ED 373 085 TM 021 974

AUTHOR	Ackerman, Terry
TITLE	Graphical Representation of Multidimensional Item Re:sponse Theory Analyses.
PUB DATE	Apr 94
NOTE	38p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Posta≥e.
DESCRIPTORS	Ability; *Achievement Tests; *Educational Assessment; *Item Response Theory; *Measurement Techniques; Scores; Test Items; *Test Use
IDENTIFIERS	Dimensionality (Tests); *Graphic Representation; *Multidimensionality (Tests)

ABSTRACT

The purpose of this paper is to demonstrate how graphical analyses can enhance the interpretation and understanding of multidimensional item-response theory (IRT) analyses. Conceptually many of the unidimensional ~~~ concepts such as item characteristic curves, information, etc., can be extended to multiple dimensions. However, as the dimensionality increases, new problems and issues arise, most notably how to represent these conceptual features within a multidimensional framework. This paper provides examples of different graphical representations including item-response surfaces, information vectors, and centroid plots of conditional two-dimensional ability distributions given the number-correct score. All of these are intended to supplement quantitative and substantive analyses and thereby help the testing practitioner determine more precisely what a test is measuring, the degree of the measurement precision, and the consistency of the assessment process. Fifteen figures illustrate the analyses. (Contains 13 references.) (Author/SLD)

Graphical Representation of Multidimensional Item Response Theory Analyses

Terry Ackerman
University of Illinois

Paper presented at the 1994 AERA Annual Meeting, New Orleans

Running head: GRAPHICAL REPRESENTATION OF MIRT ANALYSES

Abstract

The purpose of this paper is demonstrate how graphical analyses can enhance the interpretation and understanding of multidimensional item response theory (IRT) analyses. Conceptually many of the unidimensional IRT concepts such as item characteristic curves, information, etc., can be extended to multiple dimensions. However, as the dimensionality increases, new problems/issues arise, most notably how to represent these conceptual features within a multidimensional framework. This paper provides examples of different graphical representations including item response surfaces, information vectors, and centroid plots of conditional two-dimensional ability distributions given number correct score. All of which are intended to supplement quantitative and substantive analyses and thereby help the testing practitioner determine more precisely what a test is measuring, the degree of the measurement precision, and the consistency of the assessment process.

Graphical Representation of Multidimensional Item Response Theory Analysis

Most cognitive achievement tests measure, to different degrees, multiple skills or skill composites. As such, testing practitioners have the responsibility to establish the construct validity of their test and subsequently provide an interpretation for their users. If a test is measuring multiple skills some questions that need to be immediately addressed include: What composite skills are being measured? Of the skills being measured which are primary (intended-to-be-measured) and which are secondary (not-intended-to-be measured)? How accurately are each of the various composites being assessed? What is the correct interpretation of the number correct (or standard) score scale? Is this interpretation consistent throughout the entire observe score range, or do low scores reflect levels of one skill composite and high scores levels of another skill composite? Could the secondary skills result in differential performance between identifiable groups of examinees?

Typically, routine sets of item, test, and DIF analyses are conducted after every administration of a standardized test. The purpose of this paper is to present a series of graphical representation of multidimensional analyses that can supplement these analyses. The goal of pictorially representing quantitative results is to help the practitioner gain more insight into the measurement process and thus, strengthen the relationship between test construction process and the quantitative analyses of the test results. Graphical analyses serve several functions:

1. They can provide a visual perspective that can triangulate or cross validate traditional quantitative item, test, and bias analyses;

2. They help measurement specialists gain a better conceptual understanding of the principles of measurement as they apply to the their test;
3. They n help to strengthen the link between quantitative analyses and substantive interpretations of what a test is measuring and how well it is measuring;
4. They can be used to establish a "feedback loop" so that information gained from each administration can be recycled to help improve subsequent test construction procedures and provide insight about the merit of future program considerations (e.g., adaptive testing).

Background

The multidimensionality of a test is difficult to determine and always subject to interpretation. One common procedure is to construct a scree plot of the eigenvalues obtained from a principal axis factor analysis of the interitem tetrachoric correlation matrix (Reckase, 1979). One problem with this approach is trying to decide if the second (and possibly third) eigenvalues are large enough to be significant and denoted as primary dimensions or characterize random noise in the measurement process. Horn (1965) and Drasgow and Lissak (1983) suggested that interpretation could be enhanced by comparing the scree plot to one created from a factor analysis of randomly generated test data containing the same number of items.

Other approaches used to assess dimensionality include checking the assumption of local independence via the variance covariance matrix for examinees within different intervals on the ability scale (McDonald, 1981; Tucker, Humphreys, & Rosnowski, 1986). A somewhat similar approach was suggested by Stout (1987). One outgrowth of Stout's research is a computer program (DIMTEST) that allows users to apply large sample theory and statistically test the assumption that one cluster of items is dimensionally distinct from another. Recently Kim and Stout (1994) have developed a statistic, that is also based on conditional covariances, $\hat{\varepsilon}$, that has been extremely successful in identifying the correct dimensionality of generated test data with different correlational structure between the ability dimensions.

Once a multidimensional structure has been confirmed both statistically and substantively, practitioners can use one of several multidimensional item response theory programs (NOHARM, Fraser & McDonald (1988); TESTFACT, Wilson, Wood & Gibbons (1987); MIRTE, Carlson (1987)) to estimate multidimensional IRT item parameters.

This paper focuses on graphically representing item/test analyses that are performed on a set of test response data for which it has been confirmed that there are two dominant abilities being measured. Thus, all of the analyses discussed that will be below assume that the dimensionality for a given test has been thoroughly investigated and that two-dimensional item response theory (IRT) item parameters have been estimated.

As in the unidimensional case, multidimensionally there are three main characteristics of items that we try to model difficulty, discrimination, and guessing. Extending the concept of the ICC to multiple dimensions becomes somewhat complicated. Partly because in two dimensions ICC's become item characteristic surfaces. Additionally within a multidimensional framework an additional attribute must be considered for each item: the composite of multiple skills the item measuring best. Within a unidimensional framework, an item discriminates, to varying degrees, between all levels of underlying ability. However, there is a range in which the discrimination is optimal. Multidimensionally an item has the capability of distinguishing between levels of many composites, but optimally between levels of just one composite skill. The goal is to identify this composite.

Representations in this paper are based upon the compensatory two-dimensional IRT model. Using this model probability of a correct response to item i can be expressed as

$$P(X_i = 1) = c_i + (1 - c_i)\frac{e^{\sum_{k=1}^{m} a_{ik}\theta_k + d_i}}{1.0 + e^{\sum_{k=1}^{m} a_{ik}\theta_k + d_i}}, \qquad (1)$$

where X_i is the score (0,1) on item i,

$\mathbf{a}_{ik} = (a_{i1}, a_{i2}, \ldots, a_{im})$ is the vector of item discrimination parameters,

d_i is a scalar difficulty parameter of item i, and

c_i is a scalar guessing parameter of item i, and

$\boldsymbol{\theta}_k = (\theta_1, \theta_2, \ldots, \theta_m)$ is the vector of ability parameters

In this formulation there is one item discrimination parameter for each dimension being modeled, but only one overall item difficulty parameter. Because the terms in the exponent (i.e., the logit) are additive, being low on one ability can be compensated for by being high on the other abilities.

It is important to note that the characteristics that are being modeled should always be examined in concert with the substantive and contextual features of the individual items. In fact, this is the only way one can correctly interpret the meaning of the dimensional structure.

Provided below are different graphical representations designed to supplement traditional two-dimensional item/test analyses. Although, by itself each analysis may answer a different question about what the test is measuring, there is a lot of confirmation between analyses that also needs to occur.

The data that are used to illustrate each example come from a two-dimensional estimation of item parameters for two Law School Admissions

Tests (LSAT). These parameters were obtained using NOHARM, a multidimensional IRT estimation program for two forms of the LSAT (December, 1991 and October, 1992). Each calibration was done on a random sample of 5000 examinees. Because NOHARM estimates parameters for the two-dimensional normal ogive model the parameters estimates were rescaled for the logistic model in (1).

It needs to be stressed that the quality of graphical representations and subsequent interpretations demands accurate estimates. The representations are divided into four broad categories: ability estimation representation, item representation, information representation, conditional analyses, and expected score distribution. Each analysis is prompted by a question about the measurement process.

Ability estimation representation:

One logical use of the two-dimensional item parameter estimates is to obtain ability estimates for examinees or identified groups of examinees to compare their distributions. That is, how disparate are the ability distributions for black and white examinees, and does this difference lead to differential item functioning (DIF). Certain response vectors may yield abnormal ability estimates. The maximum likelihood ability estimation procedure is graphically represented in Figure 1. Represented in this diagram are the log likelihood surface, corresponding contour and the location of the intermediate ability estimates for each iteration of a Newton Raphson maximum likelihood estimation for a specified set of item parameters and given response vector. The location of the first and final ability estimates are numbered above the surface and on the contour. The final estimate is provided in the title.

Insert Figure 1 about here

Ability estimates that do not reach convergence, or have a large standard error of measurement could be investigated from a graphical perspective to provide further insight.

Item representation

Graphically, the probability of a correct response given two abilities can be graphically depicted via the item characteristic surface, ICS. Several ways are available by which we can represent a single item: construct the ICS, construct a contour of the ICS, use vector notation (Reckase, 1986).

Four different perspectives of an item characteristic surface are shown in Figure 2. In colored plots the bottom side would be represented with a different color than the top side. Although somewhat appealing such representations are not very informative, nor do they enable easy comparisons between items.

Insert Figure 2 about here

A second way to represent an item is to create the contour of the ICS. The contour lines are equiprobability contours. For the particular multidimensional model described in (1) these contours will always be parallel. Examinees whose θ_1,θ_2 ability places them on the same contour all have the same probability of responding correctly to the item. Contour plots are more informative than the surface plots because they permit one to easily note several features of the item:

1) The ability composite the item is best measuring (i.e., the composite direction orthogonal to the equiprobability contours).
2) Where in the ability plane the item is most effective at distinguishing between ability levels (i.e., is most discriminating). The greater the slope of the surface the more discriminating the item, and hence, the

closer together the equiprobability contours.

3) The difficulty of the item. Assuming a particular underlying bivariate ability distribution one could roughly estimate the proportion of examinees that would be expected to get the item correct.

The contour of the ICS of the item shown in Figure 2 is graphed in Figure 3.

Insert Figure 3 about here

The one drawback of contour plots is that, like with surface plots, only one item can be represented at a time.

The final way to represent items in a two-dimensional ability plane is to use vector notation of Reckase (1986). When represented as a vector the overall discrimination of the item is denoted by the value of MDISC:

$$MDISC - \sqrt{a_1^2 + a_2^2}$$

The amount of discrimination is indicated by the length of the vector. All vectors if extended would pass through the origin of the latent ability plane. Because the discrimination parameters are constrained to be positive vectors are located in only the first and third quadrants. The angular direction of the vector from the θ_1 axis represents the composite of θ_1,θ_2 ability that is being most accurately measured. This implies that the vector representing an item will always lie orthogonal to the equiprobability contours. The difficulty of the item is denoted by the location of the vector in the space. That is, the tail of the vector lies upon the p = .5 equiprobability contour. Thus, easy items will lie in the third quadrant, difficult items will lie in the first quadrant.

