DOCUMENT RESUME

ED 264 270 TM 850 756

AUTHOR Weiss, David J.; Suhadolnik, Debra
TITLE Robustness of Adaptive Testing to Multidimensionality.
SPONS AGENCY Air Force Human Resources Lab., Brooks AFB, Texas.; Office of Naval Research, Washington, D.C.
PUB DATE Jul 82
CONTRACT N00014-79-C-0172; NR-150-433
NOTE 34p.; In: Item Response Theory and Computerized Adaptive Testing Conference Proceedings (Wayzata, MN, July 27-30, 1982) (TM 850 744).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150) -- Statistical Data (110)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Factor Structure; *Latent Trait Theory; *Monte Carlo Methods; *Multidimensional Scaling; Postsecondary Education; Simulation; Tables (Data); Test Items
IDENTIFIERS Armed Services Vocational Aptitude Battery; *Robustness

ABSTRACT

The present monte carlo simulation study was designed to examine the effects of multidimensionality during the administration of computerized adaptive testing (CAT). It was assumed that multidimensionality existed in the individuals to whom test items were being administered, i.e., that the correct or incorrect responses given by an individual were generated from a specified multidimensional structure, rather than the unidimensional item response theory (IRT) model normally assumed to have generated the observable dichotomous test item responses. The dichotomous response was then treated for CAT item selection and ability estimation purposes as if it had been generated by the unidimensional model. To the extent that the observed item response was affected by dimensions other than the first (which corresponded to the single dimension assumed to underlie the item selection and ability estimation process) errors should be introduced into the adaptive testing process. These errors should affect the ability estimates and the efficiency of CAT. The study focused on the nature and degree of these errors under a variety of multidimensional structures, to determine how robust CAT is to the effects of multidimensionality in examinees' responses to test items. (PN)

Robustness of Adaptive Testing to Multidimensionality

David J. Weiss and Debra Suhadolnik
University of Minnesota

Before computerized adaptive testing (CAT) can be applied in various operational settings, its characteristics must be evaluated under a variety of conditions. Studies of the reliability and validity of CAT (e. g., Johnson & Weiss, 1980; Kiely, Zara, & Weiss, 1983; McBride & Martin, 1983; Moreno, Wetzel, McBride & Weiss, 1984; Sympson, Weiss, & Ree, 1982) provide important information comparing CAT to conventional tests in applied situations. Live-testing studies such as these, however, are expensive and time-consuming, provide results that are dependent on the characteristics of the sample of subjects and the specific criterion variables used, and do not permit an answer to the important questions about how well CAT measures true ability levels and whether ability is better estimated at different ability levels. Live-testing studies also incorporate a number of uncontrolled sources of error (e. g., item parameter estimation error, various errors of measurement due to idiosyncratic characteristics of examinee responses to test items) which further complicate the process of reaching generalizable conclusions.

Monte carlo simulation provides a means of systematically examining the performance of CAT under a variety of conditions and of identifying the effects of various kinds of errors on the performance of CAT strategies. Early studies were concerned with the comparison of CAT item selection strategies with conventional tests (e.g., Betz & Weiss, 1973, 1974, 1975; Larkin & Weiss, 1974; Vale & Weiss, 1975a, 1975b) and with each other (e.g., Larkin & Weiss, 1975). These studies provided global evaluations of CAT strategies that were useful in eliminating some strategies from further consideration. Later studies then concentrated on the more promising strategies, generally those that are based on item response theory (IRT), examining the performance of these testing strategies conditional on true ability levels (e.g., McBride, 1977; Vale, 1975; Weiss & McBride, 1984).

One factor that can affect the performance of CAT is the nature of the item pool from which it draws items. McBride (1977; Weiss & McBride, 1984) studied the performance of a Bayesian CAT in perfect and ideal item pools and in realistic item pools in which the IRT diffficulty and discrimination parameters were correlated. Others (e.g., Urry, 1974) also examined CAT performance in a variety of item pool configurations.

In addition to the distributions of item difficulties and discriminations in a given item pool, the degree of error in the IRT item parameter estimates in

a real item pool can affect the performance of CAT, particularly since items are selected on the basis of their IRT parameter estimates. Crichton (1981) investigated the effects of errors in item parameter estimates on the performance of maximum information and Bayesian CAT strategies in the context of the 3-parameter logistic model. Mattson (1983) extended Crichton's study to the 1- and 2-parameter logistic models, both Bayesian and maximum likelihood scoring, and to the more realistic situation in which the IRT difficulty and discrimination parameters had varying degrees of correlation. These later studies provide valuable information about the performance of CAT under the realistic situation in which adaptive testing is to be done using item pools with parameters estimated with varying degrees of error.

A second factor that is likely to have an effect on the performance of IRT-based CAT is multidimensionality. Operational IRT models used for CAT assume that unidimensionality exists at two stages: (1) in the process of estimating item parameters, and (2) in the process by which an individual generates a response to a test item with given item parameters. Presumably, any deviations from unidimensionality that exist at either of those stages in CAT could result in non-optimal performance of IRT-based CAT strategies.

While many tests of ability and achievement approximate unidimensionality, none have shown the strict unidimensionality required by operational IRT models. This motivated Drasgow and Parsons (1983) to examine the effects of deviations from unidimensionality during the item parameter estimation process on IRT item parameter estimates.

Purpose

The present monte carlo simulation study was designed to examine the effects of multidimensionality during CAT test administration. It was assumed that multidimensionality existed in the individuals to whom test items were being administered--i.e., that the correct or incorrect responses given by an individual were generated from a specified multidimensional structure, rather than the unidimensional IRT model normally assumed to have generated the observable dichotomous test item responses. The dichotomous response was then treated for CAT item selection and ability estimation purposes as if it had been generated by the unidimensional model. To the extent that the observed item response was affected by dimensions other than the first (which corresponded to the single dimension assumed to underlie the item selection and ability estimation process) errors should be introduced into the adaptive testing process. These errors should affect the ability estimates and the efficiency of CAT. The study focused on the nature and degree of these errors under a variety of multidimensional structures, to determine how robust CAT is to the effects of multidimensionality in examinees' responses to test items.

METHOD

Initial Factor Analyses

Item response vectors for forms 8A and 8B of the Armed Services Vocational

Aptitude Battery (ASVAB) were obtained for a sample of military recruits. For those subtests of the ASVAB (Mathematics Knowledge, General Science and Mechanical Comprehension) in which forms 8A and 8b were identical except for the order of the items, the response vectors for form 8B were rearranged to match the order of the items in form 8A. This resulted in datasets with sample sizes of 5,127 for these three subtests, sample sizes of 2,621 for form 8A of the other seven subtests, and sample sizes of 2,506 for form 8B of the other seven subtests.

Tetrachoric inter-item correlations were computed for eight of the ten subtests; the Numerical Operations and Coding Speed subtests were not included in further analyses due to the speeded nature of these subtests. The tetrachoric correlations for the other eight subtests were then factor-analyzed using a principal axes factor extraction method and a Varimax rotation. Of the resulting factor structures, the factor structure of the General Science subtest exhibited the greatest degree of multidimensionality. Table 1 lists the factor loadings on the first four factors for the items in this subtest. This factor structure was used as the model for generating subsequent factor structures with varying degrees of multidimensionality.

Generation of Factor Structures

The first step in creating factor structures with varying degrees of multidimensionality was to round the 25 factor loadings on the first factor of the ASVAB General Science (GS) subtest to the nearest multiple of .05. This set of 25 rounded factor loadings was then repeated six times to create a set of factor loadings for 150 items on one factor with the same configuration of loadings as the first factor for the ASVAB GS subtest. This factor, the original strength ASVAB factor (OSAF), was used as the basis for one of three sets of factor structures.

Sixteen factor structures of varying dimensionality were constructed using OSAF as the first factor. Factors other than the first factor were constructed to be proportional in strength to the first factor. These sixteen factor structures are described in Table 2. Factor structures varied from a 2-factor structure with the second factor 1/8 as strong as the first factor (Dataset 2) to a 3-factor structure with Factors 2 and 3 equal in strength to Factor 1 (Dataset 16). An additional dataset (17) consisted of the actual 4-factor ASVAB GS factor solution.

The 150 factor loadings on OSAF were then increased to yield a first factor that was approximately 1.5 times as strong as OSAF. This new first factor (1.5 OSAF) was used as the first factor in a set of sixteen different factor structures which are also described in Table 2 (Datasets 18-33). Factors other than the first factor in Datasets 18-32 were again constructed to be proportional to this strengthened first factor in all of the factor structures except the 4-factor structure (Dataset 33), where the second, third and fourth factors were the actual second, third, and fourth factors from the original factor analysis of the ASVAB GS subtest (see Table 1).

The 150 factor loadings on OSAF were then increased a second time to result

Table 1
Factor Loadings for the First Four Factors of the ASVAB General Science Subtest

Item Number	Factor 1	Factor 2	Factor 3	Factor 4
1	.540	-.215	-.250	.027
2	.624	-.205	-.018	-.303
3	.642	-.201	-.095	.026
4	.486	-.098	-.118	-.115
5	.668	-.233	.162	.069
6	.703	-.160	.066	.073
7	.572	.052	.103	-.019
8	.493	-.070	.072	-.067
9	.546	-.174	-.239	.119
10	.547	-.212	-.015	.016
11	.595	.060	.009	-.025
12	.398	.099	.058	-.006
13	.580	.096	-.120	-.233
14	.580	-.172	-.069	.124
15	.438	-.029	.043	.337
16	.543	.012	.100	.172
17	.462	.120	-.030	-.009
18	.639	.227	.054	-.072
19	.371	.208	.045	-.011
20	.473	.048	.132	-.096
21	.460	.273	.085	.006
22	.283	.224	.115	-.032
23	.480	.035	.147	-.062
24	.387	.650	-.310	.067
25	.396	.310	.101	.089
Factor Contribution	7.541	1.671	1.030	1.023

in a first factor that was approximately twice as strong as OSAF. This strengthened first factor (2.0 OSAF) was used as the first factor in a third set of twelve factor structures (Datasets 34-45), which are also described in Table 2. In Datasets 34-44, factors other than the first factor were constructed to be proportional in strength to this increased strength first factor; these additional factors were also constructed to avoid communalities greater than 1.0 for any item. For the 4-factor structure of Dataset 45, the factors other than the first factor were taken directly from the original factor analysis of the ASVAB GS subtest (see Table 1).

Generation of Response Vectors

To evaluate the effect of violation of the assumption of unidimensionality in adaptive testing, sets of dichotomous (0,1) item responses were generated using the factor structures with varying degrees of multidimensionality.

Table 2
Dataset Numbers for Datasets Based on First Factors of 1.0, 1.5, and 2.0 OSAF, and Factor Strengths of Factors 2 through 4, for Each of the Datasets

Dataset Number			Factor Strength as a Proportion of Factor 1		
1.0 OSAF	1.5 OSAF	2.0 OSAF	Factor 2	Factor 3	Factor 4
1	18	34	-	-	-
2	19	35	1/8	-	-
3	20	36	1/4	-	-
4	21	37	1/3	-	-
5	22	38	1/2	-	-
6	23	39	2/3	-	-
7	24	40	3/4	-	-
8	25	--	1.0	-	-
9	26	41	1/8	1/8	-
10	27	42	1/4	1/4	-
11	28	43	1/3	1/3	-
12	29	44	1/2	1/4	-
13	30	--	1/2	1/2	-
14	31	--	2/3	1/3	-
15	32	--	2/3	2/3	-
16	--	--	1.0	1.0	-
17	33	45	GS-2*	GS-3*	GS-4*

*Factor derived from factor analysis of ASVAB GS test.

The first step was to assign θ levels for each factor to a number of hypothetical examinees (simulees). This was done for each factor except the first factor by using a random number generator to create uniform distributions of 1,700 θ values between -3.2 and +3.2 for each factor independently of all other factors. θ levels for the first factor were assigned so that 100 simulees were assigned to each of 17 θ levels ranging from -3.2 to +3.2 in increments of .4. θ levels for the first factor were assigned in this manner in order to have a sufficient number of replications at each θ level so that indices conditional on θ could be computed.

Next, matrices of item response theory (IRT) item parameters were calculated and generated. Item discrimination parameters (as) were computed using the following formula:

$$a_{gj} = F_{gj} / \left[1 - (F_{gj}^2)\right]^{\frac{1}{2}} \qquad [1]$$

where a_{gj} = item discrimination parameter for item g and factor j, and
F_{gj} = factor loading for item g on factor j.

