DOCUMENT RESUME

ED 359 214 TM 019 841

AUTHOR Liu, Xiufeng
TITLE The Dimensionality of Test Data Generated by Compensatory and Non-Compensatory Two-Dimensional IRT Models and Its Effect on Model-Data-Fit.
PUB DATE Jun 92
NOTE 31p.; Paper presented at the Annual Meetings of the Canadian Society for the Study of Education (Charlottetown, Prince Edward Island, June 1992) and the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Chi Square; Comparative Analysis; Computer Simulation; Foreign Countries; *Goodness of Fit; *Item Response Theory; *Multidimensional Scaling; *Test Construction
IDENTIFIERS *Compensatory Models; Dimensionality (Tests); *Noncompensatory Models; Three Parameter Model; Unidimensionality (Tests); Variance (Statistical)

ABSTRACT

The difference between compensatory and non-compensatory item response theory (IRT) models in terms of the dimensionality of test data generated by them, and its effect on the model-data-fit were examined. The STRESS (proportion of variance not accounted for by the multidimensional scaling model) and RSQ (proportion of variance accounted for by the multidimensional scaling model) values in multidimensional scaling for unidimensional test data files were used as criteria for examining the violation of unidimensionality. The number of items significantly misfitting the unidimensional model and the mean chi-squares for all items in the unidimensional data files were calculated as reference indicators of model-data-fit of items. Compared to data generated by the compensatory model, test data generated by the non-compensatory model tend to be more two-dimensional and to more seriously misfit the three-parameter unidimensional model. Test data generated by the compensatory model tend to be over-unidimensional and to seriously misfit the three-parameter unidimensional model. Although there is a significant difference between the models in terms of dimensionality, the difference only has a significant effect on model-data-fit in terms of the number of items rejected, but has no significant effect in terms of mean chi-square values. Five tables and three figures illustrate the analyses. (Author/SLD)

The Dimensionality of Test Data Generated by Compensatory and Non-compensatory Two-dimensional IRT Models and Its Effect on Model-data-fit

Xiufeng Liu

St. Francis Xavier University
Department of Education
Antigonish, N. S.
CANADA B2G 1C0

e-mail: liu@essex.stfx.ca

Paper Presented at the Annual Meeting of the Canadian Society for the Study of Education, Charlottetown, June, 1992

Abstract

The difference between compensatory and non-compensatory IRT models in terms of the dimensionality of test data generated by them and its effect on the model-data-fit were examined. The STRESS and RSQ values in multidimensional scaling for unidimensional test data files were used as criteria for examining the violation of unidimensionality. The number of items significantly misfitting the unidimensional model and the mean chi-squares for all items in the unidimensional data files were calculated as reference indicators of model-data-fit of items. It has been found that the test data generated by the non-compensatory model tends to be more two-dimensional and to more seriously misfit the three-parameter unidimensional model. The test data generated by the compensatory model tends to be over-unidimensional and to seriously misfit the three-parameter unidimensional model. The correlation between two latent traits has no significant effect on dimensionality and consequently has no significant effect on the model-data-fit. Although there is a significant difference between the compensatory and non-compensatory models in terms of dimensionality, the difference only has a significant effect on model-data-fit in terms of the number of items rejected but has no significant effect in terms of mean chi-square values.

Index terms: Compensatory multidimensional IRT model, non-compensatory multidimensional IRT model, multidimensional scaling, dimensionality, model-data-fit

The Dimensionality of Test Data Generated by Compensatory and Non-compensatory Two-dimensional IRT Models and Its Effect on Model-data-fit

In order to study the robustness of unidimensional item response models to the violation of unidimensionality, simulation studies are usually conducted. In order to generate multidimensional test data, one approach is by employing a factor analytic model and another is by employing a multidimensional item response model so that an examinee-item matrix of probability answering items correctly can be obtained and used as test data. Besides the variety of factor analytic models, there are two major types of multidimensional item response models used in the literature. One is the non-compensatory model originally proposed by Sympson (1978), which takes the following form:

$$P_i(\theta_1, \theta_2, \ldots, \theta_k) = c_i + (1-c_i) + \prod_k \{1+\exp[-Da_{ik}(\theta_k-b_{ik})]\}^{-1}, \qquad (1)$$

where $\theta_1, \theta_2, \ldots, \theta_k$ are latent traits;

a_{ik} is the discrimination of item i on latent trait dimension k;

b_{ik} is the difficulty of item i on latent trait dimension k;

c_i is the guessing level of item i.

