ED 344 940 TM 018 335

AUTHOR	Chang, Yu-Wen; Davison, Mark L.
TITLE	A Comparison of Unidimensional and Multidimensional IRT Approaches to Test Information in a Test Battery.
PUB DATE	Apr 92
NOTE	21p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE	Reports - Evaluative/Feasibility (14?) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Comparative Testing; Computer Simulation; Correlation; *Error of Measurement; *Estimation (Mathematics); *Item Response Theory; Mathematical Models; Research Design; Sample Size; Scores; *Test Bias
IDENTIFIERS	Ability Estimates; *Multidimensional Models; Test Batteries; *Unidimensional Models

ABSTRACT

Standard errors and bias of unidimensional and multidimensional ability estimates were compared in a factorial, simulation design with two item response theory (IRT) approaches, two levels of test correlation (0.42 and 0.63), two sample sizes (500 and 1,000), and a hierarchical test content structure. Bias and standard errors of subtest scores (subtests within longer tests) were smaller for the multidimensional approach, but the bias of longer test scores did not improve. Sample size and test intercorrelation had a negligible effect on the results. As the number of items in a test becomes large, in an absolute sense, and relative to the total number in the battery, the advantage of the multidimensional approach seems to disappear. Two tables and 13 figures present results of the simulation. There is an 11-item list of references. (SLD)

A Comparison of Unidimensional and Multidimensional IRT Approaches to Test Information in A Test Battery

Yu-Wen Chang, and Mark L. Davison

University of Minnesota

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1992

Requests for reprints should be addressed to Mark L .Davison,
Department of Educational Psychology, University of Minnesota,
178 Pillsbury Dr. S.E., Minneapolis, MN 55455

Abstract

Standard errors and bias of unidimensional and multidimensional ability estimates were compared in a factorial, simulation design with two levels of sample size, two levels of test correlation and a hierarchical test content structure. Bias and standard errors of subtest scores (subtests within longer tests) were smaller for the multidimensional approach, but the bias of longer test scores did not improve. Sample size and test intercorrelation had a negligible effect on the results.

Introduction

Item response theory (IRT) has proven to be a very powerful and useful measurement tool. The successful application of IRT, however, is paid for by the need for strong assumptions. An assumption that is basic to most commonly used IRT models is the unidimensionality assumption, which limits their usefulness in many testing situations. One such situation is the application of IRT to a test battery. To apply unidimensional IRT to a test battery, we calibrate the item and ability parameters several times, once for each test. This approach is called the consecutive approach by Davey and Hirsch (1991).

The cost of this approach is information loss, since collateral information (Ackerman & Davey, 1991), information supplied by items keyed to other tests in the battery, is ignored. The information loss becomes substantial, when scores on subtests in test batteries with a hierarchical structure are of interest. In a hierarchical structure, there are progressively fewer items for subtests toward the bottom of the hierarchy, so the number of items keyed to a particular ability is small, relative to the total number of items in a test battery. Such a hierarchical structure typifies many aptitude and achievement test batteries. Because subtest scores are more informative, and thus more useful than a simple total test score, it is very important for a scoring procedure to be able to provide precise subtest scores.

Conceptually, the use of a multidimensional IRT model can provide more precise subtest scores, since it allows us to use the information in all items in a test a battery (Chang & Davison, 1991). However, multidimensional IRT is mathematically more complex and the estimation procedures are less precise (Broch, 1990; Luecht & Miller, in press). The basic question that arises then is: under what circumstances and to what extent can the precision of subtest scores be improved by using multidimensional IRT?

Luecht and Miller (in press) evaluated the trade-offs between using a multidimensional IRT model or using a unidimensional IRT model in the context of parameter estimation error and bias, ability estimation issues and goodness-of-fit in response pattern predictions. They found that the accuracy and stability of the unidimensional discrimination and difficulty parameters were equal to or better than their multidimensional counterparts. Further, the unidimensional IRT model tended to fit the data quite well. Finally, they pointed out that there was a variance reduction in the multidimensional ability estimates which did not occur in the unidimensional IRT model. Also, the unidimensional estimates seemed to better represent the intercorrelation of the two simulated abilities. Hence, they concluded that a unidimensional IRT approach is superior to a multidimensional IRT approach under the multidimensionality conditions simulated. Two problems exist in their study. First, the conclusions about ability estimation were questionable, for they compared the properties of the unidimensional ability estimates to multidimensional scores on unrotated dimensions, dimensions which may not have best corresponded to the abilities of interest. Furthermore, they did not directly compare the stability of ability estimates derived from the two approaches. Conceivably, the estimation errors of parameters will be larger for the multidimensional IRT approach, but the information loss will be larger for the unidimensional IRT approach, So, it is important to evaluate the parameter estimation errors as well as the accuracy of ability composites.

Two studies (Ackerman & Davey, 1991; Davey & Hirsch, 1991) have investigated the use of collateral information for improving the precision of ability composites. Ackerman and Davey (1991) studied the utility of collateral information in computerized adaptive testing. Their study compared the theoretical standard error and bias values for unidimensional and multidimensional IRT approaches. Because they did not consider parameter estimation errors, their results might not hold in practice.

