

The Interpretation and Application of Multidimensional Item Response Theory Models; and Computerized Testing in the Instructional Environment Final Report

Mark D. Reckase

Research Report ONR 89-2 September 1989



The American College Testing Program Assessment Programs Area Test Development Division Iowa City, Iowa 52243

Prepared under Contract No. N00014-85-C-0241 with the Cognitive Science Research Programs Office of Naval Research

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.



89 11 06 113

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT	OCUMENTATIO	ON PAGE Form Approved OMB No 0704-01				
1a REPORT SECURITY CLASSIFICATION		16 RESTRICTIVE MARKINGS				
Unclassified						
2a SECURITY CLASSIFICATION AUTHORITY		3 DISTRIBUTION AVAILABILITY OF REPORT				
26 DECIASSIFICATION / DOWNGRADING SCHEDU	LE	Approved for public release: distribution unlimited. Reproduction in whole or in				
4 PERFORMING ORGANIZATION REPORT NUMBE	R(S)	5 MONITORING	ORGANIZATION R	EPORT NU	MBER(S) U.S. Governme	
ONR 89-2						
6a NAME OF PERFORMING ORGANIZATION	7a NAME OF MO	DNITORING ORGA	NIZATION			
ACT	Cognitive	Science Re	esearch	1 Programs		
Contraction Contraction Contraction		Office of	Naval Res	earch		
DO BUE 100		76 ADDRESS (CIT	y, state, and zir i	Code)		
PU BOX 168		Code 1142	CC AV 0001	7 5000	f i	
Towa City, IA 52243		Ariingtor	I, VA 2221	7-3000	1	
Ba. NAME OF FUNDING / SPONSORING ORGANIZATION	8b OFFICE SYMBOL (If applicable)	9 PROCUREMEN	T INSTRUMENT ID	ENTIFICAT	ION NUMBER	
		N00014-85	5-C-0241			
8c. ADDRESS (City, State, and ZIP Code)	•	10 SOURCE OF F	UNDING NUMBER	25		
	PROGRAM ELEMENT NO	PROJECT NO PRO4 204	TASK NO	WORK UNIT ACCESSION NO		
Mark D. Reckase 13a TYPE OF REPORT Final FROM 8	DVERED 5/6TO_89/9_	14 DATE OF REPO 89/9	RT (Year, Month, }	Day) 15	PAGE COUNT	
16 SUPPLEMENTARY NOTATION 17 COSATI CODES FIELD GROUP SUB-GROUP	18 SUBJECT TERMS (Item response theory, MIRT,	Continue on revers theory, mul computerize	e if necessary and tidimensior d testing,	<i>didentify</i> nal ite medium	<i>by block number)</i> m response n effects	
19 ABSTRACT (Continue on reverse if necessary This report summarizes the	and identify by block r work in the dev	number) velopment and		n of mu	ultidimensional	
item response theory models testing in the instructional of descriptive statistics for completion, the evaluation models, and the application The work on computerized test achievement tests in a mili adaptive testing in that se	(MIRT) and in 1 environment. for items that r of calibration of MIRT process sting investiga tary training s tting.	the evaluation The work or require more programs, the dures to prace ated medium a setting and the control of the the control of the	ion of the MIRT incl than an ab ne comparis tical test and order e the feasibi	use of uded th ility f on of o ing pro ffects lity of	computerized ne definition for successful competing MIRT oblems. for classroom f computerized	
20 DISTRIBUTION AVAILABILITY OF ABSTRACT		21 ABSTRACT SEC Unclassi	CURITY CLASSIFIC	ATION		
22a NAME OF RESPONSIBLE NOIVIDUAL Dr. Charles Davis		226 TELEPHONE (202/696-4	Include Area Code 4046	0NI	R 1142 CS	
)D Form 1473, JUN 86	Previous editions are S/N=0102-LF-0	obsolete 14-6603	SEC URITY	CLASSIEG	VEON OF THIS PACE	

The Interpretation and Application of Multidimensional Item Response Theory Models

Final Report

Contents

Page	e
Introduction	
Multidimensional Item Response Theory (MIRT)1	
Conceptual Framework2	;
Characteristics of Multidimensional Test Items)
Estimation of MIRT Item Parameters14	
Evaluation of MIRT Parameter Estimation Programs	
Alternative MIRT Models22	
Procedures for Test Construction	j
Computerized Testing	,
Hardward and Software	,
Research Studies	I
Medium Effects	I
Order Effects40)
Estimation of Parameters41	
Alternatives to Fully Adaptive Testing42	!
References	,

Accession For NTIS GRA&I DTIC TAB Unannounced Justification By____ Distribution/ Availability Codes Avail and/or Special Dist

The Interpretation and Application of Multidimensional Item Response Theory Models

Final Report

The work on this research contract consisted of two distinct and unrelated parts: (1) the investigation of the characteristics of multidimensional item response theory (MIRT) models, and (2) the application and evaluation of computerized testing procedures in a military training environment. For each of these two parts of the project, the initial objectives will be specified, the research performed will be summarized, and conclusions about the research will be presented. The work on the MIRT models will be presented first followed by the work on computerized testing. References to contract publications will be given as resources for more complete presentations of the contract work.

Multidimensional Item Response Theory (MIRT)

Multidimensional item response theory (MIRT) is a theoretical framework for describing the interaction between a person and a test item when it is believed that performance on the test item is sensitive to person differences on more than one dimension. Previous ONR contract work on MIRT (Models for Multidimensional Tests and Hierarchically Structured Training Materials. NR 150-474, N00014-81-K0817: Reckase, 1985) identified a number of possible MIRT models. Work on this contract focused on bringing these models to operational status. The objectives for this part of the contract, as stated in the original proposal, are listed below.

- (KK)

- The development of a conceptual framework for studying tests composed of items sensitive to multiple cognitive dimensions.
- The description of the characteristics of multidimensional test items.
- 3. The development of a computer program to estimate the characteristics of multidimensional items.
- The evaluation of the available model parameter estimation programs.
- 5. The evaluation of alternative MIRT models.
- The development of procedures for using MIRT concepts for test construction.

Although these objectives overlap somewhat, work on each one will be summarized separately.

Conceptual Framework

Before work could be done to develop applications of MIRT to practical testing problems, a good understanding was needed of what was meant by (1) a multidimensional ability test, and (2) test items that are sensitive to differences in cognitive abilities on multiple dimensions. In order to conceptualize these issues, an ability space was first considered that encompassed all the areas of variation that exist in the cognitive skills in a large population of individuals. A person's position in this space is indicated by a vector of numbers, $\boldsymbol{\theta}$, specifying a location in a Cartesian coordinate system. The coordinates of this space are defined as orthogonal for mathematical convenience. This does not imply that the abilities are

orthogonal. Persons may vary in the space along directions that may not coincide with the coordinate axes. If these directions of variation are thought of as ability dimensions, then the axes of the ability space and the coordinate space may not be the same. They have the same dimensionality, and there is a one-to-one correspondence between a point located using one set of axes and a point located using the other, but the axes of one space do not necessarily correspond to the axes of the other. Different frames of reference are provided by each of these conceptual spaces. The term "0-space" will typically refer to the space with orthogonal coordinate axes that is the mathematically convenient frame of reference.

The common dimensionality of these two reference systems indicates the number of orthogonal measures that must be obtained to unambiguously locate a person in the space. In principle, test items can be developed to distinguish between persons on each one of the coordinate system dimensions, or combinations of the dimensions. However, a particular test may not contain items that are sensitive to differences on all of the dimensions, or examinees may not differ across all tested dimensions. Therefore, the dimensionality of the data generated by the administration of a test to the population of interest may not have the same dimensionality as the full **0**-space.

In order to more completely specify what is meant by the dimensionality of a test, the dimensionality of test items will first be addressed. In this analysis, only dichotomously scored test items will be considered. Items that yield more than two score categories greatly complicate the analysis. If a test item is administered to persons who are distributed throughout the multidimensional space, the item will divide the population into two-groups-those that responded correctly, and those that did not. If the item is sensitive to differences in the abilities of the persons in the space, the

item may divide the space into two parts that are fairly distinct with persons in one region having mostly correct responses, and persons in another region with mostly incorrect responses. Correct responses are assumed to be related to a higher level of an ability dimension than incorrect responses. Figure 1 provides an example of responses to such a test item in a two-dimensional space. The more sensitive the item is to differences in abilities in the space, the sharper will be the transition between the regions defined by the correct and incorrect responses.

Figure 1. Typical responses to an item for persons located at different points in a two-dimensional ability space.

The data generated by the interaction between one item and the persons in the population is always unidimensional. Persons will vary only on the dimension of correctness of the item. However, that dimension may not correspond to any of the axes in the θ -space or the dimensions in the ability space. The dimension corresponds to the direction of greatest rate of change from incorrect responses to correct responses. This direction may be the same for all points in the space, implying that the item requires the same combination of skills for all persons, or the direction of greatest rate of change may vary depending on the location in the space, implying that different combinations of skills are required for different persons in the space.

Examples of these two conditions are presented in Figures 1 and 2. In Figure 1, the direction of greatest rate of change is basically from lower left to upper right and is the same for persons at the upper left or at the lower right. In Figure 2, the direction of greatest rate of change is not the same for all points in the space. For persons at the upper left, the density of zeros and ones changes horizontally. For persons at the lower right, the change in density is vertical. Thus, the item shown in Figure 2 measures (i.e., is sensitive to) different combinations of skills at different points in the θ -space. However, in both cases the data generated by the interaction between the single item and the examinee population are unidimensional since persons can be ordered on a single number, the probability of a correct response, and this number summarizes the information in the zeros and ones. The implication of this conceptualization is that all one-item, dichotomously scored tests are unidimensional from a statistical perspective even though the item may require relatively high levels of skills on several different cognitive dimensions to arrive at the correct response.

Figure 2. Typical response to an item with sensitivity to different dimensions at different points in a two dimensional ability space.