One additional feature that has been added to Reckase's (1991) notation is the denoting of the size of the guessing parameter. Items with a guessing parameter less than .1 have and open arrow at the tip ( ); items with a c estimate between .1 and .2 have a completed arrowhead (); and items with large c estimates, greater than .2, will have a closed arrow.

Vectors for three items are drawn in Figure 4. Item One (a_1 =.4, a_2 =.6 , d = 2.4 , c = .15) is an easy item, moderately discriminating, and is measuring best in about a 56° direction (primarily θ_2). Item Two (a_1 =1.8, a_2 =.2 , d =-1.5, c =.25) is a more difficult item, highly discriminating, has a relatively large guessing parameter, and is measuring best in the 6° direction (predominantly θ_1). The third vector corresponds to the item represented in Figures 2 and 3.

Insert Figure 4 about here

Information Representation

In IRT the accuracy of the measurement process is discussed in terms of "information". information is inversely related to the standard error of measurement. That is, the smaller the standard error of measurement, the greater the information, and hence, the greater the measurement precision of the test.

Ackerman (1992) developed procedures to compute the multidimensional information function for the multidimensional IRT model given in (1) taking into account the lack of local independence once a direction in the latent ability space is specified. Ackerman determined that the amount of information provided by an item *i* at a specified θ_1,θ_2 in the direction α can be computed as

$$I_i(\theta_1,\theta_2) - (\cos\alpha)^2 Var(\hat{\theta}_1|\theta,\theta_2) + (\sin\alpha)^2 Var(\hat{\theta}_2|\theta_1,\theta_2) + 2(\sin 2\alpha) Cov(\hat{\theta}_1,\hat{\theta}_2|\theta_1,\theta_2) \qquad (2)$$

(The reader should refer to (Ackerman, 1992) for a more in depth discussion.)

Because items are capable of distinguishing between levels of abilities

To investigate this relationship the following procedure was performed. First, a group of examinees was randomly generated from a bivariate normal distribution (ρ_{θ_1,θ_2} = .5, a value estimated of correlation between θ_1 and θ_2 by NOHARM). The latent plane was then divided into eight "wedges" or octants. Using the generated response data for the examinees in each octant, the angular direction or θ_1,θ_2 composite that had the greatest correlation with the observed score was calculated. The results are displayed in Figure 7. In each octant the number indicating the angular direction of maximum correlation is written. The font size of the number indicates the magnitude of the linear relationship (r^2): the larger the number the stronger the relationship or correlation.

Insert Figure 7 about here

Two important features should be noted in Figure 7. First, it is important that the observed score scale have a consistent interpretation throughout the observable range. A sense of score scale consistency can be seen in each panel, especially in the first, second, and fifth octants where a majority of examinees would be located with a correlation of .5.

A second important analysis is to compare results across forms. If the different forms are truly parallel the relationship between the underlying abilities and the number correct score would be the same. That is, over time the meaning imparted to the observed score scale should remain constant within and between forms. For these two forms there appears to be a "consistent" relationship between θ_1, θ_2 and X with the largest angular composite difference within forms occurring in the fourth and sixth octants for each form.

Another question that relates to the measurement precision question is: What is the expected observed score for each examinee? This can be graphically determined by creating a true score surface plot. The expected score or true score, ξ, for any one examin . is simply the sum of the probabilities of a correct response (as given in (1))for each of the n items in the test,

$$\xi = \sum_{i=1}^{n} P_i(\theta_1, \theta_2)$$

To generate the surface, ξ, is computed for examinees at selected θ_1,θ_2 points in a 31 x 31 grid over the ability plane. A surface and corresponding contour plot is then created. The contours provide insight about how the latent ability plane is mapped onto the true score scale. Note, that the contour lines do not have to be parallel, and that some curvlinearity may occur if subtest items are measuring dimensionally distinct ability composites. The true score surfaces and their corresponding contours for the two LSAT forms are displayed in Figure 8.

Insert Figure 8 about here

The contour of the true score surface plot can be directly related to the clamshell plots. That is, the longest vectors in the clamshell plot would be orthogonal to the true score contour lines. Likewise regions which contain the longest vectors in the clamshell plot (i.e., the regions where the measurement precision is the greatest) should be the same regions where the true score contours are closest together (indicating regions in which the test is doing a better job at distinguishing between levels of true score).

After computing the true score surface for each form, the next logical question would be to compare them. That is, if two test forms are truly parallel any examinee would be expected to have the same true score no matter which form he or she was given. This implies that if two LSAT forms are for different θ_1,θ_2 composites several natural questions a: se:

1. What composite of skills is being best measured by each subtest, as well as the total test?
2. Is the same composite of skills being most accurately throughout the two-dimensional ability plane?
3. Does the number correct score scale have the same meaning or interpretation throughout the observable range of scores?
4. Which subtests provide the greatest measurement precision and for which ability composites?
5. Is the same degree of measurement precision of various composite skills displayed across forms?

An interesting way to graphically display information was developed by Reckase and McKinley (1991). Referred to as "clamshell" plots because of their shape, such plots denote the amount of information at selected points in several different directions. Specifically, a "clamshell " plot is created by computing the amount of information (using (2)) at selected θ_1,θ_2 points in a 7 x 7 grid of points. The amount of information that the test is providing for 10 different composites or measurement directions (from 0° to 90° in 10° increments) is computed for each point and represented by a vector originating from the specified point. The result of the ten vectors at each two-dimensional ability point resembles a "clamshell".

The clamshell plot for each of the two LSAT forms are displayed in Figure 5. In the top panel of Figure 5 is the information plot for the estimated item parameters for the entire LSAT December 1991 test and in the bottom panel are the results for the October 1992 test. The similarity is clearly remarkable with all composites being measured about equally well, with the 40° to 50° being measured most accurately for both forms over most of the latent ability plane.

Insert Figure 5 about here

Although clamshell plots help us to ascertain the amount of information and which composites are being best measured, one may want to know what direction is being best measured overall at each of the 49 points. This direction can found by examining all composites from 0° to 90° in 1° increments and noting the direction having the largest value. This process would be repeated at each of the selected points. One way to display these results is shown in Figure 6. At each selected point the number denoting the angular direction of the θ_1,θ_2 composite that is being measured most precisely is written. The font size of the number indicates the relative amount of information in the specified direction: the larger the font the greater the measurement precision.

Insert Figure 6 about here

Results for the entire December 1991 test are again shown in the top panel, with the results for the entire October 1992 test displayed in the bottom panel. The degree of similarity between the forms suggests that they are remarkably parallel. Samejima (1977) referred to tests that have similar information profiles as "weakly parallel tests".

Practitioners usually report test results in terms of raw scores or standard scores. Thus, there is a need to determine the relationship between the number correct score scale and the two-dimensional latent ability plane. One way to examine this relationship is to compute the linear weighting of θ_1 and θ_2 that ,rovides the greatest relationship or correlation with the number correct scale.

parallel their true score surfaces should be very similar. One way to analyze the degree of parallelism, is to graph the surface representing the difference between the two true score surfaces. Such a plot is shown in Figure 9.

Insert Figure 9 about here

In Figure 9 the zero plane is outlined. If there were no difference between the true score surfaces (i.e., the forms were strictly parallel), the difference surface would lie in this zero plane. Regions where the difference surface lies above the plane (e.g., first quadrant) indicate that examinees would have a higher true score on the December 1991 form; where the difference surface dips below the zero plane (e.g., thi. ' quadrant), indicates regions where examinees would have a higher true score on the October 1992 form. The maximum difference indicated by the contour is about two true score points -- somewhat remarkable for two 101-item tests.

If pretest sample sizes permitted accurate multidimensional IRT calibration of the item parameters this analysis could b. conducted as part of the test construction phase and before any forms are administered.

Another question still remains concerning the true score surface. Any curvilinearity of the true score contours implies that differences between levels of true score may not have the same meaning (in terms of the θ_1, θ_2-composites) throughout all regions of the ability plane. More specifically, what composite of θ_1,θ_2 is being best measured for the examinees at each true score level?

To investigate this issue further thirty different θ_1,θ_2-combinations, each having the same true score, were systematically located for each possible true score. At each of the thirty ability points, the direction of maximum information was computed, and an average direction, weighted by the density of the examinees at each of the thirty points, was then calculated. This process was repeated for each possible true score. The results for each of the two LSAT forms is displayed in Figure 10.

Insert Figure 10 about here

Consistency of scale interpretation can be seen by the narrow band containing the best measured composites at each possible true score. Again, the similarity between the two forms appears to be exceptional.

Conditional analyses

Another set of analyses focus on such measurement questions as: What is the mean θ_1, and mean θ_2 (or centroid) of the people who would be expected to achieve a particular score? How variable (in terms of their θ_1 values and θ_2 values) is the distribution of examinees who would be expected to achieve a particular score? These questions are also at the heart of the issue of score scale consistency and have strong implications for issues relating to equating and bias analyses.

To answer the first question the conditional bivariate distribution

$$h(\theta_1, \theta_2 | X=x) \ , \ 0 \le x \le n$$

for each possible raw score x, is computed. Assuming a bivariate normal distribution of ability, $f(\theta_1, \theta_2)$, and using the estimated item response function $P_i(\theta_1, \theta_2)$ given in (1), the conditional ability distributions, $h(\theta_1, \theta_2 | x)$ can be estimated by the Bayesian formula

$$h(\theta_1, \theta_2 | x) = \frac{Prob(\boldsymbol{\Theta} = \boldsymbol{\theta} \; and \; X=x)}{Prob(X=x)}$$

$$h(\theta_1, \theta_2 | x) = \frac{Prob(X=x | \boldsymbol{\Theta} = \boldsymbol{\theta}) f(\theta_1, \theta_2)}{\iint Prob(X=x | \boldsymbol{\Theta} = \boldsymbol{\theta}) f(\theta_1, \theta_2) d(\theta_1) \, d(\theta_2)} \qquad (3)$$

Using a recursive formula developed by Lord and Stocking (1983), the $\mathit{Prob}(X-x|\boldsymbol{\Theta}-\boldsymbol{\theta})$ was computed for a grid of (θ_1,θ_2) values and $h(\theta_1,\theta_2|x)$ estimated using (3). The centroids of the distributions,

$$(\bar{\theta}_1,\bar{\theta}_2) - \mathscr{E}(h(\theta_1,\theta_2)|X-x)$$

were then plotted. The results form each LSAT form are plotted for each of the three subtests Analytical Reasoning, Logical Reasoning, and Reading Comprehension and the total test in Figures 11 (December 1991) and 12 (October 1992).

Insert Figures 11 & 12 about here

These figures are truly interesting. The numbers that are plotted indicate the particular score category and the location of the number is the position of the centroid for that score category. The differences between the various subtests are quite noticeable. Such analysis helps to define substantively θ_1 as representing analytical reasoning skill, and θ_2 as reading comprehension skill. The logical reasoning items appear to be measuring an equal weighting of both skills, as does the over all total score.

Related to the plot of the centroids is another graph highlighting the second moments of the conditional distributions discussed in the previous plots. In order for the observed score scale to have a consistent interpretation in terms of θ_1 and θ_2 not only would the centroids have to be linear but the conditional θ_1 and θ_2 for each expected score must be similar. For a given score category, the ability having the smaller variance is represents the dimension that is being measured more accurately.