Davey and Hirsch (1991) compared concurrent and consecutive scoring procedures for two test scores in a test battery. The concurrent procedure estimated abilities by using all items in the battery. This required that every item have several sets of item parameters, one for each ability to be estimated. These parameters were of two types: primary and secondary. The primary parameter for each item was identical to that obtained from the unidimensional IRT approach. Secondary item parameters were obtained by calibrating an item with respect to the ability defined by items of the other tests. For the sake of discussion , items which contribute collateral information will be refered to as secondary items, whereas items which were keyed to a particular ability will be termed primary items. Specifically, consider a test battery containing tests A and B. The unidimensional IRT item parameters for each test would be estimated separately. Items in test B would then be included, one at a time, and calibrated while keeping the primary parameters of test A fixed at their original estimates. Davey and Hirsch found that concurrent estimation was more precise but also more biased than consecutive estimation. This bias finding was inconsistent with that of Ackerman and Davey (1991). Ackerman and Davey found that concurrent adaptive measurement was more precise and less biased than consecutive adaptive measurement. The inconsistency may result from the different ways the two studies calculated the secondary parameters. The secondary parameters in Ackerman and Davey (1991) were derived from the work of Wang (1986), while Davey and Hirsch (1991) used the procedure described above to obtain the secondary parameters. This procedure seemed to overestimate secondary parameters as shown in their scatter plot of primary parameters against secondary parameters. This probably made concurrent estimation more biased than consecutive estimation.

From the above discussion, two points are clear. First, the less precise parameter estimation for the multidimensional IRT model might cancel out the potential information gain. Therefore, to see whether collateral information from the multidimensional IRT model is robust against errors of parameter estimation, an empirical study which compares the measurement accuracy of multidimensional and unidimensional IRT approaches is needed. Second, the existing studies either did not consider simultaneously the precision of ability and parameter estimates or they used an unsatisfactory way to estimate secondary parameters.

This study differs from the earlier work of Ackerman and Davey (1991) and Davey and Hirsch (1991) in two aspects. While the earlier studies have used a unidimensional approach to capture collateral information, we have used multidimensional approach. In addition, the test structure in this study is hierarchical.

Multidimensional Approach

The model selected for this study was the multidimensional two-parameter logistic model proposed by Mckinley and Reckase (1983). The model is given by

$$P(\theta_j) = P(x_{ij} = 1 \mid \boldsymbol{a}_i, d_i, \theta_j) = \frac{\exp(\boldsymbol{a}_i' \theta_j + d_i)}{1 + \exp(\boldsymbol{a}_i' \theta_j + d_i)} \tag{1}$$

where x_{ij} is the score (0,1) on item i by person j,

$P_i(\theta_j)$ is the probability of a correct response to item i by person j,

θ_j is the vector of ability parameters for person j,

$\boldsymbol{a}_i$ is the vector of item discrimination parameters, and d_i is given by

$$d_i = -\sum_{k=1}^{m} a_{ik} b_{ik} \tag{2}$$

where a_{ik} is the discrimination parameter for item i on dimension k, b_{ik} is the difficulty parameter for item i on dimension k, and m is the number of dimensions. The d_i term, then, is related to item difficulty. The interpretation of the multidimensional parameters was provided by Reckase (1985) and Reckase and Mckinley (1991).

The multidimensional approach includes the following three steps; estimating item and person parameters, determining the direction in the space corresponding to each ability, and finally estimating person scores for each ability.

In step one, all items in the battery are calibrated using a multidimensional estimation procedure. Following the calibration, the direction in the space corresponding to each unidimensional ability scale can be determined using the analytical results for unidimensional approximation of a multidimensional data matrix (Wang, 1986). Specifically, the directions, $\cos\alpha_i$, will equal the first standardized eigenvector of the matrix A'A, where A is the matrix of discrimination parameters for all the primary items in the test. In the final step, the ability will be estimated by using the following equation (in a 2-dimensional solution)

$$\theta_c = \cos\alpha_1 * \theta_1 + \cos\alpha_2 * \theta_2, \tag{3}$$

where θ_c is the composite ability correspoding to the unidimensional ability estimate.

According to multidimensional information theory (Reckase & Mckinley, 1991), total test information is the sum of all the item informations in the direction corresponding to the ability. That is, the total test information includes information from primary items as well as collateral information from secondary items. In unidimensional IRT, only primary item are used to estimate a given ability. Assuming the collateral information is nonzero, then multidimensional information must exceed unidimensional information.

Method

This study employed a 2x2x2 design with two IRT approaches (uni- and multidimensional), two levels of test correlation (0.42, 0.63), and two sample sizes (500, 1,000). A hierarchical test structure--two tests, A and B each with two subtests--was used. Test A measured predominately θ_1, and test B measured predominately θ_2. The subtest correlation within each test was 0.91, which corresponds to an angle of 25 degrees. Each subtest had 25 items.

