.530	0.359	0.451	0.369	0.120	0.636	0.713	0.298	1.000	0.449	
07E	0.532	0.361	0.438	0.395	0.103	0.620	0.714	0.297	1.000	0.450	
08E	0.515	0.351	0.411	0.363	0.081	0.623	0.706	0.283	1.000	0.421	
09R	0.481	0.349	0.436	0.380	0.107	0.645	0.727	0.282	1.000	0.441	
09L	0.478	0.334	0.430	0.372	0.090	0.644	0.723	0.277	1.000	0.434	
											AVE
SDN	0.020	0.006	0.014	0.014	0.013	0.009	0.013	0.008	0.000	0.011	0.012
SD14	0.007	0.006	0.026	0.004	0.029	0.033	0.040	0.012	0.000	0.022	0.020

New: Average new (5, 6, 7, 8, 9R) form correlation
D14: Cor(Form1) – Cor(Form4)
Dn4: Cor(New) – Cor(Form4)
SDN: SD of new forms
SD14: SD of Forms 1 and 4
AveA: Average of absolute difference values
AVE: Average of SD values

Table A.10. Score Correlations for VE

	-GS-	-AR-	-WK-	-PC-	-MK-	-MC-	-EI-	-AO-	-AS-	-VE-	
01E	0.790	0.624	0.964	0.880	0.478	0.569	0.652	0.416	0.419	1.000	
04E	0.765	0.603	0.958	0.861	0.495	0.586	0.600	0.429	0.464	1.000	
New	0.788	0.624	0.967	0.892	0.504	0.644	0.651	0.425	0.440	1.000	
											AveA
D14	0.025	0.021	0.007	0.019	-0.017	-0.018	0.052	-0.013	-0.045	0.000	0.024
Dn4	0.023	0.021	0.009	0.031	0.009	0.058	0.051	-0.003	-0.024	0.000	0.025
05E	0.786	0.625	0.969	0.899	0.502	0.640	0.690	0.422	0.440	1.000	
06E	0.782	0.630	0.966	0.890	0.514	0.633	0.671	0.420	0.449	1.000	
07E	0.787	0.627	0.966	0.890	0.504	0.656	0.640	0.424	0.450	1.000	
08E	0.777	0.618	0.965	0.885	0.497	0.634	0.641	0.431	0.421	1.000	
09R	0.808	0.622	0.968	0.897	0.502	0.657	0.614	0.429	0.441	1.000	
09L	0.800	0.611	0.968	0.894	0.482	0.649	0.604	0.412	0.434	1.000	
											AVE
SDN	0.011	0.004	0.001	0.005	0.006	0.011	0.026	0.004	0.011	0.000	0.009
SD14	0.013	0.011	0.003	0.010	0.008	0.009	0.026	0.006	0.022	0.000	0.012

New: Average new (5, 6, 7, 8, 9R) form correlation
D14: Cor(Form1) – Cor(Form4)
Dn4: Cor(New) – Cor(Form4)
SDN: SD of new forms
SD14: SD of Forms 1 and 4
AveA: Average of absolute difference values
AVE: Average of SD values

Appendix B

Comparisons of Post-Equating Composite Distributions

Table B.1. Definition of Service Composites

Service	Composite	Computational Formula
Army	General Technical (GT)	AR + 2(PC + WK)
	Clerical (CL)	*
	Combat (CO)	*
	Electronics Repair (EL)	*
	Field Artillery (FA)	*
	General Maintenance (GM)	*
	Mechanical Maintenance (MM)	*
	Operators/Food (OF)	*
	Surveillance/Communication (SC)	*
	Skilled Technician (ST)	*
Navy	General Technician (GT)	2(PC + WK) + AR
	Electronics (EL)	GS + AR + MK + EI
	Basic Electricity and Electronics (BEE)	GS + AR + 2MK
	Engineering (ENG)	AS + MK
	Mechanical1 (MEC)	AR + AS + MC
	Mechanical2 (MEC2)	AR + AS + AO
	Nuclear (NUC)	2(PC + WK) + AR + MK + MC
	Operations (OPS)	2(PC + WK) + AR + MK + AO
	Hospitalman (HM)	2(PC+WK) + GS + MK
	Administrative (ADM)	2(PC + WK) + MK
Air Force (AF)	Mechanical (M)	AR + 2(PC + WK) + MC + AS
	Administrative (A)	2(PC + WK) + MK
	General (G)	2(PC + WK) + AR
	Electronic (E)	AR + MK + EI + GS
Marine Corps (MC)	Mechanical (MM)	AR + MC + AS + EI
	Clerical (CL)	2(PC + WK) + MK
	General Technician (GT)	2(PC + WK) + AR + MC
	Electrical (EL)	AR + MK + EI + GS
All	AFQT	2(VE) + AR + MK

* Computed as a non-integer weighted linear combination of the ASVAB tests GS, AR, WK, PC, MK, EI, AS, and MC.

Table B.2. Composite Moments (Mean, Standard Deviation)

	04E	05E	06E	07E	08E	09R
01 GT-ARMY	(102.60, 15.72)	(102.61, 15.84)	(102.60, 15.86)	(102.61, 15.83)	(102.59, 15.80)	(102.61, 15.83)
02 CL-ARMY	(103.32, 15.52)	(103.34, 15.74)	(103.32, 15.72)	(103.33, 15.69)	(103.32, 15.59)	(103.33, 15.65)
03 CO-ARMY	(103.82, 16.26)	(103.82, 16.45)	(103.82, 16.44)	(103.82, 16.40)	(103.82, 16.25)	(103.82, 16.35)
04 EL-ARMY	(103.59, 16.25)	(103.59, 16.48)	(103.58, 16.47)	(103.59, 16.41)	(103.59, 16.28)	(103.59, 16.35)
05 FA-ARMY	(103.93, 16.17)	(103.94, 16.40)	(103.92, 16.37)	(103.93, 16.36)	(103.93, 16.21)	(103.93, 16.30)
06 GM-ARMY	(103.61, 16.62)	(103.61, 16.78)	(103.61, 16.78)	(103.61, 16.74)	(103.61, 16.60)	(103.61, 16.67)
07 MM-ARMY	(103.48, 17.48)	(103.48, 17.49)	(103.47, 17.53)	(103.47, 17.49)	(103.48, 17.37)	(103.47, 17.47)
08 OF-ARMY	(103.62, 16.58)	(103.64, 16.74)	(103.62, 16.74)	(103.62, 16.73)	(103.63, 16.59)	(103.63, 16.67)
09 SC-ARMY	(103.70, 15.96)	(103.71, 16.24)	(103.70, 16.22)	(103.70, 16.14)	(103.71, 16.03)	(103.71, 16.07)
10 ST-ARMY	(103.60, 15.91)	(103.60, 16.17)	(103.59, 16.14)	(103.60, 16.13)	(103.59, 16.00)	(103.60, 16.08)
11 GT-NAVY	(102.43, 14.63)	(102.43, 14.73)	(102.42, 14.75)	(102.43, 14.73)	(102.42, 14.69)	(102.44, 14.72)
12 EL-NAVY	(205.84, 28.01)	(205.85, 28.60)	(205.83, 28.39)	(205.84, 28.21)	(205.84, 28.11)	(205.85, 28.10)
13 BEE-NAVY	(205.97, 28.59)	(205.99, 28.83)	(205.97, 28.60)	(205.97, 28.58)	(205.96, 28.37)	(205.98, 28.77)
14 ENG-NAVY	(102.37, 13.31)	(102.38, 13.11)	(102.37, 13.23)	(102.36, 13.13)	(102.37, 13.01)	(102.36, 13.16)
15 MEC-NAVY	(155.17, 22.26)	(155.17, 22.34)	(155.17, 22.29)	(155.17, 22.33)	(155.16, 22.25)	(155.17, 22.30)
16 MEC2-NAVY	(158.48, 21.37)	(158.51, 21.89)	(158.45, 21.53)	(158.46, 21.83)	(158.49, 21.72)	(158.48, 21.60)
17 NUC-NAVY	(207.25, 27.13)	(207.27, 27.83)	(207.24, 27.69)	(207.25, 27.81)	(207.25, 27.56)	(207.26, 27.68)
18 OPS-NAVY	(208.11, 26.51)	(208.15, 26.70)	(208.08, 26.59)	(208.09, 26.66)	(208.12, 26.56)	(208.11, 26.62)
19 HM-NAVY	(153.57, 21.06)	(153.58, 21.19)	(153.57, 21.14)	(153.57, 21.09)	(153.57, 21.00)	(153.57, 21.28)
20 ADM-NAVY	(102.68, 13.79)	(102.69, 13.84)	(102.67, 13.87)	(102.68, 13.81)	(102.68, 13.80)	(102.68, 13.82)
21 M-AF	(54.42, 26.04)	(54.39, 26.63)	(54.22, 26.48)	(54.37, 26.60)	(54.29, 26.33)	(54.35, 26.44)
22 A-AF	(54.68, 23.34)	(54.52, 23.51)	(54.42, 23.62)	(54.43, 23.67)	(54.46, 23.27)	(54.45, 23.61)
23 G-AF	(54.21, 24.88)	(54.02, 25.39)	(53.91, 25.44)	(54.05, 25.46)	(53.95, 25.29)	(54.03, 25.38)
24 E-AF	(55.33, 25.11)	(55.40, 25.74)	(55.52, 25.70)	(55.46, 25.71)	(55.49, 25.45)	(55.43, 25.45)
25 MM-MC	(104.01, 17.71)	(104.01, 17.81)	(104.00, 17.81)	(104.01, 17.80)	(104.01, 17.75)	(104.01, 17.78)
26 GT-MC	(104.14, 16.14)	(104.14, 16.54)	(104.13, 16.46)	(104.14, 16.57)	(104.14, 16.46)	(104.15, 16.51)
27 EL-MC	(103.37, 16.19)	(103.38, 16.53)	(103.37, 16.41)	(103.38, 16.31)	(103.37, 16.24)	(103.38, 16.24)
28 CL-MC	(102.92, 15.05)	(102.93, 15.10)	(102.91, 15.13)	(102.92, 15.08)	(102.92, 15.06)	(102.92, 15.08)
29 AFQT	(54.05, 24.00)	(53.88, 24.41)	(53.77, 24.47)	(53.83, 24.50)	(53.83, 24.15)	(53.86, 24.42)