In the process of computing the conditional centroids, the conditional variances were also calculated. The size of these variances are represented as an ellipse. The length of the horizontal axis of the ellipse denotes the size of the θ_1 variance. The size of the vertical axis indicates the θ_2 variance. Thus, a circle, would indicate that both abilities are measured equally well for that expected score category. Graphically the ellipses are color coded to indicate which variance is greater. When graphed each ellipse is centered about its corresponding centroid. For the sake of clarity ellipses are drawn for only selected score categories. The number indicating the particular score category is located at the centroid inside the ellipse. The conditional variance ellipses for the LSAT forms are illustrated in Figures 13 (December, 1991) and 14 (October 1992).

Insert Figures 13 & 14 about here

Th: vertically elongated ellipses in the top left panel for both forms imply that these items are best measuring θ_1. (This could be confirmed by constructing a plot like that shown in Figure 6 for each subtest individually.) A somewhat "opposite" case occurs with the Reading comprehension items. Somewhat ironical, is that the net effect when all three subtest types are combined is that there appears to be a consistent measure across the true score scale of a composite that reflects an equal weighting of both analytical reasoning and reading comprehension.

Expected score distribution

The final graphical analysis relates to the expected score distribution given a set of two-dimensional item parameters. This information, which is computed in concert with the centroid analysis, is illustrated in Figure 15. In this figure, the contour of the underlying ability distribution, and corresponding centroid is graphed in the bottom part of the plot. At the top of the plot is a relative frequency curve of the expected true score distribution. This information is truly important to furnish a check of the degree of parallelism between created test forms before they are administered. In Figure 15, the expected score distribution for the 49-item logical reasoning test for the December 1991 form is drawn.

Insert Figure 15 about here

Discussion

This paper outlines a series of graphical representations of item and test analyses that can be used to provide the testing practitioner with more insight about what a test is measuring, how well it is measuring, and how the score scale can best be interpreted. All of the analyses shown are predicated on the fact that a two-dimensional solution is the correct solution and the accurate multidimensional IRT item parameter estimates have been obtained.

There should always be a close tie between the test construction phase and the post administration analyses. The bridge between statistical analyses and item writing is an important one; even moreso because of the issue of multidimensionality. Item writers' work should not stop after a test form has been administered. Likewise psychometricians' work should not stop after the item analysis and equating have been completed. Collectively they need to perform a post hoc analysis to consider such questions as: What makes an item more discriminating or more difficult than others? What features of the text contribute to an item's potential to measure different skill composites? Once hypotheses about these issues have been developed item writers should be challenged to see if they can create items and subsequently predict certain multidimensional characteristics. Through an iterative and cooperative effort the goal should be to find the ideal magnification power provided by the multidimensional analytical lens that can accurately detect the important attributes designated by the test specifications that are used to establish the construct validity of the test.

References

Ackerman, T.A. (in press). Creating a test information profile in a two-dimensional latent space. Applied Psychological Measurement.

Carlson, J.E. (1987). Multidimensional item response theory estimation: A computer program (Research report ONR87-2). Iowa City, IA: American College Testing.

Drasgow, F. & Lissak, R.I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. Journal of Applied Psychology, 68, 363-373.

Fraser, C. & McDonald, R.P. (1988). NOHARM: Least squares item factor analysis. Multivariate Behavioral Research, 267-269.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179-185.

Kim, H.R. & Stout, W.F. (1994, April). A new index for assessing the amount of multidimensionality and or simple structure present in test data. A paper presented at the AERA Annual meeting: New Orleans.

Reckase, M.D. & McKinley, R.L. (1991). The discriminating power of items that measure more than one dimension. Applied Psychological Measurement

Reckase, M.D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Samejima, F. (1977), Weakly parallel tests in latent trait theory with some criticism of classical test theory. Psychometrika, 42, 193-198.

Stocking, M. & Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Stout, W.F. (1987). A nonparametric approach to assessing latent trait unidimensionality. Psychometrika. 52, 589-617.

Tucker, L.R., Humphreys, L.G., & Rosnowski, M.A. (1986). Comparative accuracy of five indices of dimensionality of binary items. Champaign-Urbana: University of Illinois, Department of Psychology.

Wilson, D., Wood, R., & Gibbons, R. (1987). TESTFACT. Scientific Software: Mooresville, IN.

Figure Captions

Figure 1. A graphical illustration of the Newton-Raphson MLE ability estimation in a two-dimensional ability space.

Figure 2. Four perspectives of an item characteristic surface.

Figure 3. The contour plot of an item characteristic surface.

Figure 4. Vector representation of two-dimensional items.

Figure 5. Test information vectors for ten different ability composites from 0° to 90° in ten degree increments at 49 selected θ_1,θ_2-abilities.

Figure 6. Angular composites of maximum information at 49 selected θ_1,θ_2-abilities.

Figure 7. Angular composites have the largest correlation with number correct score for generated examinees in eight regions of the two-dimensional ability plane.

Figure 8. Two-dimensional true score surface and corresponding contour plots.

Figure 9. A plot of the difference between the true score surfaces for the December 1991 LSAT test minus the true score surface for the October 1992 LSAT test.

Figure 10. Angle of average maximum information for each possible true score value.

Figure 11. Conditional centroids for selected observed score values for the three LSAT subtests and total test from the December 1991 administration.

Figure 12. Conditional centroids for selected observed score values for the three LSAT subtests and total test from the October 1992 administration.

Figure 13. Conditional variance ellipses for selected observed score values of the three LSAT subtests and total test from the December 1991 administration.

Figure 14. Conditional variance ellipses for selected observed score values of the three LSAT subtests and total test from the October 1992 administration.

Figure 15. Expected true score distribution and contour of specified underlying two-dimensional ability distribution for the December 1991 Logical Reasoning test.

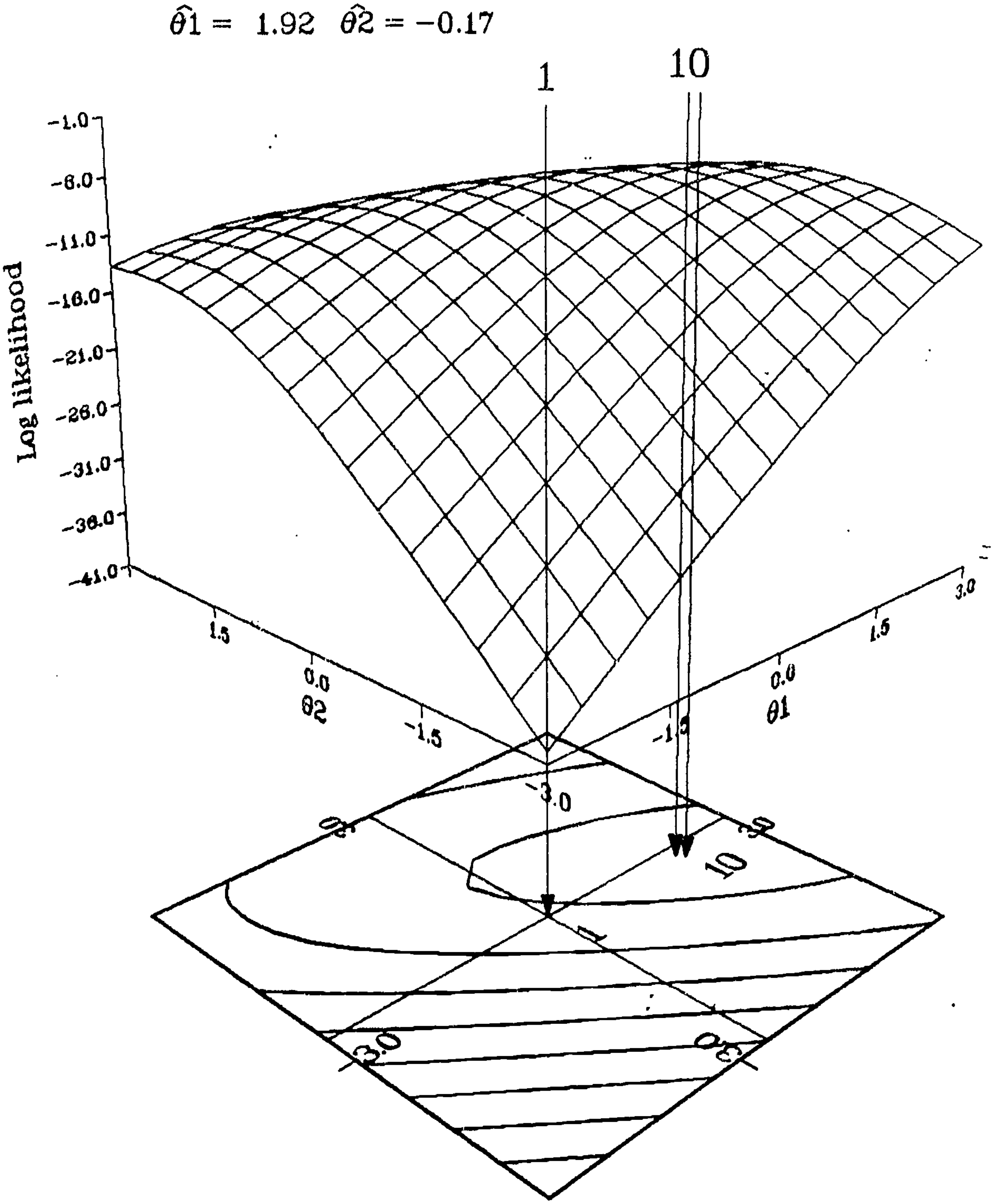

$\hat{\theta}1 = 1.92 \quad \hat{\theta}2 = -0.17$
1
10
Log likelihood
-1.0
-6.0
-11.0
-16.0
-21.0
-26.0
-31.0
-36.0
-41.0
1.5
0.0
θ2
-1.5
-3.0
3.0
1.5
0.0
θ1
-1.5
3.0
10
1
3.0
3.0