Data Generation

The θ vectors were generated to fit a bivariate normal distribution, both scaled to have mean 0.0 and standard deviation 1.0. The test correlation was manipulated by rotating the *a* vectors. The original a_1 values for test A were drawn from a uniform distribution over the interval 0.6 to 1.35, while a_2 values were all zeros. The b_1 values were drawn from a uniform distribution over the interval -2.25 to 2.25. Next, the d_i values were calculated. For a test correlation equal to .42, the *a* vectors for subtest A_2 (i.e., item 26 to 50) were rotated counterclockwise 25°, while the *a* vectors for subtest A_1 remained the same. For subtest B, the original a_1 values were all zero, while a_2 values were drawn from a uniform distributic n over the interval 0.6 to 1.35. The b_2 values were drawn from a uniform distribution over the interval -2.25 to 2.25. The *a* vectors of subtest B_2 (i.e., item 76 to 100) were rotated clockwise 25°, while *a* vectors of subtest B_1 were not changed. For a test correlation equal to 0.63, the *a* vectors for subtest A_1 and A_2 were rotated counterclockwise 7°and 32° respectively, while the *a* vectors for subtest B_1 and B_2 were rotated clockwise 7°and 32° respectively.

For each combination of θ and item parameters, the probability of a correct response was computed and the result compared to a uniform random number on the interval 0 to 1. If the random number was less than the computed probability, a response of "1" was generated, if not, a response of "0" was generated. There were 25 replications for each combination of test correlation and sample size.

Parameter estimation

Unidimensional item and ability parameters were obtained by using BILOG (Mislevy & Bock, 1986). Each test and subtest was calibrated separately. That is, each person had six unidimensional ability estimates. The ability estimates were rescaled to have mean 0 and variance 1.00 during Phase III.

Multidimensional item and ability parameters were derived from TESTMAP (Mckinley, 1991a), and THETA-PC (Mckinley, 1991b) respectively. The intitial estimates produced by THETA-PC were rotated so as to be orthogonal, and the ability estimates (θ_1,θ_2) along the orthogonal reference vectors were standardized to have mean 0 and variance 1.00. Estimates of item discrimination and difficulty parameters were rotated and rescaled so that the $P(\theta_j)$ remained the same. Then, the composite abilities were calculated by projecting θ_1 and θ_2 onto the directions of tests and subtests (Wang, 1986). Therefore, each person had six multidimensional ability estimates corresponding to the six unidimensional estimates.

Analyses

Because the amount of information gain depends on the θ points, thirty-four points in the two-dimensional ability space were selected so that the points represented 34 of the preselected 37 points (Figure 1). The 37 points were selected to cover the region with greatest density for the bivariate normal distribution. Three of the 37 preselected subjects did not correspond to any simulated subjects and were dropped.

Insert Figure 1 about here.

For each data set the empirical values of bias and standard deviation of estimated abilities were computed. Bias was computed by subtracting the true abilities from the mean of the estimated abilities. The empirical standard error was obtained by computing the standard deviation of the estimated abilities.

Results

The standard error results for four combinations of test correlation and sample size are presented in Table 1. In all conditions, subtest scores were more precise when multidimensional IRT was employed. It is interesting to note that the improvement in stability did not increase, when the test correlation increased. Also, the improvement of precision was not affected by sample size.

Insert Table 1 about here.

Since test correlation and sample size did not affect the collateral information in the study, only the results from N=1000 and r=.42 are further reported. To get a better feel for the results, let us graph the standard errors of composite abilities for the two models. Shown in Figure 2a-2f are the plots of standard errors for the two models. The amount of standard error reduction depends on θ. On the average, the standard errors for subtests A_1 and B_1 are reduced about 20% (from .44 to .35 and from .43 to .34), while the standard errors for subtests A_2 and B_2 are reduced about 30% (from .42 to .30 and from .44 to .30). The standard errors for subtest A_2 are smaller than those for subtest A_1, because of the higher correlations among subtest A_2 and subtest B_1 and B_2.

Somewhat surprisingly, the use of a multidimensional IRT model did not improve the precision of test scores A and B. The reason probably is that the amount of collateral information for test scores A and B is small relative to that for subtest scores. The multidimensional parameter estimation errors cancel out the potential collateral information gain. Also, the number of items for tests is twice that for subtests. Therefore, ability can be estimated more precisely, with the unidimensional approach. The mean standard error of test scores for the unidimensional approach is smaller than the mean standard error for subtest scores.

Insert Figure 2a-2f about here.

Based on their true composite ability scores, the thirty-four subjects were divided into two groups. One group (upper) included subjects with true abilities equal to or greater than zero, while the other (lower) consisted of subjects with true abilities less than zero. For each group, mean bias was computed. Table 2 presents the bias results. Again, the results are very similar in the four conditions. Figures 3a-3f show the plots of unidimensional vs. multidimensional bias for subtest and test scores. Multidimensional estimates of subtest scores were usually, but not always, less biased than their unidimensional counterparts. For test scores the performance of the two approaches was similar.

Insert Table 2 and Figures 3a-3f about here.

Discussion

The purpose of this paper was to investigate whether the use of multidimensional IRT can improve the accuracy of ability estimates in a test battery. The results indicate that both bias and the standard error for subtest scores were smaller when a multidimensional approach was used. The results are consistent with Ackerman and Davey (1991). This suggests that even when the parameter estimation errors are taken into account, the multidimensional approach is still superior to the unidimensional approach in estimating subtest scores in a battery. Whereas the standard errors results are also consistent with Davey and Hirsch (1991), the bias results do not agree with their findings. Thus, the increased bias in their findings seem specific to their particular estimation scheme.

