CHAPTER

11

Item Pool Development and Evaluation

Daniel O. Segall, Kathleen E. Moreno, and Rebecca D. Hetter

By the mid-1980s, an item pool had been constructed for use in the experimental CAT-ASVAB system (Chapter 9), and had been administered to a large number of subjects participating in research studies. However, this pool was ill-suited for operational use. First, many items had been taken from retired P&P-ASVAB forms (8, 9, and 10). Using these items in an operational CAT-ASVAB would degrade test security, since these items had broad exposure through the P&P testing program. In addition, the experimental CAT-ASVAB system contained only one form. For retesting purposes, it is desirable to have two parallel forms (consisting of non-overlapping item pools) to accommodate applicants who take the battery twice within a short time interval. To avoid practice and compromise effects, it is desirable for the second administered form to contain no common items with the initial form.

This chapter summarizes the procedures used to construct and evaluate the operational CAT-ASVAB item pools. Although specific reference is made to Forms 1 and 2, many of the same procedures were applied more recently to the development of other CAT-ASVAB forms. The first section describes the development of the primary and supplemental item banks. Additional sections discuss dimensionality, alternate form construction, and precision analyses. The final section summarizes important findings with general implications for CAT item pool development.

Development and Calibration

Primary Item Banks

The primary item banks for CAT-ASVAB Forms 1 and 2 were developed and calibrated by Prestwood, Vale, Massey, and Welsh (1985). The P&P-ASVAB Form 8A was used to outline the content of items written in each area. However, important differences between the development of adaptive and conventional (paper-and-pencil) item pools were noted, which led to several modifications in P&P-ASVAB test specifications:

- *Increased range of item difficulties*

Domain specifications were expanded to provide additional easy and difficult items.

- *Functionally independent items*

The Paragraph Comprehension test (as measured in P&P-ASVAB) typically contains reading passages followed by several questions referring to the same passage. Items of these types are likely to violate the assumption of local independence made by the standard unidimensional IRT model. Consequently, CAT-ASVAB items were written to have a single question per passage.

- *Unidimensionality*

In the P&P-ASVAB, auto and shop items are combined into a single test. However, to help satisfy the assumption of unidimensionality, Auto and

Shop Information were treated as separate content areas: Large non-overlapping pools were written for each, and separate item calibrations were conducted.

About 3,600 items (400 for each of the nine content areas) were written and pretested on a sample of recruits. The pretest was intended to screen about half of the items for inclusion in a large-sample item calibration study. Items administered in the pretest were assembled into 71 booklets, with each booklet containing items from a single content area. Examinees were given 50 minutes to complete all items in a booklet. Data from about 21,000 recruits were gathered, resulting in about 300 responses per item. IRT item parameters were estimated for each item using the ASCAL (Vale & Gialluca, 1985) computer program.[1]

For each content area, a subset of items with an approximately rectangular distribution of item difficulties was selected for a more extensive calibration study. This was accomplished from an examination of the IRT difficulty and discrimination parameters. Within each content area, items were divided into 20 equally spaced difficulty levels. Approximately equal numbers of items were drawn from each level, with preference given to the most highly discriminating items.

The surviving 2,118 items (about 235 items per content area) were assembled into 43 P&P test booklets, similar in construction to the pretest (each booklet containing items from a single content area; 50 minutes of testing per examinee). Data from 137,000 applicants were collected from 63 Military Entrance Processing Stations (MEPSs) and their associated Mobile Examining Team Sites (METSs) during late spring and early summer of 1983. Each examinee was given one experimental form and an operational P&P-ASVAB. After matching booklet and operational ASVAB data, about 116,000 cases remained for IRT calibration analysis (providing about 2,700 responses per item). Within each content area, all experimental and operational P&P-ASVAB items were calibrated jointly using the ASCAL computer program. This helped ensure that the item parameters were properly linked across booklets, and provided IRT estimates for several operational P&P-ASVAB forms on a common metric.

[1] ASCAL is a joint maximum-likelihood/modal-Bayesian item calibration program for the three-parameter logistic item response model.

Table 11-1 Linking Design

	P&P-ASVAB Form							
Calibration	**8A**	**8B**	**9A**	**9B**	**10A**	**10B**	**10X**	**10Y**
			Common Forms					
Primary			X	X	X	X	X	X
Supplemental	X	X	X	X	X	X		

Supplemental Item Bank

An analysis of the primary item banks (described below) indicated that two of the content areas, Arithmetic Reasoning (AR) and Word Knowledge (WK), had lower than desired precision over the middle ability range. Therefore, the item pools for these two content areas were supplemented with additional items taken from the experimental CAT-ASVAB system (166 AR items; and 195 WK items). The supplemental items were calibrated by Sympson and Hartmann (1985) using a modified version of LOGIST 2.b. Data for these calibrations were obtained from a MEPS administration of P&P booklets. Supplemental item parameters were transformed to the "primary item-metric" using the Stocking and Lord (1983) procedure. The linking design is shown in Table 11-1.

The primary calibration included six P&P-ASVAB forms; the supplemental calibration included a different but overlapping set of six P&P-ASVAB forms. The two sets of parameters were linked through the four forms common to both calibrations: 9A, 9B, 10A, and 10B. The specific procedure involved the computation of two test characteristic curves (TCCs), one based on the primary item calibration, and another based on the supplemental item calibration. The linear transformation of the supplemental scale that minimized the weighted sum of squared differences between the two TCCs was computed. The squared differences at selected ability levels were weighted by a $N(0,1)$ density function. This procedure was repeated for both AR and WK. All AR and WK supplemental IRT discrimination and difficulty parameters were transformed to the primary metric, using the appropriate transformation of scale.