Another model is the compensatory model advocated by Christoffersson (1975) and Hattie (1981). The compensatory model takes the following form:

$$P_i(\theta_1, \theta_2, ..., \theta_k) = c_i + (1-c_i)\,[1+\exp(-D\,(\sum_k (a_{ik}\,\theta_k - b_{ik})))]^{-1}. \qquad (2)$$

For compensatory models, there are some other variations. For example, Doody-Bogan and Yen (1983) represent the compensatory model as

$$P_i(\theta_1, \theta_2, ..., \theta_k) = c_i + (1-c_i)\,[1+\exp(-D\,(\sum_k a_{ik}(\theta_k - b_{ik})))]^{-1}. \qquad (3)$$

Another compensatory model is represented by Reckase (1985) as:

$$P_i(\theta_1, \theta_2, ..., \theta_k) = 1 + [1+\exp(-\,d_i \sum_k a_{ik}\,\theta_k)]^{-1}. \qquad (4)$$

Drasgow and Parsons (1983) used factor analysis models to examine the robustness of unidimensional IRT models to the violation of multidimensionality, it was found that unidimensional IRT models are robust against moderately weak prepotent of the general trait. Ansley and Forsyth (1985) used the non-compensatory model to generate multidiemensional data and concluded that the unidimensional models are not robust against the violation of unidimensionality. Since the data generated by a factor analytical model is equivalent to the data generated by a compensatory IRT model, the compensatory and non-compensatory multidimensional IRT models are thus different in terms of the robustness of unidimensional IRT models when they are used to generate data. Way, Ansley and Forsyth (1988) compared the non-compensatory and compensatory (formula 3) IRT models in terms of the closeness between the unidimensional estimates (a, b,θ) and their original multidimensional parameters (a_1,a_2; b_1,b_2; and θ_1, θ_2). Ackerman (1989) compared non-compensatory and compensatory (formula 4) IRT models in terms

of the closeness between the unidimensional estimates (a, b,θ) and their original multidimensional parameters (a_1,a_2; b_1,b_2; and θ_1, θ_2) used to generate data. The comparison between the non-compensatory and compensatory (formula 2) IRT models has not been reported. In Ackerman and Way et al's studies, the multidimensionality of data generated by the non-compensatory and compensatory IRT models are not examined and compared, it is conjectured that the characteristics of multidimensionality of data generated by the non-compensatory and compensatory IRT models are not be the same. In their studies, the unidimensional estimates are examined against original multidimensional parameters, it is difficult to decide if the estimates should be closely correlated with a_1, b_1 and θ_1, or a_2, b_2 and θ_2, or the means of them. This paper will examine the dimensionality of test data generated by the non-compensatory and compensatory (formula 2) IRT models. A comparison of non-compensatory and compensatory IRT (equation 2) models in terms of the model-data-fit between the generated responses and the predicted responses after the three-parameter unidimensional IRT model is applied will be conducted.

Methods

The generation of test data

The data files generated in this study are listed in Table 1. In this study, 5 test data sets were generated, each of which contains 10 data files reflecting 10 replications. Ten replications were used in this study because, according to a study by Stone (1991), the simulation study results for 10 replications are quite consistent with the results for 100 replications, assuming that findings for the 100 replications reflect the true effects. Test data set 1 is unidimensional test data generated by the three-parameter logistic item response model with equal discrimination power and non-guessing. Test data sets 2 to 5 are two dimensional test data with equal discrimination power and non-

guessing. Test data sets 2 and 4 were generated by the compensatory item response model, test data sets 3 and 5 were generated by the non-compensatory item response model. In test data sets 2 and 3, the correlation between the two latent traits was 0.2, in test data sets 4 and 5, the correlation between the two latent traits was 0.8. The correlational latent traits were simulated by the formula developed by Hoffman (1959). At first, two uncorrelational latent traits X_i and Y_i (normally distributed with mean of 0 and standard deviation of 1) were generated, then a third latent trait Z_i was obtained by using the following formula:

$$Z_i=X_i+(k/r)Y_i,$$

where $k=(1-r^2)^{1/2}$. The latent trait Z_i is expected to correlate with X_i with the correlational coefficient of r.