The multidimensional standard error of subtest scores did not decrease as test correlation increased. Both Ackerman and Davey (1991) and Davey and Hirsch (1991) found that the multidimensional standard error decreased as the test correlation increased, although the decrease was quite small. The difference in findings is probably due to the differences in research design. In both earlier studies, at each unidimensional θ point, 900 examinees were simulated, while in our study each examinee represented a different θ. Since the effect of correlation levels is quite small, it did not show in our study. This also explains why sample size did not influence the results. Although bias and standard errors decreased for subtests, they did not improve for tests. As the number of items in a test becomes large, in an absolute sense and relative to the total number in the battery, the advantage of the multidimensional approach seems to disappear.

References

Broch, E. (1990). *An investigation of the effect of item complexity and dimension strength on item parameter recovery in multidimensional datasets.* Unpublished dissertation, University of Minnesota, Minneapolis.

Chang, Y. & Davison, M. L. (1991, April). *On test information: a comparison of the multidimensional and the unidimensional IRT. Proceedings of the International Educational Statistics and Measurement Symposium.* Tainan, Taiwan, R.O.C..

Davey, T. & Hirsch, T. M. (1991). *Concurrent and consecutive esitmates of examinee ability profiles.* Paper presented at the annual meeting of the Psychometric Society, New Brunswick, NJ.

Luecht, R. M. & Miller, T. R. (in press). Unidimensional calibrations and interpretations of compostie abilities for multidimensional tests. *Applied Psychological Measurement.*

Mckinley, R. L. (1991a). *TestMAP version 2.0 User's Guide.* Princeton, NJ: Educational Testing Service.

Mckinley, R. L. (1991b). *User's Guide to THETA PC Version 2.0.* Princeton, NJ: Educational Testing Service.

Mckinley, R. L. & Reckase, M. D. (1983). *An extension of the two parameter logistic model to the multidimensional latent space* (Research Report 83-2). Iowa City, IA: The American College Testing Program.

Mislevy, R. L. & Bock, R. D. (1986). *PC-BILOG 1.1: Item Analysis and Test Scoring with Binary Logistic Models.* Mooresville, IN: Scientific Software, Inc.

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

Reckase, M. D. & Mckinley, R. L.(1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361-373.

Wang, M. M. (1986). *Fitting a unidimensional model to multidimensional response data.* Paper presented at the annual Office of Naval Research Contractor's Meeting. Knoxville, TN.

Table 1.

Descriptive Statistics for Standards Errors of Unidimensional (U) and Multidimensional (M) Ability Estimates.

(Standarad Errors Based on 25 Replications, 34 Subjects)

	Subtest A_1		Subtest A_2		Subtest B_1		Subtest B_2		Test A		Test B	
	M	U	M	U	M	U	M	U	M	U	M	U
N=500, r=.42												
Mean	.35	.41	.31	.40	.36	.45	.31	.46	.33	.32	.33	.35
SD	.05	.05	.04	.05	.05	.06	.04	.07	.04	.05	.05	.05
N=500, r=.63												
Mean	.34	.40	.28	.41	.38	.46	.30	.43	.31	.32	.33	.35
SD	.05	.07	.04	.07	.06	.05	.04	.06	.05	.05	.05	.04
N=1,000, r=.42												
Mean	.35	.44	.30	.42	.34	.43	.30	.44	.33	.33	.32	.33
SD	.05	.07	.05	.06	.07	.07	.04	.07	.05	.04	.05	.06
N=1,000, r=.63												
Mean	.35	.41	.29	.42	.38	.43	.31	.46	.32	.33	.34	.34
SD	.04	.06	.04	.08	.07	.07	.05	.07	.04	.05	.06	.05

Table 2

Descriptive Comparision of Biases of Unidimensional (U) and Multidimensional (M) Ability Estimates. (Biases Based upon 25 Replications, 34 Subjects)

	Subtest A_1		Subtest A_2		Subtest B_1		Subtest B_2		Test A		Test B	
	M	U	M	U	M	U	M	U	M	U	M	U
N=500, r=.42												
Upper	-.069	-.109	-.054	-.107	-.129	-.172	-.101	-.135	-.060	-.050	-.117	-.104
Lower	.068	.103	.056	.105	.094	.153	.091	.121	.073	.068	.088	.079
N=500, r=.63												
Upper	-.071	-.101	-.042	-.083	-.083	-.128	-.038	-.100	-.063	-.056	-.053	-.065
Lower	.111	.110	.083	.153	.133	.151	.103	.114	.101	.105	.107	.081
N=1,000, r=.42												
Upper	-.073	-.106	-.058	-.126	-.118	-.174	-.092	-.130	-.070	-.075	-.094	-.096
Lower	.083	.144	.037	.061	.056	.072	.045	.093	.060	.053	.043	.042
N=1,000, r=.63												
Upper	-.068	-.122	-.043	-.064	-.122	-.154	-.074	-.167	-.072	-.058	-.098	-.110
Lower	.089	.140	.034	.067	.042	.084	.031	.091	.073	.063	.028	.027

Note: Upper and Lower refer to above and below average abilities respectively.

Figure Captions

Figure 1. The preselected 37 points throughout the θ_1, θ_2 plane.

Figure 2a. Standard errors for unidimensional and multidimensional estimates of subtest A_1 plotted against true composite ability with N=1,000 and r=.42.

Figure 2b. Standard errors for unidimensional and multidimensional estimates of subtest A_2 plotted against true composite ability with N=1,000 and r=.42.

