DOCUMENT RESUME

ED 333 014 TM 016 462

AUTHOR	Ackerman, Terry A.
TITLE	A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective.
PUB DATE	Apr 91
NOTE	40p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Equations (Mathematics); *Item Bias; Item Response Theory; *Mathematical Models; Multidimensional Scaling; *Scoring; Test Items; Test Use; *Test Validity
IDENTIFIERS	Ability Estimates; Mantel Haenszel Procedure; Simultaneous Item Bias Procedure

ABSTRACT

Many researchers have suggested that the main cause of item bias is the misspecification of the latent ability space. That is, items that measure multiple abilities are scored as though they are measuring a single ability. If two different groups of examinees have different underlying multidimensional ability distributions and the test items are sensitive to these differences, any scoring scheme that does not reflect all of the skills in the interaction of the items and examinees (the complete latent space) will likely produce item bias. Insight is offered the testing practitioner concerning the difference between item bias and item impact and how each relates to item validity. These concepts are addressed from a multidimensional item response theory perspective. Two detection procedures, the Mantel-Haenszel (as modified by P. W. Holland and D. T. Thayer) and the Simultaneous Item Bias procedure of R. Shealy and W. Stout are used to illustrate the detection of item bias. It is concluded that empirically, two or more items will always produce multidimensionality, and as such, their parameters need to be estimated using multidimensional models. One table and 12 figures illustrate the discussion. (SLD)

A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective

Terry A. Ackerman

University of Illinois

Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1991

A Didactic Explanation of Item Bias, Item Impact and Item Validity from a Multidimensional IRT Perspective

Abstract

Many researchers have suggested that the main cause of item bias is the misspecification of the latent ability space. That is, items which measure multiple abilities are scored as though they are measuring a single ability. If two different groups of examinees have different underlying multidimensional ability distributions and the test items are sensitive to these differences, then any scoring scheme that does not reflect all of the skills in the interaction of the items and examinees (the complete latent space) will likely produce item bias. It is the purpose of this paper to provide the testing practitioner with insight about the difference between item bias and item impact and how they relate to item validity. These concepts will be explained from a multidimensional item response theory (IRT) perspective. Two detection procedures, the Mantel-Haenszel (as modified by Holland and Thayer) and Shealy and Stout's Simultaneous Item Bias (SIB), will be used to illustrate how practitioners can detect item bias.

Introduction

It is the purpose of most standardized tests to distinguish between levels of ability for individual examinees. These types of tests are purposely designed to rank order individuals. To rank examinees accurately requires that all of the items in a test be able to discriminate between levels of the same purported skill. Problems are encountered when a test contains items that discriminate between levels of several different abilities or several different composites of abilities. Unfortunately, because ordering is a unidimensional concept, we cannot order examinees on two or more skills at the same time, unless we base our ranking on a weighted sum of each skill being measured. Specifically, if a test is multidimensional there is no unique one-to-one mapping between an examinee's estimated unidimensional ability and the examinee's underlying composite of abilities.

To study the relationship between an examinee's latent ability and the probability of a correct response, researchers use probabilistic models of item response theory (IRT). These models describe the interaction of an examinee's level of ability and the difficulty and discrimination parameters of an item. In most cases, practitioners tend to use unidimensional IRT models, even though the score may reflect a composite of multiple abilities. Problems related to this model misspecification have plagued psychometricians for years, especially when they try to model cognitive processes (cf. Traub, 1983).

It is important that practitioners who use IRT models realize that items and examinees interact, and it is this interaction that needs to be closely examined. The interaction between a group of examinees and items on a test will be unidimensional in basically one of three ways. An item may be sensitive to, or require the application of, several skills to produce a correct response. But if a group of examinees only vary significantly on one of the requisite skills, or on the same composite of skills, the interaction can appropriately be modeled unidimensionally. The reverse scenario is also possible. Test items may be sensitive to, or capable of measuring, only a single skill, or the same composite of skills, and, although a group of examinees vary along several skill dimensions, the interaction will be unidimensional. The third situation is the degenerative case in which the test is only one item long. Considered by itself one item is always unidimensional. It can probably be argued that a test composed of two or more items is never exactly unidimensional.

The dimensionality imposed by this test-examinee interaction (i.e., the complete latent ability space dimensionality) is the intersection of the set of abilities, each of which at least one item is capable of measuring, and the set of latent abilities on which the examinees may vary. It becomes readily apparent that the process of measuring different groups has many possible interactions. In particular, because of individual differences, the dimensionality for a given test may change from one group of examinees to another.

Because the measurement process can be quite complex, researchers and practitioners estimate examinee abilities and item parameters using large groups of individuals, all assumed to be homogeneous in the skills they bring to bear on each of the items. If the groups are not homogeneous and thus have different underlying ability distributions the potential for item bias exists.

One approach used to determine if an item is biased is to compare the two sets of parameter estimates, each obtained by independent calibration of the test on different groups of inteiest. Before two sets of unidimensional IRT estimated item parameters can be compared they must first be placed on the same scale. Because of the indeterminacy of the IRT ability scale (Hambleton & Swaminathan, 1985, p. 55) the unidimensional item parameter estimates for each group may be determined from two distinct ability scales which differ in their metric. If the interaction between the examinee latent space and the item is unidimensional, the item parameter estimates should be invariant up to a simple transformation. That is, the only differences that can occur are due to group ability differences (impact, formally defined below) and these can be resolved via the linear transformation. However, if the examinee-item interaction is multidimensional and the items are calibrated to fit a unidimensional model, group ICC differences cannot be resolved by a single transformation. The problem that arises is that after a single transformation is applied and the resulting ICCs are compared, differences that appear may be due to impact, bias, or bias and impact acting together.

To examine bias within a multidimensional framework the true and the nuisance ability dimensions need to be identified and handled separately. More precisely, researchers must examine the conditional distribution of the nuisance ability for each level of the valid ability. If this conditional distribution of the nuisance ability differs across groups of interest the potential for bias exists. This is what motivates Shealy and Stout's (1991) mathematical conception of potential for bias.