The difficulty parameter b_i for the compensatory IRT model is generated from a uniform distribution in the range of -2 and 2, the difficulty parameter b_{ik} for the non-compensatory IRT model is generated from a bivariate uniform distribution in the range of -2 and 2. The ability parameter θ_{ik} for the compensatory and non-compensatory IRT models are generated from a bivariate normal distribution. The discrimination parameter a_{ik} in this study were chosen as unity to signify the effect of number of abilities (θ_{ik}) on the characteristics of dimensionality of the test data generated and the effect on the model-data-fit, since the values of discrimination will also affect the characteristic of dimensionality of the test data generated. In robustness studies of unidimensional IRT models against multidimensionality when Monte Carlo methods are used, it is plausible to fix a_{ik} and let θ_{ik} varies. The first part of this study shall study the dimensionality of test

data generated by the above specifications, and then its effects on model-data-fit will be examined.

Insert Table 1 about here

The sample size for this simulation study was selected as 1000 and the number of test items as 20. The sample of 1000 examinees and test length of 20 items are justified by some authors, such as Hambleton (1983), to be sufficient for consistent parameter estimation by LOGIST for the three-parameter logistic model.

At first, a 1000 x 20 matrix consisting of the probability of correct response of each examinee to each item was computed using an item response model by providing appropriate latent trait and item parameters; then a 1000 x 20 matrix of uniformly distributed random numbers in the range of 0 and 1 was generated and compared to the previous probability matrix. If the probability is greater than or equal to the correspondent random number, the response is coded as 1, otherwise the response is coded as 0.

Computer program used to calibrate the test data

ASCAL in the MicroCAT computer package (Assessment System Corporation, 1989) was used for this study. ASCAL uses Bayesian procedure and can be used for the three parameter calibration. According to a study by Hsu and Yu (1989), the parameter estimates provided by ASCAL are as accurate as those produced by LOGIST.

The dimensionality study on the data generated by compensatory and non-compensatory item response models

ALSCAL statistical procedure in SPSS:x was applied to each test data file to fit a multidimensional scaling model(MSM). The dimensionality in MSM was selected as two, reflecting the situation of two dimensional item response data generated by compensatory and non-compensatory models. The mean STRESS and RSQ values for unidimensional test data (data set 1) were used as criteria for comparing the violation of unidimensionality. Since STRESS can be interpreted as the proportion of variance not accounted for by the MSM, STRESS values for two dimensional data are expected to be smaller than those for unidimensional test data. Similarly, since RSQ can be interpreted as the proportion of variance accounted for by the MSM, RSQ values for two dimensional data are expected to be greater than those for unidimensional data.

The effect of dimensionality on model-data-fit of items

ASCAL in the MicroCAT computer package gives chi-square statistics for model-data-fit of items. After applying ASCAL to each test data file, a chi-square statistic and the appropriate degree of freedom for each item will be calculated by ASCAL. The number of items significantly misfitting the models according to critical chi-square values and the mean chi-squares for all item in a data file were calculated as general indication of model-data-fit of items for the data file.

Results

The dimensionality of test data generated by the compensatory and non-compensatory item response models

After applying ALSACAL to each data file in each data set (5 x 10 data files), the mean STRESS and RSQ for each data set were obtained and tested for significance. The summary results are list in Table 2.

Insert Table 2 about here

In Table 2, the STRESS and RSQ values for data set 1 (non-violation of unidimensionality) are taken as criteria for evaluating degrees of unidimensionality violation for other data sets. Table 2 shows that for data sets generated by the non-compensatory two dimensional item response model (data sets 3 and 5), there is a transition pattern from unidimensionality to two-dimensionality. When the correlation between the two latent traits is low, i.e. r=0.2, the data sets generated by the non-compensatory item response model are approximately unidimensional, since there is no significant difference of STRESS and RSQ from those for test data set 1 generated by the three-parameter unidimensional item response model (p=0.15 and 0.359 for STRESS and RSQ respectively). However, when the correlation between the latent traits is high, i.e. r=0.8, the data sets generated by the non-compensatory item response model are more two-dimensional. There is a significant difference of STRESS and RSQ from those for test data set 1 generated by the three-parameter unidimensional item response model (p=0.023 and 0.024 for STRESS and RSQ respectively). For the data sets generated by the compensatory two-dimensional model (data sets 2 and 4), there is also a transition pattern from "over-unidimensionality" to unidimensionality. The STRESS values for data sets 2 and 4 are consistently greater than those of data set 1, and RSQ values for data sets 2 and 4 are consistently smaller than those of data set 1 (unidimensional data). Assuming the data set generated by the three-parameter logistic model is unidimensional, the data sets

generated by the compensatory model are more unidimensional than data set 1 under the two dimensional scaling. When the correlation between the two latent traits is small, i.e. r=0.2, the STRESS and RSQ for the compensatory data sets tend to be significantly different from those for unidimensional data (p=0.08 and 0.007 for STRESS and RSQ respectively). When the correlation between the two latent traits is large, i.e. r=0.8, the STRESS and RSQ for the compensatory data sets tend to be close to those for unidimensional data (p=0.131 and 0.036 for STRESS and RSQ respectively). The mechanism of violating unidimensionality for compensatory and non-compensatory models are thus different with compensatory data tending to be over-unidimensional and non-compensatory data tending to be more two-dimensional.