Figure 2c. Standard errors for unidimensional and multidimensional estimates of subtest B_1 plotted against true composite ability with N=1,000 and r=.42.

Figure 2d. Standard errors for unidimensional and multidimensional estimates of subtest B_2 plotted against true composite ability with N=1,000 and r=.42.

Figure 2e. Standard errors for unidimensional and multidimensional estimates of subtest A_1 plotted against true composite ability with N=1,000 and r=.42.

Figure 2f. Standard errors for unidimensional and multidimensional estimates of subtest A_1 plotted against true composite ability with N=1,000 and r=.42.

Figure 3a. Standard errors for unidimensional and multidimensional estimates of subtest A_1 plotted against true composite ability with N=1,000 and r=.42.

Figure 3b. Biases for unidimensional and multidimensional estimates of subtest A_2 plotted against true composite ability with N=1,000 and r=.42.

Figure 3c. Biases for unidimensional and multidimensional estimates of subtest B_1 plotted against true composite ability with N=1,000 and r=.42.

Figure 3d. Biases for unidimensional and multidimensional estimates of subtest B_2 plotted against true composite ability with N=1,000 and r=.42.

Figure 3e. Biases for unidimensional and multidimensional estimates of subtest A_1 plotted against true composite ability with N=1,000 and r=.42.

Figure 3f. Biases for unidimensional and multidimensional estimates of subtest A_1 plotted against true composite ability with N=1,000 and r=.42.