Table B.3. Results of K-S Tests (Maximum CDF Difference, p-Value)

	05E	06E	07E	08E	09R	AVE
01 GT-ARMY	(0.014,0.175)	(0.015,0.108)	(0.017,0.053)	(0.015,0.096)	(0.015,0.105)	(0.015,0.107)
02 CL-ARMY	(0.014,0.184)	(0.015,0.112)	(0.016,0.077)	(0.009,0.603)	(0.016,0.073)	(0.014,0.210)
03 CO-ARMY	(0.017,0.049)	(0.014,0.139)	(0.017,0.046)	(0.012,0.329)	(0.015,0.090)	(0.015,0.130)
04 EL-ARMY	(0.015,0.108)	(0.014,0.165)	(0.017,0.043)	(0.011,0.390)	(0.013,0.229)	(0.014,0.187)
05 FA-ARMY	(0.015,0.087)	(0.016,0.083)	(0.016,0.058)	(0.011,0.454)	(0.015,0.101)	(0.015,0.157)
06 GM-ARMY	(0.017,0.048)	(0.015,0.099)	(0.021,0.008)	(0.013,0.200)	(0.016,0.060)	(0.017,0.083)
07 MM-ARMY	(0.020,0.011)	(0.024,0.001)	(0.027,0.000)	(0.020,0.013)	(0.022,0.004)	(0.023,0.006)*
08 OF-ARMY	(0.018,0.026)	(0.015,0.122)	(0.019,0.020)	(0.011,0.356)	(0.014,0.136)	(0.016,0.132)
09 SC-ARMY	(0.016,0.082)	(0.015,0.115)	(0.016,0.066)	(0.009,0.650)	(0.014,0.182)	(0.014,0.219)
10 ST-ARMY	(0.014,0.136)	(0.015,0.119)	(0.015,0.097)	(0.009,0.722)	(0.013,0.230)	(0.013,0.261)
11 GT-NAVY	(0.014,0.175)	(0.015,0.108)	(0.017,0.053)	(0.015,0.096)	(0.015,0.105)	(0.015,0.107)
12 EL-NAVY	(0.018,0.037)	(0.019,0.019)	(0.020,0.013)	(0.017,0.053)	(0.011,0.366)	(0.017,0.097)
13 BEE-NAVY	(0.006,0.975)	(0.012,0.289)	(0.008,0.817)	(0.016,0.068)	(0.014,0.175)	(0.011,0.465)
14 ENG-NAVY	(0.011,0.446)	(0.009,0.707)	(0.018,0.026)	(0.011,0.397)	(0.018,0.032)	(0.013,0.322)
15 MEC-NAVY	(0.023,0.002)	(0.025,0.001)	(0.027,0.000)	(0.025,0.001)	(0.027,0.000)	(0.025,0.001)*
16 MEC2-NAVY	(0.024,0.001)	(0.017,0.042)	(0.024,0.001)	(0.019,0.016)	(0.023,0.003)	(0.021,0.013)*
17 NUC-NAVY	(0.018,0.029)	(0.017,0.043)	(0.026,0.000)	(0.014,0.135)	(0.018,0.029)	(0.019,0.047)
18 OPS-NAVY	(0.014,0.184)	(0.015,0.115)	(0.019,0.015)	(0.010,0.523)	(0.014,0.133)	(0.014,0.194)
19 HM-NAVY	(0.006,0.957)	(0.011,0.369)	(0.012,0.265)	(0.013,0.208)	(0.011,0.377)	(0.011,0.435)
20 ADM-NAVY	(0.011,0.425)	(0.013,0.231)	(0.017,0.054)	(0.020,0.011)	(0.012,0.311)	(0.014,0.206)
21 M-AF	(0.017,0.049)	(0.015,0.105)	(0.017,0.047)	(0.009,0.640)	(0.011,0.450)	(0.014,0.258)
22 A-AF	(0.011,0.425)	(0.013,0.231)	(0.017,0.054)	(0.020,0.011)	(0.012,0.311)	(0.014,0.206)
23 G-AF	(0.014,0.175)	(0.015,0.108)	(0.017,0.053)	(0.015,0.096)	(0.015,0.105)	(0.015,0.107)
24 E-AF	(0.018,0.037)	(0.019,0.019)	(0.020,0.013)	(0.017,0.053)	(0.011,0.375)	(0.017,0.099)
25 MM-MC	(0.030,0.000)	(0.034,0.000)	(0.033,0.000)	(0.036,0.000)	(0.030,0.000)	(0.033,0.000)*
26 GT-MC	(0.023,0.002)	(0.019,0.018)	(0.023,0.002)	(0.018,0.025)	(0.020,0.014)	(0.021,0.012)*
27 EL-MC	(0.018,0.037)	(0.019,0.019)	(0.020,0.013)	(0.016,0.065)	(0.011,0.375)	(0.017,0.102)
28 CL-MC	(0.011,0.425)	(0.013,0.231)	(0.017,0.054)	(0.020,0.011)	(0.012,0.311)	(0.014,0.206)
29 AFQT	(0.013,0.196)	(0.017,0.057)	(0.019,0.020)	(0.011,0.355)	(0.013,0.194)	(0.015,0.164)

AVE: Average maximum CDF difference and p-value across the five new pools
* The average p-value is significant at the 0.01 level

Appendix C

Comparisons of Subgroup Performance

Table C.1. ANOVA Results for Females, GS

	04E	05E	06E	07E	08E	09R	F	p
N	2475	2430	2503	2453	2593	2491		
Means	47.7	47.1	47.1	47.2	47.1	48.3	9.23	0.000

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I – Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.6	-0.2	1.4
=	04E	06E	0.6	-0.2	1.4
=	04E	07E	0.5	-0.3	1.3
=	04E	08E	0.6	-0.1	1.4
=	04E	09R	-0.6	-1.4	0.2
=	05E	06E	0.0	-0.7	0.8
=	05E	07E	-0.1	-0.9	0.7
=	05E	08E	0.1	-0.7	0.9
N	05E	09R	-1.2	-2.0	-0.4
=	06E	07E	-0.1	-0.9	0.7
=	06E	08E	0.0	-0.8	0.8
N	06E	09R	-1.2	-2.0	-0.4
=	07E	08E	0.1	-0.6	0.9
N	07E	09R	-1.1	-1.9	-0.3
N	08E	09R	-1.2	-2.0	-0.5

(=) indicates confidence intervals that span zero, implying non-significance
(N) indicates confidence intervals that do not span zero

Table C.2. ANOVA Results for Females, AR

	04E	05E	06E	07E	08E	09R	F	p
N	2475	2430	2503	2453	2593	2491		
Means	49.2	49.3	48.8	49.1	48.7	49.3	2.44	0.032

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	-0.2	-0.9	0.6
=	04E	06E	0.3	-0.4	1.1
=	04E	07E	0.0	-0.7	0.8
=	04E	08E	0.4	-0.3	1.2
=	04E	09R	-0.1	-0.9	0.6
=	05E	06E	0.5	-0.3	1.3
=	05E	07E	0.2	-0.6	1.0
=	05E	08E	0.6	-0.2	1.4
=	05E	09R	0.0	-0.7	0.8
=	06E	07E	-0.3	-1.1	0.4
=	06E	08E	0.1	-0.7	0.8
=	06E	09R	-0.5	-1.2	0.3
=	07E	08E	0.4	-0.4	1.2
=	07E	09R	-0.2	-0.9	0.6
=	08E	09R	-0.6	-1.3	0.2

(=) indicates confidence intervals that span zero, implying non-significance

Table C.3. ANOVA Results for Females, MK

	04E	05E	06E	07E	08E	09R	F	p
N	2475	2430	2503	2453	2593	2491		
Means	52.0	52.0	51.5	52.0	51.9	51.9	1.37	0.233

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.0	-0.7	0.8
=	04E	06E	0.5	-0.3	1.2
=	04E	07E	0.0	-0.7	0.7
=	04E	08E	0.1	-0.7	0.8
=	04E	09R	0.1	-0.6	0.8
=	05E	06E	0.4	-0.3	1.1
=	05E	07E	0.0	-0.8	0.7
=	05E	08E	0.0	-0.7	0.7
=	05E	09R	0.1	-0.6	0.8
=	06E	07E	-0.5	-1.2	0.3
=	06E	08E	-0.4	-1.1	0.3
=	06E	09R	-0.3	-1.1	0.4
=	07E	08E	0.1	-0.6	0.8
=	07E	09R	0.1	-0.6	0.9
=	08E	09R	0.1	-0.7	0.8