Item Reviews

Primary and supplemental items were screened using several criteria. First, an Educational Testing Service (ETS) panel performed sensitivity and

quality reviews. The panel recommendations were then submitted to the Service laboratories for their comments. An Item Review Committee made up of NPRDC researchers reviewed the Service laboratories' and ETS reports and comments. When needed, the committee was augmented with additional NPRDC personnel having expertise in areas related to the item content under review. The committee reviewed the items and coded them as unacceptable, marginally unacceptable, less than optimal, and acceptable, in each of the two review categories (sensitivity and quality).

Item keys were verified by an examination of point-biserial correlations, computed for each distractor. Items with positive point-biserial correlations for incorrect options were identified and reviewed.

The display suitability of the item screens was evaluated for: (a) clutter (particularly applicable to PC), (b) legibility, (c) graphics quality, (d) congruence of text and graphics (do words and pictures match?), and (e) congruence of screen and booklet versions. In addition, items on the Hewlett Packard Integral Personal Computer (HP-IPC) screen were compared to those in the printed booklets. Displayed items were also examined for: (a) words split at the end of lines (no hyphenation allowed), (b) missing characters at the end of lines, (c) missing lines or words, (d) misspelled words, and (e) spelling discrepancies within the booklets. After the items were examined on the HP-IPC, reviewers presented their recommendations to a review group, which made final recommendations.

Options Format Study

The primary item pools for AR and WK consisted of multiple-choice items with five response alternatives, while the supplemental items had only four alternatives. If primary and supplemental items were combined in a single pool, examinees would probably receive a mixture of four- and five-choice items during the adaptive test. There was concern that mixing items with different numbers of response options within a test would cause confusion or careless errors by the examinee, and perhaps affect item difficulties.

The authors conducted a study to examine the effect of mixing four- and five-option items on computerized test performance. Examinees in this study were 1,200 male Navy recruits at the Recruit Training Center, San Diego, California. The task for each examinee was to answer a mixture of 4- and 5-option items. These included 32 WK items followed by 24 PC items administered by computer using a conventional nonadaptive strategy.

Subjects were randomly assigned to one of six conditions. Specific items administered in each condition for WK are displayed in Table 11-2. Examinees assigned to Conditions A or B received items of one type exclusively: Examinees assigned to Condition A received items 1–32 (all 5-option items), examinees assigned to Condition B received items 33–64 (all 4-option items). Items in Conditions A and B were selected to span the range of difficulty. Note that 4- and 5-option items were paired {1,33}, {2,34}, {3,35}, . . . so that items in the same position in the linear sequence would have similar item response functions (and consequently similar difficulty and discrimination levels). Examinees assigned to Condition C received alternating sequences of 5- and 4-choice items (5, 4, 5, 4, . . .). Examinees assigned to Condition D received a test in which every fourth item was a 4-option item (5, 5, 5, 4, 5, 5, 5, 4, ...). In Condition E, every 8th item administered was a 4-option item. Finally, in Condition F, an equal number of randomly selected 4- and 5-option items were administered to each examinee. The first item administered was randomly selected from {1 or 33}, the second item was selected from {2 or 34}, etc. An example assignment for this condition is given in the last column of Table 11-2. Note for this condition, assignments were generated independently for each examinee. An identical design was used for PC, except that only 24 items were administered to each examinee. Three different outcome measures were examined to assess the effects of mixing item formats: item difficulty, test difficulty, and response latency.

Item difficulty. For Conditions C, D, E, and F, item difficulties (proportion of correct responses) were compared with those of the corresponding items in the Control Conditions (A or B). For example, comparison of difficulty values in Condition C included pairs: {Condition C, Item 1} with {Condition A, Item 1}; {Condition C, Item 34} with {Condition B, Item 34}; etc. The significance of the difference between pairs of item difficulty values were tested using a 2×2 chi-square analysis. For WK, only seven of the 160 comparisons (about 4.4%) produced significant differences (at the .05 alpha level). For PC, only one of the 120 comparisons of item difficulty was significant.

Test difficulty. For examinees in Conditions C, D, and E, two number-right scores were

Table 11-2 Options Format Study: WK Item Lists Presented in Control and Experimental Conditions

Control		Experimental			
Condition A (5-Option)	Condition B (4-Option)	Condition C (Mixed: 1:1)	Condition D (Mixed: 3:1)	Condition E (Mixed: 7:1)	Condition F (Random: 1:1)
1	33	1	1	1	1
2	34	34	2	2	2
3	35	3	3	3	3
4	36	36	36	4	36
5	37	5	5	5	37
6	38	38	6	6	6
7	39	7	7	7	39
8	40	40	40	40	40
9	41	9	9	9	41
10	42	42	10	10	10
11	43	11	11	11	43
12	44	44	44	12	12
13	45	13	13	13	13
14	46	46	14	14	14
15	47	15	15	15	47
16	48	48	48	48	16
17	49	17	17	17	49
18	50	50	18	18	50
19	51	19	19	19	51
20	52	52	52	20	20
21	53	21	21	21	21
22	54	54	22	22	54
23	55	23	23	23	55
24	56	56	56	56	24
25	57	25	25	25	57
26	58	58	26	26	58
27	59	27	27	27	27
28	60	60	60	28	28
29	61	29	29	29	29
30	62	62	30	30	30
31	63	31	31	31	63
32	64	64	64	64	64

computed: One based on 4-option items, and another based on 5-option items. Number-right scores from corresponding items were computed for examinees in the Control conditions A and B. The number of items entering into each score for each condition are displayed in the second and fifth columns of Table 11-3. The significance of the difference between mean number-right scores across the Experimental and Control groups was tested using an independent groups t statistic. The results are displayed in Table 11-3. None of the comparisons displayed significant results at the .05 alpha level.