When a test is composed of more than one question, the issue of the dimensionality of the data generated by the interaction of the population of persons and the items becomes more interesting and more complex. If the direction of maximum rate of change of the density of zeros and ones is the same at each point in the space for all of the items in the test, then the test is unidimensional from a statistical perspective. However, if the direction of maximum rate of change at each point in the space differs across items and the density of correst responses increases monotonically in a particular direction, the dimensionality of the data generated by the interaction of the population and the test is equal to the dimensionality of the space needed to contain all of the directions specified.

In an idealized case, the distribution of zeros and ones over the space for an item can be shown by a surface that represents the proportion of correct responses at each point in the space. For one item, a particular proportion of correct responses, \underline{p} , will trace a contour, labeled the pcontour, on the surface. All persons at θ -points related to a particular p-contour have the same probability of correctly responding to the item. Thus they can all be mapped onto a unit line at point \underline{p} to represent their relative ability on the ability measured by the item. Since all persons in the space (all θ -points) can be mapped onto this line, the interaction between the single item and the examinee population generates data that is unidimensional. Some of the contours for the items shown in Figures 1 and 2 are given in Figures 3 and 4. The degenerate case where all persons in the population have the same probability of a correct response to the item will be ignored.

If a second test item is administered to the same population as the first item, the proportion of correct responses on this new item for all the θ -points corresponding to the original p-contour can be determined. If the resulting proportions are all the same, they define a contour for item two, a p'-contour where p' may differ from p, that is related to exactly the same set of θ -points as the p-contour. If this is true for all of the contours that can be defined for items 1 and 2, the interaction of the population and the items will generate data that are unidimensional. All persons with θ -points related to the (p, p')-contours can be mapped onto the same point on a line to show the relative position on the construct measured by these two items to the θ -points related to other pairs of contours.

In general, if <u>n</u> items have contours that are identical except for differing on the proportion of correct responses, the interaction of these <u>n</u>



Figure 3. Equiprobable contour plot corresponding to Figure 1.



<u>Figure 4</u>. Equiprobable contour plot corresponding to Figure 2.

items and the examinee population will generate data that are unidimensional. If the contours do not match, then the maximum possible dimensionality of the data is equal to the number of distinct sets of items that have matching contours, but that have contours that differ from other sets of items.

An actual matrix of dichotomous responses may not have the maximum possible dimensionality that can be generated by a test when the examinee population does not vary on all of the dimensions to which the items are sensitive. Thus, the actual dimensionality of a dichotomous data matrix will be equal to the number of dimensions to which the test items are sensitive <u>and</u> for which there is variation in the examinee population. The dimensionality of the data matrix must always be less than or equal to the smaller of the number of dimensions of variation of the examinee population and the number of dimensions of sensitivity of the items in the test.

The key points in this discussion are that:

- (1) test items may require more than one cognitive skill for successful solution but still generate a statistically unidimensional data set through the interaction with a population that varies on many dimensions.
- (2) a multidimensional ability space has many useful frames of reference -a mathematical coordinate system, cognitive ability dimensions, and
 dimensions defined by what is measured by one or more test items.
- (3) the statistical dimensionality of a data set is the result of the interaction of the dimensions of variation of the population and the dimensions of sensitivity of the test items.

(4) single test items generate data that are unidimensional.

The coordinate system specified in Figures 1 and 2 is strictly arbitrary. All of the above points hold for any monotonic transformation of the 0-space. However, leaving the metric of the 0-space undefined does not allow a very parsimonious description of the interaction between a person and an item. To achieve a parsimonious description of this interaction, it is desirable to select a metric that results in proportion-correct surfaces that have a convenient form. Many different forms are possible and the selection of one of them is usually made on the basis of theoretical beliefs about the underlying psychological process, or for mathematical convenience. Of course, once a model is selected, its usefulness depends on how well it models empirical data rather than other considerations. In the research performed on this contract, several different models were considered and an important issue is which of these best models data gener.ted by the actual interaction of persons and test items.

Characteristics of Multidimensional Test Items.

It is almost trivial to state that the purpose of a test item is to provide information about the capabilities of an examinee. A not so trivial issue is how much information is provided and for what examinees does the item provide the most information. In the unidimensional item response theory (UIRT) context, the functioning of the test item is described by a measure of location (b-parameter), a measure of discriminating power (a-parameter), and a function that indicates the amount of information provided at each point on the ability scale. These item descriptors indicate the group of examinees the

item is best at measuring according to the group location on the ability line. The intent of the research in this portion of the project was to generalize these UIRT-features for items that require multiple abilities for successful solution. Thus a measure of location of an item in the ability space, its discriminating power, and the information it provides about examinee abilities were desired.

The conceptual framework given above provides a method for describing the characteristics of an item in a way that makes use of the same concepts as UIRT. In UIRT, the measure of location indicates the ability level at which an item is most discriminating. Thus for a MIRT-based item description, it would also be useful to indicate the location of the item as the ability vector for which the item is most discriminating. However, for an item in a multidimensional space, there may be many points where the item is most discriminating. That is, there may be many locations in the space where the rate of change of the probability of a correct response is the highest that occurs for the item. The solution proposed here is to define the analog of the UIRT b-parameter as the distance and direction from the origin of the 9-space to the nearest point of maximum discrimination.

Thus, multidimensional item difficulty (MDIFF) has two component parts, the distance, \underline{D} , from the origin of the space to the point of maximum rate of change, with distance specified in terms of the metric of the ability space, and the direction specified by the vector of angles, **a**, between the coordinate axes and the line connecting the origin and the point of maximum slope.

The analog of the a-parameter of UIRT, MDISC, is defined as a function of the slope of the proportion correct surface at the point of maximum rate of change in the direction, α , from the origin. The multidimensional information function, MINF, is defined by exactly the same expression as is used in UIRT. That is

$$I_{\alpha}(\theta) = \frac{(slope_{\alpha}(\theta))^{2}}{variance(\theta)}.$$

However, for the MIRT case, the information is indexed by a particular direction, α . That is, an item does not provide a certain amount of information about a particular θ , it provides information for differentiating θ from a nearby θ' that is less than ε away in a particular direction α . An item may provide substantial information in one direction while providing no information for differentiating between persons in another direction.

These descriptive features of a test item in the multidimensional space have been derived explicitly for one of the MIRT models given by

$$P(\underline{u}_{i} = 1 | \boldsymbol{\theta}_{j}, \mathbf{a}_{i}, \underline{d}_{i}) = \frac{e^{\mathbf{a}_{i} \boldsymbol{\theta}_{j} + \underline{d}_{i}}}{1 + e^{\mathbf{a}_{i} \boldsymbol{\theta}_{j} + \underline{d}_{i}}}, \quad (1)$$

where \underline{u}_i is the item response

 $\boldsymbol{\theta}_{j}$ is a vector of abilities \boldsymbol{a}_{i} is a vector of discrimination parameters and \underline{d}_{i} is a scalar related to item difficulty.

Based on this model, the MIRT descriptive features of an item are given by;

MDISC₁ =
$$(\sum_{k=1}^{2} \sum_{k=1}^{2} k)^{\frac{1}{2}}$$
, (2)

$$\underline{\mathbf{D}}_{i} = \frac{-\underline{\mathbf{d}}_{i}}{\mathsf{MDISC}_{i}},$$
(3)

$$\cos \alpha_{ik} = \frac{\underline{a}_{ik}}{MDISC_{i}}, \qquad (4)$$

and
$$MINF_{\alpha}(\theta) = P_{i}(\theta_{j}) Q_{i}(\theta_{j}) (\sum_{k} \underline{a}_{ik} \cos \alpha_{ik})^{2}$$
, (5)

where $P_i(\boldsymbol{\theta}_j) = P(\underline{u}_i = 1 | \boldsymbol{\theta}_j, \mathbf{a}_i, \underline{d}_i)$ and $Q_i(\boldsymbol{\theta}_j) = 1 - P_i(\boldsymbol{\theta}_j)$.

The derivations of these statistics are given in

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9(4), 401-412.

and

Reckase, M. D. (1986, April). <u>The discriminating power of items that measure</u> <u>more than one dimension</u>. Paper presented at the meeting of the American Educational Research Association, San Francisco.

For a model with a nonzero lower asymptote, \underline{c}_i , given by

$$P_{i}(\boldsymbol{\theta}_{j}) = \underline{c}_{i} + (1 - \underline{c}_{i}) \frac{e^{\mathbf{a}_{i}\boldsymbol{\theta}_{j}} + \underline{d}_{i}}{1 + e}$$
(6)

the statistics given by Equations 2, 3 and 4 still hold, but MINF is given by

$$MINF_{\mathbf{a}}(\boldsymbol{\theta}_{j}) = \frac{C_{i}(\boldsymbol{\theta}_{j}) (P_{i}(\boldsymbol{\theta}_{j}) - \underline{c}_{i})}{P_{i}(\boldsymbol{\theta}_{j}) (1 - \underline{c}_{i})} (\sum_{k} \underline{a}_{ik} \cos \alpha_{ik})^{2}$$
(7)

(Carlson, 1988; personal communication).

While the concepts of MDIFF, MDISC, and MINF presented here can be applied to any functional description of the proportion of correct responses at points in the θ -space that have increasing proportions of a correct response with increases in any θ_i or combinations of θ_i , many mathematical forms for the function will not yield convenient or even unique solutions. The models given in Equations 1 and 6 have particularly nice mathematical properties.

Estimation of MIRT Item Parameters

The statistics defined in the previous section would have little usefulness if estimates of the parameters \mathbf{a}_i , \underline{d}_i , and $\mathbf{\theta}_j$ were not attainable. Therefore, a computer program for obtaining estimates of these parameters was a high priority for this project.

After reviewing several methods for obtaining estimates, a joint maximum likelihood procedure, as implemented in LOGIST (Wingersky, Barton & Lord, 1982), was selected as a basis for the program. The resulting program, labeled Multidimensional Item Response Theory Estimation (MIRTE), has been thoroughly checked out and has been implemented on numerous data sets. The following report serves as a manual for the program.

Carlson, J. E. (1987, September). <u>Multidimensional item response theory</u> <u>estimation: a computer program</u> (Research Report ONR 87-2). Iowa City, IA: ACT.