(=) indicates confidence intervals that span zero, implying non-significance

Table C.4. ANOVA Results for Females, MC

	04E	05E	06E	07E	08E	09R	F	p
N	2475	2430	2503	2453	2593	2491		
Means	47.8	47.5	47.6	47.7	47.6	47.7	0.50	0.779

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.3	-0.4	1.0
=	04E	06E	0.2	-0.5	0.9
=	04E	07E	0.1	-0.6	0.8
=	04E	08E	0.2	-0.5	0.9
=	04E	09R	0.1	-0.6	0.8
=	05E	06E	0.0	-0.8	0.7
=	05E	07E	-0.2	-0.9	0.5
=	05E	08E	0.0	-0.7	0.7
=	05E	09R	-0.2	-0.9	0.5
=	06E	07E	-0.1	-0.8	0.6
=	06E	08E	0.0	-0.7	0.7
=	06E	09R	-0.1	-0.8	0.6
=	07E	08E	0.1	-0.6	0.8
=	07E	09R	0.0	-0.7	0.7
=	08E	09R	-0.1	-0.8	0.6

(=) indicates confidence intervals that span zero, implying non-significance

Table C.5. ANOVA Results for Females, EI

	04E	05E	06E	07E	08E	09R	F	p
N	2475	2430	2503	2453	2593	2491		
Means	45.8	45.7	45.5	45.1	45.6	45.3	3.46	0.004

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.1	-0.7	0.8
=	04E	06E	0.3	-0.4	1.1
N	04E	07E	0.8	0.0	1.5
=	04E	08E	0.2	-0.5	0.9
=	04E	09R	0.5	-0.2	1.3
=	05E	06E	0.3	-0.5	1.0
=	05E	07E	0.7	-0.1	1.4
=	05E	08E	0.1	-0.6	0.9
=	05E	09R	0.4	-0.3	1.2
=	06E	07E	0.4	-0.3	1.2
=	06E	08E	-0.1	-0.8	0.6
=	06E	09R	0.2	-0.6	0.9
=	07E	08E	-0.6	-1.3	0.2
=	07E	09R	-0.3	-1.0	0.5
=	08E	09R	0.3	-0.4	1.0

(=) indicates confidence intervals that span zero, implying non-significance
(N) indicates confidence intervals that do not span zero

Table C.6. ANOVA Results for Females, AO

	04E	05E	06E	07E	08E	09R	F	p
N	2475	2430	2503	2453	2593	2491		
Means	52.4	51.9	52.3	52.2	52.5	52.6	2.57	0.025

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.5	-0.3	1.3
=	04E	06E	0.0	-0.8	0.8
=	04E	07E	0.2	-0.6	1.0
=	04E	08E	-0.2	-0.9	0.6
=	04E	09R	-0.3	-1.1	0.5
=	05E	06E	-0.5	-1.3	0.3
=	05E	07E	-0.3	-1.1	0.5
=	05E	08E	-0.6	-1.4	0.2
=	05E	09R	-0.8	-1.6	0.0
=	06E	07E	0.1	-0.6	0.9
=	06E	08E	-0.2	-1.0	0.6
=	06E	09R	-0.3	-1.1	0.5
=	07E	08E	-0.3	-1.1	0.5
=	07E	09R	-0.5	-1.2	0.3
=	08E	09R	-0.1	-0.9	0.7

(=) indicates confidence intervals that span zero, implying non-significance

Table C.7. ANOVA Results for Females, AS

	04E	05E	06E	07E	08E	09R	F	p
N	2475	2430	2503	2453	2593	2491		
Means	43.4	42.4	42.4	43.2	42.6	43.0	9.08	0.000

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
N	04E	05E	1.0	0.3	1.7
N	04E	06E	1.0	0.3	1.6
=	04E	07E	0.2	-0.5	0.9
N	04E	08E	0.7	0.1	1.4
=	04E	09R	0.4	-0.3	1.0
=	05E	06E	0.0	-0.7	0.6
N	05E	07E	-0.8	-1.5	-0.1
=	05E	08E	-0.3	-0.9	0.4
=	05E	09R	-0.6	-1.3	0.0
N	06E	07E	-0.8	-1.4	-0.1
=	06E	08E	-0.3	-0.9	0.4
=	06E	09R	-0.6	-1.3	0.1
=	07E	08E	0.5	-0.1	1.2
=	07E	09R	0.2	-0.5	0.8
=	08E	09R	-0.3	-1.0	0.3

(=) indicates confidence intervals that span zero, implying non-significance
(N) indicates confidence intervals that do not span zero

Table C.8. ANOVA Results for Females, VE

	04E	05E	06E	07E	08E	09R	F	p
N	2475	2430	2503	2453	2593	2491		
Means	49.7	49.0	49.1	49.6	49.7	49.3	4.09	0.001

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.7	-0.1	1.4
=	04E	06E	0.7	-0.1	1.4
=	04E	07E	0.1	-0.6	0.9
=	04E	08E	0.0	-0.7	0.8
=	04E	09R	0.4	-0.3	1.2
=	05E	06E	0.0	-0.8	0.7
=	05E	07E	-0.6	-1.3	0.2
=	05E	08E	-0.7	-1.4	0.1
=	05E	09R	-0.2	-1.0	0.5
=	06E	07E	-0.5	-1.3	0.2
=	06E	08E	-0.6	-1.4	0.1
=	06E	09R	-0.2	-1.0	0.5
=	07E	08E	-0.1	-0.9	0.6
=	07E	09R	0.3	-0.4	1.1
=	08E	09R	0.4	-0.3	1.2

(=) indicates confidence intervals that span zero, implying non-significance

Table C.9. ANOVA Results for Females, AFQT

	04E	05E	06E	07E	08E	09R	F	p
N	2475	2430	2503	2453	2593	2491		
Means	49.9	48.7	47.7	49.4	49.3	49.0	2.66	0.021

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	1.2	-1.1	3.5
=	04E	06E	2.2	0.0	4.5
=	04E	07E	0.5	-1.7	2.8
=	04E	08E	0.6	-1.6	2.8
=	04E	09R	0.9	-1.4	3.1
=	05E	06E	1.0	-1.2	3.3
=	05E	07E	-0.7	-3.0	1.6
=	05E	08E	-0.6	-2.9	1.6
=	05E	09R	-0.3	-2.6	1.9
=	06E	07E	-1.7	-4.0	0.5
=	06E	08E	-1.6	-3.9	0.6
=	06E	09R	-1.3	-3.6	0.9
=	07E	08E	0.1	-2.2	2.3
=	07E	09R	0.4	-1.9	2.6
=	08E	09R	0.3	-1.9	2.5

(=) indicates confidence intervals that span zero, implying non-significance

Table C.10. ANOVA Results for Blacks, GS

	04E	05E	06E	07E	08E	09R	F	p
N	1446	1488	1426	1609	1507	1525		
Means	46.1	46.0	46.0	46.4	45.9	46.4	1.18	0.317

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.1	-0.9	1.2
=	04E	06E	0.2	-0.8	1.2
=	04E	07E	-0.2	-1.2	0.8
=	04E	08E	0.2	-0.8	1.2
=	04E	09R	-0.3	-1.3	0.7
=	05E	06E	0.0	-1.0	1.1
=	05E	07E	-0.4	-1.4	0.6
=	05E	08E	0.1	-0.9	1.1
=	05E	09R	-0.4	-1.4	0.6
=	06E	07E	-0.4	-1.4	0.6
=	06E	08E	0.0	-1.0	1.1
=	06E	09R	-0.5	-1.5	0.5
=	07E	08E	0.5	-0.5	1.5
=	07E	09R	-0.1	-1.0	0.9
=	08E	09R	-0.5	-1.5	0.5

(=) indicates confidence intervals that span zero, implying non-significance

Table C.11. ANOVA Results for Blacks, AR

	04E	05E	06E	07E	08E	09R	F	p
N	1446	1488	1426	1609	1507	1525		
Means	47.3	47.4	47.3	47.4	47.6	47.8	0.65	0.659

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	-0.1	-1.1	0.9
=	04E	06E	0.0	-1.0	1.0
=	04E	07E	-0.1	-1.1	0.9
=	04E	08E	-0.3	-1.3	0.7
=	04E	09R	-0.4	-1.4	0.6
=	05E	06E	0.1	-0.9	1.1
=	05E	07E	0.0	-1.0	1.0
=	05E	08E	-0.2	-1.2	0.8
=	05E	09R	-0.3	-1.3	0.7
=	06E	07E	-0.1	-1.1	0.9
=	06E	08E	-0.3	-1.3	0.8
=	06E	09R	-0.4	-1.4	0.6
=	07E	08E	-0.2	-1.2	0.8
=	07E	09R	-0.3	-1.3	0.6
=	08E	09R	-0.2	-1.2	0.8

(=) indicates confidence intervals that span zero, implying non-significance

Table C.12. ANOVA Results for Blacks, MK

	04E	**05E**	**06E**	**07E**	**08E**	**09R**		
N	1446	1488	1426	1609	1507	1525	**F**	**p**
Means	49.5	50.2	49.9	50.0	50.2	49.8	1.69	0.134