Response latencies. For examinees in Conditions C, D, and E, two latency measures were computed: One based on 4-option items, and another based on 5-option items. Latency measures were also computed from corresponding items in the Control conditions A and B. Mean latencies were compared across the Experimental and Control groups (Table 11-3). None of the comparisons displayed significant results at the .05 alpha level.

Discussion. Mixing items with different numbers of response options produced no measurable effects on item or test performance. This result

Table 11-3 Options Format Study: Significance Tests for Test Difficulties and Response Latencies

Condition	Word Knowledge			Paragraph Comprehension		
		t-value			*t*-value	
	No. Items	Difficulty	Latency	No. Items	Difficulty	Latency
Comparison With 5-Option Control						
Condition C	16	.06	−.85	12	−.08	−1.77
Condition D	24	−1.09	.47	18	−.21	−.64
Condition E	28	−.24	−.98	21	−1.82	.67
Comparison With 4-Option Control						
Condition C	16	−1.83	1.49	12	1.30	−.72
Condition D	8	−1.35	1.84	6	−.98	−1.92
Condition E	4	1.35	−.07	3	−1.40	−.28

differed from those reported by Brittain and Vaughan (1984), who studied the effects of mixing items with different numbers of options on a P&P version of the Army Skills Qualification Test. They predicted errors would increase when an item with *n* answer options followed an item with more than *n* answer options, where errors were defined as choosing nonexistent answer options. Consistent with their hypothesis, mixing items with different numbers of answer options caused an increase in errors.

Likely explanations for the different findings between the current study and the Brittain and Vaughan (1984) study involve differences in medium (computer versus P&P). In the Brittain and Vaughan study, examinees answered questions using a standard 5-option answer sheet for all items, making the selection of a nonexistent option possible. However, in the current study, software features were employed which helped eliminate erroneous responses. (These software features are common to both the current study and the CAT-ASVAB system.)

First, after the examinee makes a selection among response alternatives, he or she is required to confirm the selection. For example, if the examinee selects option "D," the system responds with:

> If "D" is your answer press ENTER.
> Otherwise, type another answer.

That is, the examinee is informed about the selection that was made, and given an opportunity to change the selection. This process would tend to minimize the likelihood of careless errors.

A second desirable feature incorporated into the CAT-ASVAB software (and included in the options format study) was the sequence of events following an "invalid-key" press. Suppose, for example, that a particular item had only four response alternatives (A, B, C, and D) and the examinee selects "E" by mistake. The examinee would see the messages:

> You DID NOT type A, B, C, or D.
> Enter your answer (A, B, C, or D)

Note that if an examinee accidentally selects a nonexistent option (i.e., "E"), the item is not scored incorrect; instead, the examinee is given an opportunity to make another selection. This feature would also reduce the likelihood of careless errors. These software features, along with the empirical results of the options format study, addressed the major concerns about mixing four- and five-choice items.

Dimensionality

One major assumption of the IRT item selection and scoring procedures used by CAT-ASVAB is that performance on items within a given content area can be characterized by a unidimensional latent trait or ability. Earlier research showed that IRT estimation techniques are robust against minor violations of the unidimensionality assumption, and that unidimensional IRT parameter estimates have many practical applications in multidimensional item pools (Reckase, 1979; Drasgow & Parsons, 1983, Dorans & Kingston, 1985). However, violations of the unidimensional adaptive testing model may have serious implications for validity and test fairness. Because of the adaptive nature

Table 11-4 Treatment Approaches for Multidimensional Item Pools

Approach	Calibration	Item Selection	Scoring
1. Unidimensional Treatment	Combined calibration containing items of each content type	No constraints placed on item content for each examinee	A single IRT ability estimate computed across items of different content using the unidimensional scoring algorithm
2. Content Balancing	Combined calibration containing items of each content type	Constraints placed on the number of items drawn from each content area for each examinee	A single IRT ability estimate computed across items of different content using the unidimensional scoring algorithm
3. Pool Splitting	Separate calibrations for items of each content	Separate adaptively tailored tests for each content area	Separate IRT ability estimates for each content area

of the test, and the IRT scoring algorithms, multidimensionality may lead to observed scores which represent a different mixture of the underlying unidimensional constructs than intended. This could alter the validity of the test. Furthermore, the application of the unidimensional model to multidimensional item pools may produce differences in the representation of dimensions among examinees. Some examinees may receive items measuring primarily one dimension, while others receive items measuring another dimension. This raises issues of test fairness. If the pool is multidimensional, two examinees (with the same ability levels) may be administered items measuring two largely different constructs, and receive widely discrepant scores.

In principle, at least three approaches exist for dealing with multidimensional item pools (Table 11-4). These approaches differ in the item selection and scoring algorithms, and in the item calibration design:

1. *Unidimensional Treatment.* This option essentially ignores the dimensionality of the item pools in terms of item calibration, item selection, and scoring. A single item calibration containing items spanning all content areas is performed to estimate the IRT item parameters. No content constraints are placed on the selection of items during the adaptive sequence—items are selected on the basis of maximum information. Intermediate and final scoring are performed according to the unidimensional IRT model, and a single score is obtained based on items spanning all content areas.