In order to examine the effects of item response models (compensatory and non-compensatory) and the correlation between the two latent traits on the STRESS and RSQ values, a ANOVA was conducted by employing types of item response models and correlation as factors. The summary ANOVA results are listed in Table 3.

Insert Table 3 about here

From the ANOVA table, it can be seen that the STRESS and RSQ are significantly different for compensatory and non-compensatory item response models, but there is no significant difference for different correlations between the two latent traits. The interaction between the model type and correlation has no significant effect on STRESS and RSQ values.

<u>The effect of dimensionality on model-data-fit of items</u>

Since chi-square statistics for different items in different data files may have different degrees of freedom, a mean chi-square value for each data file was computed by averaging chi-square values of items in the data file over their degrees of freedom to degree of freedom 17 (in most files). Using the number of items rejected as misfitting according to critical chi-square values and the mean chi-square values as criteria, the difference between other data sets and data set 1 was tested for significance. The t-test results are listed in Table 4.

Insert Table 4 about here

Table 4 shows that for data sets 3 to 5, there are significant differences ($p<0.05$) in terms of number of items rejected and mean chi-square values from those of unidimensional data set (data set 1). For data set 2 (compensatory with correlation of 0.2), the number of items rejected and mean chi-square values also tend to be significantly different from those from data set 1 ($p=0.10$ and 0.052 respectively). For data set 4, the difference is clearly significant. For unidimensional data set 1, the average percentage of items rejected as misfitting is 24.5% (4.9 out of 20), but the percentage of items rejected for other data sets (data sets 2, 3, 4 and 5) are 32%, 46.5%, 39% and 53%. If taking the mean chi-square values as a general indictor for model-data-fit of items, data set 3 to 5 can be judged as misfitting the model (the critical value for chi-square under degree of freedom 17 is 27.59). If taking the number of items rejected as misfitting as a general indictor for mode-data-fit of items, all the four data sets except data set 2 can also be judged as misfitting the model.

An ANOVA was also conducted to examine the effects of item response model type and correlation on the model-data-fit of items. The summary ANOVA results are listed in Table 5.

Insert Table 5 about here

The ANOVA results show that there is a significant difference in the number of items rejected but no significant difference in mean chi-square values between two types of item response models (compensatory and non-compensatory item response models). The degree of correlation between the latent traits has no significant effect on number of items rejected and mean chi-square values. The ANOVA results also show that there is no interaction effect between the type of item response model and the degree of correlation on the model-data-fit in terms of number of items rejected and mean chi-square values.

Discussion

For the compensatory two dimensional data, the dimensionality of data tends to be more over-unidimensional with the increase of correlation between the two latent traits, and hence tends to more seriously misfit the three-parameter unidimensional model. In order to understand the nature and characteristics of over-unidimensionality of test data, the graphic configurations of items for all the data files were plotted by the ALSCAL procedure in SPSS:x. Three typical configurations are displayed as Figures 1 to 3.

Insert Figures 1 to 3 about here

Figure 1 is for a data file generated by the non-compensatory item response model. In Figure 1, all items are almost evenly scattered around a circle, showing that items are quite inter-coordinated in two dimensional space. On the whole, the data file fits the two dimensional scaling model well.

Figure 2 is for a data file generated by the three-parameter unidimensional item response model and Figure 3 is for a data file generated by the compensatory two-dimensional item response model. Compared to Figure 1, items in Figure 2 are roughly scattered into a circle with some items gathering into sub-groups. Items in Figure 3 are roughly scattered around a circle but not as evenly and completely as in Figure 1 and Figure 2. In Figure 3, some items are also more closer to each other than to other items. It looks that Figure 2 is closer to Figure 1 than Figure 3. This may explain why STRESS and RSQ values for data set 2 and 4 are greater than those for data set 1 which, in turn, are greater than those for data 3 and 5. Assuming the data generated by the three-parameter item response model is unidimensional, the data generated by the compensatory two dimensional item response model is over-unidimensional under two dimensional scaling.