The issue of item bias and construct validity are interrelated. That is, the number of skills being measured and the degree to which comparisons between groups are appropriate is a construct validity issue. If a test lacks construct validity it must contain items that are measuring skills other than those purported to be measured and hence the potential for item bias also exists. This bias may be realized if groups of interest differ in their underlying distribution of these extraneous skills. Simply put, items invalid in the construct sense are a necessary, but not sufficient cause of item bias. If all the items are measuring only the valid skills or constructs any group differences reflect impact, not bias. In this paper *impact* is formally defined as the difference in group performance caused by valid skill group differences (e.g., proportion correct difference between two groups of interest on a valid item).

Test creators, in establishing the construct validity of an instrument, specify what the test is measuring and what the reported scores mean. For example, if a test is purported to be measuring algebraic symbol manipulation it should contain items from the universe of algebraic symbol manipulation problems. To the extent items measure supplemental abilities (e.g., reading ability via "story problems") it decreases the degree of construct validity.

Another problem caused by lack of validity occurs when test scores change in their interpretation at different points along the reported score scale. That is, the scale composite being measured changes as a function of the reported score level. Such a concern was expressed by Davey, Ackerman, and Reckase (1989). In all reported test results the assumption is always made that scores throughout the range of the score scale represent different levels of the exact same skill or exact same composite of skills. Changing conditional standard errors for different levels along the score scale might, in part, be caused by the assessment of different skills.

The interaction between a test and a group of examinees can also disguise problems related to validity. Within a test if items are measuring several different skills and the underlying distributions for the groups of interest do not differ on these skills, there can be no bias. Yet the construct validity of the test will suffer (with equal deleterious effect on both groups) because supplemental skills are being assessed along with the valid skills.

Item bias is attributable to the degree of item validity. A test item is considered to be unbiased if all individuals having the same underlying intended-to-be-measured unidimensional ability have equal probability of getting the item conect, regardless of group membership (Pine, 1977). If two groups have different underlying multidimensional ability distributions, and the test items are capable of measuring these multiple dimensions, and the dimensions are collapsed into a single dimension (i.e., score) item bias may occur. Item impact occurs when groups of interest differ in their performance on the skills being measured by the valid items. This is an unavoidable outcome of most tests and should not be viewed in a pejorative manner. Impact simply represents results caused by true differences in the target ability. By contrast test bias (cf. Shealy & Stout, 1991) simply represents score differences caused by nuisance abilities.

It is the purpose of this paper to illustrate how the reduction of a two-dimensional latent ability space to a single score and unidimensionality-based statistical procedures such as those employed in the calibration programs LOGIST (Wood, Wingersky & Lord, 1976) or BILOG (Mislevy & Bock, 1986) can magnify the lack of validity in the form of item bias. Specifically contrived examples will be used to illustrate both impact and bias. Within the context of these examples two nonparametric DIF detection approaches, the Mantel-Haenzsel (Holland & Thayer, 1987) and Shealy and Stout's Simultaneous Item Bias (SIB) (1989) will be described. These specific approaches are used because of their strong theoretical basis for detecting bias as separate from impact.

Theoretical background

Insight about the multidimensional nature of items and examinee abilities can be easily understood through the use multidimensional item response theory (MIRT) models. The work of Reckase (1986), which formally defines MIRT item characteristics, provides a basis for examining the interaction between multidimensional items and the multidimensional ability distributions associated with groups of examinees.

Reckase's work is based upon the MIRT compensatory model (M2PL) in which the probability of a correct response to item i by examinee j is given as

$$P(X_{ij}=1 \mid a_i, d_i, \theta_j) = \frac{e^{(a_i'\theta_j+d_i)}}{1.0+e^{(a_i'\theta_j+d_i)}} \tag{1}$$

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

a_i is the vector of item discrimination parameters,

d_i is a scalar difficulty parameter of item i, and

θ_j is the vector of ability parameters for person j.

In a two dimensional latent ability space (e.g., math and verbal ability dimensions), the a_i vector designates the composite of θ_1 and θ_2 that is being measured. If $a_1 = a_2$ both dimensions would be measured equally well. However, if $a_1 = 0$ and $a_2 = 1.0$ discrimination would only occur along the θ_2 dimension. Using Reckase's notation, the amount of different composites of skill being assessed can be readily apparent. That is, if all of the items are measuring exactly the same (θ_1,θ_2) composite (i.e., the same "direction" on the (θ_1,θ_2) coordinate system) the test would be strictly unidimensional. The more varied the composites that are being assessed, the more multidimensional the test.

Using the notation of Reckase (1986), an item i that requires two abilities for a correct response can be represented in the two-dimensional latent ability space as a vector. The length of the vector for a given item i is equal to the degree of multidimensional discrimination, MDISC given as

$$MDISC_i=\sqrt{a_{1i}^2+a_{2i}^2} \quad . \tag{2}$$

MDISC is analogous to the unidimensional IRT model's discrimination parameter. The measurement direction of the vector in degrees from the positive θ_1 axis is

$$\alpha_i=\arccos\frac{a_{1i}}{a_{1i}^2+a_{2i}^2} \quad . \tag{3}$$

This angle represents the composite of the θ_1-θ_2 ability space that item i is measuring.

The item vector is graphed orthogonal to the p=.5 equiprobability contour. In the compensatory model described in (1) these equiprobability contours are always parallel. The distance, D, froin the origin to the p=.5 contour is computed as

$$D_i=\frac{-d_i}{MDISC} \tag{4}$$

D is analogous the unidimensional difficulty parameter. Because the discrimination parameters can never be negative, the item vectors can only lie in the third quadrant (representing easy items) or in the first quadrant (representing more difficult items.) Figure 1 illustrates an item vector whose M2PL parameters are given as a_1 = 1.2, a_2 = .7, and d = 1.5. Also illustrated in Figure 1 are the equiprobability contours.