2. *Content Balancing.* This approach balances the numbers of administered items from targeted content areas. A single item calibration containing items spanning all content areas is performed to estimate the IRT item parameters. During the adaptive test, items are selected from *content-specific* subpools in a fixed sequence. For example, the content balancing sequence for General Science could be LPLPLPLPLPLPLPL (L = Life Science, P = Physical Science). Accordingly, the first item administered would be selected from among the candidate Life Science items. The second item administered would be selected from the physical science items, and so forth. Within each targeted content area, items are selected on the basis of IRT item information. Intermediate and final scores are based on the unidimensional ability estimator computed from items spanning all content areas.

3. *Pool Splitting.* Item pools for different dimensions are constructed and calibrated separately. For each content area, separate adaptive tests are administered and scored. It is then usually necessary to combine final scores on the separate adaptive tests to form a single composite measure that spans the separately measured content areas.

For each item pool, a number of criteria were considered in determining the most suitable dimensionality-approach, including: (a) statistical factor significance, (b) factor interpretation, (c) item difficulties, and (d) factor intercorrelations. The relation between these criteria and the recommended approach is summarized in Table 11-5.

Table 11-5 Decision Rules for Approaches to Dimensionality

Case	Statistical Factor Sig.	Interpretable Factors	Overlapping Item Difficulties	Factor Correlations	Approach
1.	No	—	—	—	Unidimensional
2.	Yes	Yes	Yes	High	Content Bal.
3.	Yes	Yes	Yes	Low	Split Pool
4.	Yes	Yes	No	—	Unidimensional
5.	Yes	No	Yes	—	Unidimensional
6.	Yes	No	No	—	Unidimensional

Statistical Factor Significance

The first, and perhaps most important criterion for selecting the dimensionality-approach is the factor structure of the item pool. If there is empirical evidence to suggest that responses of an item pool are multidimensional, then content-balancing or pool-splitting should be considered. In the absence of such evidence, item pools should be treated as unidimensional. Such empirical evidence can be obtained from factor analytic studies of item responses using one of several available approaches, including TESTFACT (Wilson, Wood, & Gibbons, 1991) and NOHARM (Fraser, 1988). The full item-information procedure used in TESTFACT allows the statistical significance of multidimensional solutions to be tested against the unidimensional solution using a hierarchical likelihood ratio procedure.

This strong empirical emphasis recommended here is not shared by all adaptive testing programs. The adaptive item selection algorithm used in the CAT-GRE (Stocking & Swanson, 1993) incorporates both item information and test plan specifications. The test plans are based on expert judgments of content specialists. Accordingly, there is likely to be a disconnect between the test plan specifications and the empirical dimensionality of the item pools. This can lead to situations where constraints are placed on the presentation of items that are largely unidimensional. In general, overly restrictive content-based constraints on item selection will lead to the use of less informative items, and ultimately to test scores with lower precision.

Factor Interpretation

According to a strictly empirical approach, the number of factors could be determined by statistical considerations, and items could be allocated to areas based on their estimated loadings. Items could be balanced with respect to these areas defined by the empirical analysis. However, a major drawback with this approach is the likelihood of meaningless results, both in terms of the number of factors to be balanced, and in the allocation of items to content areas. Significance tests applied to large samples would almost certainly lead to high-dimensionality solutions, regardless of the strength of the factors. Furthermore, there is no guarantee that the rotated factor solution accurately describes the underlying factors.

The alternative judgmental approach noted above would divide the pool into areas on the basis of expert judgments. The major problem with this approach is that without an examination of empirical data, it is not possible to determine which content areas affect the dimensionality of the pool. Choice of content areas could be defined at several arbitrary levels. As Green et al. (1982) suggest, "There is obviously a limit to how finely the content should be subdivided. Each item is to a large extent specific."

In CAT-ASVAB development, we formed a decision rule based on a compromise between the empirical and judgmental approaches. If a pool was found to be statistically multidimensional, items loading highly on each factor were inspected for similarity of content. If agreement between factor solutions and content judgments was high, then balancing was considered, otherwise balancing was not considered.

Item Difficulties

Another important criterion for selecting among dimensionality-approaches concerns the overlap of item difficulties associated with items of each content area. The overlap of item difficulties can

provide some clues about the causes of the dimensionality, and suggest an appropriate remedy. Lord (1977) makes an important observation:

> Suppose, to take an extreme example, certain items in a test are taught to one group of students and not taught to another, while other items are taught to both groups. This way of teaching increases the dimensionality of whatever is measured by the test. If items would otherwise have been factorially unidimensional, this way of teaching will introduce additional dimensions. (p. 24)

If a pool contains some items with material exposed to the entire population (say nonacademic content), and other items are taught to a subpopulation (in school—academic content), then we would expect to find statistically significant factors with easy items loading on the nonacademic factor, and moderate to difficult items loading on the academic factor. Application of the unidimensional item selection and scoring algorithms would result in low ability test-takers receiving easy (nonacademic) items, and moderate to high ability test-takers receiving academic items. Thus the unidimensional treatment would appropriately tailor the content of the items according to the standing of the test-taker along the latent dimension. Note that content balancing in this situation could substantially reduce the precision of the test scores. For example, if an equal number of items from each content area were administered to each examinee, then low ability examinees would receive a large number of uninformative difficult items; and conversely, high ability examinees would receive a large number of uninformative easy items.

We would expect to observe a different pattern of item difficulty values if substantially non-overlapping subgroups were taught different material. In this instance, we would expect to observe two or more factors defined by items with overlapping difficulty values (falling within a common range). Here, an appropriate remedy would involve content balancing or pool-splitting, since different dimensions represent knowledge of somewhat independent domains.