These matrices of a parameters were calculated for each of the 45 factor structures.

Matrices of item difficulty parameters (bs) were generated for each of the 45 factor structures using a random number generator which generated a uniform distribution of 150 values between -3.2 and +3.2 independently for each factor in a given factor structure. Item pseudo-guessing parameters (cs) were also generated for each factor in the 45 factor structures; they were generated to yield a normal distribution of 150 values with a mean of .20 and a standard deviation of .02 for each factor.

After the item parameter matrices for each factor structure were determined, the probability of a correct response to each item for each factor was computed for each of the 1,700 simulees using the three-parameter logistic model,

$$P_{igj}(\theta_j) = c_{gj} + \frac{(1 - c_{gj})}{1 + \exp[-1.7a_{gj}(\theta_j - b_{gj})]} \qquad [2]$$

where $P_{igj}(\theta_j)$ = probability of a correct response to item g on factor j for a simulee with trait level θ_j,

c_{gj} = IRT pseudo-guessing parameter for item g on factor j,

a_{gj} = IRT discrimination parameter for item g on factor j, and

b_{gj} = IRT difficulty parameter for item g on factor j.

The probabilities for each item on each factor were then combined using Equation 3 to calculate the overall probability of a correct response for each individual on each item:

$$r_{ig} = \frac{\sum_{j=1}^{K} F_{gj}^2 P_{igj}}{\sum_{j=1}^{K} F_{gj}^2} \qquad [3]$$

where r_{ig} = overall probability of a correct response for simulee i on item g,

F_{gj} = factor loading for item g on factor j, and

P_{igj} = probability of a correct response for simulee i on factor j for item g.

Dichotomous item scores (u_{ig}) were then generated using r_{ig} and a random number generator. For each simulee and item, a random number between 0 and 1 was generated. If r_{ig} was greater than this random number, an item score u_{ig} = 1 was assigned for the response of simulee i to item g. If r_{ig} was less than the random number, an item score u_{ig} = 0 was assigned to the item for the simu-

lee. In this manner, each of the 1,700 simulees received an item score of 0 or 1 on each of the 150 items for each factor structure.

Adaptive Testing Strategy

The sets of dichotomous item responses u_{ig} generated from the factor structures with varying degrees of multidimensionality were used with a maximum information adaptive testing strategy to obtain θ estimates. Since the adaptive testing strategy used assumes a unidimensional set of item responses, the obtained θ estimates can be used to determine the effect of violation of the assumption of unidimensionality. For each factor structure:

1. $\hat{\theta}$ was set to 0.0 for each simulee.

2. Information at $\hat{\theta}$ was computed for each of the 150 items using first factor a, b, and c parameters in the following equation:

$$I_g(\hat{\theta}) = [P'_g(\hat{\theta})]^2 / P_g(\hat{\theta})Q_g(\hat{\theta}) \qquad [4]$$

where $I_g(\hat{\theta})$ = information at $\hat{\theta}$ for item g,
$P_g(\hat{\theta})$ = probability of a correct response to item g at $\hat{\theta}$,
$P'_g(\hat{\theta})$ = first derivative of $P_g(\hat{\theta})$, and
$Q_g(\hat{\theta}) = 1 - P_g(\hat{\theta})$.

3. The item with the highest level of information at $\hat{\theta}$ was selected as the next item to be administered.

4. The item responses to the item chosen to be administered were read from the generated item response matrix for each simulee.

5. A new $\hat{\theta}$ was calculated for each simulee using maximum likelihood scoring:

$$L(\theta_{i1}|u_i) = \prod_{g=1}^{K} P_{ig}(\hat{\theta}_i)^{u_{ig}} Q_{ig}(\hat{\theta}_i)^{1-u_{ig}} \qquad [5]$$

where $L(\theta_{i1}|u_i)$ = likelihood of the simulee's observed response pattern (u_i) at θ_{i1},
$P_{ig}(\hat{\theta}_i)$ = probability of a correct response to item g for simulee i with trait level estimate $\hat{\theta}_i$,
u_{ig} = 1 for a correct response to item g,
= 0 for an incorrect response to item g,
$Q_{ig}(\hat{\theta}_i) = 1 - P_{ig}(\hat{\theta}_i)$, and
K = the number of items administered.

The value of θ which had the greatest likelihood for the observed item responses was selected as the new θ estimate for a simulee ($\hat{\theta}$ was restricted to the range +4 to -4).

6. Steps 2 through 5 were repeated using the new $\hat{\theta}$s for each simulee until 30 items were administered;

7. The $\hat{\theta}$s were saved at 5, 10, 15, 20, 15 and 30 items.

Evaluative Indices

Conditional indices. Since no one optimal evaluative index was available, four different evaluative indices were used to determine the effect of violations of the assumption of unidimensionality in adaptive testing. Each of the following four indices were computed at each of the 17 θ levels on the first factor and for all six test lengths.

1. Bias:

$$\text{Bias}(\theta_{p1}) = \frac{\sum_{i=1}^{N(\theta_{p1})} (\hat{\theta}_i - \theta_{i1})}{N(\theta_{p1})} \qquad [6]$$

$\hat{\theta}_i$ = estimated θ level for simulee i,
θ_{p1} = true θ level for simulee i on factor 1, and
$N(\theta_{p1})$ = number of simulees at level p (usually 100, but occasionally smaller due to maximum likelihood convergence failures).

This index takes into account both the size and direction of the difference between true and estimated θ.

2. Inaccuracy:

$$\text{Inaccuracy}(\theta_{p1}) = \frac{\sum_{i=1}^{N(\theta_{p1})} |\hat{\theta}_i - \theta_{i1}|}{N(\theta_{p1})} \qquad [7]$$

Inaccuracy considers only the size, and not the direction, of the difference between estimated and actual θ levels for each simulee at a given θ level and test length.

3. Root Mean Square Error (RMSE). RMSE was calculated as

$$\text{RMSE}(\theta_{p1}) = \left(\frac{\sum_{i=1}^{N(\theta_{p1})} (\hat{\theta}_i - \theta_{i1})^2}{N(\theta_{p1})} \right)^{\frac{1}{2}} \qquad [8]$$

This index gives more weight to larger differences between estimated and true θ levels.

4. Efficiency. Efficiency was defined by

$$I(\theta_l) = \frac{\sum_{g=1}^{K} I_{g^*}(\theta_l)}{\sum_{g=1}^{K} I_g(\theta_l)} \qquad [9]$$

where g* indexes items actually administered and g indexes the items with the maximum levels of information at θ_l.

Thus, efficiency is the ratio of the information in the k items actually administered to the k most informative items at θ_l. It will equal 1.0 when the adaptive testing strategy administers the k items with maximum information at θ_l. Deviations from 1.0 result from the fact that, at any stage of the adaptive test, $\hat{\theta}$ is not usually exactly equal to θ_l.

Comparison of conditional multidimensional and unidimensional results. To summarize the effects of multidimensionality on each of the evaluative indices, distance measures were computed across the 17 θ_l levels between the values of each of the conditional evaluative indices for the unidimensional (UD) datasets and the multidimensional (MD) datasets for all six test lengths. Cronbach and Gleser's (1953) formulas were used for computing a distance measure, D^2, between two profiles and for decomposing D^2 into components due to mean differences, scatter differences, and shape differences. Profiles were plots of the values of an evaluative index for a given dataset and test length across all 17 θ_l levels. The formulas used were:

$$D^2_{UD,MD} = \sum_{p=1}^{17} (X_{pUD} - X_{pMD})^2 \qquad [10]$$

where $D^2_{UD,MD}$ = overall squared distance between profile UD and profile MD,
X_{pUD} = value of the evaluative index for dataset UD and θ level p, and
X_{pMD} = value of the evaluative index for dataset MD and θ level p.

$$D^{2'}_{UD,MD} = D^2_{UD,MD} - 17(\Delta^2 EL_{UD,MD}) \qquad [11]$$

where $D^{2'}_{UD,MD}$ = squared distance between profiles UD and MD after differences in mean level between the two profiles are eliminated.
$\Delta^2 EL_{UD,MD}$ = squared difference in mean level between profiles UD and MD, and

$$D^{2''}_{UD,MD} = \frac{D^{2'}_{UD,MD} - \Delta^2 S_{UD,MD}}{S_{UD} S_{MD}} \qquad [12]$$

where $D^{2''}_{UD,MD}$ = squared distance between profiles UD and MD after differences due to mean level and scatter between the two profiles are eliminated

$$S_{UD} = \left(\sum_{p=1}^{17} (X_{pUD} - \bar{X}_{UD})^2 \right)^{\frac{1}{2}} \qquad [13]$$

where S_{UD} = scatter for profile UD,

$\bar{X}_{UD}$ = mean of the 17 values of the evaluative index for profile UD,

$\bar{X}_{MD}$ and S_{MD} are defined similarly, and

$\Delta^2 S_{UD,MD}$ = squared difference between scatters for profiles MD and UD.

The presence of scatters less than 1.00 for many of the datasets resulted in values of $D^{2''}$ that were larger than the values of $D^{2'}$ for the same profiles. This made interpretation of the distance measures difficult, so the values of each of the four evaluative indices at each of the 17 θ levels were multiplied by 10. This fact should be taken into account in interpreting the magnitude of the differences between profiles and the distance measures.

To aid in interpreting the differences in profiles due to level, scatter and shape, the proportion of the squared distance (D^2) due to each of these components was computed using the following formulas:

$$\text{Level Effect}_{UD,MD} = \frac{D^2_{UD,MD} - D^{2'}_{UD,MD}}{D^2_{UD,MD}} , \qquad [14]$$

the proportion of D^2 due to differences in level between profiles UD and MD,

$$\text{Scatter Effect}_{UD,MD} = \frac{D^{2'}_{UD,MD} - D^{2''}_{UD,MD}}{D^2_{UD,MD}} , \qquad [15]$$

the proportion of D^2 due to differences in scatter between the two profiles, and

$$\text{Shape Effect}_{UD,MD} = \frac{D^{2''}_{UD,MD}}{D^2_{UD,MD}} , \qquad [16]$$

the proportion of D^2 due to differences in shape between profiles MD and UD.

Unconditional indices. In addition to examining the bias, inaccuracy, and RMSE conditional on θ level, mean values of these indices were computed across the 17 θ levels for each dataset and test length. Also computed for each condition was the fidelity correlation between $\hat{\theta}$ and θ_1. These correlations were computed for a normally distributed sample of 630 simulees selected from the 1,700 rectangularly distributed simulees in each dataset.

RESULTS

Unconditional Indices

Fidelity

Table 3 shows fidelity correlations for each of the datasets based on OSAF, 1.5 OSAF, and 2.0 OSAF, as a function of test length. For the single-factor Dataset 1, fidelity increased with increasing test length from .546, when 5 items were administered, to .928 at 30 items. For the 2-factor datasets (2-8) fidelity generally decreased with increasing strength of the second factor with two exceptions: (1) Dataset 5, which had a second factor 1/2 as strong as the first factor, had consistently higher fidelity than Dataset 4, in which the second factor was only 1/3 as strong as the first; and (2) Dataset 6, in which the second factor was 2/3 the strength of the first, had consistently lower fidelity than Dataset 7, in which the second factor was slightly stronger (3/4 of the first). In both these cases, differences between the fidelities decreased with increases in test length. For all datasets, fidelity increased with increasing test length.

For these 2-factor datasets, multidimensionality had fairly substantial effects on fidelity. For example, at the 15-item test length fidelity was .872 for the single-factor Dataset 1, but dropped to .548 when there were two equal factors (Dataset 8). When the second factor was only 1/4 the strength of the first factor (Dataset 3), fidelity for a 15-item test decreased from .872 to .784. To overcome the effect of this degree of multidimensionality, the 15-item test of Dataset 3 would need to be doubled in length, resulting in a fidelity of .880. For degrees of multidimensionality beyond those represented by Dataset 3, tests would need to be well beyond 30 items in length to equal the fidelity of the 15-item test in UD Dataset 1.

A similar pattern of results was observed for the 3-factor structures (Datasets 9-16), but the effects of multidimensionality on fidelity were even stronger. In these datasets there was, again, a general decrease in fidelity with increasing strength of the second and third factors. Fidelity also increased with test length for all datasets. In general, however, fidelities were lower for the 3-factor datasets than for those with two factors, even when the total variance accounted for by factors beyond the first was equal. For example, at the 15-item length, fidelity for Dataset 13 (with factors 2 and 3 each 1/2 of the first factor in strength) was .443; when the same amount of variance was concentrated in only the second factor (Dataset 8), fidelity was .548. Only Dataset 9, with second and third factors each 1/8 of the first factor, attained a sufficiently high fidelity at 30 items (.869) to approximate that of UD Dataset 1 at 15 items (.872).