Insert Figure 1 about here

Whereas Reckase's work is more from a geometric perspective, other researchers have approached the relationship between the multidimensional and unidimensional IRT models from an analytic perspective. Wang (1986) determined explicit algebraic relationships between the unidimensional estimates (e.g., via LOGIST) and the true multidimensional ability and item parameters for the case where the response process is modeled by the M2PL MIRT model and the unidimensional IRT model is the two-parameter logistic:

$$p(X=1|\theta,a_i,b_i)=\frac{e^{1.7a_i(\theta-b_i)}}{1.0+e^{1.7a_i(\theta-b_i)}} \tag{5}$$

where a_i and b_i are the unidimensional discrimination and difficulty parameters and θ is the unidimensional latent ability measure. Wang demonstrated that when the 2PL IRT model is used to calibrate multidimensional da‸ the resulting unidimensional scale, termed the ***reference composite***, is actually a weighted composite of the underlying multiple dimensions. The weights determining the direction (i.e., relative proportion of θ_1 to θ_2 being measured) of the reference composite are a function of the multidimensional discrimination parameters and the variance-covariance matrix of the underlying multidimensional ability distribution. The concept of the reference composite is important because it provides an interpretation of the unidimensional score scale. Using Wang's formulation, Ackerman (1989) illustrated how group performance and ultimately item bias could be predicted given the underlying ability distributions of two groups of interest and the two-dimensional item parameters.

If all of the items had the same MDISC value, the reference composite would be the average "direction" (principal component) being measured by all of the items. Note that the test could be "almost unidimensional" or "quite multidimensional" and still have the same reference composite because the reference composite is, in a sense, only an average direction.

According to Shealy and Stout (1991), all items having a particular measurement direction, α_{valid} (recall (3)) constitute the valid subtest. The intuitive idea os that this specified direction embodies exactly the composite of abilities the test is designed to measure; that is, all items measuring this composite are valid. In any test, however, items will all have different α's no matter how carefully the test has been constructed. I thus propose to formally define a ***validity sector*** as a narrow sector (and its mirror image projecting through the origin) as constituting the valid subtest items. Figure 2 displays a possible validity sector for a given test. This provides practitioners with a conceptual framework that can be used to describe the range of composite skills that a test is suppose to measure. Items which lie outside of the validity sector are measuring nuisance skills and should be deleted by the test constructor if the goal to make the test more homogeneous and to increase internal consistency. Each test's validity sector will consist of two sectors, one which lies in the first quadrant and contains difficult valid items and its mirror image which lies in the third quadrant and contains only easy valid items. The width of this sector would obviously vary depending upon how narrowly the valid construct can be defined. An example of a validity sector in shown in Figure 2.

Insert Figure 2 about here

A possible measure of item validity, called the ***construct validity index***, CVI, is a function of the direction α an item is measuring (but not the amount of discrimination (MDISC) in that direction) is

$$CVI_i = \cos^2(\mid \alpha - \alpha_{valid\,ref\,comp} \mid). \tag{6}$$

In this equation $\alpha_{valid\,ref\,comp}$ represents the angle of the reference composite of those items that are chosen as the items that best measure the purported trait (i.e., a valid subtest from the Shealy-Stout perspective or a valid sector from my perspective).

Using MIRT models, practitioners can actually visualize the degree of homogeneity of the abilities being assessed. Somewhat ironically, it is the reference composite that determines the potential for item bias. Problems can arise due to a large number of invalid items (each lying slightly outside the validity sector) or due to the large magnitude of discrimination of a few invalid items. If there are enough invalid items the reference composite can be "pulled" outside of the validity sector. Such a test is considered to be ***construct invalid*** because the meaning attached to the unidimensional score scale (i.e., the reference composite for the test) is different than intended. Unidimensional bias analyses could be misinterpreted on such tests because valid items might be determined to be biased and vice versa. Specifically, analyses which condition on the score scale provided by the **entire** test may actually be conditioning on a scale whose direction is more heavily influenced by nuisance abilities than valid skills. This potential exists in bias detection procedures that employ all of the items in the analysis, such as the Mantel-Haenzsel. Practitioners should (and can easily) perform the MH procedure by conditioning only on the examinees' scores from the valid items. The concept of conditioning on only a valid set of items is central to the Shealy-Stout procedure (1989).

One factor which determines the orientation of the reference composite is the principal axis of the underlying ability distribution. This is illustrated in Figure 3. In this diagram each pair of two ellipsoidal contours represents the density of two distinct two-dimensional latent ability distributions for groups, A and B. Also drawn on the plot are two equally discriminating items. Item 1 (a_1 = 1.3, a_2 = .4, d = -1.0) has a measurement angle of 17^{o}; Item 2 (a_1 = .4, a_2 = 1.3, d = -1.0) has a measurement angle of 78^{o}. The reference composite was computed for this two-item test for each group using Wang's (1986) formulation. The reference composites are illustrated as a dotted vectors. The reference composite for Group A, which has a much greater θ_2 than θ_1 variance ($\sigma^2_{\theta_1}$ = .5, $\sigma^2_{\theta_2}$ = 2.5) lies at an angle of 59^o above the positive θ_1 axis. Group B which has much greater θ_1 than θ_2 variance ($\sigma^2_{\theta_1}$ = 2.5, $\sigma^2_{\theta_2}$ = .5) has a reference composite whose angle is only 31^o. Thus, Figure 3 clearly shows that even though each group would be administered the exact same two multidimensional items, the undimensional interpretation (e.g., via LOGIST) of the measured skill would be different for each group because of the influence of the items' interaction with the different underlying ability distributions. The closer a multidimensional item's measurement angle, α, is to the direction of the ability group's principal axis, the larger its estimated unidimensional discrimination parameter will be. If one were to estimate the unidimensional discrimination of the Item 2, for each group in Figure 3, the estimate for group A would be larger. Thus if a test contains several invalid items and their αs are all close to the orientation of the latent ability distribution of a particular group, the reference composite for the group, which is influenced by the interaction of the multidimensional discrimination power of an item and the variance-covariance structure of the multidimensional latent ability distribution, could be pulled outside the validity sector.