Figure 1

A1 = 1.60 A2 = 0.80 D = −0.40

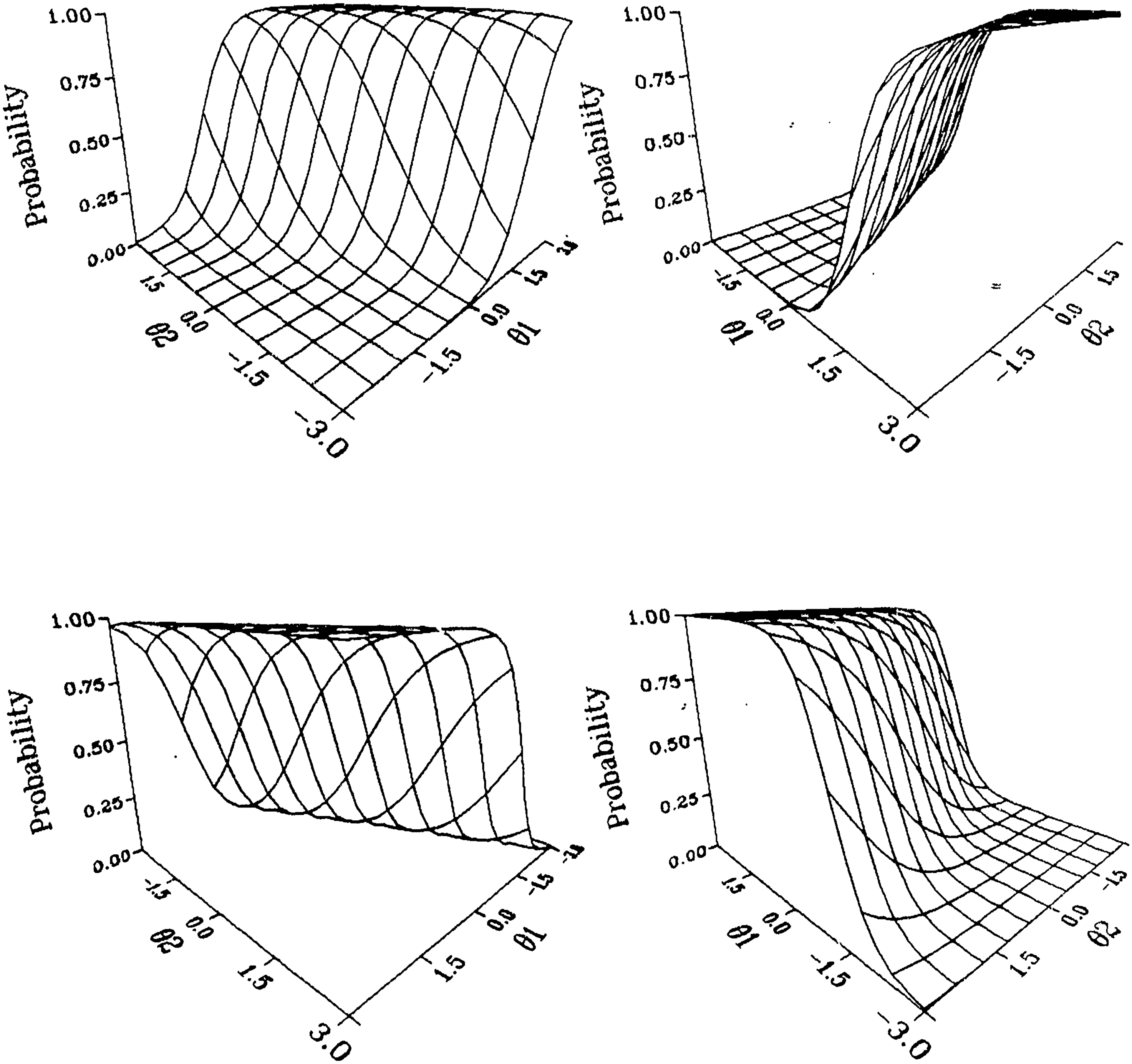


Figure 2

A1 = 1.60 A2 = 0.80 D = −0.40

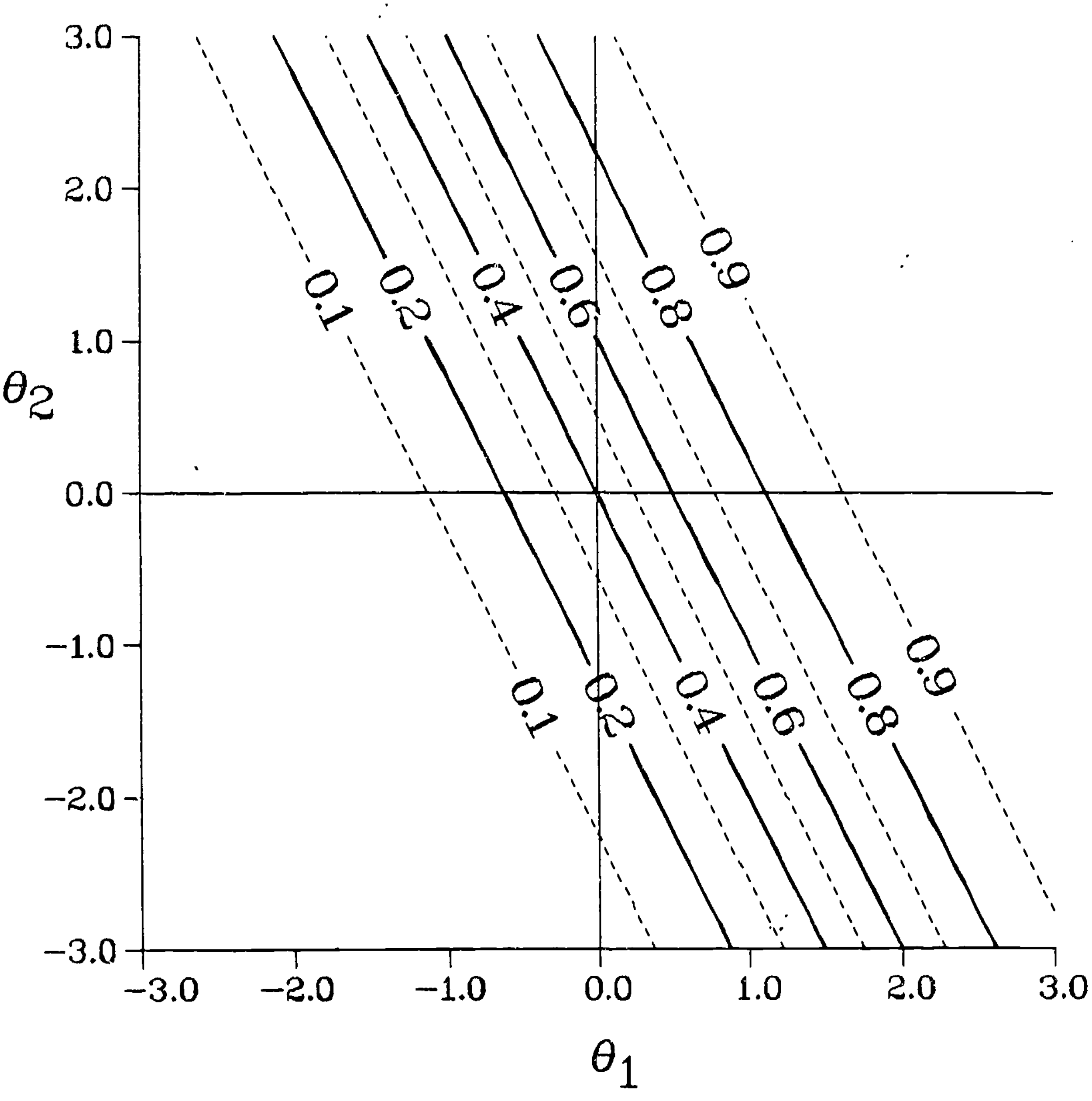


Figure 3

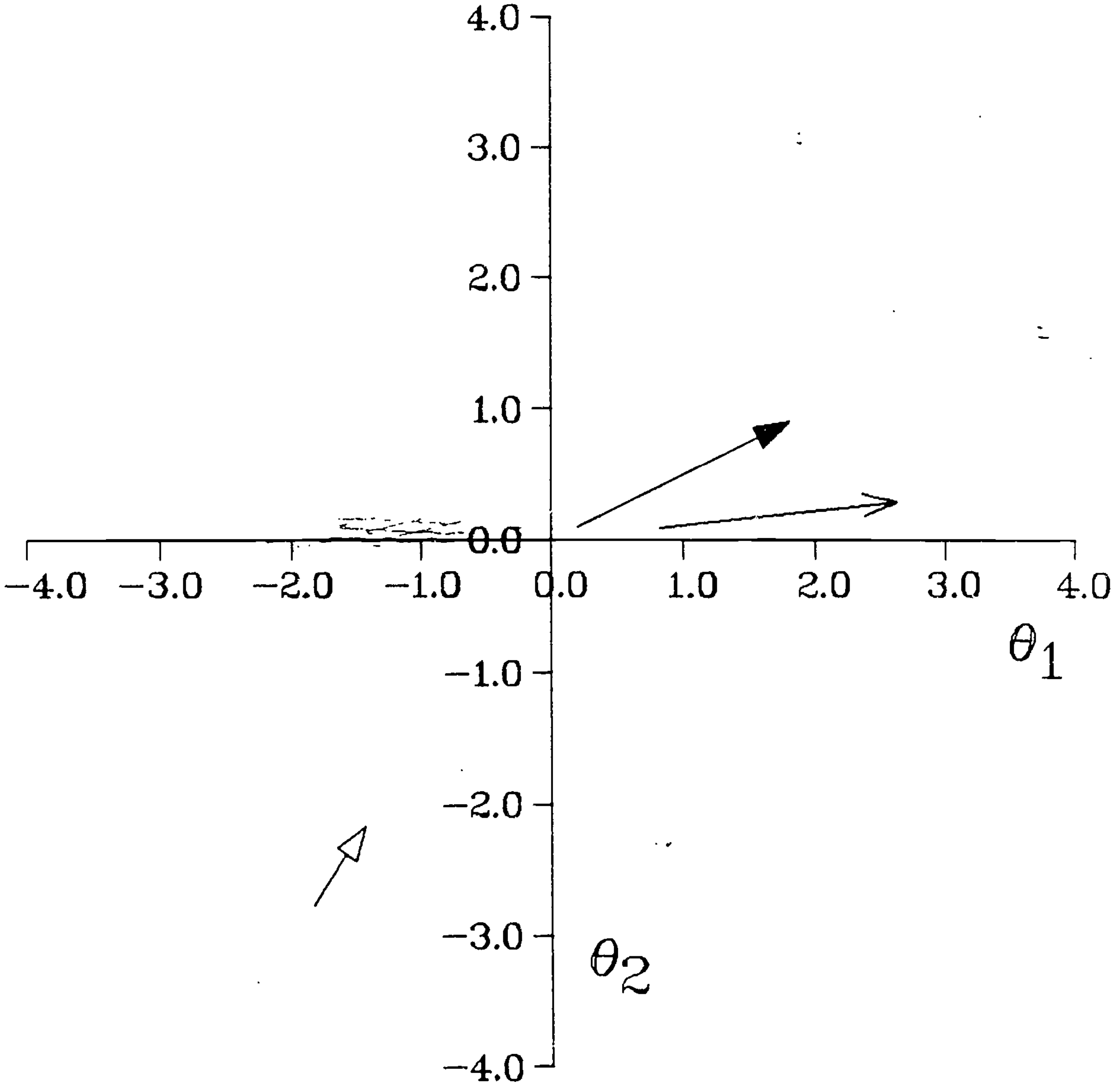
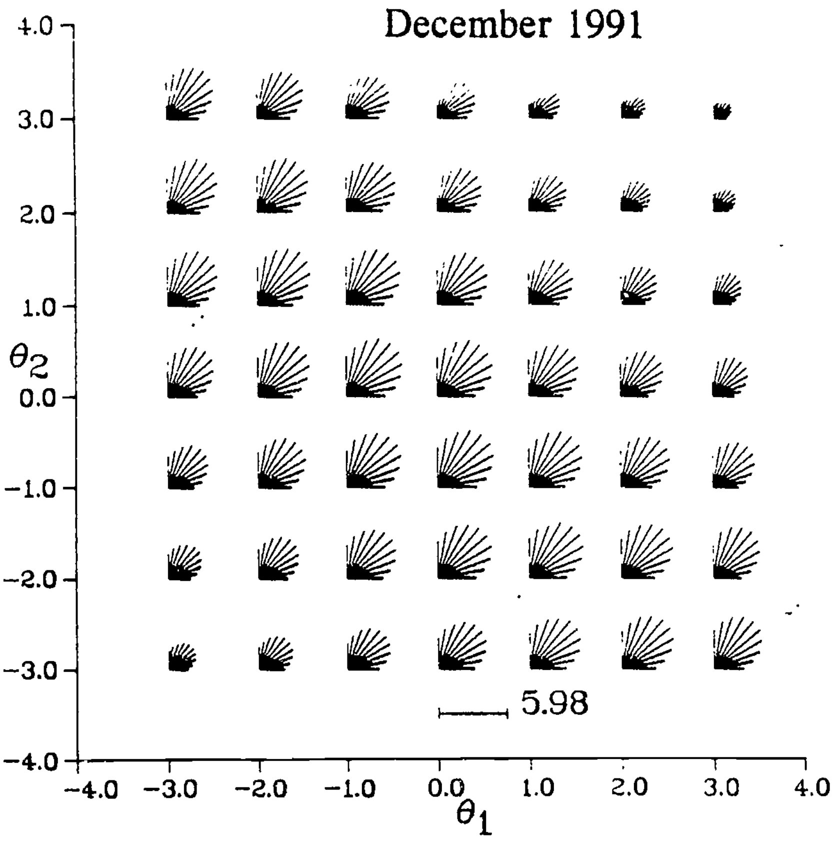