Factor Correlations

A final consideration for selecting among dimensionality-approaches concerns the magnitude of the correlation between latent factors. Different approaches might be desirable depending on the correlation between factors estimated in the item factor analysis. If factors are highly correlated, then content balancing may provide the most satisfactory results. In this instance, the unidimensional model used in conjunction with content balancing is likely to provide an adequate approximation for characterizing item information, and for estimating latent ability.

If the correlations among factors are found to be low or moderate, then the usefulness of the unidimensional model for characterizing item information and estimating latent abilities is questionable. When the factors have low correlations, pool-splitting is likely to provide the best remedy. Separate IRT calibrations should be performed for items of each factor; separate adaptive tests should be administered; and final adaptive test scores can be combined to form a composite measure representing the standing among examinees along the latent composite dimension.

Choosing Among Alternative Approaches

Table 11-5 summarized different possible outcomes and the recommended approach for each. If an item factor analysis provides no significant second, or higher order factors, then the pool should be treated as unidimensional (Case 1). If statistically significant higher order factors are identified, these factors relate to item content, and item difficulties of each content span a common range, then consideration should be given to content balancing (Case 2, if the factor intercorrelations are high), or to pool-splitting (Case 3, if the factor intercorrelations are low to moderate). For reasons given above, if the statistical factors are not interpretable (Case 5 and 6), or if the item difficulty values of each content area span non-overlapping ranges (Case 4 and 6), then unidimensional treatment may provide the most useful approach.

Results and Discussion

In earlier studies of the Auto-Shop content area, a decision was made to apply the pool-splitting approach: This content area was split into separate auto and shop item pools (Case 3, Table 11-5). As described in an earlier section, these pools were calibrated separately. The decision to split these pools was based on the moderately high correlation among the auto and shop dimensions. In the analysis described below, the auto and shop pools

were examined separately, and subjected to the same analyses as other pools.

The first step in the dimensionality analysis involved factor analyses using item data (Prestwood et al., 1985). Empirical item responses were analyzed using the TESTFACT computer program (Muraki, 1984), which employs full information item factor analysis based on IRT (Bock & Aitkin, 1981). While the program computes item difficulty and item discrimination parameters, guessing parameters are treated as known constants and must be supplied to the program. For these analyses, the guessing parameters estimated by Prestwood et al., were used. For all analyses, a maximum of four factors were extracted, using a stepwise procedure. An item pool was considered statistically multidimensional if a change in chi-square (between the one-factor solution and the two-factor solution) was statistically significant (at the .01 alpha level). If the change in chi-square for the two-factor solution was significant, the three- and four-factor solutions were also examined for significant changes in chi-square. Since items within a pool were divided into separate booklets for data collection purposes, all items within a pool could not be factor analyzed at once. Therefore, subsets of items (generally, all items in one booklet) were analyzed. The number of statistically significant factors found across booklets was not necessarily identical. In such cases, the factor solutions examined were the number found in the majority of the booklets. The number of statistically significant factors found for each item pool is summarized in Table 11-6. For those item pools showing statistical evidence of multidimensionality, items were reviewed to determine whether the pattern of factor loadings was related to content, mean difficulty parameters were computed by content area, and factor intercorrelations were examined. These results are displayed in Table 11-6.

Based on the factor analyses, PC and MC were found to be unidimensional (Case 1, Table 11-5). All other item pools were multidimensional, with GS and MK having four factors and AR, WK, AI, SI, and EI having two factors. For those areas having two factors, the pattern of factor loadings was readily apparent. Items that loaded highly on the first factor were nonacademic items (i.e., taught to the whole group through everyday experiences). Items that loaded highly on the second factor were academic items (i.e., taught to a subgroup through classroom instruction or specialized experience). Means of IRT difficulty parameters for academic and nonacademic items are displayed in Table 11-7. As indicated, the mean difficulty values for nonacademic items were much lower than those for academic items. Accordingly, AR, WK , AI, SI, and EI were treated as unidimensional item pools (Case 4, Table 11-5).

The GS pool appeared, in part, to follow a different pattern than the five pools discussed above. An examination of the factor solutions and item content provided some evidence for a four-factor solution interpreted as (a) nonacademic, (b) life science, (c) physical science, and (d) chemistry. This interpretation is supported by the fact that many

Table 11-6 Dimensionality of CAT-ASVAB Item Pools

Item Pool	No. Significant Factors	Interpretable Factors	Overlapping Item Difficulties	Factor Correlations	Case	Approach
GS	4	Yes	Yes	High	2	Content Bal.
AR	2	Yes	No	—	4	Unidimensional
WK	2	Yes	No	—	4	Unidimensional
PC	1	—	—	—	1	Unidimensional
AI	2	Yes	No	—	4	Unidimensional
SI	2	Yes	No	—	4	Unidimensional
MK	4	No	Yes	—	5	Unidimensional
MC	1	—	—	—	1	Unidimensional
EI	2	Yes	No	—	4	Unidimensional

Table 11-7 Mean IRT Item Difficulty (*b*) Parameters

Item Content	AR	WK	AI	SI	EI
Nonacademic	−2.37	−2.30	−2.28	−2.15	−1.51
Academic	.30	.47	.48	.57	.61