Insert Figure 3 about here

Bias, according to Shealy & Stout (1991), can be measured by examining the difference in the marginal item characteristic curves (ICCs) for the two groups of interest. The marginal ICC for a particular group is computed by

$$P(X_i=1 \mid \Theta=\theta)=\int P_i[\theta,\eta]f(\eta|\theta)d\eta \qquad (7)$$

where $P_i(\theta,\eta)$ is the M2PL response function defined in (1) and $f(\eta|\theta)$ is the specified group's conditional distribution of the nuisance dimension, η, given a fixed value of θ, the target ability. This is the ICC that would be obtained via calibration using LOGIST, if the test was strictly unidimensional (because there is in effect no nuisance dimension η to be marginalized out by integration). If the test is measuring two abilities this represents the ICC that would be obtained if differences in the nuisance direction are integrated out. It is important to note, that if $f(\eta|\theta)$ is the same for both groups, bias cannot occur because examinees of equal θ ability will have the same probability of getting the item right.

Bias detection methods

Although there has been a proliferation of methods to detect item bias this paper will focus only on two, the Mantel-Haenzsel (MH), Holland & Thayer (1988) and Shealy and Stout's SIB (1989). Both of these procedures are nonparametric and thus require no model calibration. However, they do have an IRT framework and as such they will be explained within the IRT context that has been already developed.

The MH procedure, when placed in an IRT framework, is analogous to examining item bias using the one-parameter Raasch model. In this model all items are believed to be equal in discrimination, a tenuous assumption at best. As such the MH procedure is only sensitive to uniform bias. An item displays uniform bias if its ICCs for the different groups differ by only a horizontal translation (i.e., they are parallel but not coincident.) It is important to note that if the response process is modeled using the 2PL or 3PL IRT models, the ICCs may be non-parallel, causing non-uniform bias. By including the suspect item in the matching criterion it can be shown (Holland & Thayer, 1988) that, when all of the items exhibit no bias except the suspect item, the procedure partials out the effect due to impact in the case of the Raasch model. Thus, the ΔMH, the theoretical index measuring the amount of bias of an item for two groups from the Mantel-Haenzsel perspective, is given by

$$\Delta MH = -2.35(b_f - b_r) \tag{8}$$

where b_f and b_r are the Rasch difficulty parameters for the marginal ICCs of the studied item given in (6) above for the focal and reference groups respectively. The ΔMH index represents the difference in the mean **horizontal** distance between the marginal ICCs and has a common log odds ratio. (Note that when ΔMH < 0 the studied item is biased against the focal group.)

The horizontal direction is considered because the MH statistic is examining differences in the odds ratio at each score level for the two groups of interest. Some studies (Shealy, 1989; Shealy & Stout, 1991) have reported that the MH chi square procedure is reasonably robust against inflated Type I error when impact is present as well as robust against loss of power when uniform bias is present, (even if the generating model is a 2PL or a 3PL IRT model).

Shealy and Stout have a similar theoretical item (and test) bias index called b_{uni} (Shealy & Stout, 1991) which, in the IRT context, is the vertical distance between the marginal ICCs of the studied item (with respect to θ, the valid subtest ability). This index has a simple empirical interpretation. It is the average difference in probability of correct response experienced by the two groups for the studied item with impact partialled out. In this sense b_{uni} is similar conceptually to the Standardization index (Dorans & Kulick, 1986). Computationally b_{uni} is expressed by

$$b_{uni} = \int_{-\infty}^{+\infty} [T_R(\theta) - T_F(\theta)] f_F(\theta) d\theta \tag{9}$$

where $T_R(\theta)$ and $T_F(\theta)$ are the marginal ICCs of the suspect item for the reference and focal groups respectively given by (7) and $f_F(\theta)$ is the θ-marginal density of the focal group. Shealy and Stout also have another index, b_{gen}, which is identical to (9) except that the absolute value of the difference between the two marginal ICCs is computed. This index is designed for cases in which nonuniform bias would occur.

One way to express the potential for bias from the Shealy-Stout perspective is by examining the difference between the expected values of the reference and focal group $\eta \mid \theta$ conditional distributions. If this difference is zero, there is no potential for bias. But by examining the expression of this difference it becomes quite clear what differences in the underlying ability distributions will produce bias. That is, the difference between the expected value of the conditional distributions for a given value of y can be expressed as

$$E[\eta_R|\theta] - E[\eta_F|\theta] = (\mu_{\eta R} - \mu_{\eta F}) + (\rho_R \frac{\sigma_{\eta R}}{\sigma_{\theta R}})(\theta - \mu_{\theta R}) - (\rho_F \frac{\sigma_{\eta F}}{\sigma_{\theta F}})(\theta - \mu_{\theta F}) \tag{10}$$

By examining ways in which this difference is not zero, one can obtain insight into what could cause bias. The difference may be nonzero (or the potential for bias may occur) if

1) the θ means are not equai;

2) The η means are different; this is one obvious source of potential bias - differences in the nuisance ability

3) The ratio $\frac{\sigma_{\eta}}{\sigma_{\theta}}$ is not the same for both groups; this is a second source of potential bias and closely relates to the above discussion of how the shape of the underlying multidimensional ability distribution affects the size of the unidimensional parameter estimates.

4) A fourth source of potential bias exists if the correlations between the valid and nuisance dimensions are not the same for both groups. Obviously there are also many possible combinations that could produce potential for bias or a combination of bias and impact. The point is that any choice of the 10 parameters determining the focal and reference ability distributions that cause $E[\eta_R|\theta] - E[\eta_F|\theta] \neq 0$ constitutes potential for bias at θ. Similarly, any choice that produces $E[\eta_R|\theta] - E[\eta_F|\theta] = 0$ rules out bias at θ.