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

				99.0% Confidence Interval	
	Group I	**Group J**	**Mean I - Mean J**	**Lower Limit**	**Upper Limit**
=	04E	05E	-0.7	-1.7	0.3
=	04E	06E	-0.4	-1.4	0.6
=	04E	07E	-0.5	-1.5	0.4
=	04E	08E	-0.7	-1.7	0.3
=	04E	09R	-0.3	-1.3	0.7
=	05E	06E	0.3	-0.7	1.3
=	05E	07E	0.2	-0.8	1.1
=	05E	08E	0.0	-1.0	1.0
=	05E	09R	0.4	-0.6	1.3
=	06E	07E	-0.1	-1.1	0.8
=	06E	08E	-0.3	-1.3	0.7
=	06E	09R	0.1	-0.9	1.1
=	07E	08E	-0.2	-1.1	0.8
=	07E	09R	0.2	-0.8	1.1
=	08E	09R	0.4	-0.6	1.4

(=) indicates confidence intervals that span zero, implying non-significance

Table C.13. ANOVA Results for Blacks, MC

	04E	05E	06E	07E	08E	09R	F	p
N	1446	1488	1426	1609	1507	1525		
Means	47.3	47.0	47.2	47.3	47.3	47.0	0.58	0.713

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.3	-0.7	1.2
=	04E	06E	0.1	-0.9	1.1
=	04E	07E	-0.1	-1.0	0.9
=	04E	08E	0.0	-1.0	0.9
=	04E	09R	0.3	-0.7	1.2
=	05E	06E	-0.2	-1.1	0.8
=	05E	07E	-0.4	-1.3	0.6
=	05E	08E	-0.3	-1.3	0.7
=	05E	09R	0.0	-1.0	0.9
=	06E	07E	-0.2	-1.2	0.8
=	06E	08E	-0.1	-1.1	0.8
=	06E	09R	0.1	-0.8	1.1
=	07E	08E	0.1	-0.9	1.0
=	07E	09R	0.3	-0.6	1.3
=	08E	09R	0.3	-0.7	1.2

(=) indicates confidence intervals that span zero, implying non-significance

Table C.14. ANOVA Results for Blacks, EI

	04E	05E	06E	07E	08E	09R	F	p
N	1446	1488	1426	1609	1507	1525		
Means	46.3	45.6	45.6	46.2	46.0	46.1	1.58	0.163

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

				99.0% Confidence Interval	
	Group I	Group J	Mean I - Mean J	Lower Limit	Upper Limit
=	04E	05E	0.7	-0.4	1.8
=	04E	06E	0.7	-0.4	1.8
=	04E	07E	0.1	-0.9	1.2
=	04E	08E	0.3	-0.8	1.4
=	04E	09R	0.2	-0.9	1.2
=	05E	06E	0.0	-1.1	1.1
=	05E	07E	-0.5	-1.6	0.5
=	05E	08E	-0.4	-1.4	0.7
=	05E	09R	-0.5	-1.6	0.6
=	06E	07E	-0.5	-1.6	0.6
=	06E	08E	-0.4	-1.4	0.7
=	06E	09R	-0.5	-1.6	0.6
=	07E	08E	0.2	-0.9	1.2
=	07E	09R	0.0	-1.0	1.1
=	08E	09R	-0.1	-1.2	0.9

(=) indicates confidence intervals that span zero, implying non-significance

Table C.15. ANOVA Results for Blacks, AO

	04E	05E	06E	07E	08E	09R	F	p
N	1446	1488	1426	1609	1507	1525		
Means	50.0	49.9	49.9	50.1	50.1	50.1	0.21	0.957

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	99.0% Confidence Interval Upper Limit
=	04E	05E	0.1	-1.0	1.2
=	04E	06E	0.1	-1.0	1.2
=	04E	07E	-0.1	-1.2	1.0
=	04E	08E	-0.1	-1.2	1.0
=	04E	09R	-0.1	-1.2	1.0
=	05E	06E	0.0	-1.1	1.1
=	05E	07E	-0.2	-1.3	0.9
=	05E	08E	-0.2	-1.3	0.9
=	05E	09R	-0.3	-1.3	0.8
=	06E	07E	-0.2	-1.3	0.9
=	06E	08E	-0.2	-1.3	0.9
=	06E	09R	-0.2	-1.3	0.9
=	07E	08E	0.0	-1.1	1.1
=	07E	09R	-0.1	-1.1	1.0
=	08E	09R	-0.1	-1.1	1.0

(=) indicates confidence intervals that span zero, implying non-significance

Table C.16. ANOVA Results for Blacks, AS

	04E	05E	06E	07E	08E	09R	F	p
N	1446	1488	1426	1609	1507	1525		
Means	43.9	43.4	43.2	43.9	43.3	43.7	2.36	0.038

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.5	-0.5	1.4
=	04E	06E	0.7	-0.2	1.7
=	04E	07E	0.0	-0.9	1.0
=	04E	08E	0.6	-0.4	1.5
=	04E	09R	0.2	-0.8	1.2
=	05E	06E	0.3	-0.7	1.2
=	05E	07E	-0.5	-1.4	0.5
=	05E	08E	0.1	-0.9	1.1
=	05E	09R	-0.3	-1.2	0.7
=	06E	07E	-0.7	-1.7	0.2
=	06E	08E	-0.2	-1.1	0.8
=	06E	09R	-0.5	-1.5	0.4
=	07E	08E	0.6	-0.4	1.5
=	07E	09R	0.2	-0.8	1.1
=	08E	09R	-0.4	-1.3	0.6

(=) indicates confidence intervals that span zero, implying non-significance

Table C.17. ANOVA Results for Blacks, VE

	04E	05E	06E	07E	08E	09R	F	p
N	1446	1488	1426	1609	1507	1525		
Means	47.6	47.1	47.6	47.7	47.8	47.7	1.72	0.125

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.5	-0.4	1.4
=	04E	06E	0.0	-1.0	0.9
=	04E	07E	-0.1	-1.1	0.8
=	04E	08E	-0.2	-1.1	0.7
=	04E	09R	-0.1	-1.1	0.8
=	05E	06E	-0.5	-1.5	0.4
=	05E	07E	-0.6	-1.6	0.3
=	05E	08E	-0.7	-1.6	0.2
=	05E	09R	-0.6	-1.6	0.3
=	06E	07E	-0.1	-1.0	0.8
=	06E	08E	-0.2	-1.1	0.8
=	06E	09R	-0.1	-1.1	0.8
=	07E	08E	-0.1	-1.0	0.9
=	07E	09R	0.0	-0.9	0.9
=	08E	09R	0.1	-0.9	1.0

(=) indicates confidence intervals that span zero, implying non-significance

Table C.18. ANOVA Results for Blacks, AFQT

	04E	05E	06E	07E	08E	09R	F	p
N	1446	1488	1426	1609	1507	1525		
Means	42.6	42.0	42.4	42.9	43.2	43.1	0.60	0.703

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.6	-2.2	3.4
=	04E	06E	0.2	-2.7	3.0
=	04E	07E	-0.3	-3.0	2.5
=	04E	08E	-0.6	-3.4	2.2
=	04E	09R	-0.5	-3.3	2.3
=	05E	06E	-0.4	-3.2	2.4
=	05E	07E	-0.9	-3.6	1.9
=	05E	08E	-1.2	-4.0	1.6
=	05E	09R	-1.1	-3.8	1.7
=	06E	07E	-0.5	-3.2	2.3
=	06E	08E	-0.8	-3.6	2.0
=	06E	09R	-0.7	-3.5	2.1
=	07E	08E	-0.4	-3.1	2.4
=	07E	09R	-0.2	-2.9	2.5
=	08E	09R	0.2	-2.6	2.9

(=) indicates confidence intervals that span zero, implying non-significance

Table C.19. ANOVA Results for Hispanics, GS

	04E	05E	06E	07E	08E	09R	F	p
N	2014	2115	2029	2028	2066	1962		
Means	47.4	46.8	47.0	46.8	47.1	47.5	2.28	0.044

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.6	-0.4	1.5
=	04E	06E	0.4	-0.6	1.3
=	04E	07E	0.6	-0.4	1.5
=	04E	08E	0.2	-0.7	1.2
=	04E	09R	-0.2	-1.1	0.8
=	05E	06E	-0.2	-1.1	0.7
=	05E	07E	0.0	-0.9	0.9
=	05E	08E	-0.3	-1.2	0.6
=	05E	09R	-0.7	-1.6	0.2
=	06E	07E	0.2	-0.7	1.1
=	06E	08E	-0.1	-1.0	0.8
=	06E	09R	-0.5	-1.4	0.4
=	07E	08E	-0.3	-1.2	0.6
=	07E	09R	-0.7	-1.6	0.2
=	08E	09R	-0.4	-1.3	0.5

(=) indicates confidence intervals that span zero, implying non-significance

Table C.20. ANOVA Results for Hispanics, AR

	04E	05E	06E	07E	08E	09R		
N	2014	2115	2029	2028	2066	1962	**F**	**p**
Means	49.5	49.3	49.4	49.1	49.4	49.6	0.74	0.596