Evaluation of MIRT Parameter Estimation Programs

In order to apply MIRT, good estimates must be obtained for the MIRT model parameters. As part of this project, MIRTE was produced to yield these

estimates. However, there are many estimation procedures that can be used to obtain the parameter estimates and it is not clear which procedure, or which particular coding of a procedure, will yield the best estimates of the parameters. Therefore, as part of this project, four estimation programs were evaluated for use in obtaining MIRT model parameter estimates. These programs were MIRTE (Carlson, 1987), TESTFACT (Wilson, Wood, & Gibbons, 1984), MULTDIM (McKinley, 1987), and LISCOMP (Muthén, 1987).

In evaluating these programs, two sets of criteria were used. The first were practical criteria concerning whether the programs could yield the required estimates and how easy the programs were to use. The second criteria were based on how well the programs recovered the parameters of two sets of simulated test data that were specifically designed for benchmark testing.

The practical criteria used for evaluating the programs are listed in Table 1. Table 2 provides a summary of information concerning these criteria. From the summary, it can be seen that no program had a particular advantage on many of the criteria. However, LISCOMP is notable in being the least expensive to use. MULTDIM seems to be the easiest to use, at least initially. Beyond these considerations the programs would have to be selected on the basis of preference for a model, or based on the particular needs of an analysis. The analysis of the benchmark data may help clarify the issue of program choice.

Table 1

Practical Criteria for Evaluating MIRT Estimation Programs

- 1. What limits are placed on the maximum number of dimensions that can be estimated with the program?
- 2. Can a non-zero lower asymptote be specified?
- 3. What type of estimation algorithm is used?
- 4. How well is the implementation of the algorithm described?
- 5. Can item- and/or ability-parameters be fixed to allow the estimation of only specified parameters?
- 6. Can the procedure be used to analyze a sparse matrix?
- 7. What MIRT model is specified?
- 8. What type of goodness of fit statistics are provided?
- 9. What constraints are placed on parameter estimates?
- 10. Are multiple group analyses possible?
- 11. What summary statistics are reported?
- 12. What limits are placed on the maximum number of persons and items?
- 13. How difficult is program set up?
- 14. How much does it cost to perform an analysis?

Table 2 Evaluation of MIRT Estimation Programs on Practical Criteria

Program MIRTE TESTFACT MULTDIM L13COMP	L1 3COMP	2	No	Generalized least squares	Good description	Yes	No	Normal Ogive	Chi-square	None	Yes	Extensive	ż	Some difficulty	Under \$10
	MULTDIM	5	Yés	Marginal maximum likelihood	Good description	No	Yes	Logistic	Likelihood ratio chi square	Set by user	N	Extensive	600 items	Easy	Over \$100
	TE STF ACT	۲.	Yes	Murginal maximum likelihood	Good description	No	No	Normal Ogive	Likelihood ratio chi square residual correlations	Set by priors	Yes	Extensive	150 items	Some difficulty	Over \$100
	MIRTE	3	Yes	Joint maximum likelihood	Good description	Yes	NO	Logistic	Residual covariances	Set by user	NO	Extensive	3,000 people - 100 items	Some difficulty	Over \$100
Criterion		-	2	3	4	2	Q	۱	Ø	6	01	11	12	13	14

The benchmark data sets that were developed to evaluate the characteristics of the programs were purposefully produced to yield an easy test of the programs. For the programs to yield good estimates of the person-parameter vectors, it was believed that the information provided by the test questions should be fairly uniform over the region of the ability space where the persons were located. Therefore, a set of item parameters was selected to yield approximately uniform information in the region within two standard deviations from the origin of a two-dimensional space. The θ -distribution was generated to be standard bivariate normal with $\rho = 0$ or $\rho = .5$. Thus, two data sets were produced using the set of item parameters, both with 2000 response strings on 50 items. The first was based on uncorrelated θ s and the second on correlated θ s. The item parameters used to generate the data sets are given in Table 3. The sample size and number of items were selected to be reasonable in magnitude, but large enough that usable parameter estimates were expected.

The four programs were evaluated using these benchmark data sets on their ability to reconstruct the original data matrix. This criterion was used rather than a comparison of the item-and person-parameter estimates to the values used to generate the data because of the rotational indeterminacy of the solutions. While MIRTE and MULTDIM imposed a positive monotonicity constraint on their solutions limiting the possible orientations of the coordinate axes somewhat, TESTFACT did not impose such a constraint. The TESTFACT solution would have to rotate to match the solutions provided by the other programs and the results would be dependent, in part, on the rotation selected.

		Ite	m Paramete	rs			
Item No.	<u>a</u> 1	<u>a</u> 2	D	d	MDISC	a	
1	1.35	0.27	-2.50	3.44	1.38	11.31	
2	0.65	1.14	0.01	-0.01	1.31	60.10	
3	1.36	0.02	-0.79	1.08	1.37	1.15	
4	0.30	1.45	2.48	-3.68	1.48	78.39	
5	1.39	1.17	2.50	-4.54	1.82	40.09	
6	1.83	0.00	0.47	-0.86	1.83	0.00	
7	1.80	0.01	-0.98	1.77	1.80	0.36	
8	1.47	0.01	2.00	-2.95	1.47	0.64	
9	0.01	1.42	-1.50	-0.82	1.42	89.52	
10	0.15	1.34	2.49	- 3.35	1.34	83.46	
11	1.32	0.29	2.07	-2.81	1.36	12.15	
12	1.68	0.22	-0.10	0.16	1.69	7.54	
13	1.42	0.00	-2.50	3.56	1.42	0.04	
14	0.12	1.81	0.87	-1.57	1.81	86.29	
15	0.18	1.29	-0.44	0.58	1.31	82.25	
16	1.41	0.04	-2.22	3.14	1.42	1.61	
17	1.35	0.00	2.39	-3.23	1.34	0.00	
18	0.24	1.74	-2.04	3.59	1.76	82.28	
19	1.11	0.84	-0.24	0.33	1.39	37.11	
20	0.00	1.44	1.31	-1.88	1.44	90.00	
21	0.01	1.52	1.75	-2.66	1.52	89.58	
22	1.40	0.06	1.94	-2.72	1.40	2.58	
23	0.35	1.38	-0.25	0.36	1.42	75.69	
24	0.00	1.57	1.36	-2.13	1.57	89.99	
25	0.09	1.38	2.38	-3.29	1.38	86.13	
26	0.21	1.48	-1.50	-1.51	1.50	82.08	
27	1.54	0.43	0.89	-1.43	1.60	15.55	
28	0.40	1.34	-2.36	3.30	1.40	73.20	
29	0.81	1.52	-0.93	1.61	1.72	61.94	
30	1.46	0.13	2.05	-3.00	1.46	5.19	

Table 3 Item Parameters for the Benchmark Data sets

<u></u>		Ite	m Paramete	ers	<u>. </u>	
Item No.	<u>a</u> 1	<u>a</u> 2	D	<u>d</u>	MDISC	α
31	0.61	2.12	-2.22	4.90	2.21	74.06
32	1.38	0.00	2.00	-2.75	1.38	0.08
33	0.09	1.64	-1.98	3.24	1.64	86.74
34	0.16	1.50	2.50	-3.78	1.51	84.00
35	0.00	1.34	2.34	-3.14	1.34	90.00
36	1.45	0.29	-0.22	0.32	1.48	11.24
37	1.89	0.12	-2.43	4.60	1.90	3.55
38	0.03	1.38	-1.17	1.62	1.38	88.91
39	0.40	1.35	0.06	-0.08	1.41	73.71
40	2.17	0.01	-0.71	1.54	2.17	0.15
41	0.06	1.36	1.57	-2.12	1.36	87.60
42	0.68	1.28	-0.86	1.25	1.45	61.77
43	0.06	1.47	2.49	-3.67	1.47	87.50
44	1.27	0.82	2.49	-3.76	1.51	32.62
45	0.44	1.41	-1.40	2.08	1.48	72.73
46	1.45	0.27	0.98	-1.45	1.48	10.39
47	0.08	1.42	-0.34	0.49	1.43	85.89
48	1.32	0.04	-2.39	3.15	1.32	1.56
49	1.41	0.00	-2.50	3.52	1.41	0.01
50	1.40	0.00	0.40	0.56	1.40	0.00

Table 3 (Continued) Item Parameters for the Benchmark Data sets

The actual criteria used to evaluate the programs was the standardized residual computed by subtracting the expected response from the actual response and dividing by the standard error of the expected response. The average standardized residual over the entire response matrix was computed for the solution obtained for each program for each of the two data sets. These values are summarized in Table 4. LISCOMP was not included in this analysis because it did not provide θ -estimates.

		Table ⁴	4	
Avera	ge St	andardiz	ed	Residuals
for	each	Program	by	Dataset

Dataset	Program						
	MIRTE	TESTFACT	MULTDIM				
ρ = 0	.251	.001	026				
ρ = .5	.253	.000	024				

From these values it is clear that TESTFACT was the best at estimating person- and item-parameters that could reproduce the data matrix. MULTDIM was next best, but with a slight negative bias in the expected responses. MIRTE was by far the worst of the three programs. An analysis of the parameter estimates from MIRTE and the estimated responses showed that the program sometimes gave extreme values of θ -estimates that inflated the residuals. While the number of these extreme estimates was fairly small, they had a large effect on the residuals. Clearly, if MIRTE is to be used as an operational program, a means must be found to stabilize the θ -estimation process.

Alternative MIRT Models.

The MIRT model that has been emphasized up to this point is a model that has a linear combination of the elements of the θ -vector in the exponent of <u>e</u>. In general, this type of model will be called a linear MIRT model, or LMIRT model. Both logistic and normal ogive versions of LMIRT models are common. These models are also called compensatory models because a low value on one θ -dimension can be compensated for by a high value on a different θ -dimension. Also, a low θ -value on one dimension does not limit the magnitude of the probability of a correct response. With sufficiently high θ -values on other dimensions, probabilities approaching 1.0 can still be obtained.