Results for the 1.5 and 2.0 OSAF datasets were similar to those for 1.0 OSAF, with a general increase in fidelity with increasing strength of the first factor. For example, for a 15-item test based on a 2-factor structure with the second factor 3/4 the strength of the first factor, fidelity was .628 for 1.0 OSAF (Dataset 7), .685 for 1.5 OSAF (Dataset 24), and .789 for 2.0 OSAF (Dataset 40). For the 3-factor datasets with the second and third factors each 1/3 of

Table 3
Fidelity as a Function of Test Length for Unidimensional (UD) and Multidimensional Datasets Based on First Factors 1.0, 1.5, and 2.0 Times as Strong as the ASVAB General Science Factor

Dataset	No. of Factors	Test Length (Number of Items) 5	10	15	20	25	30
1.0 OSAF							
1 (UD)	1	.646	.799	.872	.903	.914	.928
2	2	.592	.762	.823	.866	.896	.909
3	2	.519	.692	.784	.833	.863	.880
4	2	.461	.592	.672	.718	.765	.790
5	2	.534	.648	.711	.780	.813	.826
6	2	.404	.543	.616	.658	.677	.705
7	2	.431	.572	.628	.662	.694	.715
8	2	.429	.510	.548	.580	.616	.631
9	3	.522	.665	.760	.821	.847	.869
10	3	.423	.567	.655	.706	.737	.763
11	3	.375	.477	.567	.633	.678	.710
12	3	.340	.467	.559	.614	.652	.679
13	3	.320	.386	.443	.499	.548	.584
14	3	.350	.467	.529	.574	.618	.645
15	3	.313	.383	.418	.434	.473	.490
16	3	.267	.339	.371	.400	.415	.438
17	4	.577	.723	.802	.847	.871	.893
1.5 OSAF							
18 (UD)	1	.691	.842	.916	.937	.949	.955
19	2	.660	.822	.881	.912	.931	.945
20	2	.587	.753	.848	.892	.914	.924
21	2	.560	.740	.828	.877	.904	.912
22	2	.569	.737	.808	.842	.867	.878
23	2	.462	.616	.724	.772	.802	.812
24	2	.478	.623	.685	.713	.748	.763
25	2	.387	.510	.607	.651	.675	.697
26	3	.590	.740	.816	.863	.892	.911
27	3	.446	.596	.702	.752	.782	.801
28	3	.439	.569	.654	.710	.755	.776
29	3	.442	.578	.650	.702	.742	.759
30	3	.447	.554	.637	.695	.731	.745
31	3	.455	.589	.690	.732	.756	.771
32	3	.415	.525	.610	.653	.681	.700
33	4	.581	.765	.858	.892	.918	.932
2.0 OSAF							
34 (UD)	1	.733	.867	.930	.953	.961	.965
35	2	.585	.775	.888	.932	.955	.964
36	2	.599	.749	.850	.911	.927	.937
37	2	.524	.694	.817	.860	.902	.924
38	2	.604	.710	.807	.853	.866	.888
39	2	.547	.655	.756	.816	.843	.849
40	2	.542	.689	.789	.813	.836	.844
41	3	.519	.690	.804	.875	.923	.929
42	3	.542	.647	.744	.813	.841	.868
43	3	.499	.631	.758	.831	.855	.874
44	3	.534	.674	.777	.819	.840	.857
45	4	.379	.482	.546	.618	.664	.700

the first factor, fidelity for a 15-item test in the 1.0 OSAF data was .567 (Dataset 11), rising to .654 when factor 1 was 1.5 OSAF (Dataset 28) and to .758 with 2.0 OSAF (Dataset 43). As in the 1.0 OSAF data, a single factor beyond the first had less effect on fidelity than did two factors equaling the strength of the single factor, though the effect diminished substantially with the stronger first factor. For example, in the 1.5 OSAF structures for a 15-item test with a second factor 2/3 of the first factor (Dataset 23), fidelity was .724 versus .654 when there were two factors beyond the first, each comprising 1/3 of the first factor (Dataset 28); comparable factor structures with 2.0 OSAF resulted in fidelities of .756 (Dataset 39) and .758 (Dataset 43).

Datasets 17, 33, and 45 provide results based on factors derived from the ASVAB 4-factor structure, in which factors 2, 3, and 4 accounted for 22.2%, 13.6%, and 13.5%, respectively, of OSAF. Table 3 shows that there were relatively small effects on fidelity for the 1.0 and 1.5 OSAF datasets, particularly for tests of 20 or more items. For example, in Dataset 17 fidelity for a 25-item test was .871 versus .914 for UD Dataset 1. Comparable results for the 1.5 OSAF data were .918 (Dataset 33) and .949 (Dataset 18). In the 2.0 OSAF data, however, the 4-factor ASVAB structure (Dataset 45) resulted in the lowest observed fidelities for those datasets; fidelity dropped from .953 (UD Dataset 34) to .618 for ASVAB at 20 items, and from .965 to .700 at 30 items.

Bias, Inaccuracy, RMSE

Table 4 provides data on mean bias, inaccuracy, and RMSE for the datasets based on 1.0 OSAF. For UD Dataset 1, bias decreased from .282 at 5 items to .010 at 30 items. Each of the 2-factor datasets (2-8) showed lower levels of positive bias and higher levels of negative bias than did Dataset 1, with bias becoming increasingly negative as the strength of the second factor increased. Thus, in 2-factor data structures $\hat{\theta}$ underestimated θ, on the average, as both test length and strength of multidimensionality increased. A similar trend was observed for most of the 3-factor datasets (9-16), with a few exceptions. In these datasets bias tended to become less positive and increasingly negative for all test lengths for Datasets 9-12, in which the sum of the variance accounted for by the second and third factors was less than that of the first factor. In Dataset 13, which had second and third factors each 1/2 of the first factor, bias was again positive for tests of 15 items or less, but this effect was reversed for Dataset 14 (factor 2 = 2/3 of factor 1, and factor 3 = 1/3 of factor 1). However, for tests of 5 or 10 items, bias then again became positive for Datasets 15 and 16, which had very strong second and third factors. There was also a slight trend toward positive mean bias in Dataset 16. As Table 4 also shows, there was a slight effect on bias when data were generated from the 4-factor ASVAB structure (Dataset 17). For these data the ASVAB structure resulted in a slight mean underestimation of θ at test lengths of 20 to 30 items with a mean bias of .006 at 15 items compared with .038 for Dataset 1.

Both inaccuracy and RMSE tended to increase with increasing strength of factors beyond the first, and to decrease with increasing test length; this held true for both the 2- and 3-factor datasets. An exception occurred for Dataset 14 (factor 2 = 2/3 of factor 1, and factor 3 = 1/3 of factor 1) for both inaccuracy and RMSE. For this dataset inaccuracy and RMSE values were lower than

Table 4
Mean Bias, Inaccuracy, and RMSE as a Function of Test Length for Unidimensional (UD) and Multidimensional Datasets, Based on SVAB General Science Factor

Dataset	No. of Factors	Test Length (Number of Items) 5	10	15	20	25	30
Bias							
1 (UD)	1	.282	.107	.038	.024	.015	.010
2	2	.247	.100	.031	-.015	-.028	-.026
3	2	.163	.056	-.004	-.026	-.042	-.051
4	2	.189	.060	-.022	-.052	-.072	-.084
5	2	.164	.065	-.017	-.053	-.075	-.085
6	2	.170	.038	-.023	-.070	-.099	-.107
7	2	.023	-.071	-.125	-.136	-.147	-.153
8	2	.057	-.026	-.103	-.135	-.171	-.190
9	3	.306	.133	.033	.012	-.020	-.028
10	3	.113	-.007	-.080	-.090	-.101	-.115
11	3	.173	.007	-.039	-.086	-.104	-.115
12	3	.128	.022	-.046	-.068	-.096	-.111
13	3	.397	.231	.061	-.040	-.089	-.131
14	3	-.051	-.097	-.127	-.174	-.193	-.201
15	3	.139	.059	-.060	-.122	-.171	-.210
16	3	.379	.240	.101	.006	-.068	-.142
17	4	.214	.075	.006	-.028	-.034	-.032
Inaccuracy							
1 (UD)	1	.906	.587	.451	.388	.357	.332
2	2	.982	.657	.517	.446	.402	.370
3	2	1.055	.713	.570	.493	.447	.413
4	2	1.251	.941	.770	.676	.617	.573
5	2	1.183	.870	.715	.604	.549	.521
6	2	1.247	.948	.791	.713	.664	.638
7	2	1.308	1.012	.861	.789	.733	.698
8	2	1.387	1.112	.997	.915	.863	.841
9	3	1.146	.781	.603	.512	.455	.418
10	3	1.373	1.027	.848	.731	.661	.612
11	3	1.424	1.100	.915	.801	.725	.675
12	3	1.455	1.135	.948	.837	.765	.720
13	3	1.622	1.292	1.113	1.016	.941	.888
14	3	1.549	1.244	1.062	.954	.875	.829
15	3	1.643	1.419	1.298	1.200	1.131	1.092
16	3	1.733	1.492	1.371	1.280	1.222	1.179
17	4	1.055	.734	.581	.489	.435	.399
RMSE							
1 (UD)	1	1.211	.785	.603	.514	.461	.425
2	2	1.328	.904	.694	.591	.521	.474
3	2	1.417	.980	.773	.658	.587	.539
4	2	1.659	1.296	1.090	.958	.868	.805
5	2	1.574	1.193	.984	.824	.757	.704
6	2	1.659	1.309	1.116	1.014	.934	.884
7	2	1.734	1.407	1.214	1.120	1.050	.999
8	2	1.809	1.498	1.356	1.258	1.201	1.162
9	3	1.539	1.069	.844	.702	.613	.547
10	3	1.800	1.401	1.198	1.043	.948	.882
11	3	1.855	1.500	1.276	1.122	1.021	.950
12	3	1.897	1.550	1.333	1.188	1.094	1.026
13	3	2.055	1.723	1.516	1.393	1.290	1.220
14	3	1.971	1.649	1.447	1.312	1.214	1.157
15	3	2.095	1.865	1.726	1.616	1.538	1.488
16	3	2.179	1.940	1.809	1.712	1.639	1.588
17	4	1.430	1.005	.797	.648	.572	.519

those for Dataset 13, in which the amount of variance accounted for by factors 2 and 3 was the same as in Dataset 14, but the factors were of equal strength; there was a trend for the difference between inaccuracies for the two datasets to increase as test length increased, with Dataset 14 resulting in lower mean inaccuracy.

As in the bias data, small effects on inaccuracy and RMSE were observed for the 4-factor ASVAB structure (Dataset 17). Both inaccuracy and RMSE decreased with increasing test length. For a 15-item test, inaccuracy was .581 for Dataset 17 versus .451 for Dataset 1; corresponding RMSE values were .797 and .603.

Although not shown here, similar trends for bias, inaccuracy, and RMSE were observed in the 1.5 and 2.0 OSAF datasets. That is, mean bias became increasingly negative with increasing multidimensionality and test length, whereas mean inaccuracy and RMSE tended to decrease with those variables. In general, however, the magnitudes of the evaluative indices were lower, indicating less effect of multidimensionality with a stronger first factor.

Conditional Indices

Effect of Test Length

Bias. Figures 1a through 1c show values of mean bias at each of 17 θ levels. Each figure compares the mean bias level across four different test lengths (10, 15, 20, and 25 items) for datasets derived from a 2-factor structure with the second factor 1/3 as strong as the first factor. The first factor for the datasets in Figure 1a was OSAF; in Figure 1b the first factor was 1.5 OSAF; and in Figure 1c it was 2.0 OSAF.

In each of these three figures the mean bias level generally decreases with increasing test length. This pattern is disrupted somewhat between θ levels of -.80 to +.80, where the bias fluctuates around 0.0 and no test length consistently shows a smaller mean bias level. Bias is most variable for the 10-item test length and least variable for the 25-item test. Regardless of the strength of the first factor, the mean bias values at θ levels greater than .80 converge for all four test lengths. Similar patterns of bias across test lengths were observed for the other datasets. In general, bias was negative for θs below the mean and positive for θs above the mean, although this effect was much less pronounced for the 1.5 OSAF datasets (Figure 1b) than for the 1.0 or 2.0 OSAF datasets (Figures 1a and 1c).