Figure 1. Preselected 37 points

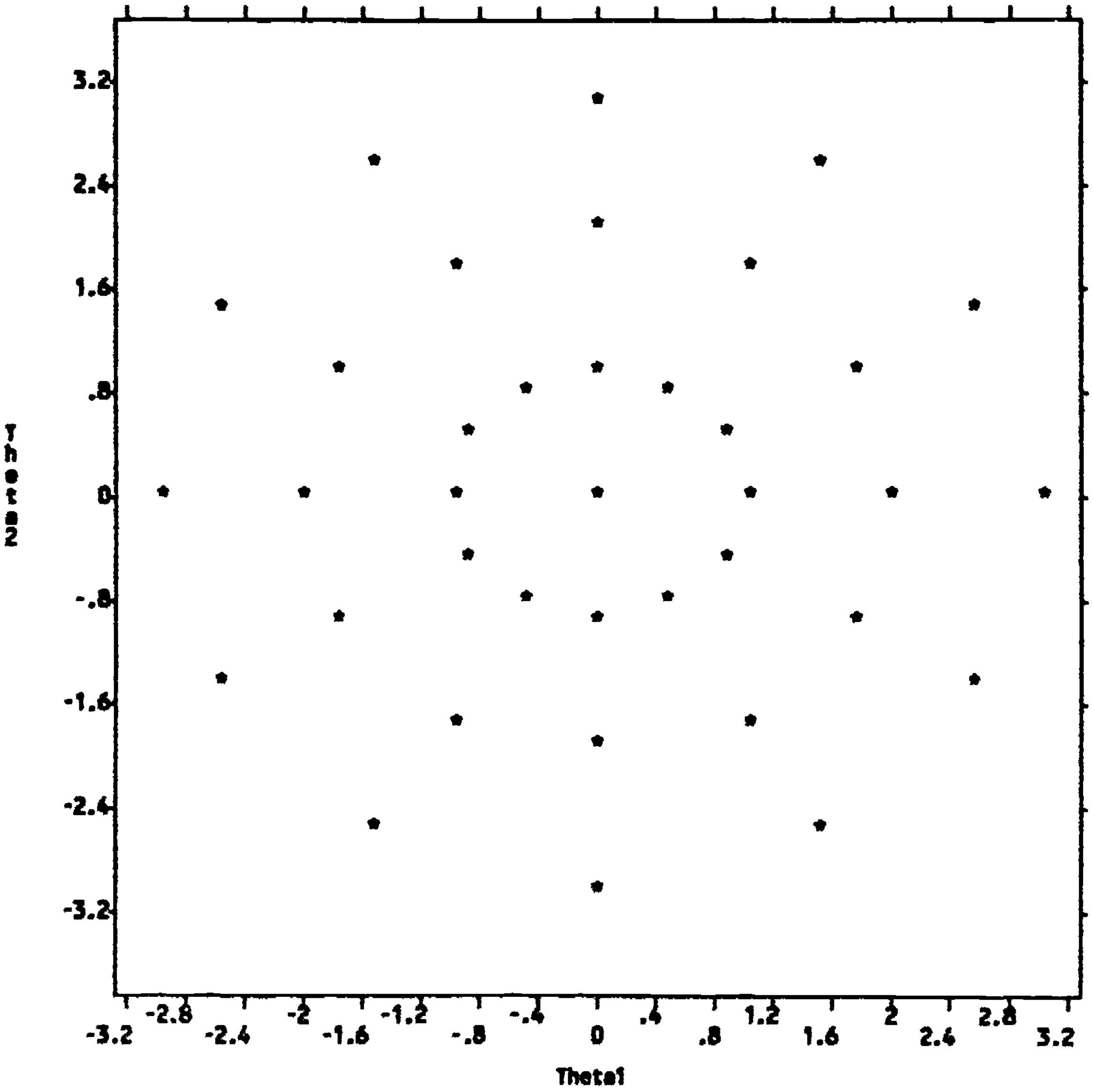
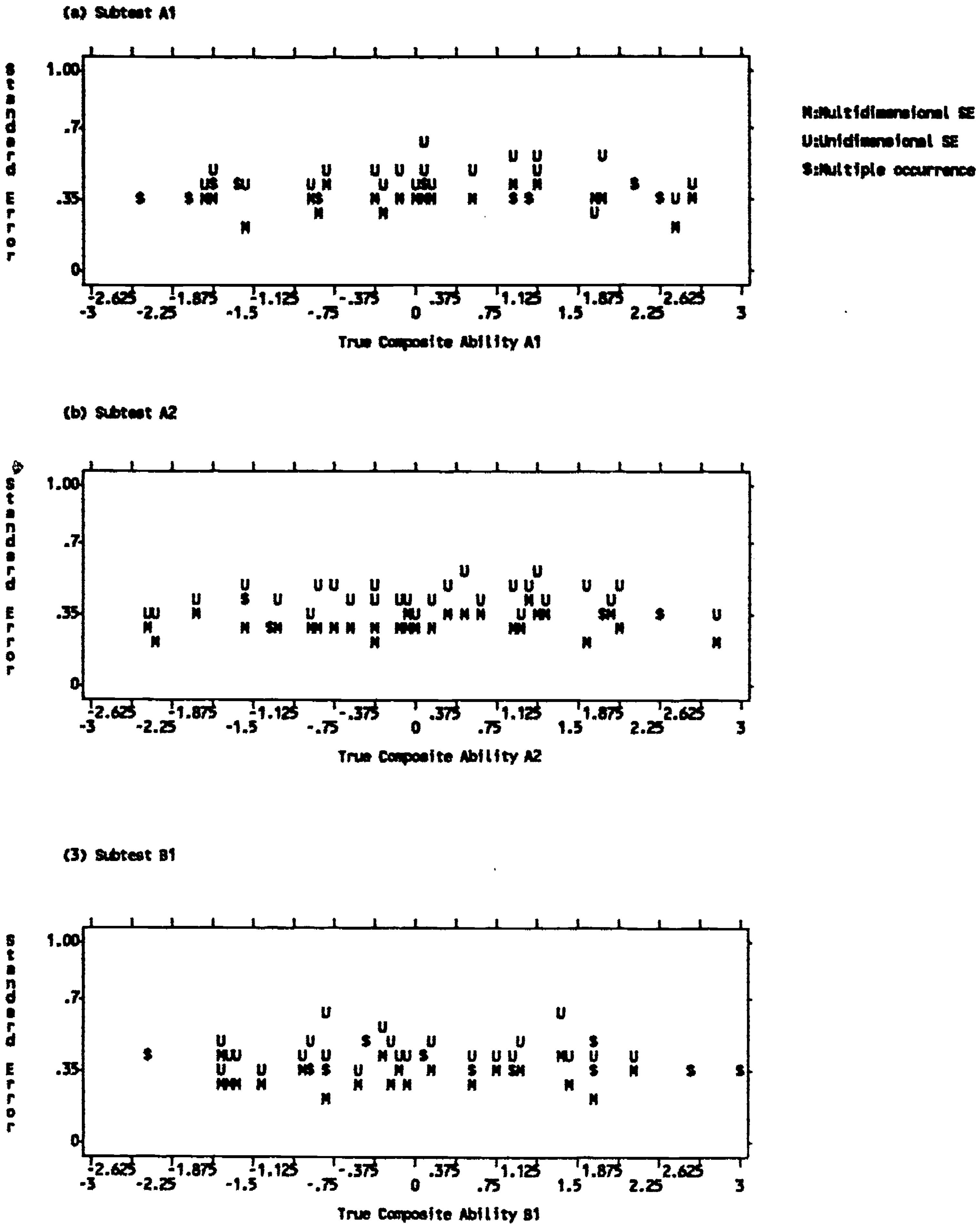


Figure 2.