One alternative to the LMIRT models is a model based on the product of a series of probability-like terms (Sympson, 1978; Embretson, 1984). This model is generally of the form

$$P(\underline{u}_{ij} = 1 | \mathbf{a}_{i}, \mathbf{b}_{i}, \underline{c}_{i}, \mathbf{\theta}_{j}) = \underline{c}_{i} + (1 - \underline{c}_{i}) \prod_{k=1}^{M} \frac{e^{\underline{a}_{ik}(\theta_{jk} - \underline{b}_{ik})}}{1 + e^{\underline{a}_{ik}(\theta_{jk} - \underline{b}_{ik})}}$$
(8)

where $\underline{u}_{i,j}$ is the item response,

- $\boldsymbol{\theta}_{1}$ is the vector of ability parameters
- **a**_i, **b**_i are vectors of item parameters

and \underline{c}_i is a scalar lower asymptote parameter. Since this model derives its properties from the multiplication of terms, it will be called a multiplicative MIRT model here, or MMIRT model. This class of models has also been labeled as noncompensatory models because a low θ -value on one dimension cannot be totally compensated for by a high θ -value on another dimension. There is some level of compensation, so a better label would be partially compensatory, but the term MMIRT is generally more descriptive.

The upper limit on the probability of a correct response to an item modeled by a MMIRT model is set by the lowest θ -value. This θ -value and the corresponding item parameters set a probability of correct response that is reached at a limit as the values of θ on the other dimensions increase.

The equiprobable contours of the MMIRT model are similar to a hyperbola in shape. This is a result of the multiplicative form of the model. An example of the contours is given in Figure 4.

Both of these types of MIRT models have advantages and disadvantages. The LMIRT model is mathematically convenient and estimation programs exist for simultaneously estimating the person- and item-parameters of the models. The MMIRT model is argued to be more psychologically realistic in that real test questions probably do not allow the total compensation of abilities. However, it is a mathematically intractable model, and no estimation program exists for the simultaneous estimation of person- and item-parameters.

Because of the interest in both models, a logical question is: Which model yields a better representation of real item response data? Because of the lack of an estimation program for the MMIRT model, this question could not be addressed in the straightforward manner of estimating parameters and checking relative fit to the data. Instead, an approach was taken that derived parameters for LMIRT and MMIRT items that were matched on the proportion of correct responses for a specified population and then analyzed item response matrices generated from the different models to determin= whether there were noticeable differences in the data matrices.

The first task was to select the item parameters from the LMIRT model given in Equation 1 that would produce "reasonable" proportion-correct indices. Therefore, a target set of p-values for a 20-item test was established based on actual ACT tests, and parameter values for the LMIRT

model were selected by trial-and-error until the expected p-value for a standard bivariate normal examinee population with $\rho = 0$ were approximately equal to the target values. Table 5 gives the set of LMIRT item parameters for the 20 items in columns 2, 3, and 4. The table also gives the expected proportion correct in the last column.

-	L. Pa	MIRT I rameters		MMIRT Parameters				LMIRT II Parameters			
Item	<u>a</u> 1	<u>a</u> 2 <u>c</u>	<u>a</u> 1	<u>a</u> 2	<u></u> ₽1	<u>b</u> 2	<u>a</u> 1	<u>a</u> 2	d	E(<u>P</u>)	
1	2.00	2.25 -1.	00 1.97	2.41	63	19	.90	1.31	67	.38	
2	3.00	1.20 -1.	50 3.00	1.52	.27	-1.34	2.10	.50	-1.13	.34	
3	2.00	2.25 1.	00 1.89	2.09 -	1.09	88	.89	1.10	.52	.60	
4	1.78	1.78	55 1.92	1.95	52	49	.99	1.00	44	.42	
5	1.80	3.00 1.	38 1.61	2.60 -	-1.63	68	.58	1.65	.78	.63	
6	2.50	3.00 1.	00 2.02	2.42 -	1.05	.73	.91	1.27	.42	.58	
7	2.50	2.50 2.	20 2.04	2.05 -	1.23	-1.29	1.03	. 95	1.08	.70	
8	1.10	3.00 .	50 1.28	2.95 -	2.21	29	.32	2.27	.38	.54	
9	1.00	1.10 2.	00 1.25	1.35 -	2.21	-1.99	.61	.72	1.63	.80	
10	1.35	1.80 .	90 1.53	1.91 -	1.41	87	.67	1.12	.60	.62	
11	1.50	1.50	20 1.73	1.75	67	64	.91	.91	21	.47	
12	1.75	3.00 .	00 1.66	2.76 -	1.29	29	.64	1.72	05	.49	
13	2.10	.90 .	50 2.25	1.25	43	-1.99	1.65	. 38	.40	.56	
14	.60	1.80 2.	00 1.04	1.91 -	3.37	-1.19	. 18	1.61	1.84	.78	
15	1.50	1.70 .	20 1.69	1.87	93	69	.82	1.02	.09	.52	
16	2.87	2.00	30 2.60	1.90	32	91	1.45	.81	24	.46	
17	3.10	2.00 1.	55 2.63	1.73	- .72	-1.56	1.64	.62	.85	.64	
18	1.00	1.00 -1.	00 1.34	1.36	28	27	.77	.76	91	.32	
19	2.30	1.40 .	20 2.35	1.57	42	-1.30	1.46	.62	. 10	.52	
20	.80	1.70 .	40 1.17	1.90 -	1.97	48	. 39	1.37	. 32	.56	

Table 5 Original LMIRT Item Parameters and their MMIRT and LMIRT Estimates In order to produce a comparable or "matched" set of MMIRT item parameters, estimates of these item parameters were obtained by minimizing

$$\sum_{j=i}^{N} \left\{ P_{L}(\underline{u}_{ij} = 1 | \mathbf{a}_{i}, \underline{d}_{i}, \mathbf{\theta}_{j}) - P_{M}(\underline{u}_{ij} = 1 | \mathbf{a}_{i}, \mathbf{b}_{i}, \mathbf{\theta}_{j}) \right\}^{2}$$

for N = 2000 randomly selected examinees with $\boldsymbol{\theta}$ distributed as before, where P_L and P_M represent LMIRT and MMIRT models respectively. This process was repeated for each item yielding the noncompensatory item parameter estimates listed in columns 5-8 of Table 4.

Classical item statistics were calculated from 0/1 item response data generated from the 2000 examinees randomly selected earlier. Although the pvalue distributions were similar for the 20-item tests from each model, the classical discrimination indices were significantly higher for the LMIRT set. Therefore, the process of estimation was repeated but this time, the MMIRT item parameters were considered known and the LMIRT item parameters, listed as LMIRT II in Tables 5, were estimated. These new matched item parameters produced closely equivalent tests, as shown by the similar item characteristics given in Table 6.

Three sets of analyses were performed using these matched sets of parameters to determine whether data generated from them were noticeably different. First, the number-correct score distributions were compared for the two models. Second, the estimated-true-score surfaces were compared. Third, the information provided about abilities in the two dimensional space was compared. Finally, the characteristics of the data were evaluated by a general MIRT model that encompassed both the LMIRT and MMIRT models.

	[MIRT	I		MMIRT		L	MIRT I	<u> </u>
Item	P	<u>r</u> pb	<u>r</u> b	p	<u>г</u> рь	<u>r</u> b	P	<u>r</u> pb	rb
1	36	70	90	37	61	78	37	60	77
2	32	63	82	31	52	68	32	53	69
3	60	74	94	59	60	76	58	55	69
4	43	69	87	43	59	74	42	54	68
5	62	75	96	61	61	77	62	60	76
6	57	78	99	56	61	77	57	56	71
7	70	72	95	68	54	71	68	51	66
8	55	69	87	55	58	63	55	56	71
9	80	47	68	80	41	59	80	38	54
10	61	65	83	62	53	68	61	51	64
11	45	63	79	45	54	68	44	51	64
12	49	74	92	48	61	76	49	58	73
13	54	60	76	55	49	62	55	47	59
14	78	52	73	78	47	66	78	45	63
15	52	68	85	50	59	73	50	54	67
16	44	74	93	45	61	76	45	55	70
17	64	73	93	62	56	71	63	50	64
18	31	50	66	31	45	59	31	43	56
19	51	68	85	49	58	73	50	55	69
20	55	63	79	55	52	65	55	52	65
. <u></u>						- 11, - -	<u> </u>	_ <u></u>	
x		10.81			10.70)		10.71	ľ
S.D.		6.40)		5.32	2		5.04	ł
KR20		.93	5		.88	3		.86)

Table 6 Item Analysis Summary LMIRT and MMIRT Models

The number-correct score densities were estimated assuming a standard bivariate normal distribution of ability with $\mu = 0$ and $\rho = 0$. These densities were estimated by applying a recursive procedure described by Lord and Wingersky (1984) to the following expression

$$h_{\ell}(\underline{y}) = \int_{-\infty}^{\infty} f_{\ell}(\underline{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta},$$

where $h_{\varrho}(\underline{y})$ is the density of score \underline{y} for model $\underline{\ell}$,

 $f_{\underline{t}}(\underline{y}|\theta) \text{ the probability of score } \underline{y} \text{ for ability vector } \theta,$ and $g(\theta)$ is the bivariate normal density.

Table 7 gives the estimated densities for the two models.

The differences between the two densities appeared to be negligible. Both distributions were negatively skewed and platykurtic with the MMIRT density slightly greater in the left tail. For large samples of examinees, such a difference in number-correct scores might be significant. These differences at the low end of the number-correct score scale are consistent with the MMIRT model's tendency to produce lower scores when either θ_1 or θ_2 is low.

The second set of analyses, the comparison of the estimated true score surface for the two models, supports this interpretation. The difference in the estimated true score surface is shown in Figure 5. The difference in the surfaces shows that the estimated true scores are fairly similar for θ -vectors with elements within one unit of each other. Large differences in the surfaces occur when the difference in the elements of the θ -vector is greater than 1.0. In these cases the LMIRT model yields higher expected true scores. Because most of the bivariate density is within one unit of the origin of the space, large differences in the true score distribution are not expected as confirmed by the data in Table 7.