Figure 4


December 1991
4.0
3.0
2.0
1.0
θ_2
0.0
−1.0
−2.0
−3.0
−4.0
5.98
−4.0
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
4.0
θ_1

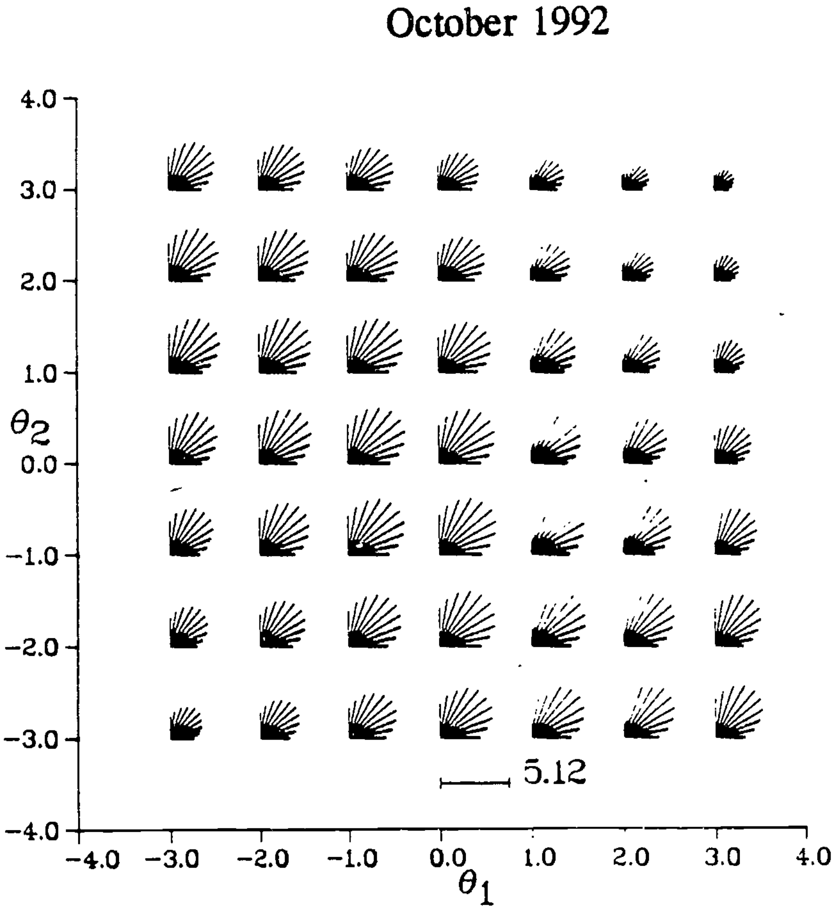

October 1992
4.0
3.0
2.0
1.0
θ_2
0.0
−1.0
−2.0
−3.0
−4.0
5.12
−4.0
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
4.0
θ_1

Figure 5

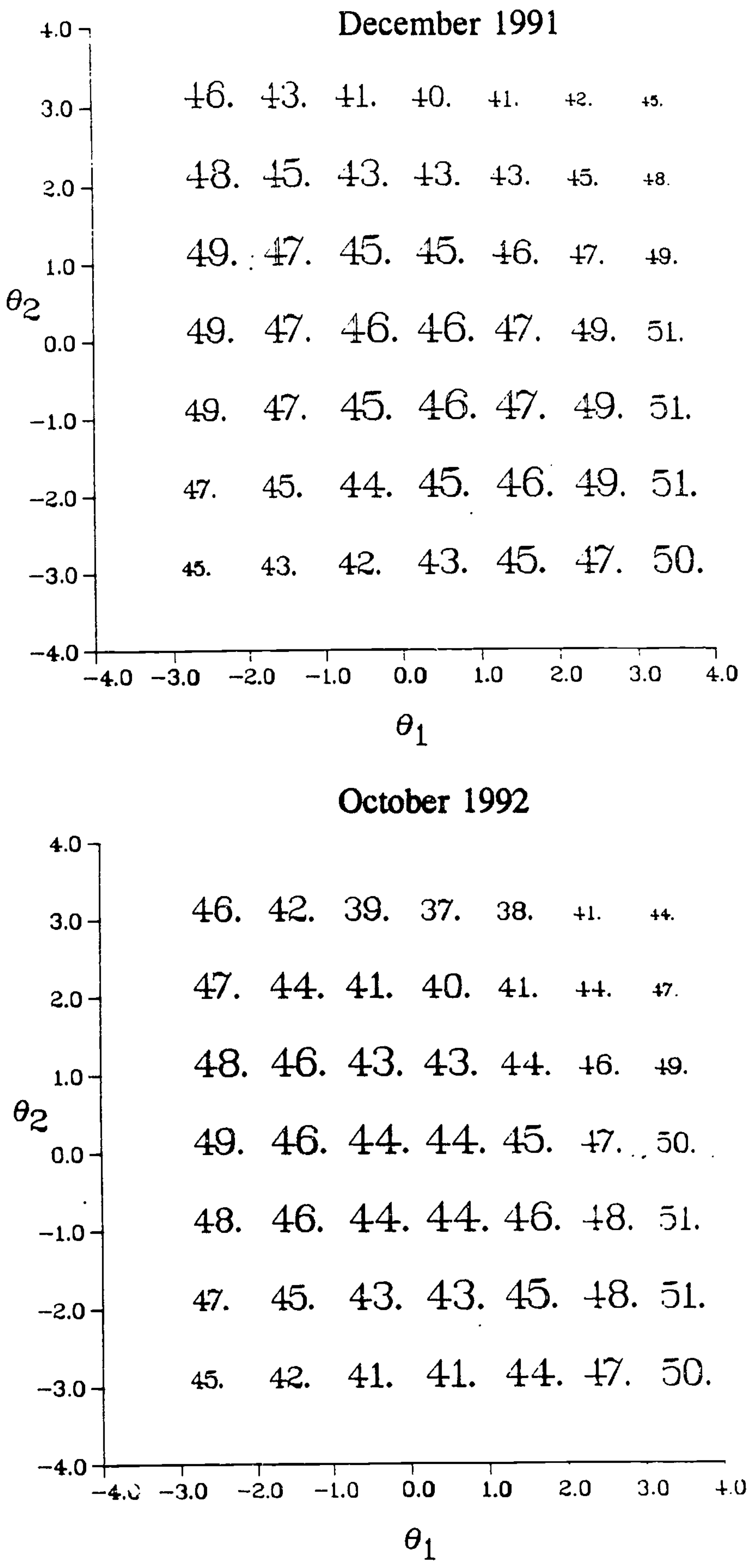
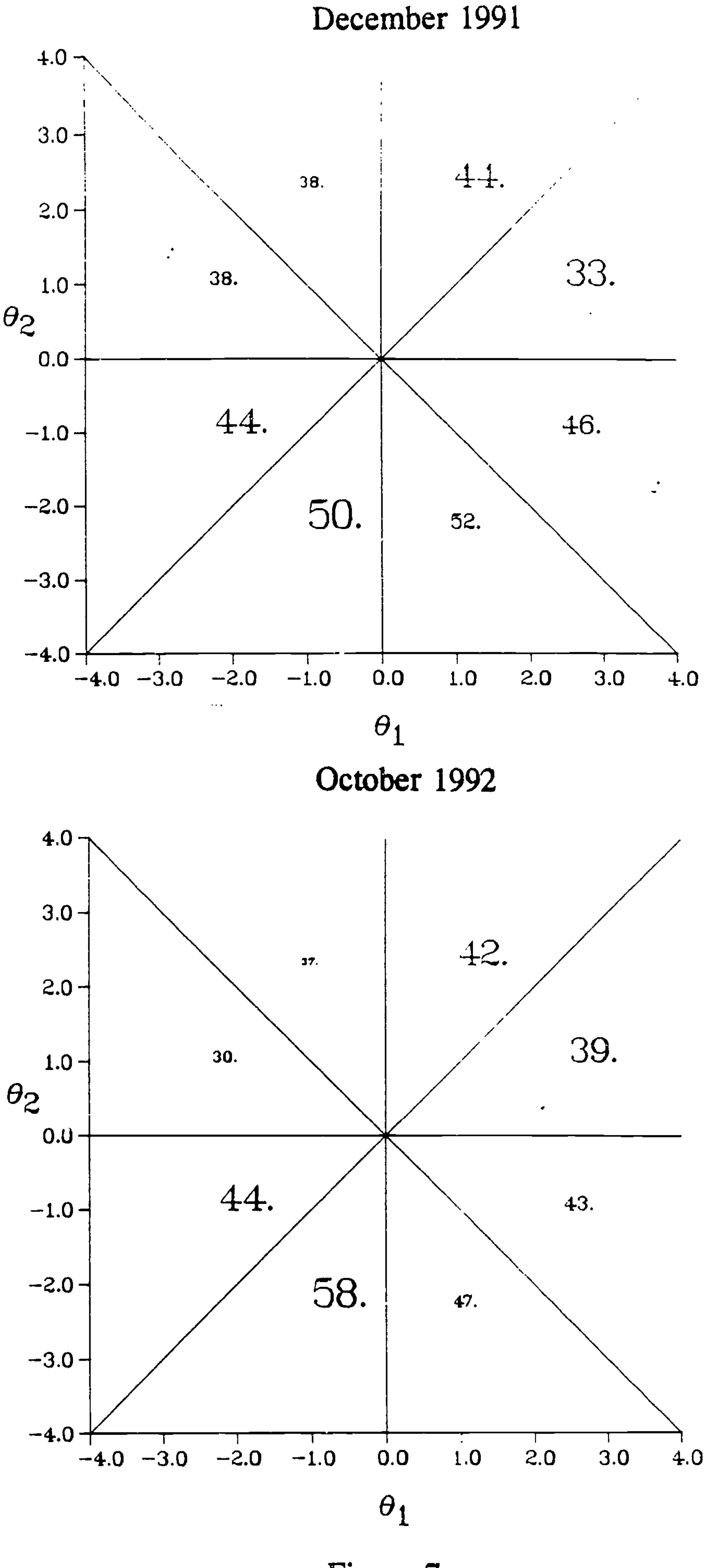