Table 11-8 Item Pools Evaluated in Precision Analyses

Condition	Content Area	Label	Form	Supplemented	Target Exposure Rate
1	GS	GS-1	1	No	1/3
2	GS	GS-2	2	No	1/3
3	AR	AR-1	1	No	1/6
4	AR	AR-2	2	No	1/6
5	AR	AR_s-1	1	Yes	1/6
6	AR	AR_s-2	2	Yes	1/6
7	WK	WK-1	1	No	1/6
8	WK	WK-2	2	No	1/6
9	WK	WK_s-1	1	Yes	1/6
10	WK	WK_s-2	2	Yes	1/6
11	PC	PC-1	1	No	1/6
12	PC	PC-2	2	No	1/6
13	AI	AI-1	1	No	1/3
14	AI	AI-2	2	No	1/3
15	SI	SI-1	1	No	1/3
16	SI	SI-2	2	No	1/3
17	MC	MC-1	1	No	1/3
18	MC	MC-2	2	No	1/3
19	MK	MK-1	1	No	1/6
20	MK	MK-2	2	No	1/6
21	EI	EI-1	1	No	1/3
22	EI	EI-2	2	No	1/3

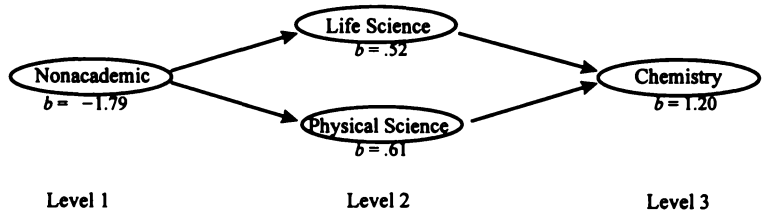

Figure 11-1
General Science Dual Track Instruction.

high schools offer a multiple-track science program (Figure 11-1). At Level 1, students have little or no formal instruction. At Level 2, some students receive training in life science, while others receive physical science training. Finally, at Level 3, some members of both groups are instructed in chemistry. Notice that each higher level contains only a subset of students contained in the levels directly below it. For example, not everyone completing a life science or a physical science course will receive instruction in chemistry. The mean IRT item difficulty values (displayed in Figure 11-1) also support this interpretation of dimensionality. The life science and physical science items are of moderate (and approximately equal) difficulty. The chemistry items appear to be the most difficult, and nonacademic items least difficult. These findings are supportive balancing content among life and physical science items (Case 2, Table 11-5). Nonacademic and chemistry items should be administered to examinees of appropriate ability levels. (See Chapter 12 for additional details on the GS content balancing algorithm.)

For MK, the pattern of factor loadings associated with the two-, three-, or four-factor solutions could not be associated with item content. Conse-

quently, the MK item pool was treated as unidimensional (Case 5, Table 11-5).

Alternate Forms

In developing the item pools for CAT-ASVAB, it was necessary to create two alternate test forms so that applicants could be retested on another form of CAT-ASVAB. Once the item screening procedures were completed, items within each content area were assigned to alternate pools. Pairs of items with similar information functions were identified, and assigned to alternate pools. The primary goal of the alternate form assignment was to minimize the weighted sum-of-squared differences between the two pool information functions. (A pool information function was computed from the sum of the item information functions.) The squared differences between pool information functions were weighted by a $N(0,1)$ density.

The procedure used to create the GS alternate forms differed slightly from the other content areas because of the content balancing requirement. GS items were first divided into physical, life, and chemistry content areas. Domain specifications provided by Prestwood, Vale, Massey, & Welsh (1985) were used for assignment to these content areas. Once items had been assigned to a content area, alternate forms were created separately for each of the three areas.

Precision Analyses

Precision is an important criterion for judging the adequacy of the items pools, since it depends in large part on the quality of the pools. Precision analyses were conducted separately for the 22 item pools displayed in Table 11-8. The content area and form are listed in columns two and four. The target exposure rate (for the battery, i.e, across the two forms) is provided in the last column. This target was used to compute exposure control parameters according to the Sympson-Hetter algorithm (Chapter 13). The fifth column shows whether the pool included supplemental items. The third column provides a descriptive label for each condition used in the text and tables.

As would be expected, the results of any precision analysis would show various degrees of precision among the CAT-ASVAB tests. But how much precision is enough? The precision of the P&P-ASVAB offers a useful baseline. It is desirable for CAT-ASVAB to match or exceed P&P-ASVAB precision. Accordingly, precision criteria were computed for both P&P-ASVAB and CAT-ASVAB.

It is important to evaluate the impact of using the CAT-ASVAB item selection and scoring algorithm on precision, since the precision of adaptive test scores depends on both, the quality of the item pools, and on the adaptive testing procedures. The specific item selection and scoring procedures used are described in Chapter 12. For each adaptively administered test, the precision of the Bayesian modal estimate was evaluated. For each item pool, two measures of precision were examined: (a) score information, and (b) reliability.

Score Information

Score information functions provide one criterion for comparing the relative precision of the CAT-ASVAB with the P&P-ASVAB. Birnbaum (1968, Section 17.7) defines the information function for any score y to be

$$I\{\theta, y\} \equiv \frac{\left(\frac{d}{d\theta}\mu_{y|\theta}\right)^2}{\mathrm{Var}(y|\theta)}. \qquad (11\text{-}1)$$

This function is by definition inversely proportional to the square of the length of the asymptotic confidence interval for estimating ability θ from score y. For each content area, information functions can be compared between the CAT-ASVAB and the P&P-ASVAB. The test with greater information at a given ability level will possess a smaller asymptotic confidence interval for estimating θ.