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval	
				Lower Limit	Upper Limit
=	04E	05E	0.2	-0.7	1.1
=	04E	06E	0.0	-0.9	0.9
=	04E	07E	0.3	-0.6	1.2
=	04E	08E	0.1	-0.8	1.0
=	04E	09R	-0.1	-1.0	0.8
=	05E	06E	-0.1	-1.0	0.7
=	05E	07E	0.2	-0.7	1.0
=	05E	08E	-0.1	-0.9	0.8
=	05E	09R	-0.3	-1.2	0.6
=	06E	07E	0.3	-0.6	1.2
=	06E	08E	0.1	-0.8	1.0
=	06E	09R	-0.2	-1.1	0.7
=	07E	08E	-0.2	-1.1	0.7
=	07E	09R	-0.5	-1.4	0.4
=	08E	09R	-0.2	-1.1	0.7

(=) indicates confidence intervals that span zero, implying non-significance

Table C.21. ANOVA Results for Hispanics, MK

	04E	05E	06E	07E	08E	09R	F	p
N	2014	2115	2029	2028	2066	1962		
Means	50.6	50.2	50.6	50.4	50.4	50.6	0.79	0.559

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.4	-0.5	1.3
=	04E	06E	0.0	-0.9	0.9
=	04E	07E	0.3	-0.6	1.2
=	04E	08E	0.2	-0.6	1.1
=	04E	09R	0.0	-0.9	0.9
=	05E	06E	-0.4	-1.2	0.5
=	05E	07E	-0.1	-1.0	0.8
=	05E	08E	-0.2	-1.0	0.7
=	05E	09R	-0.4	-1.2	0.5
=	06E	07E	0.2	-0.7	1.1
=	06E	08E	0.2	-0.7	1.1
=	06E	09R	0.0	-0.9	0.9
=	07E	08E	0.0	-0.9	0.9
=	07E	09R	-0.2	-1.1	0.7
=	08E	09R	-0.2	-1.1	0.7

(=) indicates confidence intervals that span zero, implying non-significance

Table C.22. ANOVA Results for Hispanics, MC

	04E	05E	06E	07E	08E	09R	F	p
N	2014	2115	2029	2028	2066	1962		
Means	50.0	49.7	50.0	49.7	49.9	49.7	0.96	0.441

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.4	-0.5	1.2
=	04E	06E	0.1	-0.8	0.9
=	04E	07E	0.4	-0.5	1.3
=	04E	08E	0.1	-0.8	1.0
=	04E	09R	0.4	-0.5	1.3
=	05E	06E	-0.3	-1.1	0.6
=	05E	07E	0.0	-0.8	0.9
=	05E	08E	-0.2	-1.1	0.6
=	05E	09R	0.0	-0.8	0.9
=	06E	07E	0.3	-0.6	1.2
=	06E	08E	0.0	-0.8	0.9
=	06E	09R	0.3	-0.6	1.2
=	07E	08E	-0.3	-1.1	0.6
=	07E	09R	0.0	-0.9	0.9
=	08E	09R	0.3	-0.6	1.2

(=) indicates confidence intervals that span zero, implying non-significance

Table C.23. ANOVA Results for Females, EI

	04E	05E	06E	07E	08E	09R	F	p
N	2014	2115	2029	2028	2066	1962		
Means	48.1	47.6	47.4	47.6	47.6	48.2	2.49	0.029

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.4	-0.6	1.4
=	04E	06E	0.7	-0.3	1.7
=	04E	07E	0.4	-0.6	1.4
=	04E	08E	0.4	-0.6	1.4
=	04E	09R	-0.2	-1.2	0.8
=	05E	06E	0.3	-0.7	1.3
=	05E	07E	0.0	-1.0	1.0
=	05E	08E	0.0	-1.0	1.0
=	05E	09R	-0.6	-1.6	0.4
=	06E	07E	-0.3	-1.3	0.7
=	06E	08E	-0.3	-1.3	0.7
=	06E	09R	-0.9	-1.9	0.1
=	07E	08E	0.0	-1.0	1.0
=	07E	09R	-0.6	-1.6	0.4
=	08E	09R	-0.6	-1.6	0.4

(=) indicates confidence intervals that span zero, implying non-significance

Table C.24. ANOVA Results for Hispanics, AO

	04E	05E	06E	07E	08E	09R	F	p
N	2014	2115	2029	2028	2066	1962		
Means	53.3	53.5	53.5	53.1	53.1	53.6	1.82	0.105

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	99.0% Confidence Interval Upper Limit
=	04E	05E	-0.2	-1.1	0.6
=	04E	06E	-0.2	-1.1	0.7
=	04E	07E	0.2	-0.6	1.1
=	04E	08E	0.3	-0.6	1.1
=	04E	09R	-0.3	-1.2	0.6
=	05E	06E	0.0	-0.9	0.9
=	05E	07E	0.5	-0.4	1.3
=	05E	08E	0.5	-0.4	1.3
=	05E	09R	-0.1	-0.9	0.8
=	06E	07E	0.5	-0.4	1.3
=	06E	08E	0.5	-0.4	1.3
=	06E	09R	-0.1	-0.9	0.8
=	07E	08E	0.0	-0.8	0.9
=	07E	09R	-0.5	-1.4	0.3
=	08E	09R	-0.5	-1.4	0.3

(=) indicates confidence intervals that span zero, implying non-significance

Table C.25. ANOVA Results for Hispanics, AS

	04E	05E	06E	07E	08E	09R	F	p
N	2014	2115	2029	2028	2066	1962		
Means	45.9	46.5	46.3	46.2	47.0	46.4	3.55	0.003

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	-0.6	-1.5	0.3
=	04E	06E	-0.4	-1.3	0.5
=	04E	07E	-0.3	-1.2	0.7
N	04E	08E	-1.1	-2.0	-0.1
=	04E	09R	-0.5	-1.4	0.4
=	05E	06E	0.2	-0.7	1.2
=	05E	07E	0.4	-0.5	1.3
=	05E	08E	-0.4	-1.3	0.5
=	05E	09R	0.1	-0.8	1.0
=	06E	07E	0.1	-0.8	1.1
=	06E	08E	-0.7	-1.6	0.3
=	06E	09R	-0.1	-1.1	0.8
=	07E	08E	-0.8	-1.7	0.1
=	07E	09R	-0.3	-1.2	0.7
=	08E	09R	0.5	-0.4	1.5

(=) indicates confidence intervals that span zero, implying non-significance
(N) indicates confidence intervals that do not span zero

Table C.26. ANOVA Results for Hispanics, VE

	04E	05E	06E	07E	08E	09R	F	p
N	2014	2115	2029	2028	2066	1962		
Means	47.7	47.3	47.0	47.2	47.0	47.1	2.15	0.056

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	0.4	-0.5	1.3
=	04E	06E	0.7	-0.2	1.6
=	04E	07E	0.5	-0.4	1.4
=	04E	08E	0.7	-0.1	1.6
=	04E	09R	0.6	-0.3	1.5
=	05E	06E	0.3	-0.6	1.2
=	05E	07E	0.1	-0.8	1.0
=	05E	08E	0.3	-0.5	1.2
=	05E	09R	0.2	-0.7	1.1
=	06E	07E	-0.2	-1.1	0.7
=	06E	08E	0.0	-0.9	0.9
=	06E	09R	-0.1	-1.0	0.8
=	07E	08E	0.3	-0.6	1.1
=	07E	09R	0.1	-0.8	1.0
=	08E	09R	-0.2	-1.0	0.7

(=) indicates confidence intervals that span zero, implying non-significance

Table C.27. ANOVA Results for Hispanics, AFQT

	04E	05E	06E	07E	08E	09R	F	p
N	2014	2115	2029	2028	2066	1962		
Means	46.0	44.6	44.5	44.3	44.2	44.7	1.43	0.209

Simultaneous Confidence Intervals for All Pairwise Differences of Means (Dunn-Sidak Method)

	Group I	Group J	Mean I - Mean J	99.0% Confidence Interval Lower Limit	Upper Limit
=	04E	05E	1.3	-1.2	3.9
=	04E	06E	1.5	-1.1	4.0
=	04E	07E	1.7	-0.9	4.3
=	04E	08E	1.7	-0.8	4.3
=	04E	09R	1.3	-1.3	3.9
=	05E	06E	0.1	-2.4	2.7
=	05E	07E	0.4	-2.2	2.9
=	05E	08E	0.4	-2.1	2.9
=	05E	09R	-0.1	-2.6	2.5
=	06E	07E	0.2	-2.3	2.8
=	06E	08E	0.3	-2.3	2.8
=	06E	09R	-0.2	-2.8	2.4
=	07E	08E	0.0	-2.5	2.6
=	07E	09R	-0.4	-3.0	2.2
=	08E	09R	-0.4	-3.0	2.1

(=) indicates confidence intervals that span zero, implying non-significance

Appendix D

Evaluation of Prior Distributions

Table D.1. μ and σ^2 estimates by Fiscal Year

		Form 1		Form 2	
Test	FY	M	σ	μ	σ
GS	2004	.257	.751	.393	.832
	2002	.302	.707	.434	.774
	2001	.223	.732	.378	.770
AR	2004	.288	.820	.272	.891
	2002	.295	.763	.263	.838
	2001	.220	.761	.212	.862
WK	2004	.170	.731	.166	.819
	2002	.219	.707	.213	.788
	2001	.133	.701	.152	.772
PC	2004	.295	.709	.262	.806
	2002	.308	.671	.295	.780
	2001	.249	.661	.244	.781
MK	2004	.588	.716	.549	.724
	2002	.456	.701	.417	.689
	2001	.366	.723	.346	.717
EI	2004	–.130	.631	–.243	.970
	2002	–.062	.654	–.172	.967
	2001	–.133	.675	–.239	.940
AI	2004	–.482	.797	–.592	.881
	2002	–.395	.753	–.492	.873
	2001	–.428	.781	–.518	.855
SI	2004	–.539	.936	–.504	.832
	2002	–.409	.921	–.378	.856
	2001	–.458	.920	–.406	.841
MC	2004	–.159	.957	–.154	.861
	2002	–.142	.851	–.138	.806
	2001	–.227	.862	–.214	.840
AO	2004	.420	.954	.447	.922
	2002	.263	.869	.280	.898
	2001	.147	.919	.179	.931