December 1991
46. 43. 41. 40. 41. 42. 45.
48. 45. 43. 43. 43. 45. 48.
49. 47. 45. 45. 46. 47. 49.
49. 47. 46. 46. 47. 49. 51.
49. 47. 45. 46. 47. 49. 51.
47. 45. 44. 45. 46. 49. 51.
45. 43. 42. 43. 45. 47. 50.
4.0 3.0 2.0 1.0 0.0 −1.0 −2.0 −3.0 −4.0
−4.0 −3.0 −2.0 −1.0 0.0 1.0 2.0 3.0 4.0
θ_2
θ_1
October 1992
46. 42. 39. 37. 38. 41. 44.
47. 44. 41. 40. 41. 44. 47.
48. 46. 43. 43. 44. 46. 49.
49. 46. 44. 44. 45. 47. 50.
48. 46. 44. 44. 46. 48. 51.
47. 45. 43. 43. 45. 48. 51.
45. 42. 41. 41. 44. 47. 50.
4.0 3.0 2.0 1.0 0.0 −1.0 −2.0 −3.0 −4.0
−4.0 −3.0 −2.0 −1.0 0.0 1.0 2.0 3.0 4.0
θ_2
θ_1

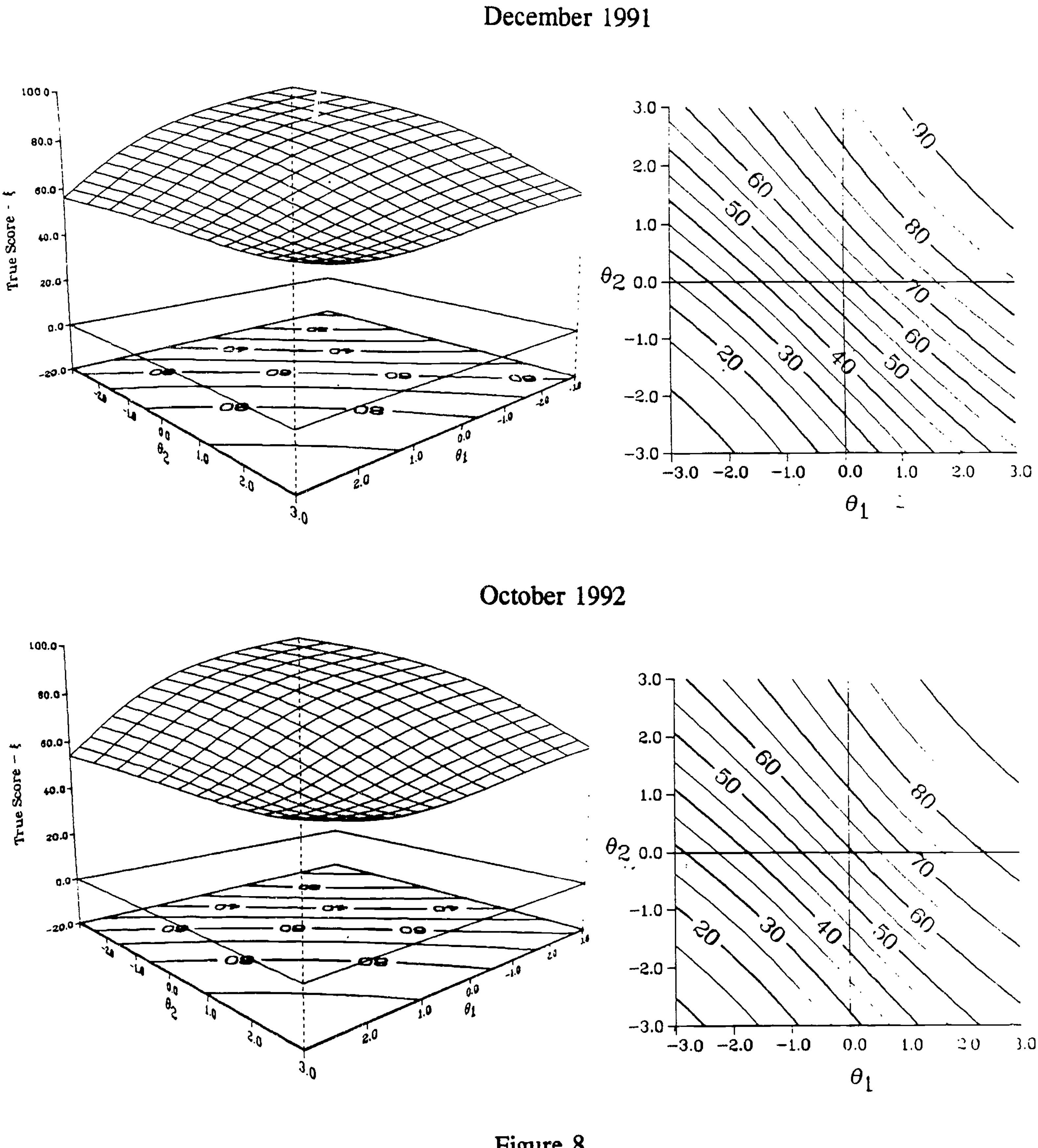
Figure 6



December 1991

October 1992

Figure 7



December 1991

October 1992

Figure 8

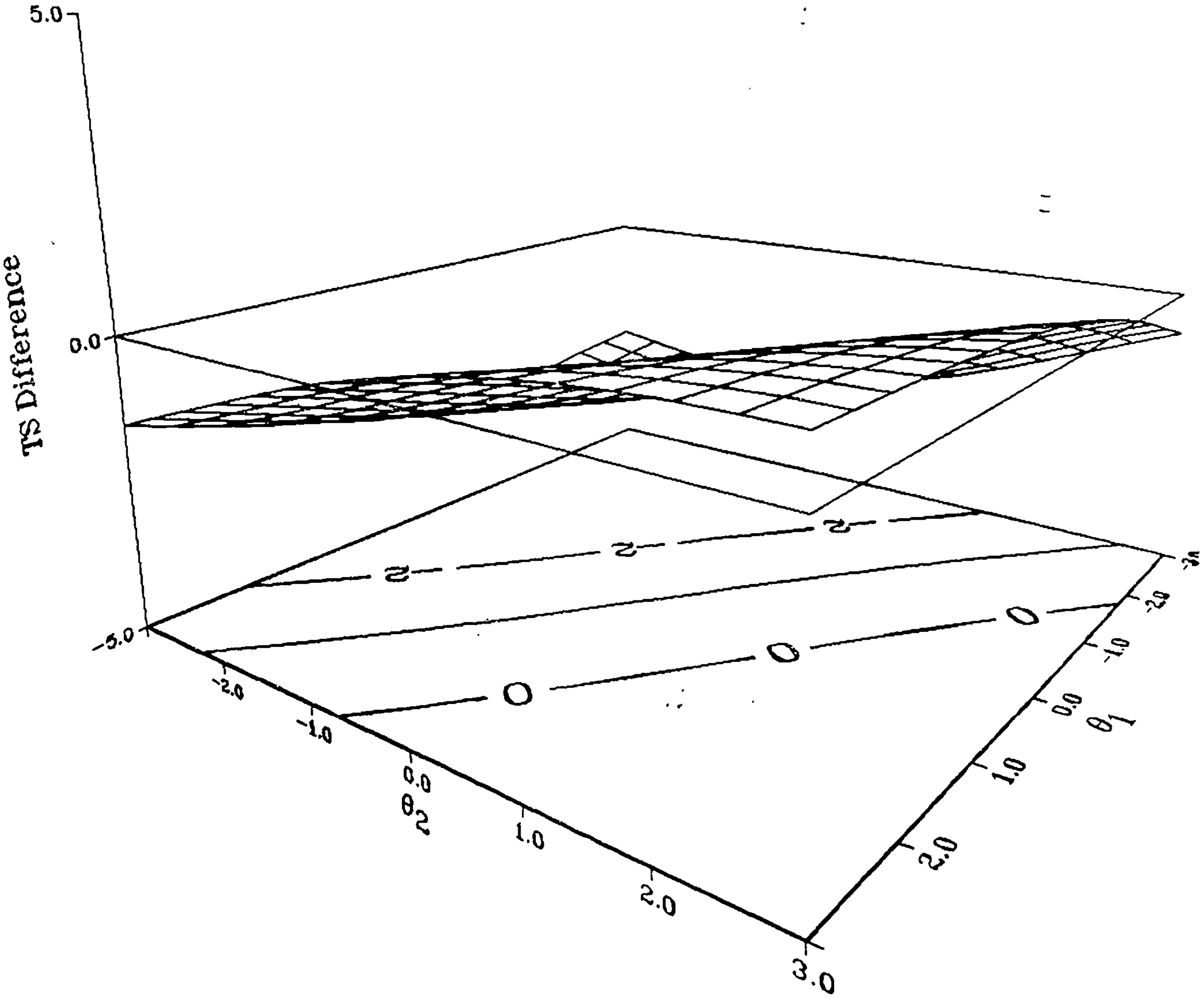
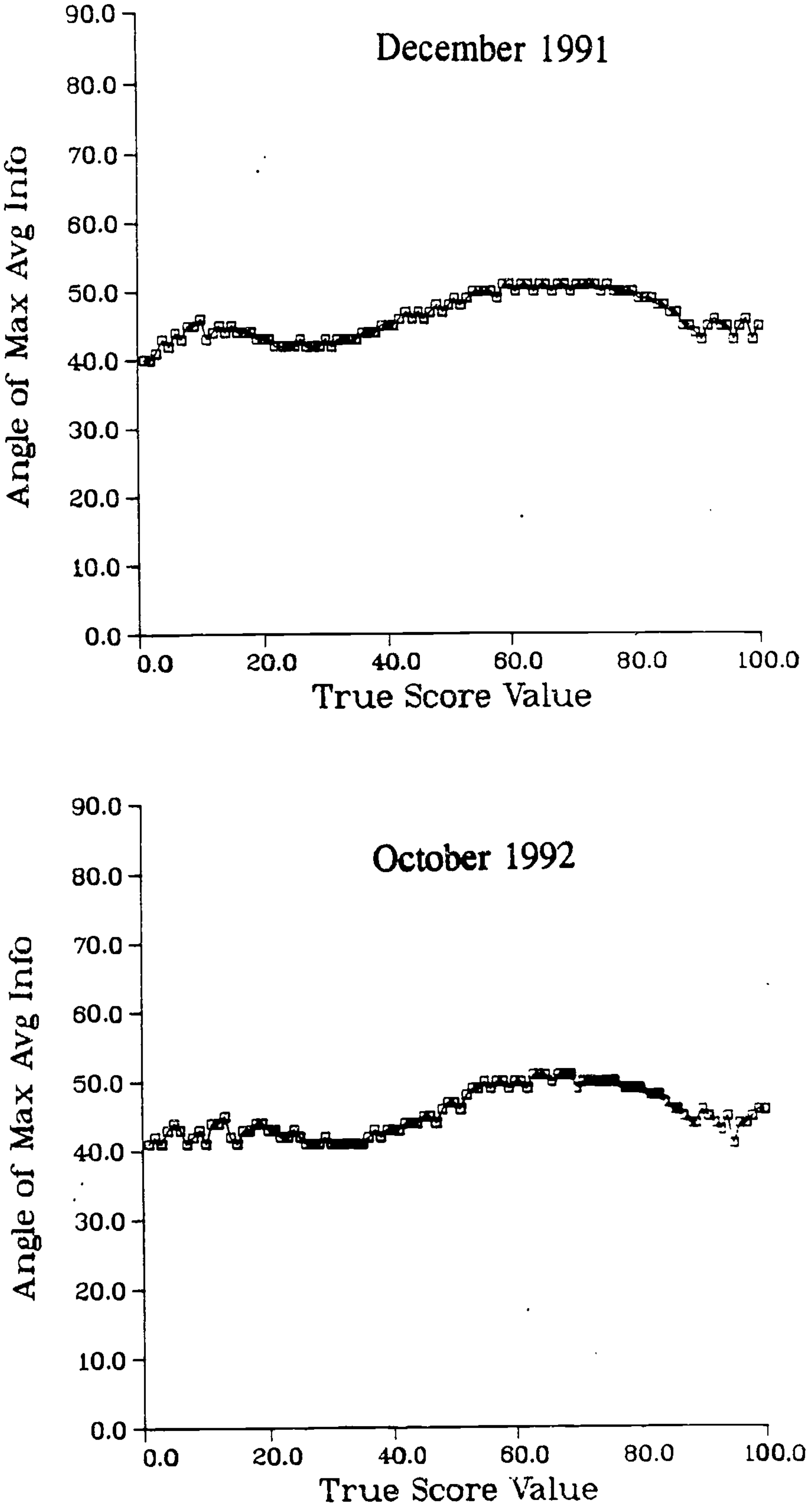