CAT-ASVAB score information functions. The score information functions (SIFs) for each CAT-ASVAB item pool were approximated from simulated test sessions. For a given pool, simulations were repeated independently for 500 examinees at each of 31 different θ levels. These θ levels were equally spaced along the $[-3, +3]$ interval. At each θ level, the mean m and variance s^2 of the 500 final scores were computed. The information function at each selected level of θ can be approximated from these results, using (Lord, 1980a, eq. 10-7)

$$I\{\theta, \hat{\theta}\} \approx \frac{[m(\hat{\theta}|\theta_{+1}) - m(\hat{\theta}|\theta_{-1})]^2}{(\theta_{+1} - \theta_{-1})^2 s^2(\hat{\theta}|\theta_0)}, \qquad (11\text{-}2)$$

where θ_{-1}, θ_0, θ_{+1} represent the successive levels of θ. However, the curve produced by this approxi-

mation often appears jagged, with many local variations. To reduce this problem, information was approximated by

$$I\{\theta, \hat{\theta}\} \approx \frac{\left[\frac{m(\hat{\theta}|\theta_{+1}) + m(\hat{\theta}|\theta_{+2})}{2} - \frac{m(\hat{\theta}|\theta_{-1}) + m(\hat{\theta}|\theta_{-2})}{2}\right]^2}{\left[\frac{\theta_{+1} + \theta_{+2}}{2} - \frac{\theta_{-1} + \theta_{-2}}{2}\right]^2 \left[\frac{1}{5}\sum_{k=-2}^{+2} s(\hat{\theta}|\theta)\right]^2} \tag{11-3}$$

$$= \frac{25[m(\hat{\theta}|\theta_{+2}) + m(\hat{\theta}|\theta_{+1}) - m(\hat{\theta}|\theta_{-1}) - m(\hat{\theta}|\theta_{-2})]^2}{(\theta_{+2} + \theta_{+1} - \theta_{-1} - \theta_{-2})^2 \left[\sum_{k=-2}^{+2} s(\hat{\theta}|\theta_k)\right]^2} \tag{11-4}$$

where θ_{-2}, θ_{-1}, θ_0, θ_{+1}, θ_{+2} represent successive levels of θ. This approximation results in a moderately smoothed curve with small local differences.

P&P-ASVAB Score information functions. The P&P-SIF for a number right score x was computed by (Lord, 1980a, eq. 5-13)

$$I\{\theta, x\} = \frac{\left[\sum_{i=1}^{n} P_i'(\theta)\right]^2}{\sum_{i=1}^{n} P_i(\theta) Q_i(\theta)} \tag{11-5}$$

This function was computed for each content area by substituting the estimated P&P-ASVAB (9A) parameters for those assumed to be known in Equation (11-5).

A special procedure was used to compute SIF for AS since this test is represented by two tests in CAT-ASVAB. The AS-P&P (9A) test was divided into AI and SI items. SIFs (eq. 11-5) were computed separately for these AI-P&P and SI-P&P items to simplify comparisons with the corresponding CAT-ASVAB SIFs. Parameters used in the computation of these SIFs were taken from the joint calibrations of P&P-ASVAB and CAT-ASVAB items. In these calibrations, AS-P&P items were separated and calibrated among CAT-ASVAB items of corresponding content (i.e., AI-P&P items were calibrated with AI-CAT, and SI-P&P with SI-CAT items). However, two AS-P&P (9A) items appeared to overlap in AI/SI content, and appeared in both AI and SI calibrations. For computations of score information, these two items were included in both AI-P&P and SI-P&P information functions.This represents a conservative approach (favoring the P&P-ASVAB), since we are counting these two items twice in the computations of the P&P-ASVAB SIFs.

Score information results. CAT-ASVAB SIFs were computed for each of the 22 conditions listed in Table 11-8. For comparison, the P&P-ASVAB SIF (for 9A) was computed. The SIFs for the CAT-ASVAB equaled or exceeded the P&P-ASVAB SIFs for all but four conditions: 3, 4, 7, and 8. These four exceptions involved the two pools of AR and WK that consisted of only primary items. When these pools were supplemented with additional items (see conditions 5, 6, 9, and 10) the resulting SIFs equaled or exceeded the corresponding P&P-ASVAB SIFs.

Table 11-9 lists the number of items used in selected SIF analyses. The number of times (across simulees) that an item was administered was recorded for each SIF simulation. The values in Table 11-9 represent the number of items that were administered at least once during the 15,500 simulated test sessions. A separate count for pri-

Table 11-9 Number of Used Items in CAT-ASVAB Item Pools

Content Area	Exposure Rate	Number of Used Items					
		Form 1			Form 2		
		Primary	Supp.	Total	Primary	Supp.	Total
GS	1/3	72	—	72	67	—	67
AR	1/6	62	32	94	53	41	94
WK	1/6	61	34	95	55	44	99
PC	1/6	50	—	50	52	—	52
AI	1/3	53	—	53	53	—	53
SI	1/3	51	—	51	49	—	49
MK	1/6	84	—	84	85	—	85
MC	1/3	64	—	64	64	—	64
EI	1/3	61	—	61	61	—	61

mary and supplemental items is provided for AR and WK.

Reliability

A reliability index provides another criterion for comparing the relative precision of the CAT-ASVAB with the P&P-ASVAB. These indices were computed for each pool and for one form (9A) of the P&P-ASVAB. The reliabilities were estimated from simulated test sessions: 1,900 values were sampled from a $N(0,1)$ distribution. Each value represented the ability level of a simulated examinee (simulee). The simulated tests were administered twice to each of the 1,900 simulees. The reliability index was the correlation between the pairs of Bayesian modal estimates of ability from the two simulated administrations. The CAT-ASVAB reliabilities were computed separately for each pool. The item selection and scoring procedures match those used in CAT-ASVAB (Chapter 12).