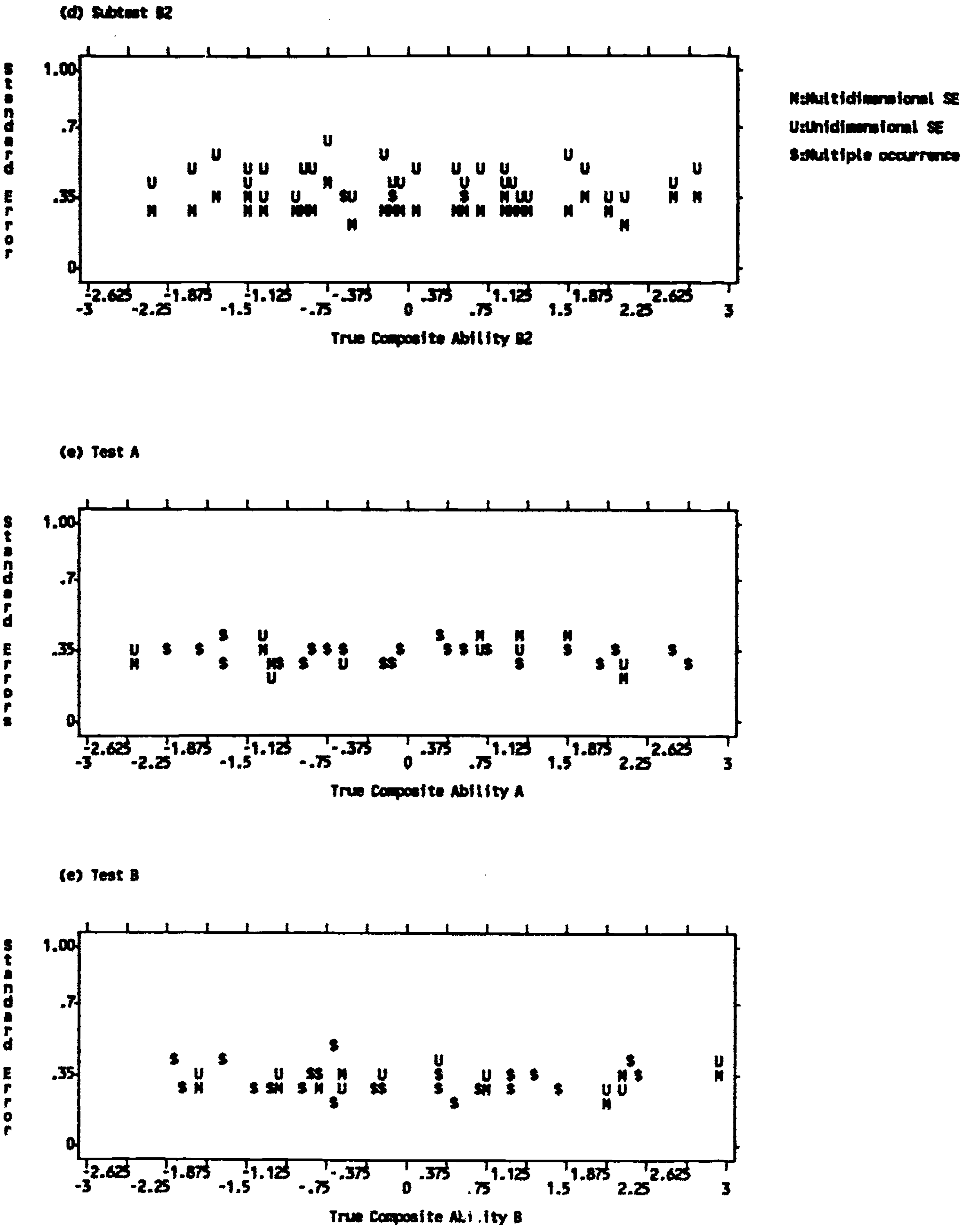


(a) Subtest A1

(b) Subtest A2

(3) Subtest B1

Figure 2.