Inaccuracy. Figure 2 compares the mean inaccuracy levels at each of four different test lengths (10, 15, 20, and 25 items) across all 17 θ levels for Dataset 29, in which the first factor is 1.5 OSAF, the second factor is 1/2 as strong as the first factor, and the third factor is 1/4 as strong as the first factor. Inaccuracy tended to decrease with increasing test length. Inaccuracy levels for the 10-item test length varied across θ levels and were most constant for the 25-item test. This same pattern held for the comparisons across test length of the mean inaccuracy values for each of the 45 datasets.

RMSE. Comparison of the conditional RMSE values for the same dataset

Figure 1
Conditional Bias of θ Estimates for Tests of 10, 15, 20, and 25 Items
for Datasets with Factor 1 of 1.0, 1.5 and 2.5 OSAF
and Factor 2 One-Third the Strength of Factor 1

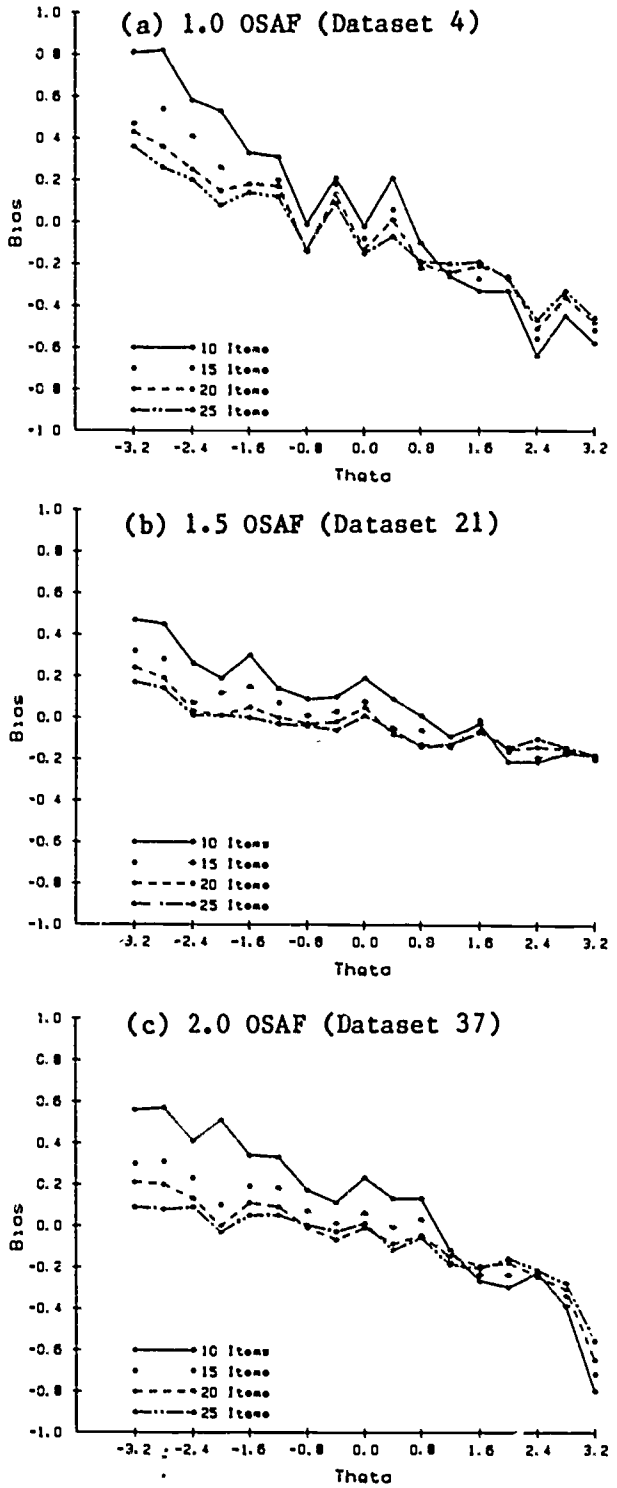


Figure 2
Conditional Inaccuracy of θ Estimates for Tests of 10, 15, 20, and 25 Items for Dataset 29

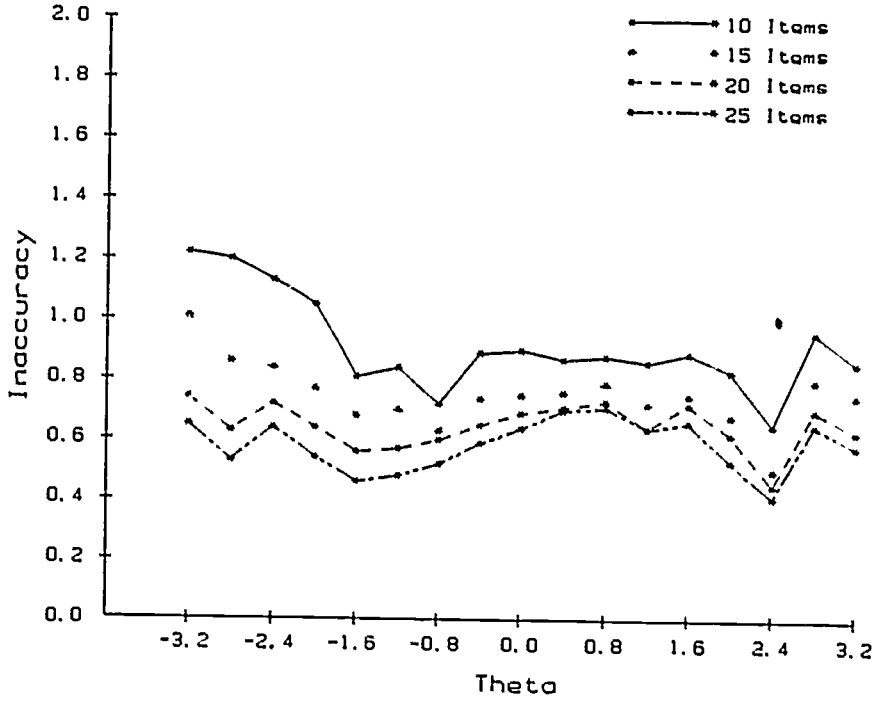
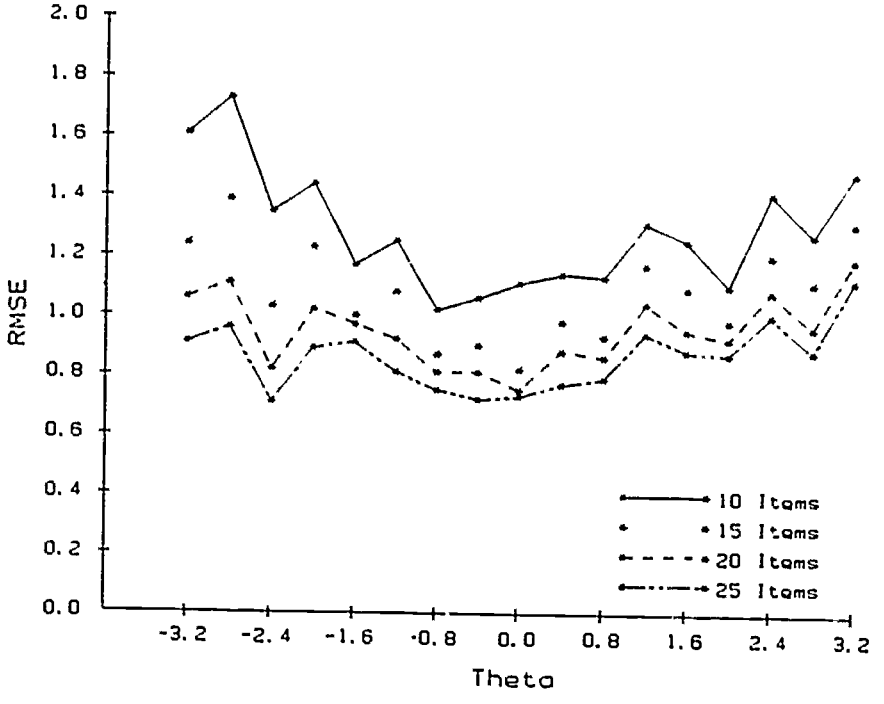


Figure 3
Conditional RMSE of θ Estimates for Tests of 10, 15, 20, and 25 Items for Dataset 4



across different test lengths yielded the same results as for inaccuracy. An example is shown in Figure 3 for Dataset 4. RMSE decreases with increasing test length, and the RMSE values for the shorter test lengths (10 and 15 items) vary more across θ levels than those for the longer tests.

Efficiency. Comparison of the mean efficiency levels for a given dataset across θ levels for a number of test lengths indicated that the efficiency levels increased and followed the same pattern, as test length increased. Figure 4 provides an example of these comparisons for Dataset 29 at 10-, 15-, 20-, and 25-item test lengths.

Since the results for all four conditional indices showed relatively systematic trends as a function of test length, the remainder of the results reported are only for the 15-item test length.

Effect of Multidimensionality

Tables 5 through 8 contain values of the distance measures across 17 θ levels for conditional values of each of the evaluative indices between each UD dataset and each of the MD datasets with the same strength first factor, for tests of 15 items in length. These tables also contain the proportions of the distance measure due to level, scatter, and shape effects.

Bias. Table 5 shows results of the D^2 profile analysis for bias. For the datasets based on OSAF (Datasets 1-17), the UD dataset (Dataset 1) generally had a higher mean bias (.38) and a lower variability (scatter) of bias (2.60) than did the MD datasets (2-17). When a second factor was added to the data (Datasets 2-8), D^2 values tended to increase with increasing strength of the second factor; the exception to this is Dataset 5, in which D^2 values were uniformly lower than in Dataset 4 even though the second factor in Dataset 5 was stronger. The effect proportions show that in all these datasets the vast majority of the differences in bias values as a result of multidimensionality was due to increased scatter; in Datasets 2-8 at least 87% of the differences in bias values from the UD dataset was due to scatter. Level effects accounted for most of the remaining effect for most of these datasets, with the exception of Dataset 2, in which the shape effect was slightly stronger than the level effect.

Similar results were observed for the 2-factor structure in which the first factor was strengthened. For Datasets 19-25, based on 1.5 OSAF, overall D^2 values increased regularly with increasing multidimensionality, but the absolute values of D^2 were smaller than for the 1.0 OSAF data. For Datasets 35-40 a similar but more irregular trend is evident, with smaller values of D^2 than for 1.0 OSAF or 1.5 OSAF, particularly for the higher strength second factors (Datasets 37-40). The effect proportions for these datasets are similar to those for the 1.0 OSAF data, though there is a tendency for multidimensionality to result in slightly greater differences in level, with consequent reductions in the scatter effect.