The Shealy and Stout test statistic and the corresponding estimator $\hat{b}_{uni}$ of the amount of bias, are comparably new and do offer the researcher several advantages over the MH test statistic and corresponding estimator $\hat{\Delta}MH$. They were developed from a multidimensional modeling perspective and emphasize the examination of bias at the test level rather than the item level. They can be used to look at several items simultaneously whereas the MH approach examines each item individually. The Shealy and Stout (SIB) procedure also offers the flexibility of letting the practitioner decide which items compose the "valid" subtest. Thus, by forcing the practitioner to identify the valid items, there is less risk that the reference composite will lie outside the validity sectors than in the MH approach in which users typically condition on all of the test items. A third advantage is that the SIB approach encourages researchers to examine the item bias cancellation effect described by Rosnowski (1987), which can occur if there are two or more nuisance dimensions.

To help the practitioner better understand what item bias is, what impact is, and how he/she can detect bias, several examples will be illustrated. Although they are contrived for illustration purposes they are still thought to be realistic. It is believed that underlying group ability differences can result quite easily due to curricular or instructional differences. In all of the subsequent examples bias will be demonstrated from a two-dimensional perspective, in which the horizontal axis (θ) is the assumed to be the valid test direction and the vertical axis (η) represents the nuisance dimension. (In reality there are likely to be several nuisance dimensions.) In each example the MH and SIB theoretical indices measuring the amount of bias will be computed.

Assume further that it is our task to examine the same two items, (Item 1 and Item 2) from a given test to see if either item is biased against either of two groups, a Reference group and a Focal group. The M2PL parameters for Item 1 are $a_1 = 1.5$, $a_2 = 0.0$, and $d = -1.5$. Thus, this item is measuring only the valid dimension and hence sensitive to impact but not bias. Item 2's parameters are $a_1 = 1.06$, $a_2 = 1.06$, and $d = -1.5$. Item 2 is measuring both dimensions equally well and thus is sensitive to bias. It should also be noted that MDISC and d values are the same for both items.

Obviously there are many ways in which the two groups of interest can differ in underlying ability distributions but for didactic purposes only a few are illustrated here. The examples chosen follow directly from the discussion of Equation 10; in each case Eq. 10 will be evaluated. (Furthermore, for the sake of simplicity, calculations will be worked with as though the reference composite was the same for each group. Although in reality if the underlying ability distributions are different the references composites for two groups will likely also be different.)

CASE 1 - Equal θ, η distributions: no bias, no impact

Assume that the two-dimensional ability distribution for Reference group and the Focal group can be described in the following manner:

Group	Ref	Foc
Mean vector $(\bar{\theta},\bar{\eta})$	(0.0,0.0)	(0.0,0.0)
Variance-covariance	$\begin{vmatrix} 1.0 & .5 \\ .5 & 1.0 \end{vmatrix}$	$\begin{vmatrix} 1.0 & .5 \\ .5 & 1.0 \end{vmatrix}$

Any differences in the conditional distribution of $\eta \mid \theta$ represents a potential for item bias that will be realized only if an item is measuring the η dimension to any degree. However, in this case, because the underlying distributions are identical the θ- and η-marginal ICCs are coincidental for each item. Also, the conditional distributions are equal. Thus, it is easily seen from (10) that $E[\eta_R|\theta] - E[\eta_F|\theta] - 0$ at every θ (i.e., no potential for bias exists.) The theoretical values of the 2PL IRT model difficulty and discrimination parameters were approximated from numerically derived the ICCs (cf. Wang, 1986). For each group b = .99 and a = 1.43 for item 1 and b = .93 and a = 1.11 for item 2. These ICCs are displayed in Figure 4 along with the η marginal distribution for each group. Because the underlying distributions are identical both groups perform identically and the ΔMH and b_{uni} bias indices both equal zero. Likewise the η marginal distributions are coincident.

Insert Figure 4 about here

CASE 2 - Unequal θ means: uniform bias

Assume that the two-dimensional ability distributions are the same as in Case 1 except for the mean vectors. Let the Reference group $(\bar{\theta},\bar{\eta})$ vector be (1.0,0.0) and for Focal group, (-1.0,0.0). Contour plots of the two-dimensional ability distributions and their marginals are illustrated in Figure 4. Values along the contours represent the densities for each group multiplied by 100.

Insert Figure 5 about here

From (18) it can be seen that from (10) that $E[\eta_R|\theta] - E[\eta_F|\theta] - -2$ for every θ.

Despite the fact that the η marginals are identical, as seen in Figure 5, the potential for bias exists and in fact is against the reference group. Comparing the θ-marginal ICCs (Figure 6) reveals a noticeable difference in probability of a correct response for both groups for the second item.

Insert Figure 6 about here

The theoretical marginal ICCs for item 1 are identical, with 2PL parameters of a = 1.43 and b = .99. Somewhat bewildering but consistent with (10) is that item 2 actually favors the Focal group! Why this occurs can also seen in Figure 5 by examining the expected value of an η "conditional slice" at θ = 0.0. Because of the positive correlation between θ and η the Focal group will have the higher expected value. For item 2 the Reference group's parameters are a = 1.10 and b = 1.25 and for the Focal group, a = 1.10 and b = .60. This case exemplifies uniform bias. The theoretical b_{uni} and ΔMH indices for item 2 are -.06 and +1.53, respectively. (It should be noted that the b_{uni} value is much smaller because it is weighted by the density of the Focal group which does not "overlap" the ICC appreciably.)

CASE 3 - Unequal η-means: uniform bias

In this case the variance-covariance structure is the same as in the previous two cases. The difference between the focal and reference groups in their mean ability vectors. The $(\bar{\theta},\bar{\eta})$ vector for the reference group is (0.0,1.0) and for the focal group, (0.0,-1.0). Thus (10) yields $E[\eta_R|\theta] - E[\eta_F|\theta] - 2$ for every θ. This is shown in Figure 7. Because no differences exist between the θ-marginal distributions there can be no impact.