Figure D.1. Estimated Prior Distributions for GS

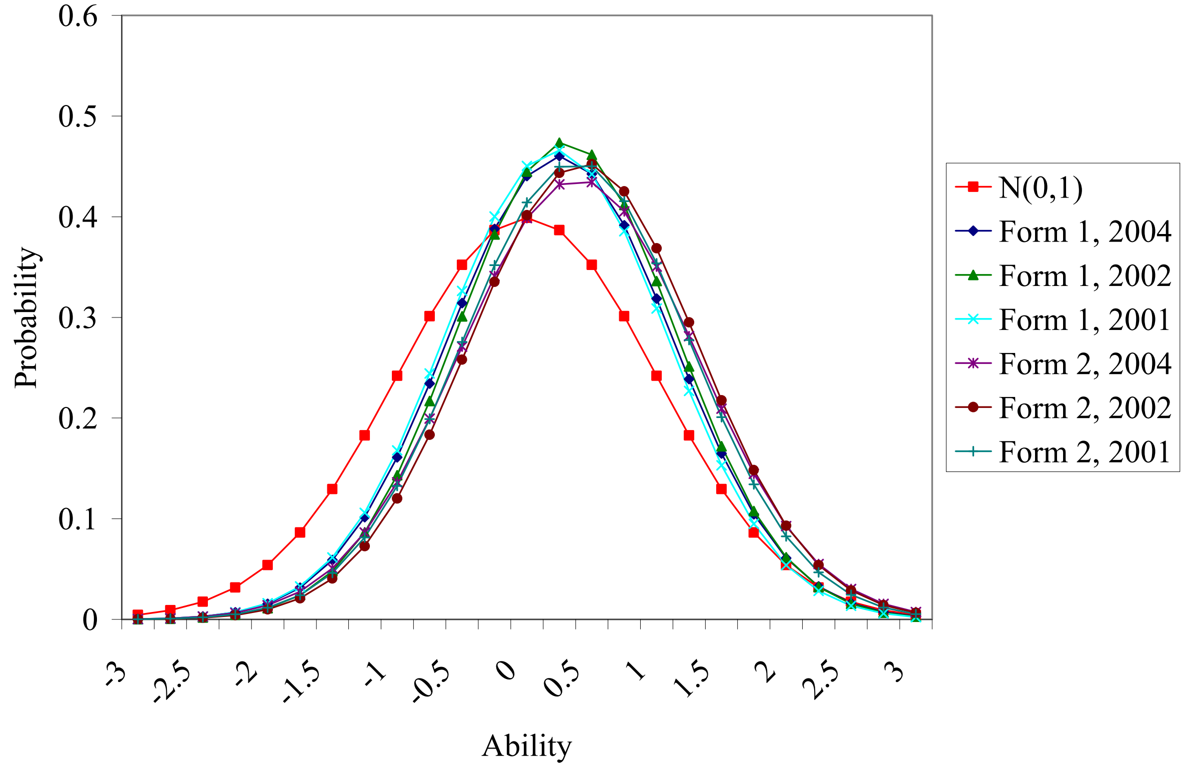


Figure D.2. Estimated Prior Distributions for AR

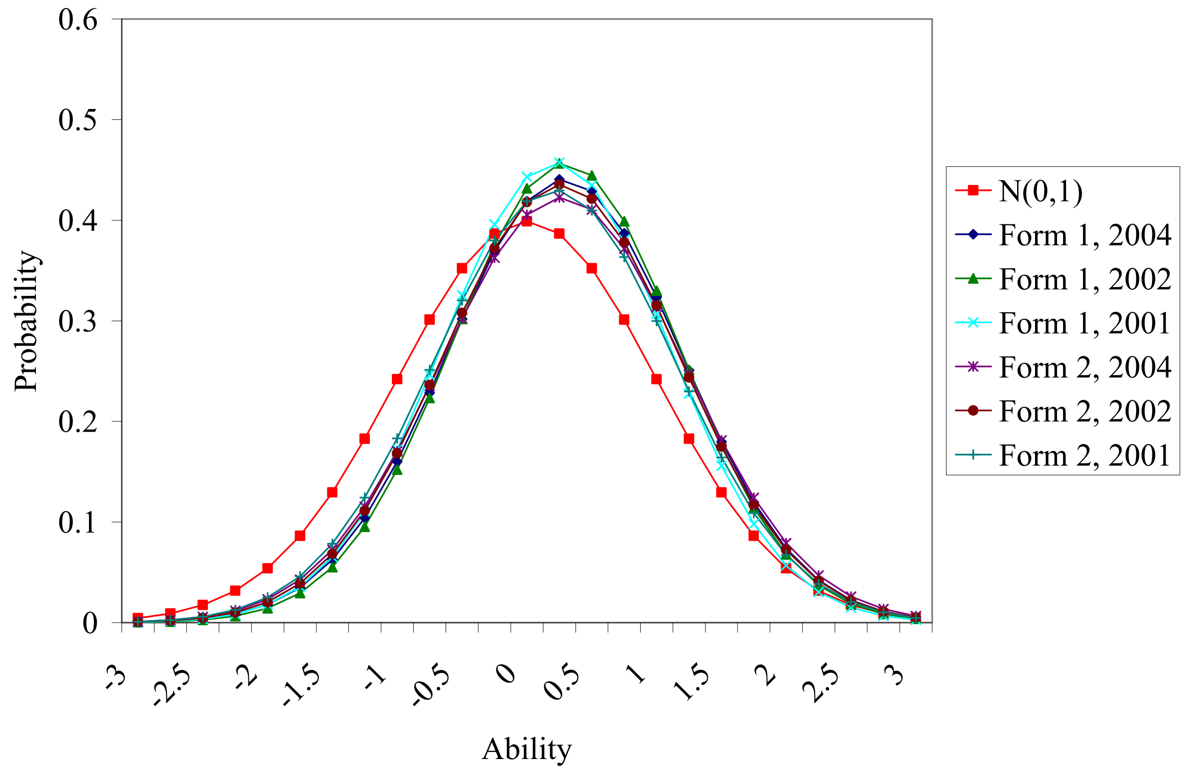


Figure D.3. Estimated Prior Distributions for WK

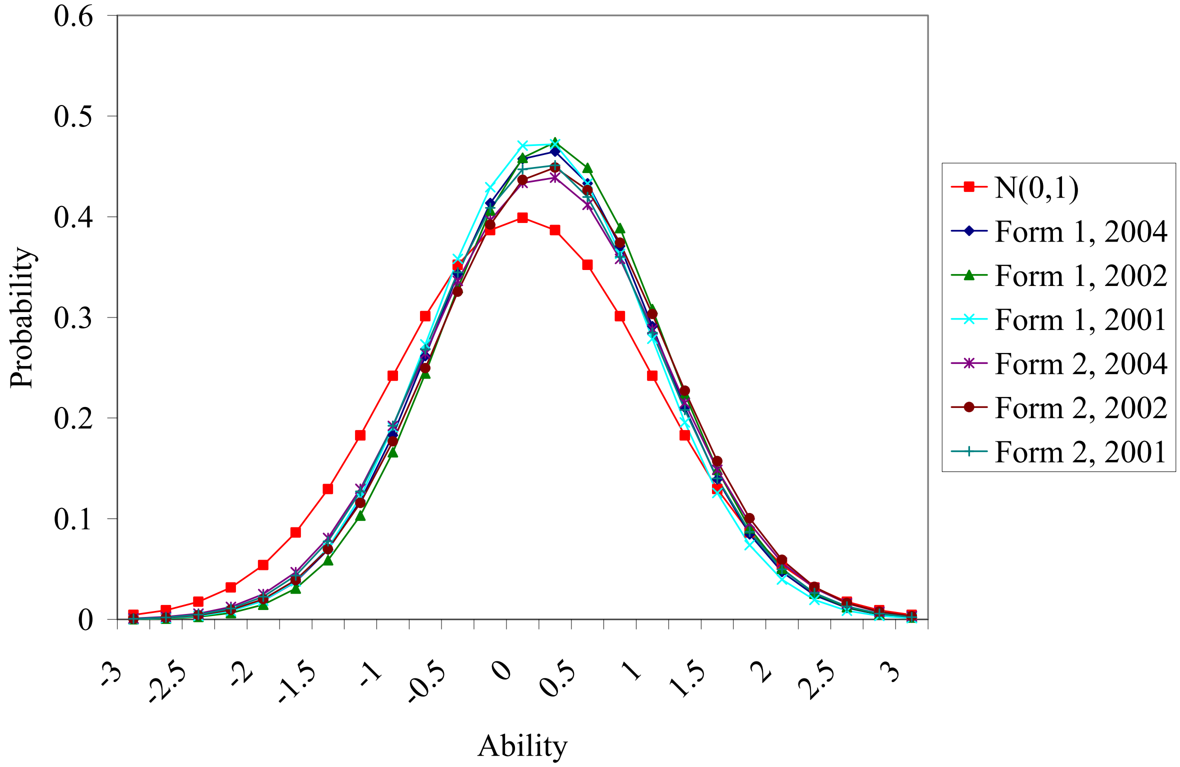