Figure 9


December 1991

October 1992

Figure 10

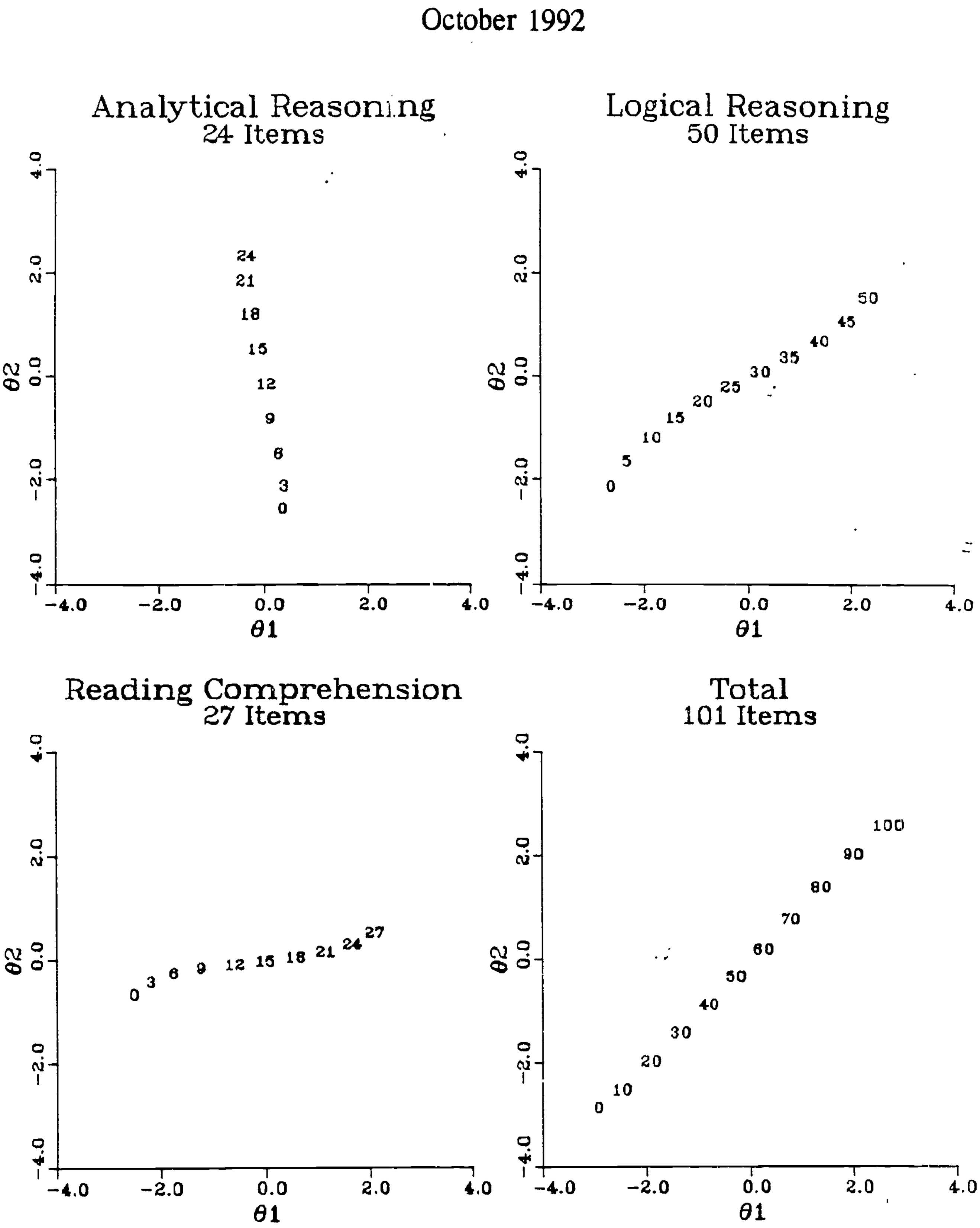

October 1992
Analytical Reasoning
24 Items
0
3
6
9
12
15
18
21
24
Logical Reasoning
50 Items
0
5
10
15
20
25
30
35
40
45
50
Reading Comprehension
27 Items
0
3
6
9
12
15
18
21
24
27
Total
101 Items
0
10
20
30
40
50
60
70
80
90
100
θ2
θ1
4.0
2.0
0.0
−2.0
−4.0

Figure 11

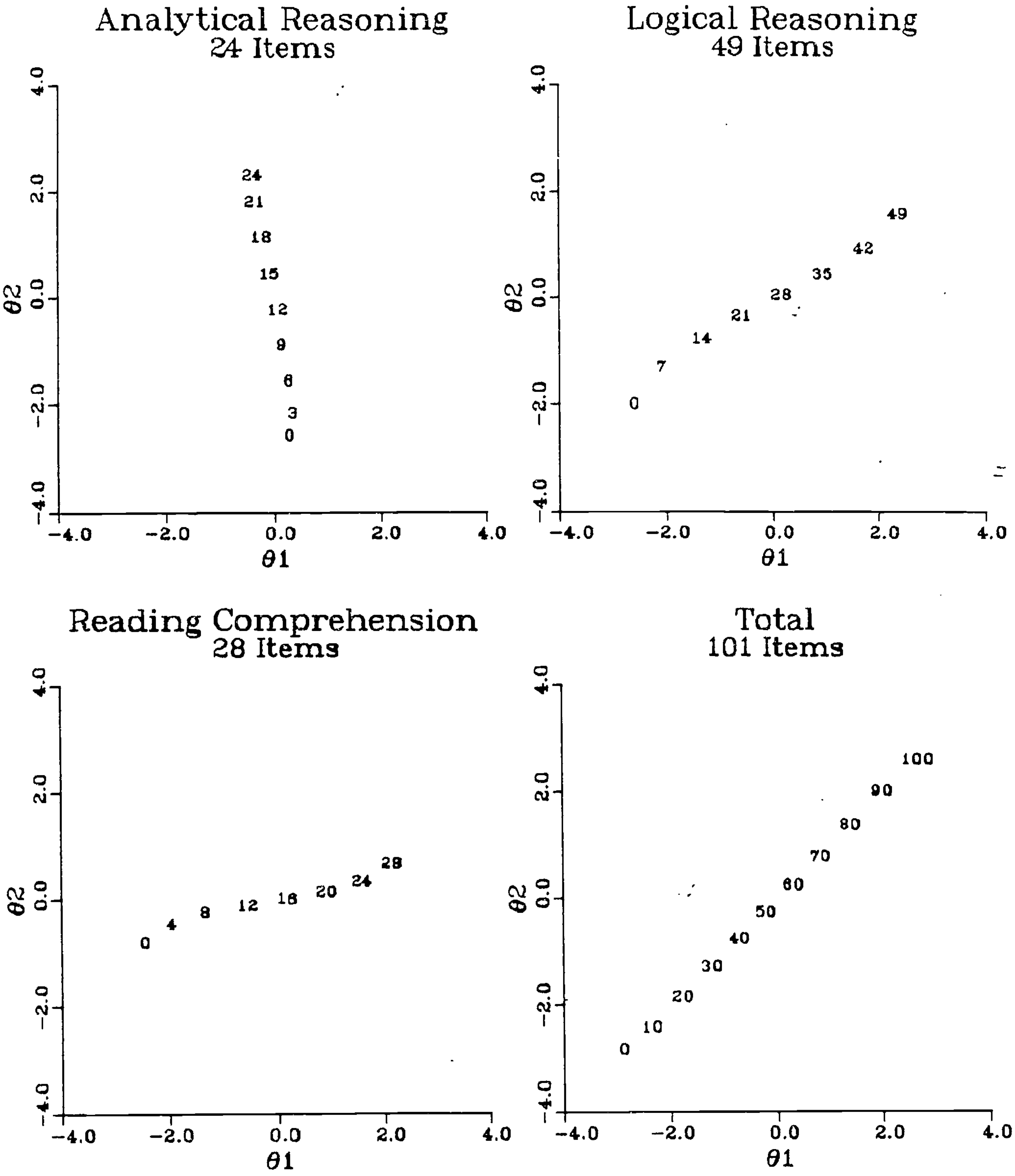

Analytical Reasoning
24 Items
θ2
θ1
4.0
2.0
0.0
−2.0
−4.0
−4.0
−2.0
0.0
2.0
4.0
24
21
18
15
12
9
6
3
0
Logical Reasoning
49 Items
θ2
θ1
4.0
2.0
0.0
−2.0
−4.0
−4.0
−2.0
0.0
2.0
4.0
49
42
35
28
21
14
7
0
Reading Comprehension
28 Items
θ2
θ1
4.0
2.0
0.0
−2.0
−4.0
−4.0
−2.0
0.0
2.0
4.0
0
4
8
12
16
20
24
28
Total
101 Items
θ2
θ1
4.0
2.0
0.0
−2.0
−4.0
−4.0
−2.0
0.0
2.0
4.0
100
90
80
70
60
50
40
30
20
10
0