Insert Figure 7 about here

The θ marginal ICCs for the two groups for item 2 are shown in Figure 8. These ICCs represent still another example of uniform bias. This will occur if the only difference between the underlying ability distributions is between the η means.

Insert Figure 8 about here

The amount of bias is a function of the degree to which an item measures the η-dimension, as well as the amount of potential for bias as expressed in (18). That is, with all other relevant factors kept constant, if an item had an α-angle (Eq. 2) greater than Item 2's 45° the θ-marginal ICCs would be even further apart than illustrated in Figure 8.

The theoretical unidimensional item parameters are a= 1.43 and b = .99 for item 1 for both groups. The a for item 2 is the same for each group, 1.11, but the b's are different: .28 for the Reference group and 1.57 for the Focal group. The Δ MH value and the b_{uni} values are -3.00 and .29, respectively.

Case 4:Unequal η variances-nonuniform bias

In this case the reference group and the focal group have identical mean vectors, (0.0,0.0), identical Θ variances, 1.0, and identical θ,η correlations, .5. The only difference between the two underlying distributions is that the η variance for the Reference group is .5 and for the Focal group 2.5. These distributions and their marginals are illustrated in Figure 9.

By (10) the potential for bias is given by $E[\eta_R|\theta] - E[\eta_F|\theta] - (\frac{1-\sqrt{5}}{\sqrt{2}})\theta$. Interestingly, if θ is negative the item will favor the Focal group; if positive the item will favor the Reference group. This result is an artifact of nonuniform bias.

Insert Figure 9 about here

The theoretical unidimensionai parameters are the same for each group for item 1: a = 1.43 and b=.99. However, nonuniform bias is illustrated in the theoretical item parameters for item 2. For this item, a = 1.14 and b = 1.02 for the Reference group and a = 1.02 and b = .77 for the Focal group. The greater the difference between the η variances the greater the differenc̃e in discrimination parameters. The θmarginal ICCs for items 1 and 2 are shown in Figure 10. Although not appropriate for this situation the ΔMH value is -.60. The b_{gen} value for this item is .19.

Insert Figure 10 about here

Case 5 - Unequal $\rho_{\theta,\eta}$'s : nonuniform bias

In the last case the underlying ability distributions have the same $(\bar{\theta},\bar{\eta})$ vectors, (0.0,0.0) and have unit variance for both the η and Θ dimensions. The only difference between the two groups is that the correlation between the valid and nuisance abilities is .8 in the Reference group and .2 in the Focal group. These distributions and their marginals are shown in Figure 11. In this case the θ and η marginal distributions are identical for both groups. According to (10) $E[\eta_R|\theta] - E[\eta_F|\theta] - .6\theta$, again indicating nonuniform bias.

Insert Figure 11 about here

The θ-marginal ICCs are illustı ated in Figure 12. Differences only exist in the item which would be sensitive to correlational differences, item 2. The theoretical parameters for item 1 are, as in the first four cases, a = 1.43 and b = .99 for both groups. Because item 1 measures in the direction of the computed reference composite it can never be biased, despite differences in the ability distributions for the two groups. However, for item 2 a = 1.48 and b = .78 for the Reference group, and a = .84 and b = 1.14 for the Focal group. Because the Reference group has a higher correlation its underlying distribution (Figure 11) its levels of ability can be more clearly distinguished by item 2 than can the Focal group. As in the previous case the Δ MH index of .86 is misleading because the ICCs violate the assumption of equal a's. Unlike the previous case, however, the ICCs actually cross. The b_{gen} value is .16.

Insert Figure 12 about here

Empirical example

To further illustrate the two bias detection procedures from a more realistic perspective a Monte Carlo study was performed. In this study estimated M2PL item parameters were used as a simulation model to generate two data sets, each having a different underlying two-dimensional ability distribution. The estimated parameters were from a calibration of Form 26A of the ACT Assessment Programs Math Usage Test. These parameters were reported by Reckase (1985). A subset of 25 items was selected. For two of the items the reported difficulty parameter estimates were altered to make the items more easy. The vector plot of the 25 items is shown in Figure 2. This specific subset was selected to illustrate how a practitioner, after calibrating items using the M2PL model could select a valid sector. Items lying outside the sector are invalid and would be suspect of being biased. Seven items in this case fall outside the validity sector. To simulate bias, response vectors using two distinct distributions of ability were generated. Using this simulation model 1000 reference group examinee response vectors were generated using a (θ,η) mean of (1.0, 0.0). In this group the two dimensions were uncorrelated with the θ-variance being 1.5 and the η-variance being .5. The same number of subject responses were generated for a focal group which had a $(\bar{\theta},\bar{\eta})$ vector of (0.0, 1.0). The two ability dimensions were also uncorrelated for this group and the θ-variance was .5 and the η-

variance was 1.5. The response matrices were then analyzed using the $\hat{\Delta}$MH and $\hat{b}_{uni}$ bias estimation procedures.

Because of the differences between the underlying two-dimensional ability differences the reference composite was noticeably different for each group. For the Reference group the reference composite direction was 17.55^{o} and for the focal group it was 43.05^{o}. This huge differences in reference composites is a strong signal that the unidimensional score scales (and also the number correct score scale) for each group are **not** representing the same skill in the two groups. Thus caution should be used in interpreting the results of the $\hat{\Delta}$MH estimator applied to the entire test, because although the statistic is computed by conditioning on number correct, the same number correct score does **not** mean the same thing in both groups. This problem would not have arisen if the data was truly unidimensional. Thus the $\hat{\Delta}$MH estimator was calculated twice, once conditioning on the entire test and once conditioning more appropriately only on the valid (and essentially unidimensional) test (i.e., the first 18 items, those lying in the validity sector.)