Figure D.4. Estimated Prior Distributions for PC

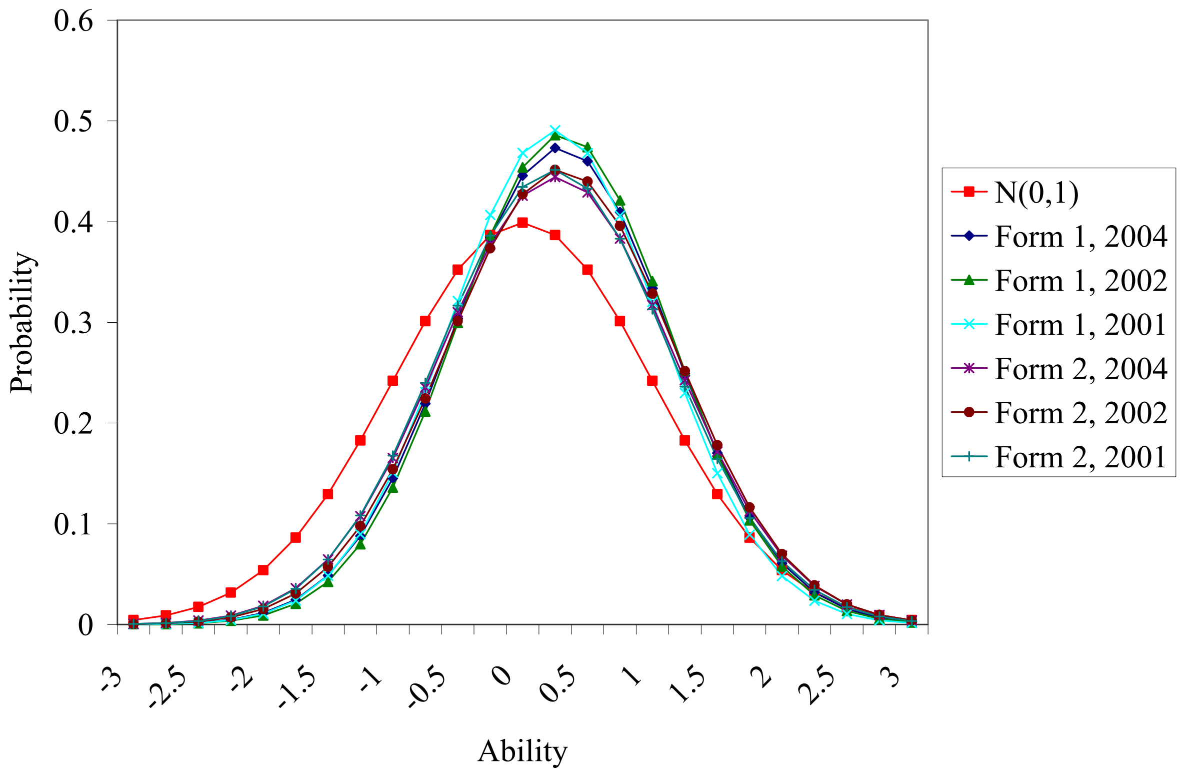


Figure D.5. Estimated Prior Distributions for MK

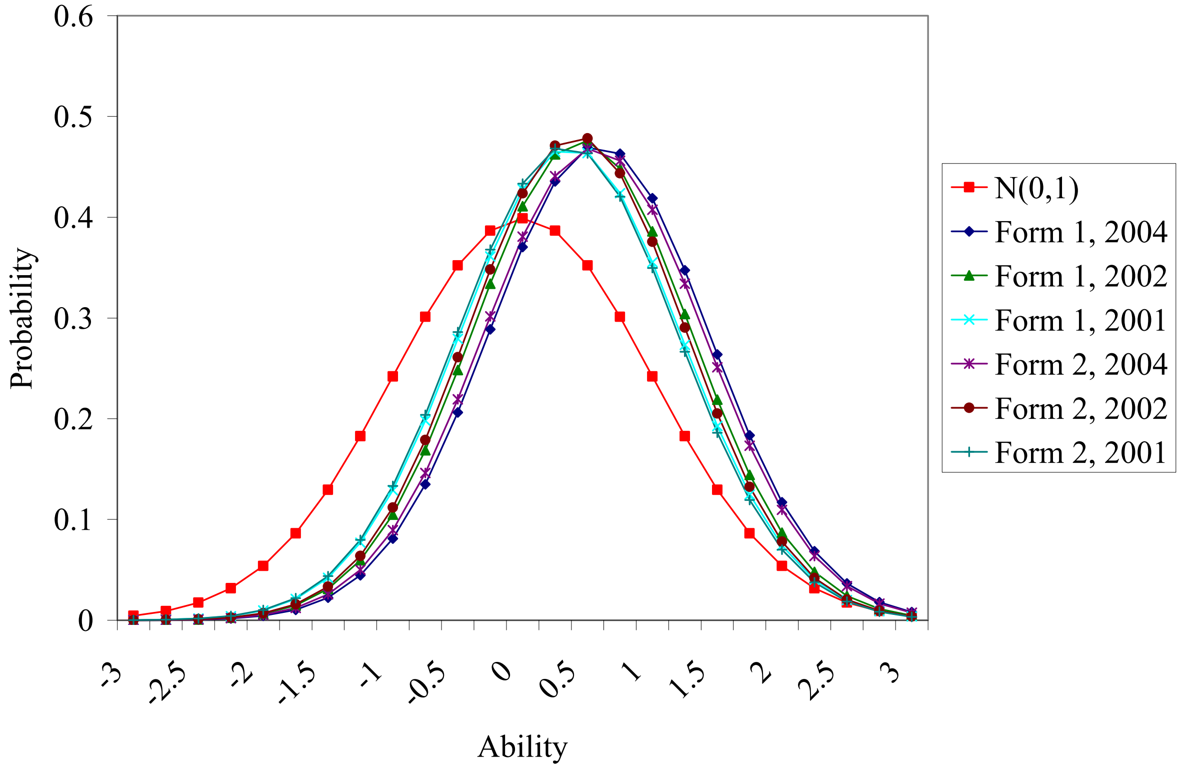


Figure D.6. Estimated Prior Distributions for EI

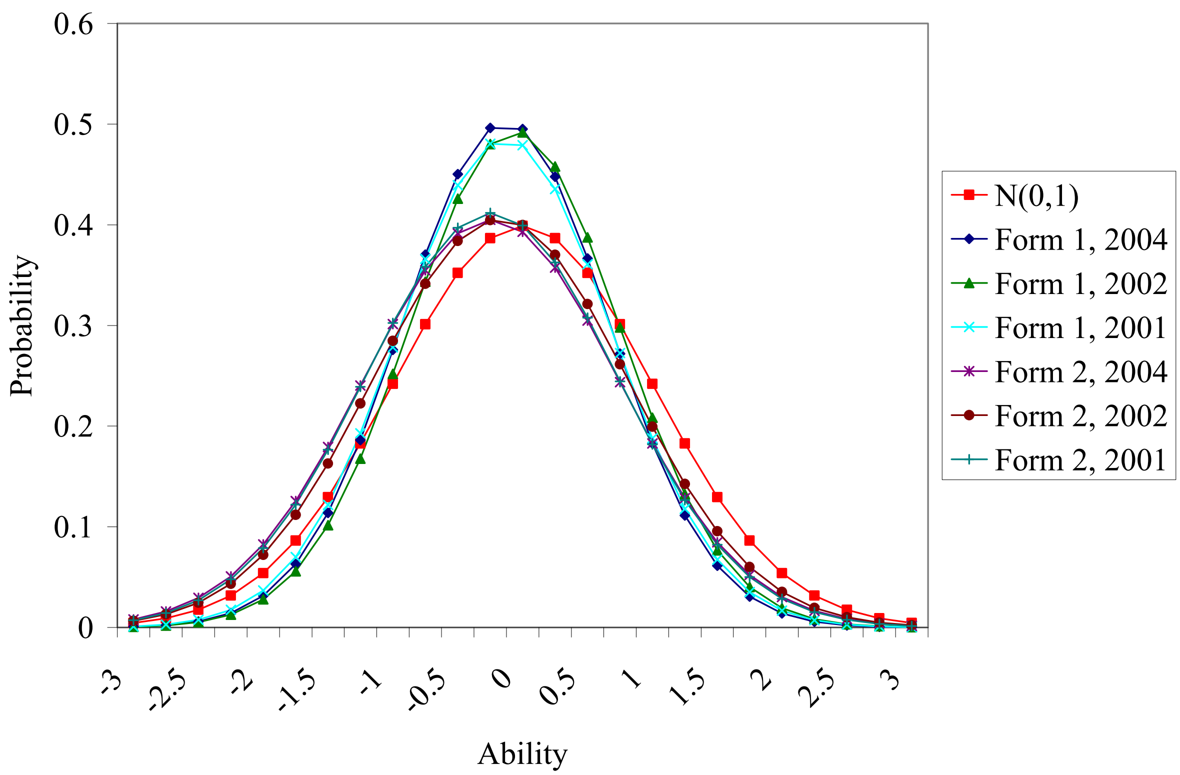


Figure D.7. Estimated Prior Distributions for AI