The P&P-ASVAB reliabilities were computed from simulated administrations of Form 9A. The following procedure was used to generate number right scores for each of the 1,900 simulees:

STEP 1: The probability of a correct response to a given item was obtained for a simulee by substituting the (9A) item parameter estimates and the simulee's ability level into the three-parameter logistic model.

STEP 2: A random uniform value in the interval [0,1] was generated and compared to the probability of a correct response. If the random number was less than the probability value, the item was scored correct; otherwise it was scored incorrect.

STEP 3: Steps 1 and 2 were repeated across test items for each simulee. The number right score was the sum of the responses scored correct.

Steps 1 through 3 were repeated twice to obtain two number-right scores for each simulee. The reliability index for the P&P-ASVAB was the correlation between the two number-right scores.

A special procedure was used to compute reliability indices for AS. These items on the P&P version (9A) were divided into two components: AI and SI. This split corresponded to the assignment made in the item calibration of these content areas. A reliability index was computed separately for each component.

Reliability indices were computed for each of the 22 conditions and are listed in Table 11-10. For comparison, the P&P-ASVAB reliability (for 9A) was computed and displayed in the same table. Exposure rates and test lengths are also provided. The estimated CAT-ASVAB reliability indices exceeded the corresponding P&P-ASVAB (9A) values for all 22 conditions.

Table 11-10 Simulated Reliabilities (N = 1,900)

Test	Form	Test Length	Exposure Rate	Reliability r
GS	CAT-1	15	1/3	.902
	CAT-2	15	1/3	.900
	ASVAB-9A	25		.835
AR	CAT_a-1	15	1/6	.924
	CAT_a-2	15	1/6	.924
	CAT-1	15	1/6	.904
	CAT-2	15	1/6	.903
	ASVAB-9A	30		.891
WK	CAT_a-1	15	1/6	.934
	CAT_a-2	15	1/6	.936
	CAT-1	15	1/6	.912
	CAT-2	15	1/6	.913
	ASVAB-9A	35		.902
PC	CAT-1	10	1/6	.847
	CAT-2	10	1/6	.855
	ASVAB-9A	15		.758
AI	CAT-1	10	1/3	.894
	CAT-2	10	1/3	.904
	ASVAB-9A	17		.821
SI	CAT-1	10	1/3	.874
	CAT-2	10	1/3	.873
	ASVAB-9A	10		.651
MK	CAT-1	15	1/6	.933
	CAT-2	15	1/6	.935
	ASVAB-9A	25		.854
MC	CAT-1	15	1/3	.886
	CAT-2	15	1/3	.897
	ASVAB-9A	25		.807
EI	CAT-1	15	1/3	.875
	CAT-2	15	1/3	.873
	ASVAB-9A	20		.768

Summary

The procedures described in this chapter formed the basis of the item pool construction and evaluation procedures. Large item pools were pretested and calibrated in large samples of applicants. Two item pools (WK and AR) were supplemented with additional items, and a special study was conducted to evaluate adverse consequences of mixing 4-option supplemental items with other 5-option items. Extensive analyses were conducted to evaluate each pool's dimensionality. For pools found to

be multidimensional, these analyses aided in selecting the most appropriate approach for item selection and scoring. Finally, extensive precision analyses were conducted to evaluate the conditional and unconditional precision levels of the item pools, and to compare these precision levels with the P&P-ASVAB.

Based on the score information analyses, the precision for the primary AR and WK pools over the middle ranges of ability was inadequate. By supplementing these pools with experimental CAT-ASVAB items, the precision was raised to an acceptable level. Why was it necessary to supplement these pools, and what lessons can be applied to the construction of future pools?

One clue comes from the distribution of difficulty parameters obtained from surviving items (those items in the pools that have a greater than zero probability of administration). An examination of this distribution indicates a bell shaped distribution, with a larger number of difficulty values appearing over the middle ranges, and fewer values appearing in the extremes. Note that the target difficulty distribution for item writing and for inclusion in the calibration study was a uniform distribution. This suggests that there were actually an excess of items in the extremes (which had zero probabilities of administration), and for WK and AR, a deficiency of items over the middle ranges. Future development efforts should attempt to construct banks of items with bell shaped distributions of item difficulty values, similar to those constructed for P&P tests.

A bell shaped distribution of item difficulties has at least two desirable properties for CAT. First, larger numbers of items with moderate difficulty values are likely to lead to higher precision over the middle range, since the adaptive algorithm is likely to have more highly discriminating items to choose from. This may be especially desirable if it is important to match the precision of a P&P test which peaks in information over the middle ability ranges. Second, the Sympson-Hetter exposure control algorithm (Chapter 13) places demands on moderately difficult items, since the administration of these items is restricted. Because of the restrictions placed on these items, more highly informative items of moderate difficulty are necessary to maintain high levels of precision.

Although CAT-ASVAB precision analyses indicated favorable comparisons with the P&P-ASVAB, many strong assumptions were made in the simulation analyses which may limit applicability of these findings to operational administrations with real test-takers. Such assumptions (including unidimensionality, local independence, and knowledge of true item functioning) are almost certainly violated to some extent in applied testing situations. Therefore, it is important to examine the precision of these pools with live test-takers who are administered tests using the same adaptive item selection and scoring algorithms evaluated here. Such an evaluation is described in Chapter 17.