Figure 5 shows a typical result for bias with increasing multidimensionality for the 1.5 OSAF data. (The values plotted in this figure and in the other figures following are the untransformed values, so that the means and scatters

Figure 4
Conditional Efficiency of θ Estimates for Tests of 10, 15, 20, and 25 Items for Dataset 29

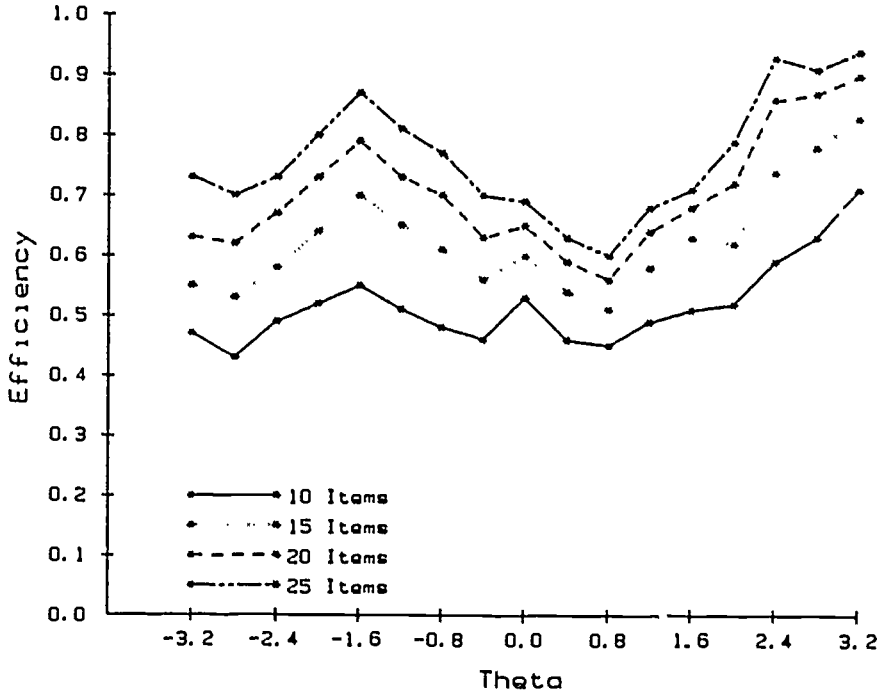


Figure 5
Conditional Bias of θ Estimates for Datasets 18, 21, 23, and 25

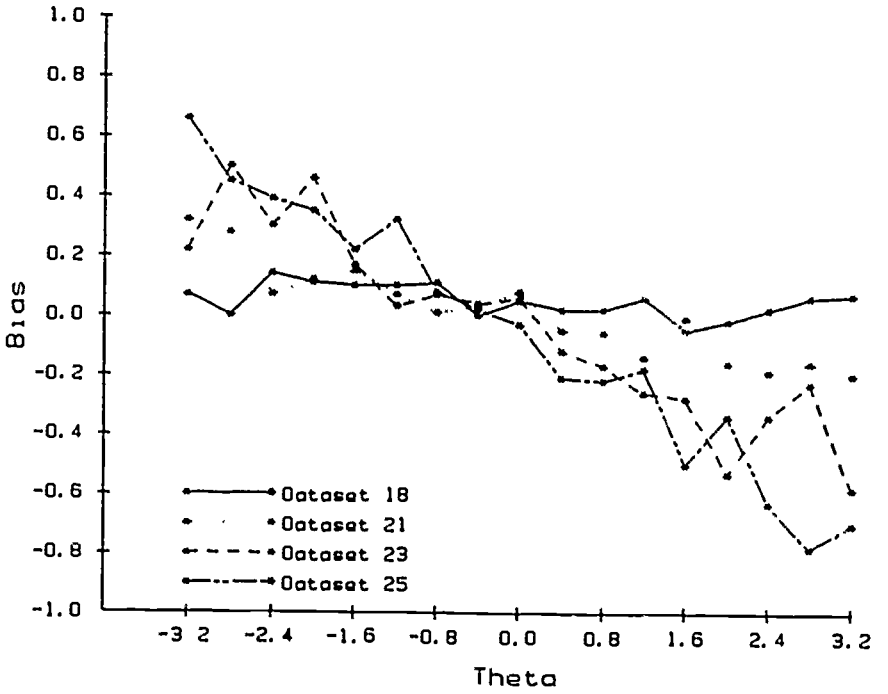


Table 5
Elevation (Mean) and Scatter of Bias (× 10) for Unidimensional (UD) and Multidimensional Datasets, Differences Between Elevation and Scatter, Total D^2 Index, D^2 with Elevation Removed ($D^{2\prime}$), D^2 with Elevation and Scatter Removed ($D^{2\prime\prime}$), and Proportion of D^2 Due to Level, Scatter, and Shape, for Tests of 15 Items

Dataset	Mean	Scatter	Difference Between		D^2	$D^{2\prime}$	$D^{2\prime\prime}$	Effect Proportion		
			Means	Scatter				Level	Scatter	Shape
1 (UD)	.38	2.60								
2	.31	3.76	.07	-1.17	25.524	25.439	2.468	.003	.900	.097
3	-.03	5.11	.41	-2.51	39.584	36.733	2.295	.072	.870	.058
4	-.20	13.57	.58	-10.98	212.407	206.674	2.445	.027	.962	.012
5	-.14	9.56	.52	-6.96	110.142	105.592	2.304	.041	.938	.021
6	-.22	14.13	.60	-11.53	225.007	218.948	2.345	.027	.963	.010
7	-1.19	18.04	1.57	-15.44	388.634	346.573	2.309	.108	.886	.006
8	-.98	17.97	1.36	-15.37	378.507	346.969	2.371	.083	.910	.006
9	.34	5.01	.04	-2.42	29.444	29.419	1.811	.001	.938	.062
10	-.79	12.39	1.17	-10.00	206.466	183.387	2.552	.112	.876	.012
11	-.37	18.31	.75	-15.72	354.696	345.152	2.063	.027	.967	.006
12	-.43	20.52	.81	-17.93	438.638	427.464	1.990	.025	.970	.005
13	.62	22.51	-.24	-19.92	535.791	534.840	2.365	.002	.994	.004
14	-1.23	22.27	1.60	-19.67	565.182	521.402	2.328	.077	.918	.004
15	-.56	28.55	.94	-25.96	852.904	838.016	2.216	.017	.980	.003
16	1.03	31.53	-.65	-28.94	1044.31	1037.20	2.441	.007	.991	.002
17	.06	3.84	.32	-1.25	22.926	21.212	1.971	.075	.839	.086
18 (UD)	.50	2.09								
19	.41	2.66	.09	-.57	4.011	3.876	.640	.034	.807	.159
20	.20	5.60	.30	-3.52	29.204	27.642	1.306	.054	.902	.045
21	.09	6.23	.41	-4.15	40.071	37.153	1.536	.073	.889	.038
22	-.09	9.04	.60	-6.95	79.711	73.661	1.343	.076	.907	.017
23	-.39	12.62	.89	-10.54	154.682	141.234	1.146	.087	.906	.007
24	-1.03	16.58	1.53	-14.49	296.347	256.390	1.343	.135	.861	.005
25	-.65	17.28	1.16	-15.19	293.890	271.188	1.121	.077	.919	.004
26	.39	5.07	.12	-2.98	22.341	22.108	1.252	.010	.934	.056
27	.38	14.73	.12	-12.64	192.998	192.752	1.073	.001	.993	.006
28	.03	15.61	.47	-13.53	229.319	225.539	1.307	.016	.978	.006
29	-.14	14.84	.64	-12.76	206.511	199.462	1.187	.034	.960	.006
30	.03	17.84	.48	-15.75	301.909	298.055	1.341	.013	.983	.004
31	.50	10.25	.00	-8.17	88.514	88.514	1.019	.000	.988	.012
32	-.28	18.49	.78	-16.40	332.691	322.277	1.379	.031	.965	.004
33	.41	3.89	.09	-1.80	12.352	12.208	1.105	.012	.899	.089
34 (UD)	.76	2.08								
35	.64	5.34	.12	-3.26	23.025	22.777	1.095	.011	.942	.048
36	-.02	8.11	.78	-6.03	55.720	45.257	.528	.188	.803	.009
37	-.28	10.74	1.05	-8.66	108.624	89.974	.671	.172	.822	.006
38	.07	9.65	.70	-7.57	77.603	69.378	.604	.106	.886	.008
39	.29	13.08	.48	-11.00	139.271	134.398	.494	.028	.968	.004
40	.21	11.62	.55	-9.54	109.813	104.641	.562	.047	.948	.005
41	.68	8.38	.08	-6.30	48.956	48.835	.527	.002	.987	.011
42	.46	9.94	.30	-7.86	71.616	70.057	.400	.022	.973	.006
43	.69	13.02	.08	-10.94	138.069	137.966	.677	.001	.994	.005
44	.75	13.94	.01	-11.86	155.258	155.254	.503	.000	.997	.003
45	-1.58	25.93	2.35	-23.85	687.320	593.785	.463	.136	.863	.001

are 1/10 of the comparable values in Tables 5 through 8.) This figure shows the effect of the strength of the second factor increasing from 1/3 of the first factor (Dataset 21) to 2/3 (Dataset 23) to 1.0 (Dataset 25). Bias for the UD dataset (Dataset 18) is close to zero throughout the θ range. For the MD datasets bias is close to zero for θ values close to 0.0, but it increases as the levels progress toward either extr me, resulting in the increased scatter due to increasing multidimensionality. Bias values are generally positive for θ values less than 0.0 and negative for θ values greater than 0.0. For Dataset 21 with the smallest second factor ('/3) bias is not substantially different from the UD dataset, except at extreme θ values; the major effect on bias for these datasets seems to occur for Dataset 23 (factor 2 = 2/3), with the additional 1/3 added to factor 2 in Dataset 25 resulting in generally little additional bias.

Results for the 3-factor datasets (9-16, 26-32, and 41-44) are also in Table 5. For the 1.0 OSAF data, overall D^2 increased regularly with increasing strength of the second and third factors; for the 1.5 OSAF data, values of D^2 were considerably lower, indicating less effect of increased strength of the second and third factors with the stronger first factor; this trend is further supported by Datasets 41-44 (2.0 OSAF), in which overall D^2 values were the lowest for all the 3-factor datasets. For all but one of the 3-factor datasets over 90% of the difference in bias values between the UD and MD datasets was due to scatter (the exception being Dataset 10 with .876), with secondary effects generally attributable to level effects.

Increasing dimensionality from two to three factors while holding constant total proportion of variance accounted for by the factors resulted in increased scatter of bias in most cases. For example, Dataset 6 was a 2-factor structure with the second factor 2/3 of the first, whereas in Dataset 11 both the second and third factors were 1/3 of the first factor. For Dataset 6 overall D^2 was 225, whereas Dataset 11 obtained a D^2 value of 355; in both cases the proportion of D^2 due to scatter was about .96. A similar effect was observed with the 1.5 OSAF data--overall D^2 for Dataset 23 was 155, whereas that for Dataset 28 was 229. The 2.0 OSAF data did not, however, exhibit this effect since overall D^2 for Datasets 39 and 43 were 138 and 139, respectively.

When results from the ASVAB 4-factor structure were compared to those of the relevant UD datasets, very minor effects on bias were observed when OSAF was used (Dataset 17) or when the first factor was increased to 1.5 its original strength (Dataset 33). In both cases mean bias was lower for the ASVAB structure than for the UD structure, though the scatter of the bias was slightly higher. The minor differences in bias for these datasets were, like the other MD structures, primarily due to scatter (.839 for Dataset 17 and .899 for Dataset 33). In contrast to the other MD structures, however, secondary effects were more important for shape than for level, indicating that the ASVAB structure changed the ordering of bias values across the 17 θ levels in comparison to the datasets. However, since there were very small effects on bias due to the ASVAB structure (overall D^2 values of 23 and 12), the shape effects are likely not important.

Using the ASVAB structure with the 2.0 OSAF data (Dataset 45) resulted in the largest overall D^2 for Datasets 35-45, a result considerably different than

that observed for Datasets 17 and 33. These data indicate that bias increased substantially both in overall level and variability from the comparable UD dataset, with 86% of the differences in bias due to scatter and 14% due to level. Since factors 2-4 were the same in all three ASVAB datasets, this difference can be attributed only to the increased absolute strength of the first factor in Dataset 45.

Inaccuracy. Table 6 contains the distance measures computed between the inaccuracy profiles of the UD datasets and each of the MD datasets with the same strength first factor. For the 2-factor structures overall, D^2 generally increases with increasing strength of the second factor in both the datasets based on 1.0 OSAF (2-8) and those based on 1.5 OSAF (19-25), with a similar but more irregular trend in the datasets based on 2.0 OSAF. As for the bias criterion, the value of D^2 tends to decrease as the strength of the first factor increases--even though the relative strength of the second factor is the same--indicating less effect on inaccuracy as the strength of the first factor increases. The effect proportions for these data show that differences in inaccuracy values were primarily the result of level effects that tended to increase with increased strength of the second factor. This increasing level effect occurred at the expense primarily of the scatter effect which, with a few exceptions, tended to decrease with increasing strength of the second factor. The only exception to the predominance of the level effect occurred when the second factor was 1/8 as strong as the first factor in Dataset 2, in which case the scatter effect was .547 and the level effect was .382; in the comparable datasets (19 and 35) with similar strength second factors but stronger first factors, the scatter effect was also relatively large. However, in all three of these datasets, D^2 was relatively small, indicating little effect on inaccuracy with a weak second factor.

A similar pattern was observed for the 3-factor structures (Datasets 9-16, 26-32, and 41-44). D^2 tended to increase with increasing strength of the second and third factors, although the trend was more irregular for the 1.5 and 2.0 OSAF data. In all cases level accounted for a minimum of 86% of the squared difference between inaccuracy values for the UD and MD datasets. There was also a marked tendency for the effect of the second and third factors to diminish substantially as the first factor increased in strength. For example, in Dataset 11 based on 1.0 OSAF and second and third factors each 1/3 as strong as the first factor, D^2 was 382 with 96% due to level; in Dataset 28 based on 1.5 OSAF D^2 was 326 with 98% due to level, and in Dataset 43 based on 2.0 OSAF D^2 was 141 with 88% due to level.