Number-Correct	LM	IRT	Μ	MIRT
Score	Relative Frequency	Cumulative Relative Frequency	Relative Frequency	Cumulative Relative Frequency
0	.007	.007	.012	.012
1	.018	.025	.025	.037
2	.027	.052	.034	.071
3	.035	.087	.040	.111
4	.042	.119	.044	. 155
5	.047	. 176	.048	.203
6	.051	.227	.050	.253
7	.055	.282	.052	.305
8	.058	.340	.054	.359
9	.060	.400	.055	.414
10	.062	.462	.056	.470
11	.063	.525	.056	.526
12	.064	.589	.057	.583
13	.064	.653	.058	.641
14	.063	.716	.058	.699
15	.061	.777	. 058	.757
16	.059	.836	.058	.815
17	.055	.891	.057	.872
18	.048	.939	.053	.925
19	.038	.977	.046	.971
20	.022	.999	.029	1.000
x		0.89	10	.82
ariance	2'	5.68	29	.09
Skewness		12	-	. 12
lurtosis		2.05	1	.95

Table 7 Number-Correct Relative Frequencies Under Each MIRT Model



Figure 5. Difference in estimated true score surfaces for the two MIRT models (LMIRT - MMIRT).

The third set of analyses was based on a general MIRT model, GMIRT, that included both MMIRT and LMIRT as special cases. This model is given by

$$P_{G}(\boldsymbol{\theta}_{j}) = \underline{c}_{i} + (1 - \underline{c}_{i}) \frac{\exp(\underline{f}_{ij^{1}} + \underline{f}_{ij^{2}})}{\{1 + \exp(\underline{f}_{ij^{1}} + \underline{f}_{ij^{2}}) + \mu\{\exp(\underline{f}_{ij^{1}}) + \exp(\underline{f}_{ij^{2}})\}}$$

where $f_{ijm} = a_{im} (\theta_{jm} - b_{im})$ and $\mu = 0$ for LMIRT and 1 for MMIRT. The key question to be answered by these analyses was whether the μ -parameter could be used to identify the special case, LMIRT or MMIRT, that was underlying the matrix of observed response data. To investigate the value of μ for this purpose, five sets of data were generated by the LMIRT parameters specified earlier and five sets of data were generated for the matching MMIRT parameters. In all cases the same set of 2000 true 0s were used. These were sampled from a standard bivariate normal with $\rho=0$.

For these ten data sets, the item parameters of GMIRT were estimated using the known 0s. When μ was close to zero, the item parameters of the LMIRT model were obtained from f_{ijm} by setting $d = -\underline{a}_1\underline{b}_1 - \underline{a}_2\underline{b}_2$. When μ was close to 1.0, the MMIRT parameters were estimated. In all cases, the estimation of the μ -parameter identified the proper model. The generating parameters and the mean and standard deviation of the estimates from the five replications of each model are given in Tables 8 and 9 for each item.

Mean estimates of the model identification parameter, μ , showed that the parameter was estimated fairly accurately under each underlying model condition. Table 8 shows that the largest estimate of the LMIRT condition (i.e., $\mu = 0$) was item #18 with $\mu = .07$ followed by item #13, $\mu = .06$. Relatively small standard deviations of these estimates for five replications on each item indicated the precision of these least squares estimates.

Table 8 Parameter Estimates from the GMIRT Model for the LMIRT data

0.0.0.4 002 03 0 ς Ω, 63 6 6 7 G 052025 020000 02 02 02 03 03 03 0.03 < 1 Estimated Paremeters* .07 .55 .14 .17 .18 .13 .13 .28 Pas . 13 . 06 . 13 . 13 -14 -14 -18 -18 -.60 -1.07 -.32 -.32 .86 -.14 1.00 -.78 .22 .49 .49 1.18 .84 1.72 .71 -.16 .06 .92 1.81 .17 ς s_{D_a} 12 08 12 01.00.000 1.30 .52 1.10 1.05 1.35 1.01 2.35 .74 1.13 .84 .74 .83 .83 .75 .75 ິສໍ sD² 10,035 .06 .07 .07 1.000 11.000.000 1.47 1.71 .85 1.52 1.52 ۔ م --21 --05 --05 --05 --05 --05 --05 -.67 -1.13 -.44 -.44 .42 1.08 1.63 .60 -.24 -.91 .32.32 σ True Parameters .81 .62 .76 .62 .62 .91 .72 .38 .38 1.61 1.31 .50 1.10 1.00 1.27 295 2.27 1.12 az .91 .64 .165 .18 .82 2.10 .89 .58 .58 1.45 1.64 1.46 1.46 .91 .32 .61 .67 aj Item No. 16 117 118 119 20 - α m ≠ n 6 9 10 10 15235

*Based on five replications

Table 9 Paremeter estimates from the GMIRT Model for the MMIRT Data

1		
	SD [°] μ	14 12 12 12 12 12 12 12 12 12 12 12 12 12
	د ۹	
	s_{b_2}	00 00 00 00 00 00 00 00 00 00 00 00 00
emeters*	b ₂	
ted Pare	sd _{b1}	00 00 00 00 00 00 00 00 00 00
Estima	ٍ ٩,	
	$s_{D_{a_2}}^{\circ}$	
	a, a	2.37 2.37 2.94 2.71 2.71 2.71 2.72 2.73 2.74 2.73 2.74 2.73 2.74 2.73 2.74 2.73 2.74 2.74 2.74 2.74 2.74 2.74 2.74 2.74
	sd ^a 1	16 17 17 17 17 17 17 17 17 17 17 17 17 17
	ื่ รั	1.97 2.93 2.93 2.93 2.93 2.93 2.94 2.92 2.93 2.92 2.93 2.92 2.92 2.92 2.92 2.92 2.92 2.92 2.92 2.92 2.92 2.92 2.92 2.92 2.92 2.92 2.92 2.92 2.92 <t< td=""></t<>
	b ₂	
ameters	p 1	
True Par	a ₂	2.41 2.60 2.60 2.60 1.95 1.95 1.91 1.90 1.90 1.90 1.90 1.90
	aı	2.25 2.25 2.25 2.25 2.25 2.25 2.25 2.25
Item	No.	-0040 00000 -00070 900000 000000 -00070 900000

*Based on five replications

Table 9 shows slightly poorer estimates of μ when the true generating model was MMIRT (i.e., $\mu = 1$). However, the LMIRT and MMIRT estimation simulations tended to support a conclusion that data generated from these two models are noticeably different, at least under the conditions reported here.

Procedures for Test Construction

In order to determine how MIRT procedures could be used to improve the test construction process, the procedures were first applied to existing data to determine their characteristics. Two major findings were the result of these analyses. First, it was determined that the data generated by the interaction of an examinee sample and a set of cognitively complex items could meet the assumptions to be considered statistically unidimensional. That is, even though each test item required many different cognitive skills, the responses to the collection of items on the test could still form a unidimensional data set. To generate unidimensional data, the items on the test must all measure the same composite of skills. In the language of MIRT, this means that all items are best at measuring in the same direction, \mathbf{a} , in the θ -space.

This finding was described in the following two convention papers.

- Reckase, M. D. (1985, August). <u>Trait estimates from multidimensional items</u>. Paper presented at the meeting of the American Psychological Association, Los Angeles.
- Reckase, M. D. (1985, November). <u>The true versus the observed dimensionality</u> <u>of test data</u>. Paper presented at the meeting of the Society for <u>Multivariate Experimental Psychology</u>, Berkeley, CA.

The procedure for constructing unidimensional tests was described in the article

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. <u>Journal of Educational</u> <u>Measurement</u>, <u>25</u> (3), 193 - 204.

The second major finding was that the substantive meaning of points on the score scale generated by a set of test items is not constant throughout the score range when the difficulty of the items is related to the content of the items. This finding applies both to the number-correct score scale and the θ -estimates obtained from unidimensional IRT models. Further, the differences in the meaning of the score scale at different points cannot be detected using global procedures such as factor analysis. Procedures like MIRT, that focus on the item/person interaction, are needed. These results were reported in the following series of convention papers:

- Reckase, M. D., Carlson, J. E., & Ackerman, T. A. (1985, June). <u>When</u> <u>unidimensional data are not unidimensional</u>. Paper presented at the meeting of the Psychometric Society, Nashville.
- Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986, June). <u>The interpretation of unidimensional IRT parameters when</u> <u>estimated from multidimensional data</u>. Paper presented at the meeting of the Psychometric Society, Toronto.

Reckase, M. D. (1987, April). <u>A comparison of the results of applying several</u> <u>different unidimensional IRT procedures to multidimensional item response</u> <u>data</u>. Paper presented at the meeting of the American Educational Research Association, Washington.

- Reckase, M. D. (1989, August). <u>Controlling the psychometric snake: or, how I</u> <u>learned to love multidimensionality</u>. Invited address at the meeting of the American Psychological Association, New Orleans.
- Davey, T., Ackerman, T. A., Reckase, M. D., & Spray, J. A. (1989, July). Interpreting score differences when item difficulty and discrimination are confounded. Paper presented at the meeting of the Psychometric Society, Los Angeles.

Because it was determined that scores on a test at different points on the score scale may have different meaning, it became of concern whether test forms that were constructed to be parallel on the basis of content specifications and traditional statistics would continue to look parallel when analyzed from a MIRT perspective. Five forms of the ACT Assessment Mathematics Usage Test were analyzed using MIRT to address that concern. In general, the forms were found to be multidimensionally parallel, but a twodimensional solution was needed to describe the item/person interactions and the score points did vary in their meaning. These results were reported in:

Reckase, M. D., Davey, T., & Ackerman, T. A. (1989, March). <u>Similarity of the</u> <u>multidimensional space defined by parallel forms of a mathematics test</u>. Paper presented at the meeting of the American Educational Research Association, San Francisco. The conclusions of this line of research were: (a) tests could be constructed to meet the unidimensional assumptions of IRT even if individual test items required more than one cognitive skill for their successful solution; (b) for the meaning of the score scales defined by parallel forms of a test to be truly parallel, the multidimensional structure of the test must be considered rather than merely meeting unidimensional statistical criteria; and (c) if the difficulty of test items is related to their content, the points on the score scale for the test will not have the same meaning throughout their range.

Computerized Testing

The goal of the computerized testing component of this project was to determine the feasibility of using computerized tests, and possibly adaptive tests, for the assessment of the skills acquired in the military training environment. To achieve this goal, several specific tasks and research studies were planned. These included:

- The selection of hardware and the development of software for the computerized testing system.
- The investigation of the effect of the medium of item presentation on the statistical characteristics of the items.
- The investigation of the effect of the order of item presentation on the statistical characteristics of items.
- The evaluation of the item pool for its adequacy for use with adaptive testing.