Figure 12

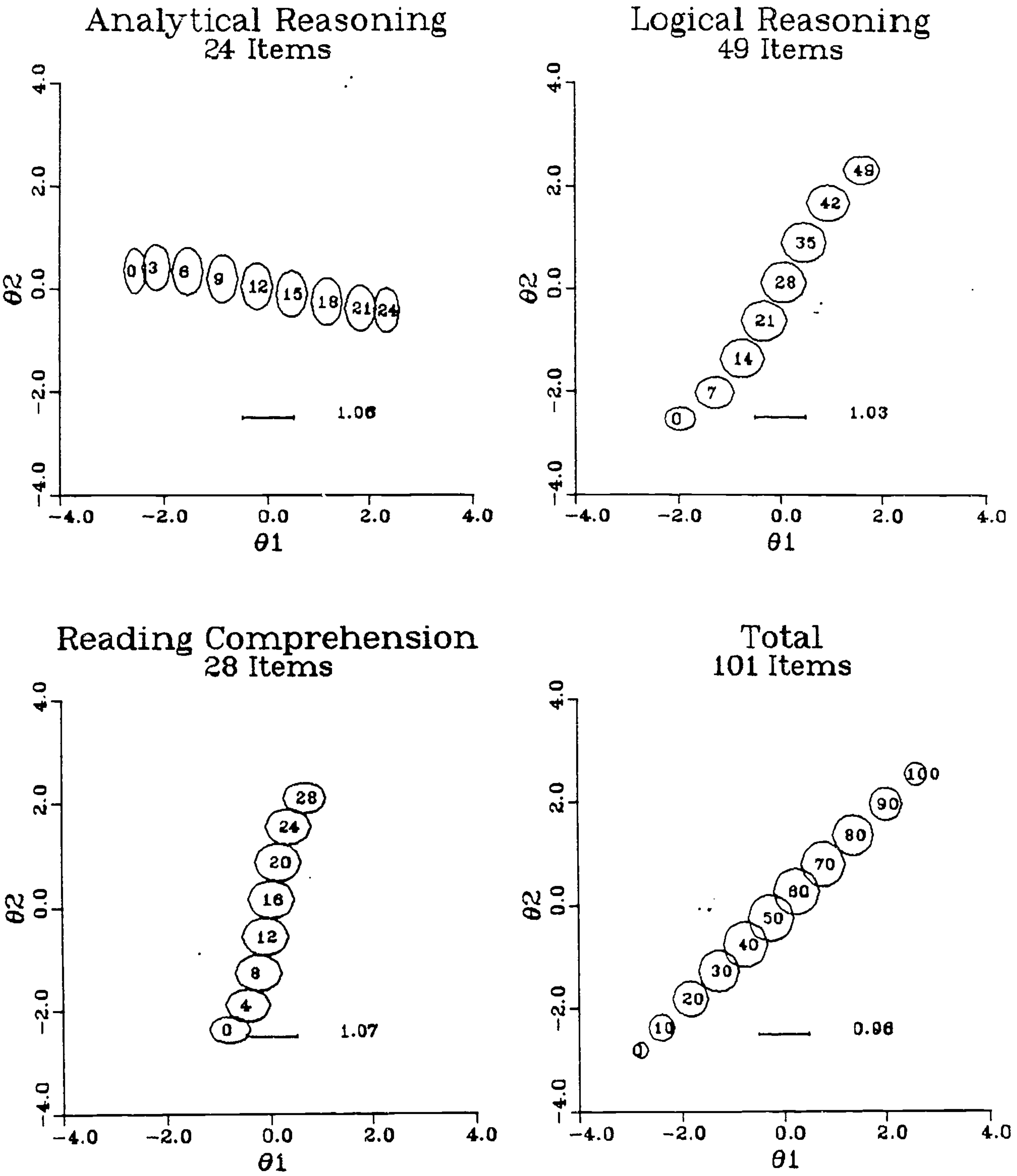

Analytical Reasoning
24 Items
0
3
6
9
12
15
18
21
24
1.06
Logical Reasoning
49 Items
0
7
14
21
28
35
42
49
1.03
Reading Comprehension
28 Items
0
4
8
12
16
20
24
28
1.07
Total
101 Items
0
10
20
30
40
50
60
70
80
90
100
0.96
θ2
θ1
4.0
2.0
0.0
−2.0
−4.0

Figure 13

October 1992

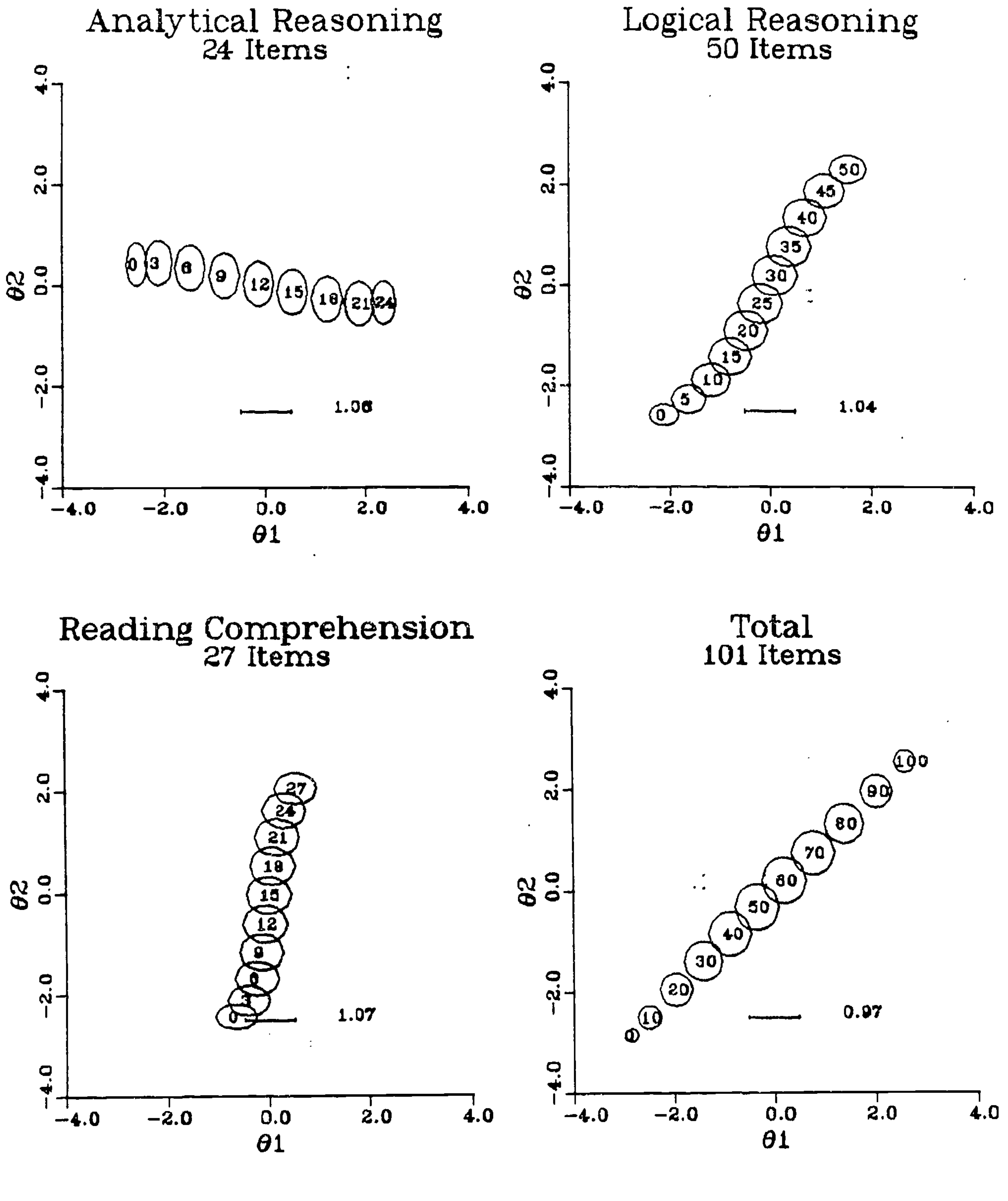


Figure 14

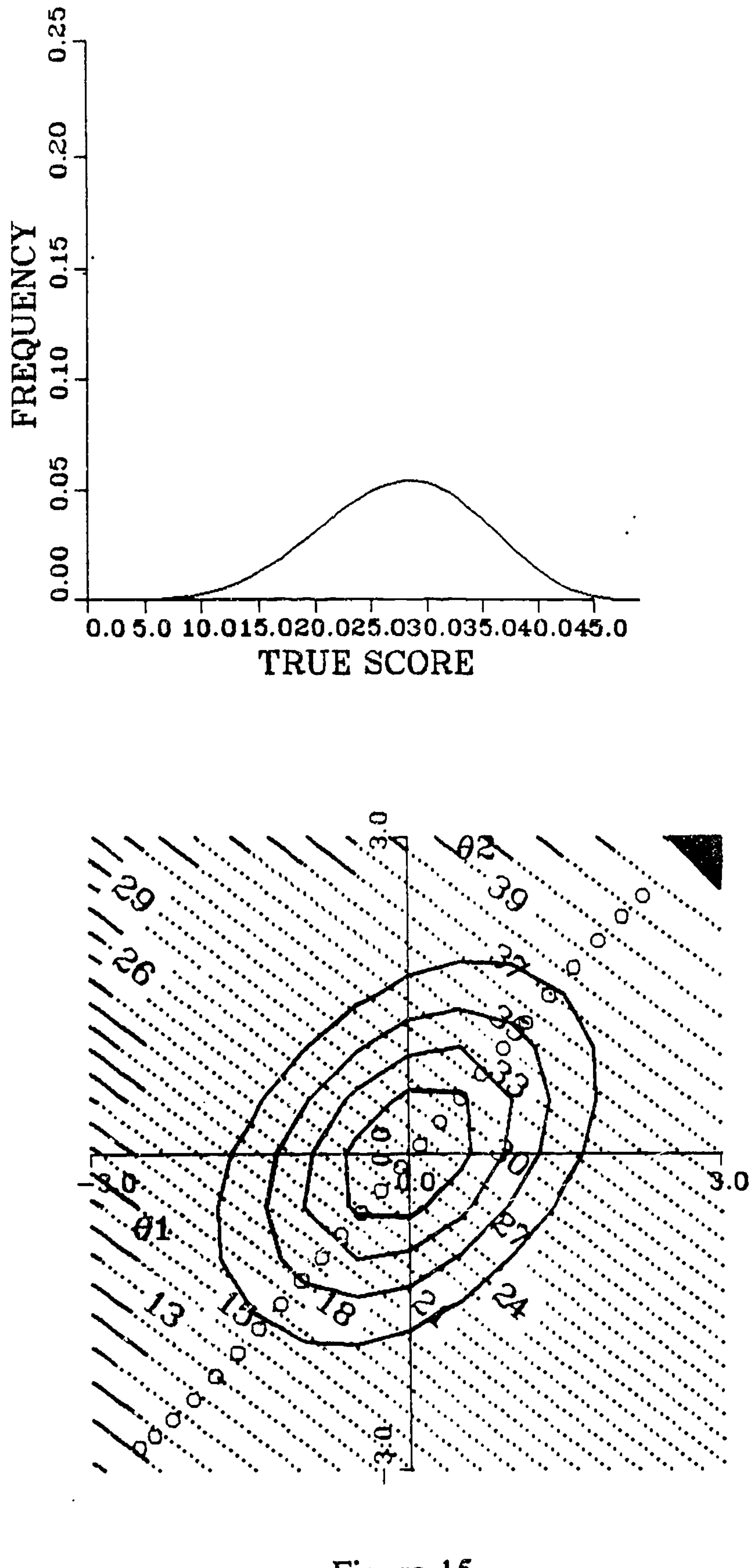

0.25
0.20
0.15
0.10
0.05
0.00
FREQUENCY
0.0 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0 45.0
TRUE SCORE
3.0
θ2
39
29
26
37
36
33
30
27
24
21
18
15
13
θ1
-3.0
0.0
3.0
-3.0

Figure 15