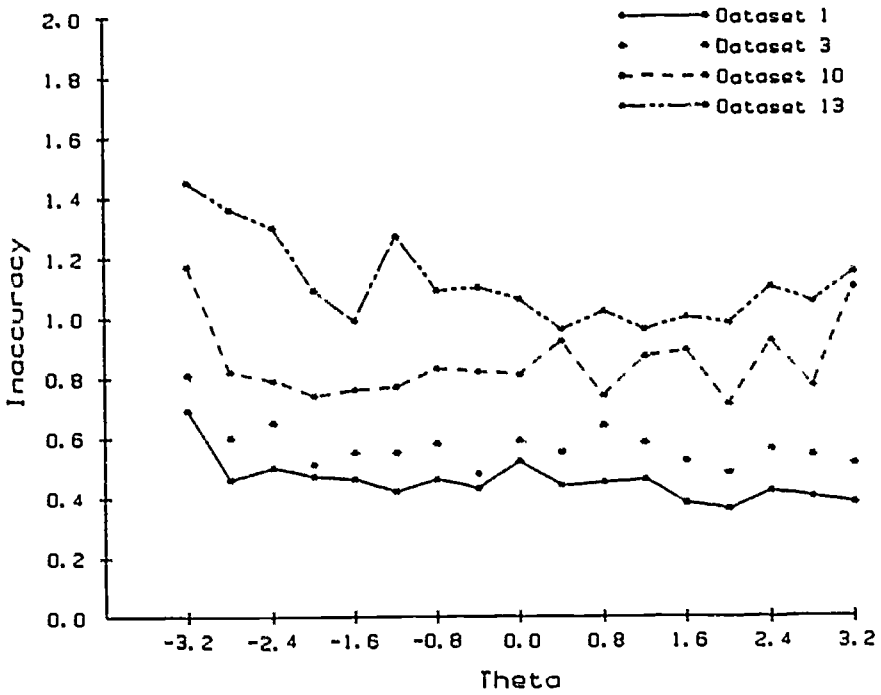
When the number of factors was increased from 2 to 3 while holding constant the proportion of variance accounted for by factors beyond the first, D^2 tended to increase, indicating a greater effect on inaccuracy for a larger number of factors. For example, in the 1.0 OSAF data, D^2 for Dataset 5 (2 factors, second factor 1/2 of first factor) was 130, whereas in Dataset 10 (3 factors, second and third factors each 1/4 of first factor) D^2 was 288; similar effects were observed in the 1.5 OSAF data for Datasets 22 versus 27 (D^2 = 91 vs. 275) and in the 2.0 OSAF data for Datasets 38 and 42 (D^2 = 71 vs. 98). Figure 6 illustrates the typical level effect found for inaccuracy within the 1.0 OSAF data. Dataset 3 with a weak (1/4) second factor results in inaccuracy values close to UD Dataset 1, whereas inaccuracy increases for Dataset 10 with two factors each 1/4 of

Table 6
Elevation (Mean) and Scatter of Inaccuracy (× 10) for Unidimensional (UD) and Multidimensional Datasets, Differences Between Elevation and Scatter, Total D^2 Index, D^2 with Elevation Removed ($D^{2\prime}$), D^2 with Elevation and Scatter Removed ($D^{2\prime\prime}$), and Proportion of D^2 Due to Level, Scatter, and Shape, for Tests of 15 Items

Dataset	Mean	Scatter	Difference Between		D^2	$D^{2\prime}$	$D^{2\prime\prime}$	Effect Proportion		
			Means	Scatter				Level	Scatter	Shape
1 (UD)	4.52	2.97								
2	5.17	2.95	-.65	.02	18.998	11.735	1.336	.382	.547	.070
3	5.71	3.12	-1.19	-.15	26.202	2.245	.240	.914	.077	.009
4	7.70	3.55	-3.19	-.57	190.049	17.495	1.628	.908	.083	.009
5	7.17	3.84	-2.65	-.86	129.530	10.312	.839	.920	.073	.006
6	7.91	3.19	-3.39	-.21	213.406	17.736	1.868	.917	.074	.009
7	8.61	6.10	-4.09	-3.13	329.511	44.879	1.932	.864	.130	.006
8	9.98	5.11	-5.46	-2.13	539.101	31.407	1.768	.942	.055	.003
9	6.03	2.92	-1.51	.06	45.036	6.235	.719	.862	.122	.016
10	8.48	4.95	-3.96	-1.97	287.725	20.615	1.138	.928	.068	.004
11	9.16	4.67	-4.64	-1.70	381.703	15.607	.916	.959	.038	.002
12	9.48	5.67	-4.96	-2.70	456.103	38.004	1.822	.917	.079	.004
13	11.13	5.87	-6.61	-2.89	766.013	22.639	.818	.970	.028	.001
14	10.62	6.32	-6.10	-3.35	698.322	65.996	2.914	.905	.090	.004
15	12.99	5.29	-8.47	-2.31	1256.73	37.017	2.014	.971	.028	.002
16	13.71	6.01	-9.19	-3.03	1459.01	22.300	.733	.985	.015	.001
17	5.81	2.47	-1.29	.51	36.439	8.198	1.083	.775	.195	.030
18 (UD)	3.47	2.24								
19	3.92	2.25	-.46	-.00	5.919	2.375	.471	.599	.322	.080
20	4.65	2.34	-1.18	-.10	26.112	2.344	.445	.910	.073	.017
21	4.69	2.48	-1.22	-.24	29.550	4.181	.741	.859	.116	.025
22	5.67	3.61	-2.20	-1.36	91.193	8.891	.868	.903	.088	.010
23	6.52	4.84	-3.05	-2.60	176.205	18.184	1.054	.897	.097	.006
24	7.37	6.08	-3.90	-3.84	286.368	27.907	.964	.903	.094	.003
25	8.30	5.40	-4.84	-3.16	424.959	27.536	1.450	.935	.061	.003
26	4.92	3.41	-1.45	-1.17	39.773	3.970	.339	.900	.091	.009
27	7.36	5.70	-3.89	-3.46	275.196	17.818	.458	.935	.063	.002
28	7.79	3.38	-4.32	-1.13	325.967	8.300	.925	.975	.023	.003
29	7.49	4.26	-4.02	-2.02	284.848	9.615	.579	.966	.032	.002
30	8.28	4.64	-4.81	-2.40	413.759	20.685	1.436	.950	.047	.003
31	6.87	3.08	-3.40	-.83	203.233	6.271	.808	.969	.027	.004
32	8.64	3.89	-5.18	-1.65	474.291	18.741	1.836	.960	.036	.004
33	4.35	3.06	-.88	-.82	17.426	4.166	.509	.761	.210	.029
34 (UD)	2.78	3.46								
35	3.59	3.62	-.80	-.16	19.489	8.578	.683	.560	.405	.035
36	4.20	3.76	-1.42	-.30	41.527	7.347	.558	.823	.163	.013
37	4.80	4.32	-2.02	-.86	89.533	20.142	1.298	.775	.210	.014
38	4.67	3.50	-1.88	-.04	70.768	10.465	.864	.852	.136	.012
39	5.70	4.25	-2.92	-.79	158.053	13.192	.855	.917	.078	.005
40	5.42	4.70	-2.64	-1.24	127.626	9.232	.474	.928	.069	.004
41	4.73	5.21	-1.95	-1.75	74.287	9.840	.377	.868	.127	.005
42	5.11	3.10	-2.33	.36	97.648	5.296	.482	.946	.049	.005
43	5.48	6.02	-2.70	-2.56	140.984	17.415	.522	.876	.120	.004
44	5.91	5.85	-3.13	-2.39	180.958	14.843	.450	.918	.080	.002
45	8.94	9.07	-6.15	-5.61	736.603	92.771	1.953	.874	.123	.003

the first factor, and increases again in Dataset 13 as factors 2 and 3 are again increased to 1/2 of the first factor.

Figure 6
Conditional Inaccuracy of θ Estimates for Datasets 1, 3, 10, and 13



The ASVAB factor structure (Datasets 17, 33, and 45) had slightly greater effects on overall D^2 for inaccuracy (Table 6) than it did for bias (Table 5). Similar to the bias data, however, the ASVAB structure resulted in lowest D^2 for the 1.5 OSAF data (Dataset 33) and a very high value of D^2 in the 2.0 OSAF data (Dataset 45). For all three datasets D^2 was primarily attributable to differences in level of conditional inaccuracy, with a secondary effect due to scatter of the inaccuracy values.

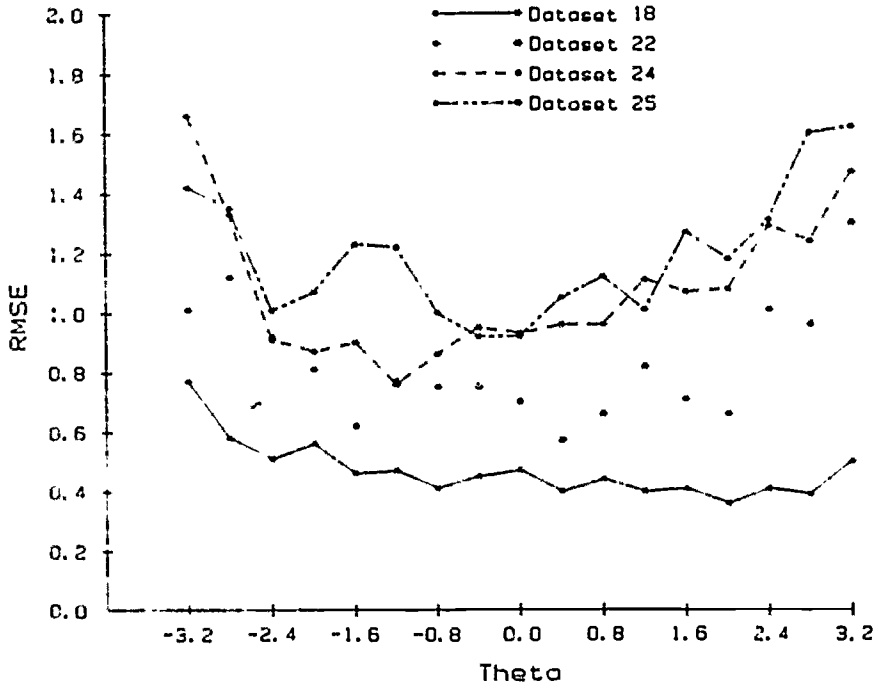
RMSE. The results for RMSE, shown in Table 7, have some similarity to those for inaccuracy. That is, for both the 2- and 3-factor structures D^2 generally increased as the strength of factors beyond the first increased. In addition, the magnitude of D^2 decreased with increasing strength of the first factor, indicating that the effect of factors beyond the first factor on RMSE was less with a stronger first factor, even though succeeding factors were proportionally as strong. In contrast to the inaccuracy results, however, for the 2-factor structures (Datasets 2-8, 19-25, 35-40), MD datasets resulted in RMSE values that were more variable than the UD datasets, as indicated by D^2 scatter proportions in the range of .10 to .20 for most of the 1.0 and 1.5 OSAF structures, and above .20 for many of the 2.0 OSAF datasets (35 to 38). With only one exception (Dataset 2), however, the predominant effect of multidimensionality was to increase the level of RMSE in all datasets, with the greatest level effects observed in the 1.5 OSAF data.