The SIB procedure circumvents the above problem by allowing the practitioner to select the valid items. In this example items which fell within a validity sector centered at 30^{o} from the θ_1 axis were selected as being the most valid items. Using these items as a basis of comparison establishes a reference composite of $11\ 92^{o}$ for the reference group and 11.03^{o} for the focal group, essentially the **same** angle for both groups. By selecting these items as the valid test the score scales thus now represent different levels of the **same** skill for both groups.

For each item the angles of measurement, α, the validity index (eq. 15), the unidimensional analog of discrimination, MDISC, and the $\hat{b}_{uni}$ estimator are presented in Table 1. The items were arranged in increasing order of α.

Insert Table 1 about here

The $\hat{b}_{uni}$ clearly indicate the bias of items 20 through 25, but fail to suggest that the invalid item 19 is biased. However, item 19 was one of the items for which the difficulty parameter was changed to make it considerably more easy (see Figure 2). Thus the inability to identify it as biased might be expected.

Two sets of Mantel-Haenzsel results for the same 25 items are reported in Table 2. The first column of $\hat{\Delta}$ MH values is when the conditioning score was only the 18 valid items plus the suspect item. Six MH analyses were computed, one for each of the nonvalid items 19 - 25. Each analysis yields a $\hat{\Delta}MH$ for the first 18 items and a $\hat{\Delta}MH$ for the particular suspect item of the run. The results of these runs were averaged for $\hat{\Delta}MH$ for items 1 - 18. The $\hat{\Delta}MH$ indicate that the last five items are clearly biased. The high index for item 2 cannot be explained, but recall this item was also given a high $\hat{b}_{uni}$ index as well. As with $\hat{b}_{uni}$ item 19 fails to be flagged as biased. It is interesting to note that the two procedures are in almost total agreement.

Insert Table 2 about here

To illustrate how the MH procedure can be misused the $\hat{\Delta}$ MH value was computed for all 25 items. As was mentioned above, this would be inappropriate because the reference composite for all 25 items is **not** the same for each group. The result, somewhat analogous to comparing "apples to oranges" would suggest that items 1 - 8, 11 and 14 are biased items as well as the final five items. In this case, items which are clearly valid, are indicated to be biased.

CONCLUSION

The main purpose of this paper is to provide the testing practitioner with insight about what causes items to be biased. Obviously the issue of bias tied closely to the concept of construct validity. Using a multidimensional IRT perspective practitioners can easily identify the composite of skills items are measuring and thereby establish a test validity sector. The task of identifying invalid items becomes quite simple. Once identified these invalid items should be checked for pɔssible bias using either the MH or SIB approach or possibly some other approach..

A secondary purpose of this paper was to encourage both the test constructor and the test user to make a conscientious effort to clearly identify what skills are being measured by a particular test. Items which measure the target cı purported skills need to be identified and those which do not need to be "weeded" out. Without doing so could render a bias analysis uninterpretable as was illustrated in Table 2.

This paper takes the view that empirically two or more items will always produce multidimensionality, and as such their parameters need to be estimated ɹsing multidi:mensional models. Such modeling requires extremely large sample sizes which may limit the number of testing situations that can use such an approach. Future research needs to examine what multidimensional analyses can be conducted with smaller sample sizes. Additionally researchers need to provide practitioners with clear guidelines about what approaches to use and in what situations to use them, as well as what approaches need to be avoided. It siould be apparent that, as much thought must go into the analysis of bias for a test as went into the original construction of the test.

Table 1

Summary Shealy-Stout indices for the simulated 25-item test.

Item	α	CVI	MDISC	$\hat{b}_{uni}$
1	0	.97	.87	.05
2	0	.97	1.92	.10
3	0	.97	2.00	.07
4	1	.98	1.22	.03
5	2	.98	1.41	.07
6	6	.99	1.21	.03
7	6	.99	1.73	.03
8	7	.99	1.23	-.01
9	12	.99	.88	-.01
10	13	.99	.98	-.06
11	13	.99	1.61	.02
12	14	.99	1.37	-.03
13	15	.99	.72	-.03
14	17	.99	1.60	-.04
15	23	.96	.59	-.09
16	23	.96	1.15	-.05
17	25	.95	2.00	-.02
18	25	.95	.77	-.03
19	48	.64	.58	-.03
20	66	.42	1.20	-.25
21	67	.41	1.31	-.28
22	75	.29	1.13	-.28
23	78	.25	1.16	-.29
24	78	.25	1.81	-.34
25	87	.06	1.94	-.37

References

Ackerman, T. A. (1988) An explanation of d'fferential item functioning from a multidimensional perspective. A paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA.

Davey, T.D., Ackerman, T. A., & Reckase, M.D. (1989) The interpretation of score differences when item difficulty and dimensionality are confounded. Faper presented at the Annual meeting of the Psychometric Society, Los Angeles.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: principles and applications. Boston: Kluwer-Nijhoff Publishing.

Holland, P.W. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H.I. Braun (Eds.), Test Validity (pp. 129-145). Hillsdale. NJ: Lawrence Erlbaum Associates.

Mislevy, R.J. & Bock, R.D.(1983). BILOG: Item analysis and test scoring with binary logistic models [Computer Program]. Mooresville, IN: Scientific Software.

Pine, S.M. (1977). Applications of item response theory to the problem of test bias. In D.J. Weiss (ED.), Applications of computerized adaptive testing (Research Report 77-1). Minneapolis: University of Minnesota, Psychometric Methods Program, Department of Psychology.

Reckase, M.D. (1985, April). The difficulty of test items that measure more than one ability. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: implications for measurement bias. Journal of Applied Psychology, 72 480-483.

Shealy, R. & Stout, W. (1989, April). A procedure to detect test bias present simultaneously in several items. Paper prsented at the annual meeting of the American Educational Research Association: San Franciscc.