5. The evaluation of adaptive testing procedures as alternatives to current testing procedures.

Each of these topics will now be discussed in turn.

Hardware and Software

In selecting a hardware configuration, several factors were considered. These included equipment reliability, security, computing capability, and ease of use. Reliability was perhaps the most important factor because the computerized tests were replacing paper-and-pencil tests which seldom suffer from equipment failure. In order to be assured of a reliable system, common, easily serviced equipment was desired with some redundancy built in to cover for any unanticipated equipment failures.

Security was a concern because of the need to control access to the item pools and examinee performance records. Sufficient computing capacity was needed to support the background computation needed for adaptive testing and to store item pools, test results, and programs for administration, scoring, and report generation. Finally, the system should be easy to program and easy for instructional staff to use.

The system configuration selected for this purpose was a network of 28 IBM PCs with two IBM PC/ATs serving as network servers. Two IBM PC/ATs were used as servers as a redundancy feature, because they were critical to the operation of the system. They contained the item pools, test results, and programs. Thus if one of these machines failed during a test session, the other would automatically take over all of the operations of the systems. The IBM PCs served as testing stations. Programs and appropriate item pools were downloaded to a testing station for a test, but would not reside there at

other times. The system also included a backup power supply and a printer for report generation.

IBM equipment was selected for its reliability and the availability of convenient servicing. It was also selected because the IBM PC-DOS environment was familiar to all persons involved.

The servers and testing stations were connected using Novell Advanced Netware-G LAN. This network allowed very tight controls over access to the item pools, records, and programs.

The software for the system was custom-developed for the project. It included modules for item pool editing and management, test assembly, test administration, and report generation. Many of the characteristics of the software were a result of the needs of the particular school that was the host site for the project. The staff of the school developed tests by randomly sampling items from sets of items constructed to measure specific objectives. Therefore, the software was required to emulate that process and produce printed forms to match the computer presented version. Thus, the software was developed to not only produce paper-and-pencil forms, but also to give the computer administered tests as many of the options as possible that are available to persons taking paper-and-pencil tests. These included the ability to review and preview the test, change answers, skip items, and review responses. Software that closely matched the paper-and-pencil administration process was needed to determine whether differences in item performance were due to medium of administration procedures, or to restrictions placed on the test-taking process.

The resulting software had the capability of randomly selecting items from the pool by objectives, administering items on the computer screen and printing matching paper-and-pencil forms, scrambling items from the base

version and administering the items in the computer screen, and printing individual and group reports of results. The software design was modular so that other administration options could be added as needed.

Research Studies

The goal of this project was to determine whether computerized adaptive testing could be profitably implemented in a military training environment. In order for such a system to work, several technical requirements had to be satisfied by the item pool and the data obtained from examinees. First, because items are usually calibrated based on the results of paper-and-pencil administrations, it was necessary that the calibrations should also apply to items administered by computer. Second, because adaptive tests administer items in different orders, the effect of presentation order must be minimal. Finally, it must be possible to obtain accurate parameter estimates from the data available from the training programs in question. The studies which investigated each of these issues will be discussed in turn.

The data for these studies were obtained with the cooperation of the staff of the Ground Radio Repair Course in the Marine Corps Communication-Electronics School at the Marine Corps Air-Ground Combat Center, Twenty-nine Palms, California. The hardware and software were installed in a room at the school and operational tests and quizzes were administered on the system.

<u>Medium Effects.</u> The first study was designed to determine whether the test items used to assess achievement in the Ground Radio Repair Course functioned differently when presented on the computer screen than when given in paper-and-pencil form. The research design used was to randomly split each class in half and administer the same items in the same order to each half of the class, but one group received the items in paper-and-pencil form and the

other on the computer screen. Three different course exams were used for this purpose. The results of the study will be reported in the following journal article.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effects of the medium of item presentation on examinee performance and item characteristics. Journal of Educational Measurement, in press.

The study showed that on two of the tests, no medium effects could be detected while on the third test, a small difference in performance, with higher scores on the paper-and-pencil version, was noted. Because the detected differences were small, it was concluded that data from paper-and-pencil administrations of items could be used to obtain parameter estimates for the computerized tests.

Order Effects. The design of the order effect study was similar to that of the medium effect study in that classes were randomly split into two groups where one group received one treatment, and the other group another. The first treatment group received the test items in a fixed order. The second treatment was to present the items in an order that was a random rearrangement of the fixed order with a different order for each examinee. Four different tests were used for this study. The results of this study were presented in:

Ackerman, T. A., Spray, J. A., Reckase, M. D., & Carlson, J. E. (1989,

February). <u>A comparison of the effects of random versus fixed order of</u> <u>item presentation via the computer</u> (Research Report ONR89-1). Iowa City, IA: ACT.

The study found seven items out of 308 that differed on difficulty and five out of 308 that differed on discrimination as a result of the different presentation orders. Because no pattern could be identified to explain the difference in performance of the items that were found to be significant, the effect was considered to be extremely weak and of little consequence.

Estimation of Parameters. Because both the medium effects and order effects studies supported pursuing adaptive testing, a study was planned to determine whether item parameter estimates of sufficient quality could be obtained using the data collected during the operational administration of the tests. Course tests were constructed on a weekly basis by sampling without replacement from the item pools. Three nonoverlapping forms were produced before items were again eligible for selection. For one of the tests in the Ground Radio Repair Course, each form was composed of 25 items randomly selected from a pool of 75 items. Thus, the entire pool was used for each of three forms that were constructed.

Over a period of several months, data on 351 examinees were collected on these 75 items. However, the 75 x 351 matrix of responses was very sparse because the items had been administered in blocks of 25. In order to determine whether these data, or a more extensive set administered to 2000 examinees, could be used to get good parameter estimates, a matrix of 75 x 2000 was generated to match the characteristics of the items and the random sampling plan for forms. This matrix was analyzed using both LOGIST and BILOG with missing data coded as not reached. Neither program yielded usable item parameter estimates.

Several procedures were evaluated for improving the estimates, such as imputing the missing data values and using more robust estimates (e.g., inverse normal transforms of p). In all cases, parameter estimates of

sufficient accuracy for use in adaptive testing could not be obtained. This was a result of both the sparse data matrix and the fact that the test items were fairly easy. As a result of these analyses, it was concluded that a fully adaptive computerized test could not be supported by the data available from this administration pattern. Unfortunately, it was not possible to modify the data collection design or the item pool to permit an evaluation of adaptive testing in this environment.

Alternatives to Fully Adaptive Testing

The tests used in the Ground Radio Repair Course are designed to make a pass/fail decision at a particular score level. They have been designed to be predominantly mastery tests. For this type of application it may not be necessary or desirable to use a fully adaptive testing model because that model is best for obtaining equal measurement precision over a wide range of achievement. Instead a model that adapts only the length of the test and focuses the test precision at the decision point may better meet the needs of the course.

A procedure that adapts the length of the test can be made very efficient if the characteristics of the test items relative to the decision point can be taken into account. Such a procedure would require that the item parameter estimates were obtained and that they be used by the procedure to determine whether a person should be classified above or below the decision point. Because parameter estimates always contain error, a concern is whether the error will affect the accuracy of decisions made using such a procedure.

To determine the effect of parameter estimation error on decision accuracy, a series of computer simulations were performed using a sequential probability ratio test (SPRT) to make the decisions. The details of this study were described in the following report:

Spray, J. A., & Reckase, M. D. (1987, September). <u>The effect of item</u> <u>parameter estimation error on decisions made using the sequential</u> <u>probability ratio test</u> (Research Report ONR87-1). Iowa City, IA: ACT.

The results showed that when parameter estimates had the amount of error typical of a calibration based on 2500 examinees, the misclassification rates were actually lower than those obtained using true parameters, but that the average number of items needed to make a decision was greater. Overall, the differences observed were negligible.

References

Ackerman, T. A., Spray, J. A., Reckase, M. D. & Carlson, J. E. (1989, February). <u>A comparison of the effects of random versus fixed order of</u> <u>item presentation via the computer</u> (Research Report ONR 89-1). Iowa City, IA: ACT.

- Carlson, J. E. (1987, September). <u>Multidimensional item response theory</u> <u>estimation: a computer program</u>. (Research Report ONR 87-2). Iowa City, IA: ACT.
- Davey, T., Ackerman, T. A., Reckase, M. D. & Spray, J. A. (1989, July). <u>Interpreting score differences when item difficulty and discrimination</u> <u>are confounded</u>. Paper presented at the meeting of the Psychometric <u>Society</u>, Los Angeles.
- Embretson, S. E. (1984). A general latent trait model for response processes. <u>Psychometrika</u>, <u>49</u>, 175-186.
- Lord, F. M. & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". <u>Applied Psychological</u> <u>Measurement</u>, 8 (4), 453-461.
- McKinley, R. L. (1987). <u>User's guide to MULTIDIM</u>. Princeton, NJ: Educational Testing Service.
- Muthén, B. O. (1987). <u>LISCOMP: Analysis of linear structural equations using</u> <u>a comprehensive measurement model</u>. Mooresville, IN: Scientific Software.
- Reckase, M. D. (1985, May). <u>Final report: models for multidimensional tests</u> <u>and hierarchically structured training materials</u> (Research Report ONR 85-1). Iowa City, IA: ACT.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. <u>Applied Psychological Measurement</u>, 9 (4), 401-412.