Table 7
Elevation (Mean) and Scatter of RMSE (× 10) for Unidimensional (UD) and Multidimensional Datasets, Differences Between Elevation and Scatter, Total D^2 Index, D^2 with Elevation Removed ($D^{2\prime}$), D^2 with Elevation and Scatter Removed ($D^{2\prime\prime}$), and Proportion of D^2 Due to Level, Scatter, and Shape, for Tests of 15 Items

Dataset	Mean	Scatter	Difference Between		D^2	$D^{2\prime}$	$D^{2\prime\prime}$	Effect Proportion		
			Means	Scatter				Level	Scatter	Shape
1 (UD)	5.92	5.04								
2	6.85	4.74	-.93	.29	52.258	37.554	1.568	.281	.689	.030
3	7.63	5.06	-1.71	-.02	57.435	7.568	.297	.868	.127	.005
4	10.79	5.43	-4.87	-1.39	460.096	56.266	1.677	.878	.119	.004
5	9.74	6.68	-3.82	-1.64	284.922	36.997	1.020	.870	.126	.004
6	11.05	6.57	-5.13	-1.53	508.037	60.475	1.755	.881	.116	.003
7	11.92	9.60	-6.00	-4.56	728.676	116.171	1.972	.841	.157	.003
8	13.43	8.02	-7.51	-2.98	1038.69	78.628	1.727	.924	.074	.002
9	8.36	5.07	-2.44	-.03	117.990	16.686	.653	.859	.136	.006
10	11.82	8.00	-5.91	-2.96	642.629	49.761	1.018	.923	.076	.002
11	12.64	7.41	-6.72	-2.38	811.771	43.702	1.019	.946	.053	.001
12	13.12	9.74	-7.21	-4.71	994.308	111.601	1.822	.888	.110	.002
13	15.02	8.76	-9.10	-3.72	1467.72	59.222	1.029	.960	.040	.001
14	14.29	9.25	-8.37	-4.21	1326.61	135.073	2.518	.898	.100	.002
15	17.16	8.18	-11.24	-3.14	2229.70	80.075	1.704	.964	.035	.001
16	17.95	9.31	-12.03	-4.27	2525.26	63.754	.971	.975	.025	.000
17	7.90	4.31	-1.98	.72	86.201	19.562	.876	.773	.217	.010
18 (UD)	4.70	3.92								
19	5.39	4.53	-.69	-.61	13.904	5.817	.306	.582	.396	.022
20	6.79	5.29	-2.08	-1.37	84.610	10.949	.438	.871	.124	.005
21	6.84	5.32	-2.14	-1.41	97.021	19.364	.834	.800	.191	.009
22	8.32	7.85	-3.62	-3.93	271.126	48.945	1.089	.819	.177	.004
23	9.68	8.32	-4.98	-4.41	479.630	58.372	1.195	.878	.119	.002
24	10.80	9.70	-6.10	-5.78	702.920	71.289	.995	.899	.100	.001
25	11.94	8.62	-7.23	-4.70	962.968	73.388	1.521	.924	.075	.002
26	7.15	6.21	-2.44	-2.29	112.695	11.181	.244	.901	.097	.002
27	10.80	9.14	-6.09	-5.22	678.314	46.987	.551	.931	.068	.001
28	11.28	6.41	-6.58	-2.50	770.580	34.305	1.118	.955	.043	.001
29	10.99	6.89	-6.28	-2.97	694.135	22.728	.515	.967	.032	.001
30	12.21	8.97	-7.50	-5.05	1025.00	68.198	1.215	.933	.065	.001
31	9.97	5.31	-5.27	-1.39	487.982	15.980	.676	.967	.031	.001
32	12.65	7.08	-7.95	-3.16	1122.39	49.268	1.417	.956	.043	.001
33	6.29	6.25	-1.59	-2.33	55.572	12.693	.297	.772	.223	.005
34 (UD)	4.09	6.65								
35	5.65	7.47	-1.56	-.81	78.437	37.022	.732	.528	.463	.009
36	6.70	7.21	-2.62	-.56	163.897	47.642	.987	.709	.285	.006
37	7.79	8.43	-3.70	-1.78	322.126	89.233	1.534	.723	.272	.005
38	7.61	8.06	-3.52	-1.41	276.660	65.480	1.184	.763	.232	.004
39	9.07	8.79	-4.98	-2.14	485.606	64.437	1.024	.867	.131	.002
40	8.68	8.70	-4.59	-2.04	405.497	47.229	.744	.884	.115	.002
41	7.75	9.10	-3.66	-2.44	264.622	37.109	.515	.860	.138	.002
42	8.32	6.12	-4.23	.53	323.970	20.145	.488	.938	.061	.002
43	8.97	10.26	-4.89	-3.61	462.204	56.519	.638	.878	.121	.001
44	9.48	10.08	-5.39	-3.43	546.413	53.325	.620	.902	.096	.001
45	13.12	12.92	-9.03	-6.27	1585.95	200.102	1.871	.874	.125	.001

Figure 7 shows a typical example of the RMSE results. This figure displays RMSE values for the 1.5 OSAF UD dataset (18) and MD Datasets 22, 24, and 25, in which the strength of the second factor increased respectively from 1/2 to 3/4 to 1.0 of the first factor. As can be seen, values of RMSE increased with increasing strength of the second factor, with only minor changes in their variability.

Figure 7
Conditional RMSE of θ Estimates for Datasets 18, 22, 23, and 25



The patterns of RMSE results for the ASVAB data structures were similar to those for inaccuracy. Lowest D^2 was observed for the 1.5 OSAF data (Dataset 33), whereas highest occurred for 2.0 OSAF. Even though the ASVAB structure included four factors, D^2 values for the 1.0 and 1.5 OSAF structures were in the range of those observed for 2-factor structures with second factors 1/8 to 1/4 those of the first factor (e.g., Datasets 2, 3, 19, 20). The ASVAB structure tended to result in D^2 values with a higher scatter effect for the 1.0 and 1.5 OSAF datasets, in comparison to most of the other MD datasets, indicating more variability in RMSE values as a function of θ levels than was evident in the corresponding UD datasets.

Efficiency. D^2 values for efficiency are in Table 8. With the exception of Dataset 2, the predominant difference in efficiency between the MD and UD datasets in the 2-factor data for 1.0 OSAF (Datasets 2-8) and 1.5 OSAF (Datasets 2-8 and Datasets 19-25) was due to level; MD structures resulted in fairly constant levels of lower efficiency in comparison to UD structures. In the 1.0 OSAF datasets the scatter/variability of observed efficiency values tended to

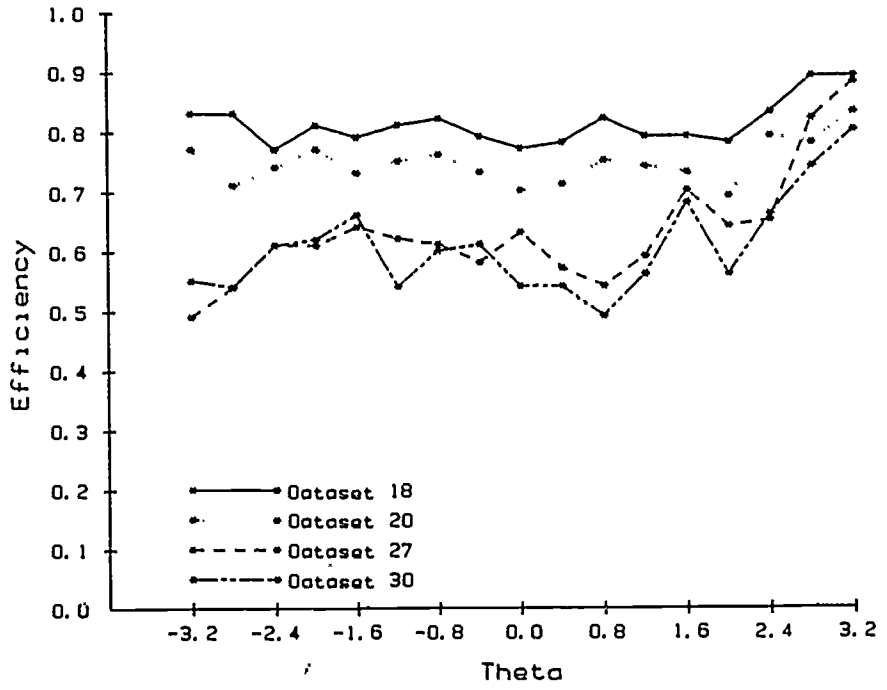
Table 8
Elevation (Mean) and Scatter of Efficiency (× 10) for Unidimensional (UD) and Multidimensional Datasets, Differences Between Elevation and Scatter, Total D^2 Index, D^2 with Elevation Removed ($D^{2\prime}$), D^2 with Elevation and Scatter Removed ($D^{2\prime\prime}$), and Proportion of D^2 Due to Level, Scatter, and Shape, for Tests of 15 Items

Dataset	Mean	Scatter	Difference Between		D^2	$D^{2\prime}$	$D^{2\prime\prime}$	Effect Proportion		
			Means	Scatter				Level	Scatter	Shape
1 (UD)	8.18	1.96								
2	7.92	1.98	.26	-.02	2.620	1.481	.383	.435	.419	.146
3	7.75	1.63	.42	.33	4.080	1.031	.289	.747	.182	.071
4	7.08	1.33	1.09	.63	22.520	2.169	.682	.904	.066	.030
5	7.31	1.37	.86	.59	13.890	1.179	.311	.915	.062	.022
6	7.05	1.16	1.12	.80	23.450	1.991	.597	.915	.059	.025
7	6.79	1.56	1.39	.40	37.060	4.298	1.355	.884	.079	.037
8	6.65	1.32	1.53	.64	42.840	3.075	1.033	.928	.048	.024
9	7.53	2.03	.65	-.08	8.060	.942	.235	.883	.088	.029
10	6.74	2.22	1.44	-.27	37.140	2.119	.471	.943	.044	.013
11	6.51	2.07	1.67	-.11	48.680	1.235	.302	.975	.019	.006
12	6.46	2.42	1.71	-.47	53.550	3.738	.742	.930	.056	.014
13	6.15	2.24	2.03	-.28	72.690	2.675	.593	.963	.029	.008
14	6.27	1.84	1.91	.11	68.460	6.709	1.857	.902	.071	.027
15	5.91	1.08	2.27	.88	89.980	2.335	.741	.974	.018	.008
16	5.64	2.12	2.54	-.16	113.460	3.681	.880	.968	.025	.008
17	7.71	1.55	.47	.40	4.680	.915	.247	.804	.143	.053
18 (UD)	8.11	1.44								
19	7.92	1.28	.19	.16	1.250	.609	.316	.512	.235	.253
20	7.46	1.42	.65	.02	8.110	.862	.421	.894	.054	.052
21	7.52	1.56	.59	-.12	7.010	1.009	.441	.856	.081	.063
22	7.15	1.58	.96	-.14	17.860	2.039	.886	.886	.065	.050
23	6.81	2.12	1.30	-.67	31.390	2.660	.723	.915	.062	.023
24	6.45	3.01	1.66	-1.57	55.270	8.159	1.317	.852	.124	.024
25	6.28	1.32	1.84	.12	60.100	2.839	1.481	.953	.023	.025
26	7.35	1.79	.76	-.35	11.390	1.601	.571	.859	.090	.050
27	6.31	3.86	1.81	-2.42	65.930	10.489	.835	.841	.146	.013
28	6.14	3.06	1.98	-1.62	72.320	5.911	.747	.918	.071	.010
29	6.26	3.61	1.85	-2.16	66.580	8.582	.750	.871	.118	.011
30	6.06	3.25	2.05	-1.81	78.710	7.062	.813	.910	.079	.010
31	6.66	2.29	1.45	-.85	37.580	1.982	.383	.947	.043	.010
32	6.18	1.92	1.94	-.48	65.990	2.319	.754	.965	.024	.011
33	7.62	1.81	.49	-.37	5.690	1.638	.573	.712	.187	.101
34 (UD)	7.80	2.06								
35	7.29	6.49	.51	-4.42	39.220	34.869	1.143	.111	.860	.029
36	7.04	6.15	.76	-4.08	43.320	33.379	1.318	.229	.740	.030
37	6.82	5.74	.98	-3.68	47.000	30.791	1.458	.345	.624	.031
38	6.90	6.49	.90	-4.42	50.650	36.880	1.294	.272	.703	.026
39	6.46	5.56	1.34	-3.50	57.280	26.701	1.258	.534	.444	.022
40	6.59	6.17	1.21	-4.11	58.540	33.578	1.311	.426	.551	.022
41	6.76	6.57	1.04	-4.50	53.820	35.599	1.130	.339	.640	.021
42	6.66	6.49	1.14	-4.43	56.380	34.241	1.091	.393	.588	.019
43	6.61	6.79	1.19	-4.72	63.490	39.249	1.210	.382	.599	.019
44	6.42	6.65	1.38	-4.58	69.750	37.265	1.184	.466	.517	.017
45	5.46	4.89	2.34	-2.83	122.740	29.561	2.135	.759	.223	.017

decrease with increasing strength of the second factor, with a somewhat more irregular trend observed for the comparable 1.5 OSAF datasets. For the 3-factor structures in the 1.0 and 1.5 OSAF datasets (Datasets 9-16 and 26-32), the predominant result was an overall reduction of efficiency values as the strength of the second and third factors increased. The level effect for these datasets tended to be in the high .80s and low .90s with a minor secondary effect due to scatter. In both the 1.0 and 1.5 OSAF structures, an increase from 2 to 3 factors while maintaining the same proportion of variance in the factors beyond the first led to decreases in efficiency, as shown by D^2 values of 23 for Dataset 6 (second factor 2/3 of the first) and 44 for Dataset 11 (second and third factors each 1/3 of the first).