Reckase, Ackerman and Carlson (1988) demonstrated that items that require more than one ability can still be unidimensional under the two dimensional compensatory model if items have the same discrimination structure (a_{ik}). The conclusion was empirically demonstrated under the situation of $\rho(\theta_1,\theta_2)=0$. In this study, a_{ik} are the same for all the items and as it has been shown above, the test data generated by the compensatory IRT model is over-unidimensional and its dimensionality changes from over-unidimensionality to unidimensionality as the correlation between the two latent traits increases.

Ackerman (1989) and Way et. al. (1988) found that correlation between θ_1 and θ_2 did not affect the correlation between true parameters (θ_1 and θ_2) and their parameter estimates θ, same result was true for a and b estimates in Ackerman's study but not in Way et al's study. In this study, it was found that correlations did not affect the compensatory and non-compensatory IRT models when compared to unidiemensional model in terms of dimensionality and model-data-fit of items. The compensatory and non-compensatory IRT models differ significantly in terms of dimensionality, and model-data-fit in terms of the number of items rejected as misfit, they do not differ significantly in terms of mean chi-square values at the total test level.

Conclusion

For the non-compensatory two dimensional data, the dimensionality of data is more two-dimensional and the data generated by the non-compensatory model significantly misfits the three-parameter unidimensional model. For the data generated by the compensatory model, the dimensionality of data is over-unidimensional and the data significantly misfits the three-parameter unidimensional model. The correlation between two latent traits has no significant effect on the dimensionality of test data generated by two models and consequently has no effect on the model-data-fit. Although there is a significant difference between the compensatory and non-compensatory item response models in terms of dimensionality, the difference has significant effects on model-data-fit in terms of the number of items rejected, but has no significant effect in terms of mean chi-square values. More research to compare the three representations of compensatory models at the same time, and compare compensatory and non-compensatory IRT models in practical testing situations, such as test equating, may be needed.

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and non-compensatory multidimensional items. Applied Psychological Measurement, 2, 113-127.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement. 9, 37-48.

Assessment System Corporation (1989). User's manual for MicroCAT testing system. St. Paul, Minnesota.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-22.

Doody-Bogan, E. N., & Yen, W. M. (1983). Detecting multidimensionality and examining its effects on vertical equating with the three parameter logistic models. Paper presented at the annual meeting of the American Educational Research Association. Montreal.

Drasgow, F., Parson, C. K. (1983). Application of unidimensional item response theory model to multidimensional data. Applied Psychological Measurement, 2, 189-199.

Hambleton, R. K. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In Weiss, David J. (ed.) New horizons in testing. Academia Press, Inc.

Hattie, H. A. (1981). Decision criteria for determining unidimensionality. Multivariate behavioral research, 19, 49-78.

Hoffman, P. J. (1959). Generating variables with arbitrary properties. Psychometrika, 24, 265-267.

Hsu, T., & Yu, L. (1989). Using computers to analyze item response data. Educational Measurement: Issues and Practice. Fall, 21-28.

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

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building unidimensional tests using multidimensional items. Journal of Educational Measurement, 3, 193-203.

Stone, Clement A. (1991). IRT based Monte Carlo research: How many replications. Paper presented at the Annual meeting of the National Council on Measurement in Education. Chicago, April, 1991.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.) Proceedings of the 1977 computerized adaptive testing conference. Minneapolis, MN: University of Minnesota. 82-98.

Way, W. D., Ansley, T. N., & Forsyth, R. (1988). The comparative effects of compensatory and non-compensatory two-dimensional data on unidimensional IRT estimates. Applied Psychological Measurement, 3, 239-252.

Table 1. Data files generated in this study

data set	model	a_{ik}	b_i(or b_{ik})	θ_k	$\rho(\theta_1,\theta_2)$
1	A	unity	uniform	normal	
2	B	unity	uniform	normal	.2
3	C	unity	uniform	normal	.2
4	B	unity	uniform	normal	.8
5	C	unity	uniform	normal	.8

note: model A ____ the three-parameter item response model;

model B ____ the compensatory two dimensional item response model;

model C ____ the non-compensatory two dimensional item response model.