- Reckase, M. D. (1985, August). <u>Trait estimates from multidimensional items</u>. Paper presented at the meeting of the American Psychological Association, Los Angeles.
- Reckase, M. D. (1985, November). <u>The true versus the observed dimensionality</u> <u>of test data</u>. Paper presented at the meeting of the Society for Multivariate Experimental Psychology, Berkeley, CA.
- Reckase, M. D. (1986, April). <u>The discriminating power of items that measure</u> <u>more than one dimension</u>. Paper presented at the meeting of the American Educational Research Association, San Francisco.
- Reckase, M. D. (1987, April). <u>A comparison of the results of applying several</u> <u>different unidimensional IRT procedures to multidimensional item response</u> <u>data</u>. Paper presented at the meeting of the American Educational Research Association, Washington.
- Reckase, M. D. (1989, August). <u>Controlling the psychometric snake: or, how I</u> <u>learned to love multidimensionality</u>. Invited address at the meeting of the American Psychological Association, New Orleans.
- Reckase, M. D., Ackerman, T. A. & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. <u>Journal of Educational</u> <u>Measurement</u>, <u>25</u> (3), 193-204.
- Reckase, M. D., Carlson, J. E. & Ackerman, T. A. (1985, June). <u>When</u> <u>unidimensional data are not unidimensional</u>. Paper presented at the meeting of the Psychometric Society, Nashville.
- Reckase, M. D., Carlson, J. E., Ackerman, T. A. & Spray, J. A. (1986, June). <u>The interpretation of unidimensional IRT parameters when estimated from</u> <u>multidimensional data</u>. Paper presented at the meeting of the Psychometric Society, Toronto.

- Reckase, M. D., Davey, T. & Ackerman, T. A. (1989, March). <u>Similarity of the</u> <u>multidimensional space defined by parallel forms of a mathematics test</u>. Paper presented at the meeting of the American Educational Research Association, San Francisco.
- Spray, J. A., Ackerman, T. A., Reckase, M. D. & Carlson, J. E. (1989). Effects of the medium of item presentation on examinee performance and item characteristics. <u>Journal of Educational Measurement</u>, in press.
- Spray, J. A. & Reckase, M. D. (1987, September). <u>The effect of item parameter</u> estimation error on decisions made using the sequential probability ratio <u>test</u>, (Research Report ONR 87-1). Iowa City, IA: ACT.
- Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), <u>Proceedings of the 1977 Computerized Adaptive Testing</u> <u>Conference</u>. Minneapolis: University of Minnesota.
- Wilson, D., Wood, R., & Gibbons, R. D. (1984). <u>TESTFACT: Test scoring and</u> <u>full-information item factor analysis</u>. Mooresville, IN: Scientific Software.
- Wingersky, M. S., Barton, M. A. & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

American College Testing Program/Reckase

Dr. Terry Ackerman Educational Psychology 210 Education Bldg. University of Illinois Champaign, IL 61801

Dr. Robert Ahlers Code N711 Human Factors Laboratory Naval Training Systems Center Orlando, FL 32813

Dr. James Algina 1403 Norman Hall University of Florida Gainesville, FL 32605

Dr. Erling B. Andersen Department of Statistics Studiestraede 6 1455 Copenhagen DENMARK

Dr. Eva L. Baker UCLA Center for the Study of Evaluation 145 Moore Hall University of California Los Angeles, CA 90024

Dr. Laura L. Barnes College of Education University of Toledo 2801 W. Bancroft Street Toledo, OH 43606

Dr. Isaac Bejar Mail Stop: 10-R Educational Testing Service Rosedale Road Princeton, NJ 08541

Dr. Menucha Birenbaum School of Education Tel Aviv University Ramat Aviv 69978 ISRAEL

Dr. Arthur S. Blaiwes Code N712 Naval Training Systems Center Orlando, FL 32813-7100 Dr. Bruce Bloxom Defense Manpower Data Center 99 Pacific St. Suite 155A Monterey, CA 93943-3231

Pr. R. Darrell Bock University of Chicago NORC 6030 South Ellis Chicago, IL 60637

Cdt. Arnold Bohrer Sectie Psychologisch Onderzoek Rekruterings-En Selectiecentrum Kwartier Koningen Astrid Bruijnstraat 1120 Brussels, BELGIUM

Dr. Robert Breaux Code 7B Naval Training Systems Center Orlando, FL 32813-7100

Dr. Robert Brennan American College Testing Programs P. O. Box 168 Iowa City, IA 52243

Dr. John B. Carroll 409 Elliott Rd., North Chapel Hill, NC 27514

Dr. Robert M. Carroll Chief of Naval Operations OP-01B2 Washington, DC 20350

Dr. Raymond E. Christal UES LAMP Science Advisor AFHRL/MOEL Brooks AFB, TX 78235

Mr. Hua Hua Chung University of Illinois Department of Statistics 101 Illini Hall 725 South Wright St. Champaign, IL 61820

American College Testing Program/Reckase

Dr. Norman Cliff Department of Psychology Univ. of So. California Los Angeles, CA 90089-1061

Director, Manpower Support and Readiness Program Center for Naval Analysis 2000 North Beauregard Street Alexandria, VA 22311

Dr. Stanley Collyer Office of Naval Technology Code 222 800 N. Quincy Street Arlington, VA 22217-5000

Dr. Hans F. Crombag Faculty of Law University of Limburg P.O. Box 616 Maastricht The NETHERLANDS 6200 MD

Ms. Carolyn R. Crone Johns Hopkins University Department of Psychology Charles & 34th Street Baltimore, MD 21218

Dr. Timothy Davey American College Testing Program P.O. Box 168 Iowa City, IA 52243

Dr. C. M. Dayton Department of Measurement Statistics & Evaluation College of Education University of Maryland College Park, MD 20742

Dr. Ralph J. DeAyala Measurement, Statistics, and Evaluation Benjamin Bldg., Rm. 4112 University of Maryland College Park, MD 20742 Dr. Dattprasad Divgi Center for Naval Analysis 4401 Ford Avenue P.O. Box 16268 Alexandria, VA 22302-0268

Dr. Hei-Kı Dong Bell Communications Research 6 Corporate Place PYA-1K226 Piscataway, NJ 08854

Dr. Fritz Drasgow University of Illinois Department of Psychology 603 E. Daniel St. Champaign, IL 61820

Defense Technical Information Center Cameron Station, Bldg 5 Alexandria, VA 22314 Attn: TC (12 Copies)

Dr. Stephen Dunbar 224B Lindquist Center for Measurement University of Iowa Iowa City, IA 52242

Dr. James A. Earles Air Force Human Resources Lab Brooks AFB, TX 78235

Dr. Susan Embretson University of Kansas Psychology Department 426 Fraser Lawrence, KS 66045

Dr. George Englehard, Jr. Division of Educational Studies Emory University 210 Fishburne Bldg. Atlanta, GA 30322

Dr. Benjamin A. Fairbank Performance Metrics, Inc. 5825 Callaghan Suite 225 San Antonio, TX 78228

American College Testing Program/Reckase

Dr. P-A. Federico Code 51 NPRDC San Diego, CA 92152-6800

Dr. Leonard Feldt Lindquist Center for Measurement University of Iowa Iowa City, IA 52242

Dr. Richard L. Ferguson American College Testing P.O. Box 168 Iowa City, IA 52243

Dr. Gerhard Fischer Liebiggasse 5/3 A 1010 Vienna AUSTRIA

Dr. Myron Fischl U.S. Army Headquarters DAPE-MRR The Pentagon Washington, DC 20310-0300

Prof. Donald Fitzgerald University of New England Department of Psychology Armidale, New South Wales 2351 AUSTRALIA

Mr. Paul Foley Navy Personnel R&D Center San Diego, CA 92152-6800

Dr. Alfred R. Fregly AFOSR/NL, Bldg. 410 Bolling AFB, DC 20332-6448

Dr. Robert D. Gibbons Illinois State Psychiatric Inst. Rm 529W 1601 W. Taylor Street Chicago, IL 60612

Dr. Janice Gifford University of Massachusetts School of Education Amherst, MA 01003 Dr. Drew Gitomer Educational Testing Service Princeton, NJ 08541

Dr. Robert Glaser Learning Research & Development Center University of Pittsburgh 3939 O'Hara Street Pittsburgh, PA 15260

Dr. Bert Green Johns Hopkins University Department of Psychology Charles & 34th Street Baltimore, MD 21218

DORNIER GMBH P.O. Box 1420 D-7990 Friedrichshafen 1 WEST GERMANY

Prof. Edward Haertel School of Education Stanford University Stanford, CA 94305

Dr. Ronald K. Hambleton University of Massachusetts Laboratory of Psychometric and Evaluative Research Hills South, Room 152 Amherst, MA 01003

Dr. Delwyn Harnisch University of Illinois 51 Gerty Drive Champeign, IL 61820

Dr. Grant Henning Senior Research Scientist Division of Measurement Research and Services Educational Testing Service Princeton, NJ 08541

Ms. Rebecca Hetter Navy Personnel R&D Center Code 63 San Diego, CA 92152-6800

American College Testing Program/Reckase

Dr. Paul W. Holland Educational Testing Service, 21–T Rosedale Road Princeton, NJ 08541

Prof. Lutz F. Hornke Institut fur Psychologie RWTH Aachen Jaegerstrasse 17/19 D-5100 Aachen WEST GERMANY

Dr. Paul Horst 677 G Street, #184 Chula Vista, CA 92010

Dr. Lloyd Humphreys University of Illinois Department of Psychology 603 East Daniel Street Champaign, IL 61820

Dr. Steven Hunka 3-104 Educ. N. University of Alberta Edmonton, Alberta CANADA T6G 2G5

Dr. Huynh Huynh College of Education Univ. of South Carolina Columbia, SC 29208

Dr. Robert Jannarone Elec. and Computer Eng. Dept. University of South Carolina Columbia, SC 29208

Dr. Douglas H. Jones Thatcher Jones Associates P.O. Box 6640 10 Trafalgar Court Lawrenceville, NJ 08648

Dr. Brian Junker University of Illinois Department of Statistics 101 Illini Hall 725 South Wright St. Champaign, IL 61820

Dr. Milton S. Katz European Science Coordination Office U.S. Army Research Institute Box 65 FPO New York 09510-1500 Prof. John A. Keats Department of Psychology University of Newcastle N.S.W. 2308 AUSTRALIA Dr. G. Gage Kingsbury Portland Public Schools Research and Evaluation Department 501 North Dixon Street P. O. Box 3107 Portland, OR 97209-3107 Dr. William Koch Box 7246, Meas. and Eval. Ctr. University of Texas-Austin Austin, TX 78703 Dr. Richard J. Koubek Department of Biomedical & Human Factors 139 Engineering & Math Bldg. Wright State University Dayton, UH 45435 Dr. Leonard Kroeker Navy Personnel R&D Center Code 62 San Diego, CA 92152-6800 Dr. Jerry Lehnus Defense Manpower Data Center Suite 400 1600 Wilson Blvd Rosslyn, VA 22209 Dr. Thomas Leonard University of Wisconsin Department of Statistics 1210 West Dayton Street Madison, WI 53705