Figure 8 shows the typical pattern of results for the 1.0 and 1.5 OSAF data. The UD data structure (Dataset 18) shows a fairly flat and high pattern of efficiency with a mean of .811. When a second factor 1/4 the strength of the first factor is added in Dataset 20, mean efficiency drops to .75 with little change in variability or shape. Datasets 27 and 30 show strong effects on efficiency through most of the θ range when two factors are added to the first. However, the strength of the second and third factors seems to have little effect on efficiency since factors 2 and 3 in Dataset 27 were each 1/4 of the first factor, whereas these factors each accounted for 1/2 the variance of the first factor in Dataset 30. The trend observed in Figure 8 for Datasets 27 and 30 appeared for most of the efficiency data--there was a tendency for strong second and third factor structures to have a greater effect for lower θ levels than for higher θ levels. This asymmetry was not evident in the bias, inaccuracy, or RMSE results.

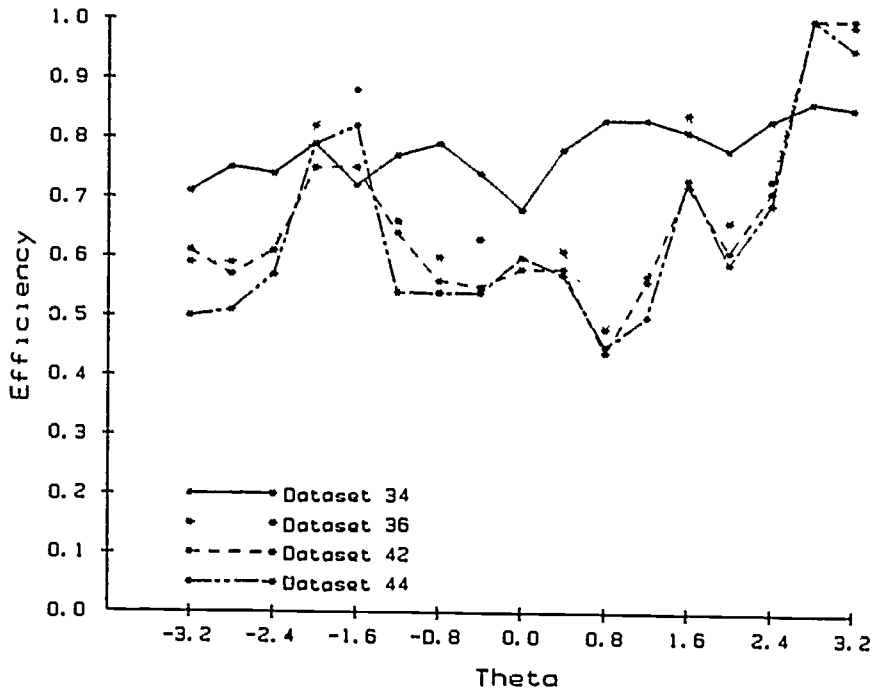
Figure 8
Conditional Efficiency of θ Estimates for Datasets 18, 20, 27, and 30



A different pattern of results emerged for the 2.0 OSAF data. For the UD Dataset 34, mean efficiency (.78) was slightly lower and its scatter higher than for UD Datasets 1 and 18. As the strength of the second and third factors increased, overall D^2 values increased to about the same levels as those observed in comparable 1.0 and 1.5 OSAF data, indicating similar overall reductions in efficiency; for example, D^2 in Dataset 40 (with a second factor 3/4 of the first factor) was 59, whereas the same structure in the 1.5 OSAF data (Dataset 24) resulted in a D^2 of 55. The difference in the 2.0 OSAF data versus the 1.5 and 1.0 OSAF was in the pattern of the efficiency results. Whereas in the latter data structures the predominant D^2 effect was for level, in the 2.0 OSAF data the majority of the change in efficiency due to multidimensionality was due to scatter, with proportions ranging from .86 for Dataset 35 to .44 for Dataset 39.

Figure 9 displays the typical pattern of results for the 2.0 OSAF data structures. UD Dataset 34 has the flattest and generally highest efficiency levels of the datasets plotted. The remainder of the datasets resulted in similar patterns of highly variable efficiency values, all following a similar pattern and differing little, even though Dataset 36 had only two factors with the second factor only 1/4 the strength of the first, whereas Datasets 42 and 44 were 3-factor structures with the second and third factors combined accounting for 1/2 and 3/4 the variance of the first factor, respectively. For all three of these datasets, efficiency values for the MD structures exceeded those of the UD structure for θ values in the range of -1.6 to -2.0 and above about 2.8.

Figure 9
Conditional Efficiency of θ Estimates for Datasets 34, 36, 42, and 44



Results for the ASVAB structures show small reductions in mean efficiency from .82 to .77 in the 1.0 OSAF data (Datasets 1 vs. 17), with a reduction in scatter; a similar small mean effect in the 1.5 OSAF data (.81 vs. .76); and a slight increase in scatter for Datasets 18 versus 33. When the first factor was increased to twice its original strength, addition of the three ASVAB factors resulted in a substantial decrease in mean efficiency and in a substantial increase in the variability of efficiency values; in Dataset 45 (the ASVAB structure) mean efficiency was .55 with scatter of .49, in comparison to values of .78 and .21 for the UD 2.0 OSAF structure (Dataset 34). However, for all three comparisons, level effects accounted for more than 70% of the differences between conditional efficiency levels for the ASVAB data and the comparable UD datasets.

CONCLUSIONS

As the overall degree of multidimensionality (as measured by the sum of the eigenvalues for each factor) in the generated item responses increased, the estimated θ values at each of the seventeen θ levels evaluated deviated further from the true (first factor) θ values. This effect was evident in the comparisons of overall bias, inaccuracy, and root mean square (RMSE) values for datasets with differing degrees of multidimensionality, and in all the conditional indices. These comparisons showed increasing levels of each of these evaluative indices as the multidimensionality of the underlying factor structure increased. The effect was also evident in the decreased efficiencies of datasets when compared to datasets with underlying factor structures that were more unidimensional. Individual θ estimates also ordered individuals differently from the true values, as reflected in the fidelity correlations. The pattern of results, therefore, suggests that maximum information adaptive testing is sensitive to changes in the dimensionality of the responses.

While all degrees of multidimensionality had effects on all the evaluative indices, effects were generally a function of the number of items administered. Thus, for the overall indices in all multidimensional datasets, fidelities increased with increasing test length, and inaccuracy and RMSE decreased, while overall bias tended to change from fairly high positive values for short test lengths to low negative values for the majority of multidimensional structures. For the conditional indices, very similar patterns of results were observed for different test lengths, with level effects (as opposed to scatter or shape effects) predominant for all but the bias index. Even for conditional bias, however, test length effects were roughly proportional for a given θ level. Consequently, while maximum information adaptive testing is affected by deviations from unidimensionality, the data suggest that in many cases, at least for relatively small degrees of multidimensionality, the effects of multidimensionality can be overcome simply by increasing test length. For example, the ASVAB factor structure resulted in a fidelity of .802 for a 15-item test compared to .872 for the UD case. When the multidimensional ASVAB structure was increased to 25 items in length, the fidelity of .871 was essentially the same as that of the 15-item unidimensional test. The same pattern was observed when the first factor of the ASVAB structure was strengthened by 50%.

The overall indices showed, in general, that increasing test length to twice the length of the multidimensional tests will overcome the effects of multidimensionality for multidimensional structures with one or two factors beyond the first that account for up to one-fourth the variance of the first factor. This finding held regardless of the strength of the first factor. Since a similar result was observed for the ASVAB structure (in which factors 2, 3, and 4 accounted for 22%, 13%, and 13% of the first factor, respectively) in the 1.0 and 1.5 OSAF data, the results suggest that the effects on maximum information adaptive testing of multidimensional factor structures in which up to one-third of the variance of the first factor appears in second and third factors, can be overcome by doubling adaptive test length. For degrees of multidimensionality beyond these levels, however, adaptive test lengths would need to be increased well beyond double to overcome the effects of multidimensionality. This conclusion must be qualified, however, when bias of the θ estimates is of concern, since the degree of bias differed at different θ levels.

There was some evidence to suggest that the number of factors (2 vs. 3), and not simply the overall strength of the underlying factor structure, affected θ estimates. For example, a single factor beyond the first had less effect on fidelity than did two factors that accounted for the same amount of variance. In addition, there was more scatter of conditional bias with three factors than with two, even though the proportion of variance in the second and third factors was equal in the two structures. Thus, the more complex factor structures seemed to affect the θ estimates more than the simpler structures. This finding, however, did not appear to extend to the 4-factor ASVAB structures.

Several factors affect the generality of the conclusions drawn from this research. First, the results are limited to the particular multidimensional model used to generate the multidimensional response vectors. Use of other models, such as those reviewed by Reckase and McKinley (1985), may yield different results. The results are also limited to maximum information adaptive testing with maximum likelihood scoring. Third, different factor structures might result in different findings, since only one basic first factor was used in this study. Thus, the study should be replicated varying these factors to further evaluate the robustness of adaptive testing to deviations from the unidimensional item response theory model used to select and to score test items.

REFERENCES

Betz, N. E., & Weiss, D. J. An empirical study of computer-administered two-stage ability testing (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1973.

Betz, N. E., & Weiss, D. J. Simulation studies of two-stage ability testing (Research Report 74-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1974.

Betz, N. E., & Weiss, D. J. Empirical and simulation studies of flexilevel ability testing (Research Report 75-3). Minneapolis: University of Minne-

sota, Department of Psychology, Psychometric Methods Program, July 1975.

Crichton, L. I. Effects of error in item parameter estimates on adaptive testing. Unpublished doctoral dissertation, University of Minnesota, 1981.

Cronbach, L. J., & Gleser, G. C. Assessing similarity between profiles. Psychological Bulletin, 1953, 50, 456-473.

Drasgow, F., & Parsons, C. K. Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement, 1983, 7, 189-199. (Also this volume, pp. 218-232.)

Johnson, M. F., & Weiss, D. J. Parallel forms reliability and measurement accuracy comparison of adaptive and conventional testing strategies. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Kiely, G. L., Zara, A. R., & Weiss, D. J. Alternate forms reliability and concurrent validity of adaptive and conventional tests with military recruits (Draft Final Report of Contract N00123-79-C-1273, submitted to Navy Personnel Research and Development Center). University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, January 1983.

Larkin, K. C., & Weiss, D. J. An empirical investigation of computer-administered pyramidal ability testing (Research Report 74-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, July 1974.

Larkin, K. C., & Weiss, D. J. An empirical comparison of two-stage and pyramidal adaptive ability testing (Research Report 75-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1975.

Mattson, J.D. Effects of item parameter error and other factors on trait estimation in latent trait-based adaptive testing. Unpublished doctoral dissertation, University of Minnesota, 1983.

McBride, J. R. Some properties of a Bayesian adaptive ability testing strategy. Applied Psychological Measurement, 1977, 1, 121-140.

McBride, J. R., & Martin, J. T. Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press, 1983.

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

Reckase, M. D., & McKinley, R. L. Some latent trait theory in a multidimensional space. In D. J. Weiss (Ed.), Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference. Minnepolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1985, pp. 151-177.

Sympson, J. B., Weiss D. J., Ree, M. J. Predictive validity of conventional and adaptive tests in an Air Force training environment (AFHRL-TR-81-40), Brooks Air Force Base TX: Air Force Systems Command, Manpower and Personnel Division, August 1982.

Urry, V. W. Computer-assisted testing: The calibration and evaluation of the verbal ability bank (Technical Study 74-3). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, December 1974.

Vale, C. D., & Weiss, D. J. A study of computer-administered stradaptive ability testing (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1975. (a)

Vale, C. D., & Weiss, D. J. A simulation study of stradaptive ability testing (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1975. (b)

Vale, C. D. Problem: Strategies of branching through an item pool. In D. J. Weiss (Ed.), Computerized adaptive trait measurement: Problems and prospects (Research Report 75-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1975.

Weiss, D. J., & McBride, J. R. Bias and information of Bayesian adaptive testing. Applied Psychological Measurement, 1984, 8, 273-285.

ACKNOWLEDGMENT

This research was supported by Contract NOOO14-79-C-0172, NR 150-433, with the Office of Naval Research, with additional funding from the Air Force Human Resources Laboratory, Army Research Institute, and the Air Force Office of Scientific Research.