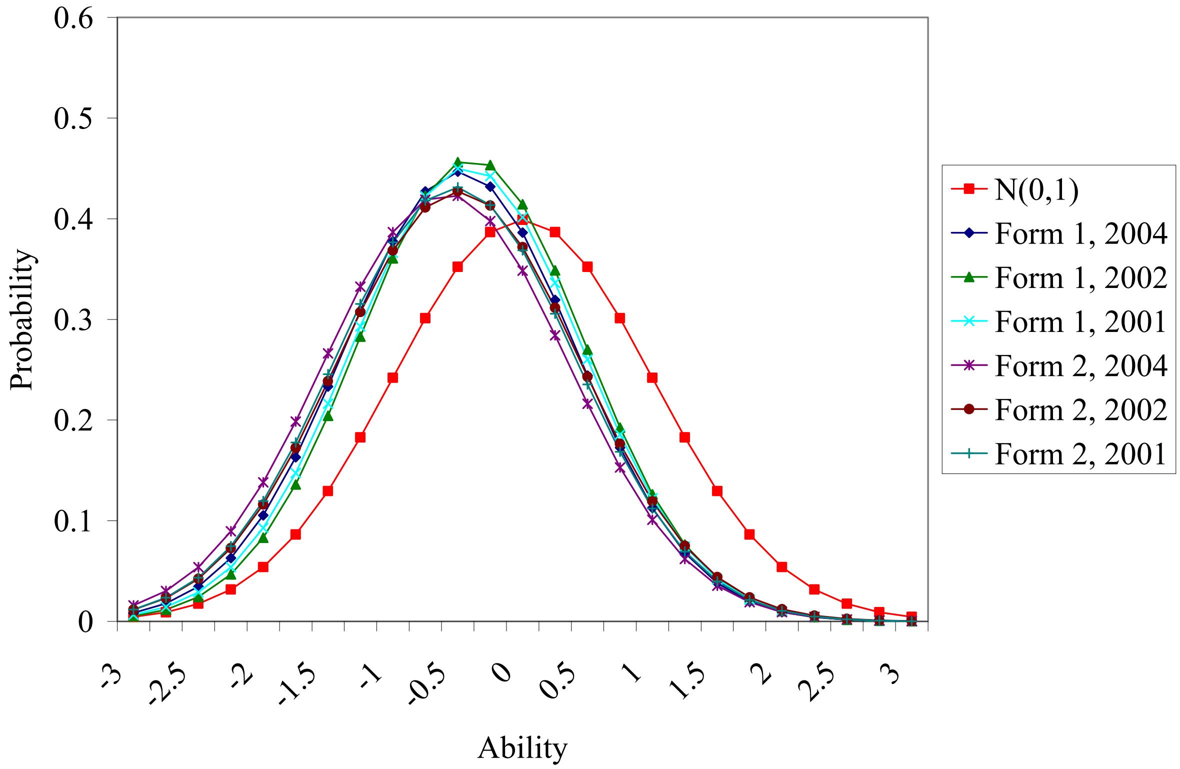


Figure D.8. Estimated Prior Distributions for SI

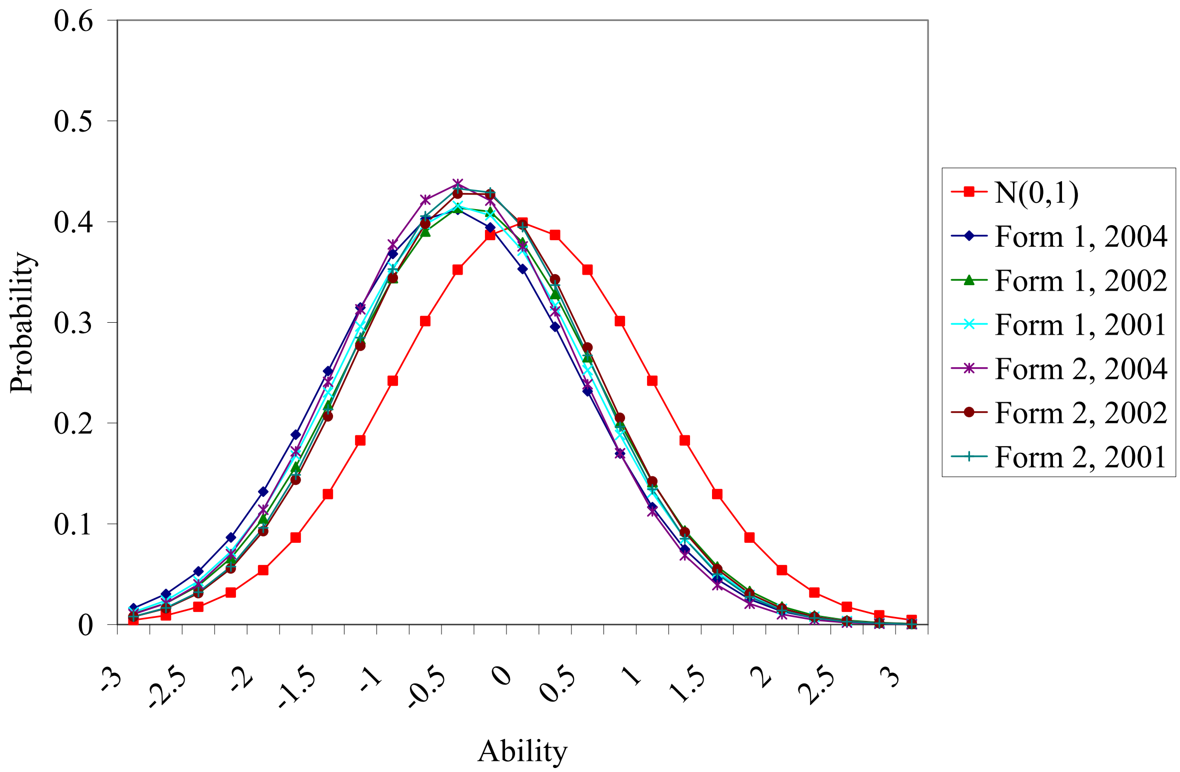


Figure D.9. Estimated Prior Distributions for MC

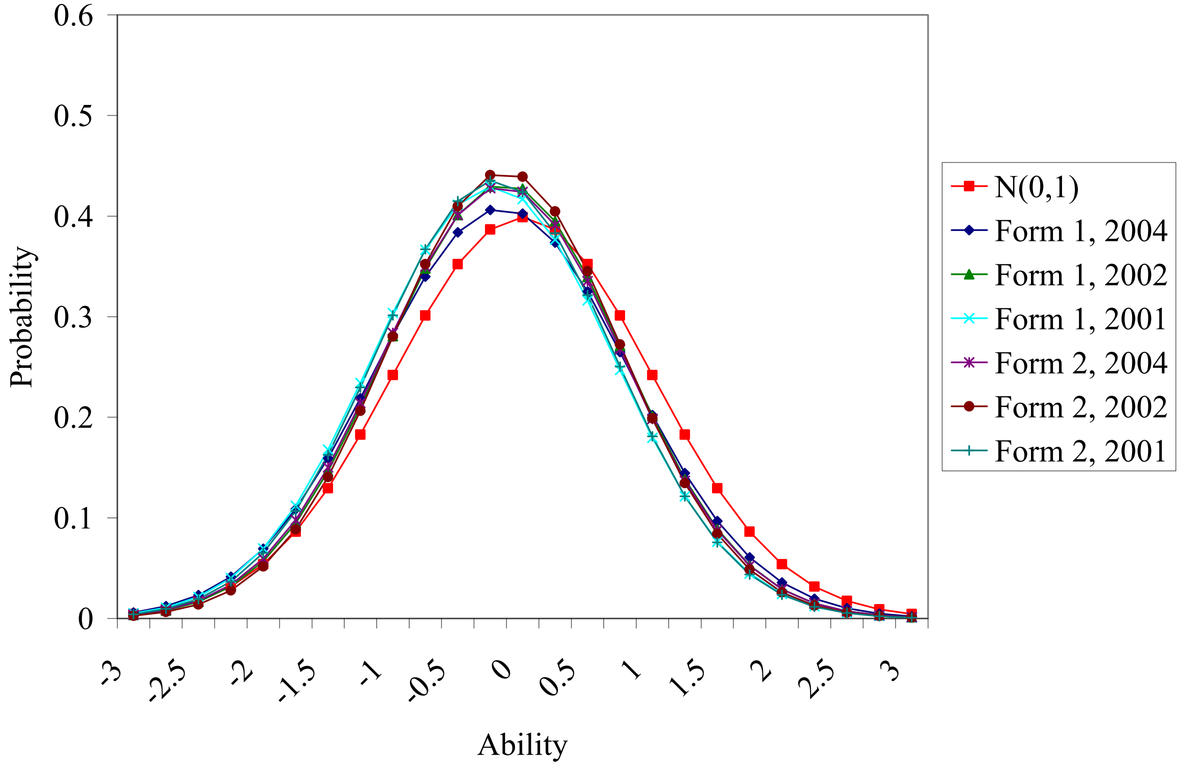


Figure D.10. Estimated Prior Distributions for AO