Table 2. t-test results on STRESS and RSQ between data set 1 and other data sets

data set	STRESS(Prob.)	RSQ(Prob.)
1	.3197	.7497
2	.3335(.08)	.6543(.007*)
3	.3063(.15)	.7858(.359)
4	.3302(.131)	.6747(.036*)
5	.2996(.023*)	.8210(.024*)

*P<.05

**P<.01

Table 3. Summary ANOVA results on STRESS and RSQ

STRESS			
effects	Mean Square	F	Sig. of F
type of model	.008	23.34	.000**
correlation	.000	.699	.409
interaction	.000	.081	.778
RSQ			
type of model	.193	30.58	.000**
correlation	.008	1.225	.276
interaction	.001	.087	.770

**P<.01

Table 4. t-test results on the differences of model-data-fit between data set 1 and other data sets

data set	# of items (Prob.)	mean X^2(Prob.)
1	4.9	24.25
2	6.4(.10)	29.83(.052)
3	9.3(.00**)	35.64(.021*)
4	7.8(.002**)	33.53(.028*)
5	10.6(.00**)	48.39(.048*)

*p<.05

**P<.01

Table 5. Summary ANOVA results on the model-data-fit

number of items			
effects	MS	F	sig. of F
type of model	81.230	10.500	.003**
correlation	18.230	2.356	.134
interaction	.025	.003	.955
mean chi-square			
type of model	1069.700	2.509	.122
correlation	676.500	1.587	.216
interaction	204.200	.479	.493