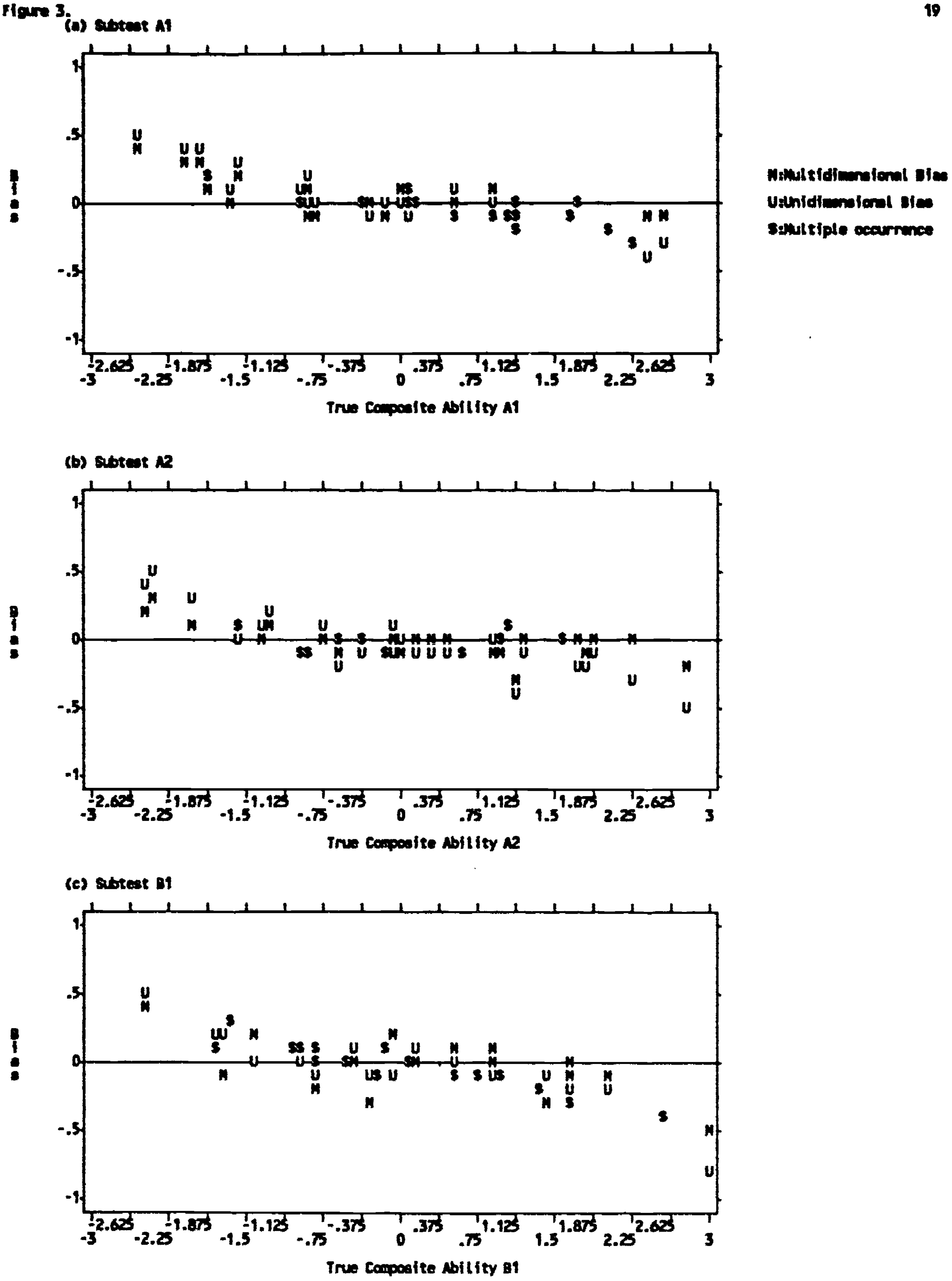


(d) Subtest B2

(e) Test A

(e) Test B

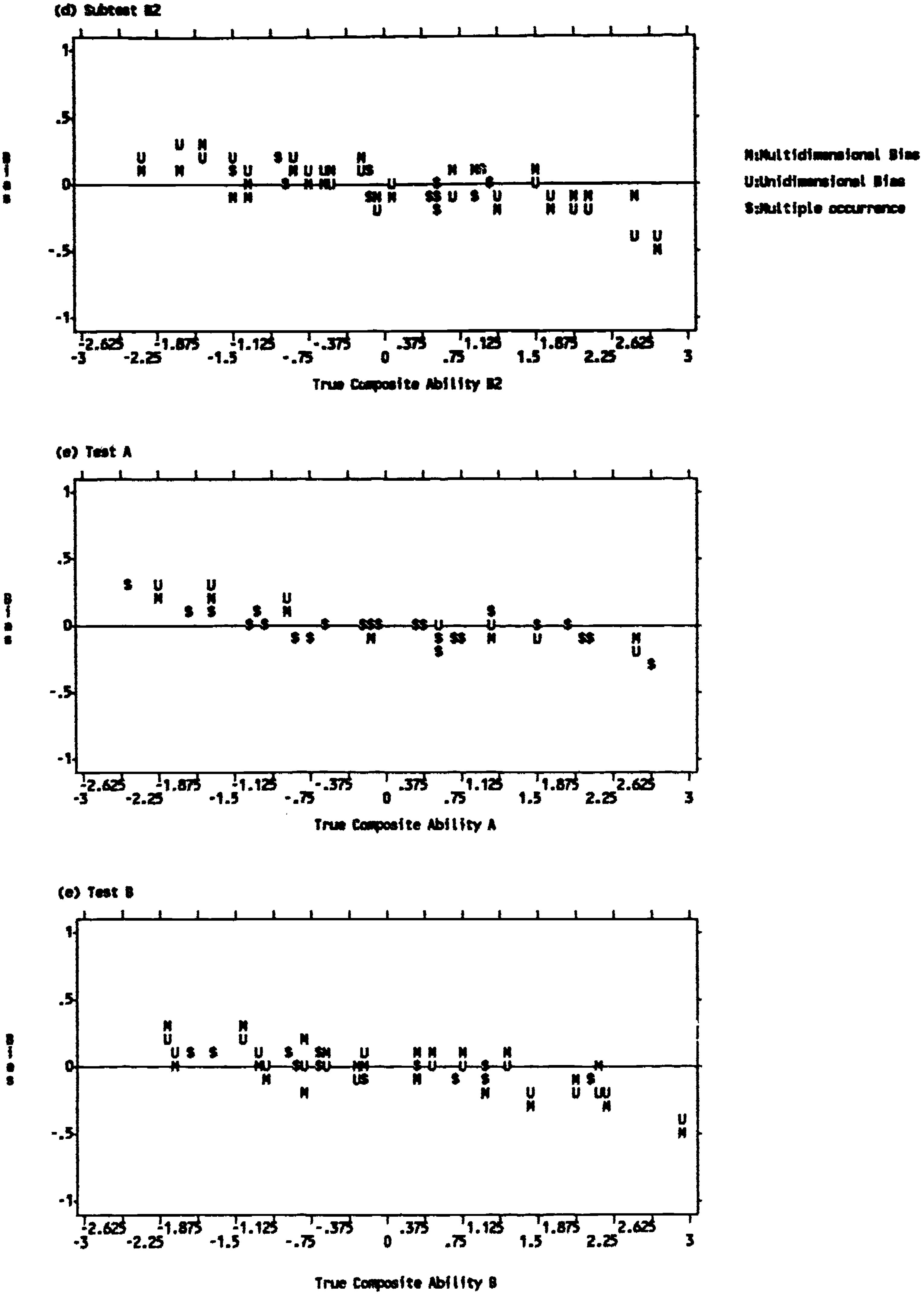
Figure 3.


(a) Subtest A1

(b) Subtest A2

(c) Subtest B1

Figure 3.



(d) Subtest B2

(e) Test A

(e) Test B