American College Testing Program/Reckase

Dr. Michael Levine Educational Psychology 210 Education Bldg. University of Illinois Champaign, IL 61801

Dr. Charles Lewis Educational Testing Service Princeton, NJ 08541-0001

Mr. Rodney Lim University of Illinois Department of Psychology 603 E. Daniel St. Champaign, IL 61820

Dr. Robert L. Linn Campus Box 249 University of Colorado Boulder, CO 80309-0249

Dr. Robert Lockman Center for Naval Analysis 4401 Ford Avenue P.O. Box 16268 Alexandria, VA 22302-0268

Dr. Frederic M. Lord Educational Testing Service Princeton, NJ 08541

Dr. George B. Macready Department of Measurement Statistics & Evaluation College of Education University of Maryland College Park, MD 20742

Dr. Gary Marco Stop 31-E Educational Testing Service Princeton, NJ 08451

Dr. James R. McBride The Psychological Corporation 1250 Sixth Avenue San Diego, CA 92101

Dr. Clarence C. McCormick HQ, USMEPCOM/MEPCT 2500 Green Bay Road North Chicago, IL 60064 Mr. Christopher McCusker University of Illinois Department of Psychology 603 E. Daniel St. Champaign, IL 61820

Dr. Robert McKinley Law School Admission Services Box 40 Newtown, PA 18940

Dr. James McMichael Technical Director Navy Personnel R&D Center San Diego, CA 92152-6800

Mr. Alan Mead c/o Dr. Michael Levine Educational Psychology 210 Education Bldg. University of Illinois Champaign, IL 61801

Dr. Robert Mislevy Educational Testing Service Princeton, NJ 08541

Dr. William Montague NPRDC Code 13 San Diego, CA 92152-6800

Ms. Kathleen Moreno Navy Personnel R&D Center Code 62 San Diego, CA 92152-6800

Headquarters Marine Corps Code MPI-20 Washington, DC 20380

Dr. Ratna Nandakumar Dept. of Educational Studies Willard Hall, Room 213 University of Deleware Newark, DE 19716

Dr. W. Alan Nicewander University of Oklahoma Department of Psychology Norman, OK 73071

American College Testing Program/Reckase

Deputy Technical Director NPRDC Code 01A San Diego, CA 92152-6800 Director, Training Laboratory,

NPRDC (Code 05) San Diego, CA 92152-6800

Director, Manpower and Personnel Laboratory, NPRDC (Code 06) San Diego, CA 92152-6800

Director, Human Factors & Organizational Systems Lab, NPRDC (Code 07) San Diego, CA 92152-6800

Library, NPRDC Code P201L San Diego, CA 92152-6800

Commanding Officer, Naval Research Laboratory Code 2627 Washington, DC 20390

Dr. Harold F. O'Neil, Jr. School of Education - WPH 801 Department of Educational Psychology & Technology University of Southern California Los Angeles, CA 90089-0031

Dr. James B. Olsen WICAT Systems 1875 South State Street Orem, UT 84058

Office of Naval Research, Code 1142CS 800 N. Quincy Street Arlington, VA 22217-5000 (6 Copies)

Office of Naval Research, Code 125 800 N. Quincy Street Arlington, VA 22217-5000 Assistant for MPT Research, Development and Studies OP 01B7 Washington, DC 20370

Dr. Judith Orasanu Basic Research Office Army Research Institute 5001 Eisenhower Avenue Alexandria, VA 22333

Dr. Jesse Orlansky Institute for Defense Analyses 1801 N. Beauregard St. Alexandria, VA 22311

Dr. Peter J. Pashley Educational Testing Service Rosedale Road Princeton, NJ 08541

Wayne M. Patience American Council on Education GED Testing Service, Suite 20 One Dupont Circle, NW Washington, DC 20036

Dr. James Paulson Department of Psychology Portland State University P.O. Box 751 Portland, OR 97207

Dept. of Administrative Sciences Code 54 Naval Postgraduate School Monterey, CA 93943-5026

Department of Operations Research, Naval Postgraduate School Monterey, CA 93940

Dr. Mark D. Reckase ACT P. O. Box 168 Iowa City, IA 52243

Dr. Malcolm Ree AFHRL/MOA Brooks AFB, TX 78235

American College Testing Program/Reckase

Mr. Steve Reiss N660 Elliott Hall University of Minnesota 75 E. River Road Minneapolis, MN 55455-0344

Dr. Carl Ross CNET-PDCD Building 90 Great Lakes NTC, 1L 60088

Dr. J. Ryan Department of Education University of South Carolina Columbia, SC 29208

Dr. Fumiko Samejima Department of Psychology University of Tennessee 310B Austin Peay Bldg. Knoxville, TN 37916-0900

Mr. Drew Sands NPRDC Code 62 San Diego, CA 92152-6800

Lowell Schoer Psychological & Quantitative Foundations College of Education University of Iowa Iowa City, IA 52242

Dr. Mary Schratz 905 Orchid Way Carlsbad, CA 92009

Dr. Dan Segall Navy Personnel R&D Center San Diego, CA 92152

Dr. W. Steve Sellman OASD(MRA&L) 2B269 The Pentagon Washington, DC 20301

Dr. Robin Shealy University of Illinois Department of Statistics 101 Illini Hall 725 South Wright St. Champaign, IL 61820 Dr. Kazuo Shigemasu 7-9-24 Kugenuma-Kaigan Fujisawa 251 JAPAN

Dr. William Sims Center for Naval Analysis 4401 Ford Avenue P.D. Box 16268 Alexandria, VA 22302-0268

Dr. H. Wallace Sinaiko Manpower Research and Advisory Services Smithsonian Institution 801 North Pitt Street, Suite 120 Alexandria, VA 22314-1713

Dr. Richard E. Snow School of Education Stanford University Stanford, CA 94305

Dr. Richard C. Sorensen Navy Personnel R&D Center San Diego, CA 92152-6800

Dr. Judy Spray ACT P.O. Box 168 Iowa City, IA 52243

Dr. Martha Stocking Educational Testing Service Princeton, NJ 08541

Dr. Peter Stoloff Center for Naval Analysis 4401 Ford Avenue P.O. Box 16268 Alexandria, VA 22302-0268

Dr. William Stout University of Illinois Department of Statistics 101 Illini Hall 725 South Wright St. Champaign, IL 61820

American College Testing Program/Reckase

Dr. Hariharan Swaminathan Laboratory of Psychometric and Evaluation Research School of Education University of Massachusetts Amherst, MA 01003

Mr. Brad Sympson Navy Personnel R&D Center Code-131 San Diego, CA 92152-6800

Dr. John Tangney AFOSR/NL, Bldg. 410 Bolling AFB, DC 20332-6448

Dr. Kikumi Tatsuoka Educational Testing Service Mail Stop 03-T Princeton, NJ 08541

Dr. Maurice Tatsuoka University of Illinois Department of Psychology 603 E. Daniel St. Champaign, IL 61820

Dr. David Thissen Department of Psychology University of Kansas Lawrence, KS 66044

Mr. Thomas J. Thomas Johns Hopkins University Department of Psychology Charles & 34th Street Baltimore, MD 21218

Mr. Gary Thomasson University of Illinois Educational Psychology Champaign, IL 61820

Dr. Robert Tsutakawa University of Missouri Department of Statistics 222 Math. Sciences Bldg. Columbia, MO 65211 Dr. Ledyard Tucker University of Illinois Department of Psychology 603 E. Daniel Street Champaign, IL 61820

Dr. David Vale Assessment Systems Corp. 2233 University Avenue Suite 440 St. Paul, MN 55114

Dr. Frank L. Vicino Navy Personnel R&D Center San Diego, CA 92152-6800

Dr. Howard Wainer Educational Testing Service Princeton, NJ 08541

Dr. Ming-Mei Wang Educational Testing Service Mail Stop 03-T Princeton, NJ 08541

Dr. Thomas A. Warm FAA Academy AAC934D P.O. Box 25082 Oklahoma City, OK 73125

Dr. Brian Waters HumRRO 12908 Argyle Circle Alexandria, VA 22314

Dr. David J. Weiss N660 Elliott Hall University of Minnesota 75 E. River Road Minneapolis, MN 55455-0344

Dr. Ronald A. Weitzman Box 146 Carmel, CA 93921

Major John Weish AFHRL/MOAN Brooks AFB, TX 78223

Dr. Douglas Wetzel Code 51 Navy Personnel R&D Center San Diego, CA 92152-6800

American College Testing Program/Reckase

Dr. Rand R. Wilcox University of Southern California Department of Psychology Los Angeles, CA 90089-1061

German Military Representative ATTN: Wolfgang Wildgrube Streitkraefteamt D-5300 Bonn 2 4000 Brandywine Street, NW Washington, DC 20016

Dr. Bruce Williams Department of Educational Psychology University of Illinois Urbana, IL 61801

Dr. Hilda Wing NRC MH-176 2101 Constitution Ave. Washington, DC 20418

Mr. John H. Wolfe Navy Personnel R&D Center San Diego, CA 92152-6800

Dr. George Wong Biostatistics Laboratory Memorial Sloan-Kettering Cancer Center 1275 York Avenue New York, NY 10021

Dr. Wallace Wulfeck, III Navy Personnel R&D Center Code 51 San Diego, CA 92152-6800

Dr. Kentaro Yamamoto 03-T Educational Testing Service Rosedale Road Princeton, NJ 08541

Dr. Wendy Yen CTB/McGraw Hill Del Monte Research Park Monterey, CA 93940 Dr. Joseph L. Young National Science Foundation Room 320 1800 G Street, N.W. Washington, DC 20550

Mr. Anthony R. Zara National Council of State Boards of Nursing, Inc. 625 North Michigan Avenue Suite 1544 Chicago, IL 60611