**P<.01

Figure 1. The configuration of non-compensatory items

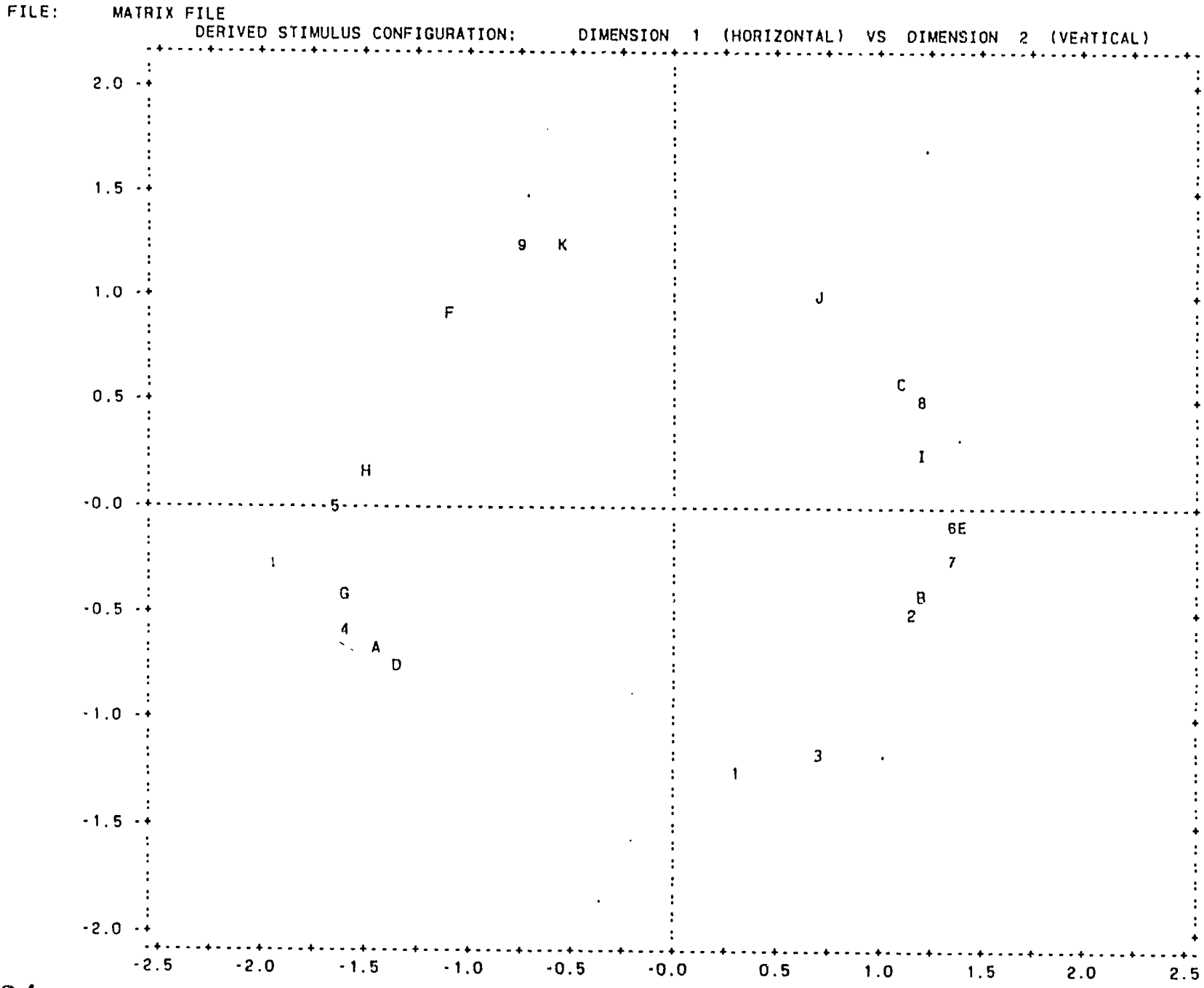

Figure 1. The configuration of non-compensatory items

Figure 2. The configuration of unidimensional items

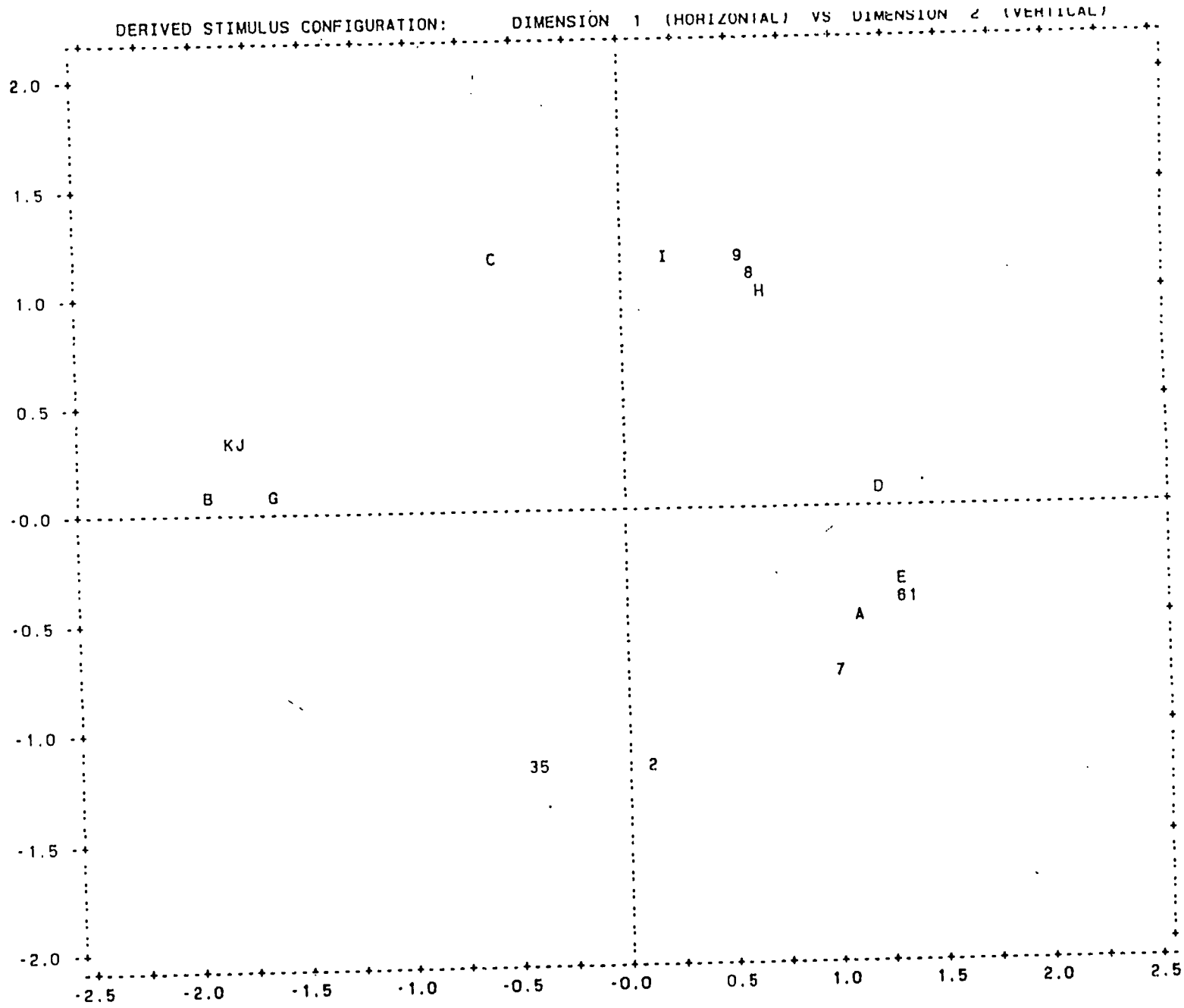

Figure 2. Configuration of Unidimensional Items

Figure 3. The configuration of compensatory items

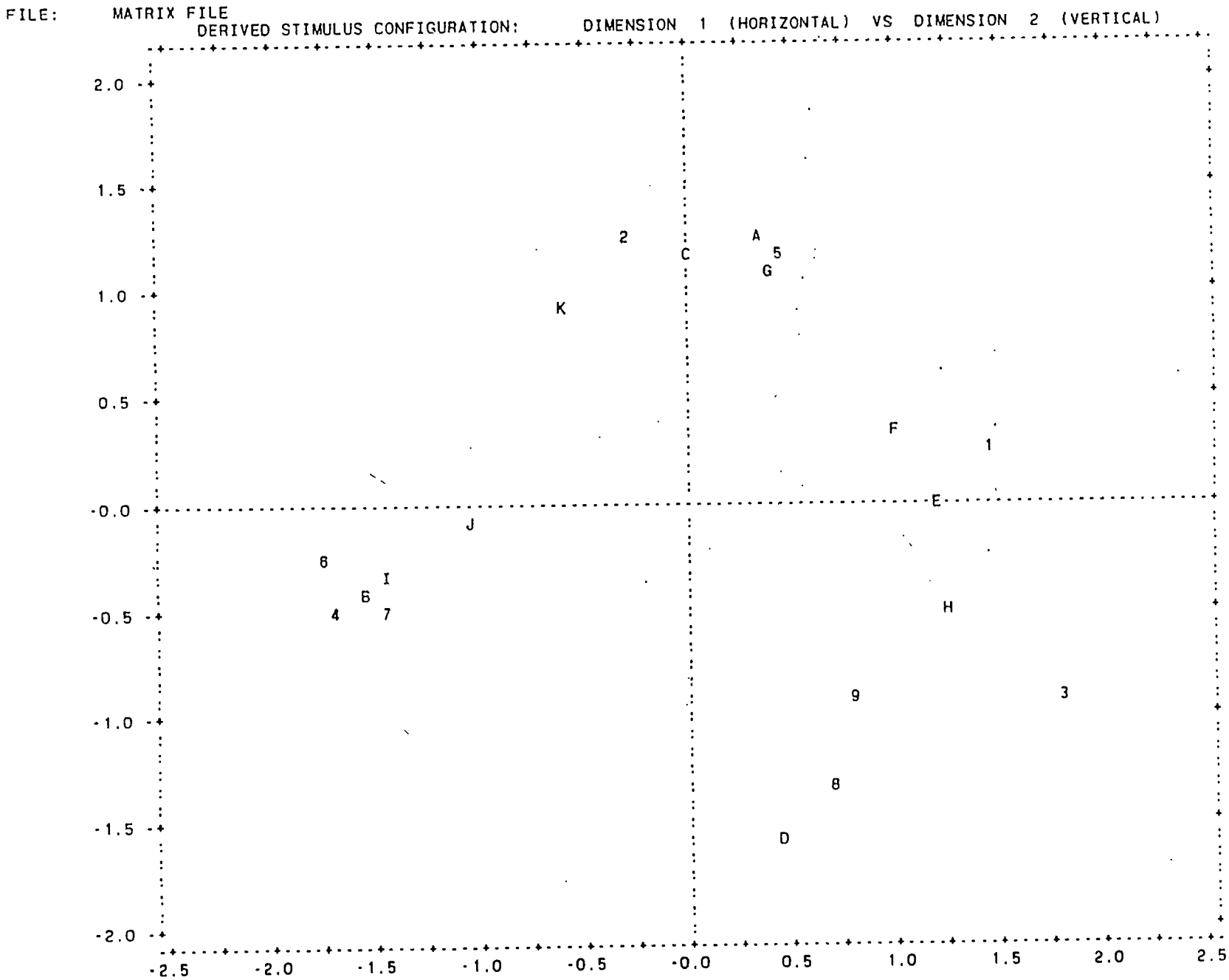

Figure 3. Configuration of Compensatory Items