Shealy, R. & Stout, W. (1991). An item response theory model for test bias. ONR Technical Report; To appear in Differential Item Functioning, Theory and Practice L. Erlbaum (in press)

Traub, R.E. (1983). A priori considerations in choosing an item response model. In R.K. Hambleton (Ed.), Applications of item response theory. Vancover: Educational Research Institute of British Columbia, 57-70.

Wang, M. (1986, April). Fitting a unidimensional model to multidimensional item response data. Paper presented at the ONR Contractors Conference. Gatlinburg, TN.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Figure Captions

Figure 1. An equiprobability contour plot of the response surface for a two-dimensional item.

Figure 2. Validity sectors for a 25-item test.

Figure 3. Two distinct two-dimensional underlying ability distributions and their reference composites for the two indicated items.

Figure 4. Coincident marginal ICCs for items 1 and 2 for Case 1 plotted along with the coincident θ-marginal distribution for each group.

Figure 5. Contour plot of the underlying two-dimensional ability distributions for the Reference and Focal Groups for Case 2.

Figure 6. Marginal ICCs for item 1 and item 2 and the θ-marginal distributions for the Reference and Focal Groups for Case 2.

Figure 7. Contour plot of the underlying two-dimensional ability distributions for the Reference and Focal Groups for Case 3.

Figure 8. Marginal ICCs for item 1 and item 2 and the θ-marginal distributions for the Reference and Focal Groups for Case 3.

Figure 9. Contour plot of the underlying two-dimensional ability distributions for the Reference and Focal Groups for Case 4.

Figure 10. Marginal ICCs for item 1 and item 2 and the θ-marginal distributions for the Reference and Focal Groups for Case 4.

Figure 11. Contour plot of the underlying two-dimensional ability distributions for the Reference and Focal Groups for Case 5.

Figure 12. Marginal ICCs for item 1 and item 2 and the θ-marginal distributions for the Reference and Focal Groups for Case 5.

Figure 1

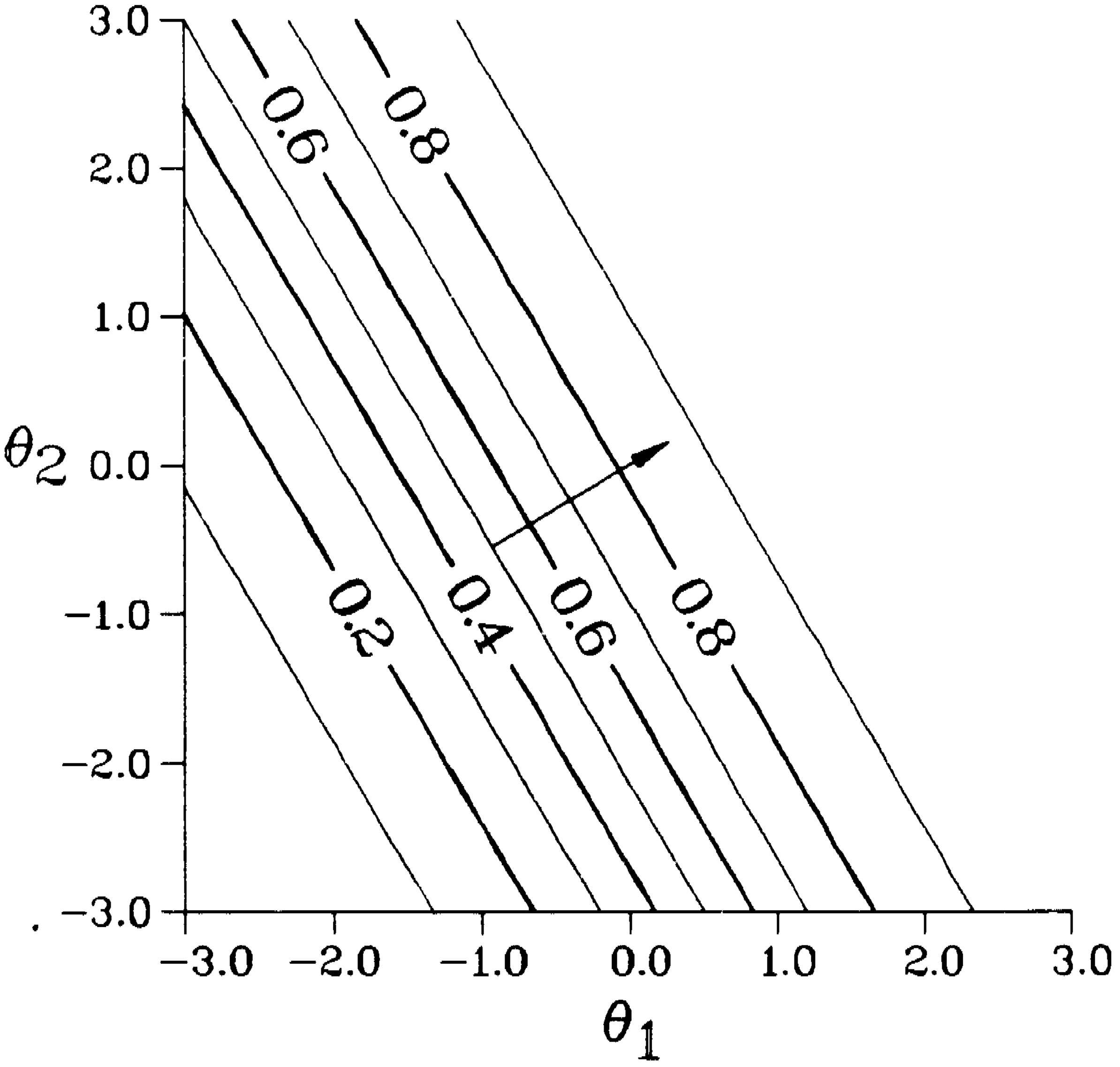

3.0
2.0
1.0
θ_2
0.0
−1.0
−2.0
−3.0
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ_1
0.6
0.8
0.2
0.4
0.6
0.8

Figure 2

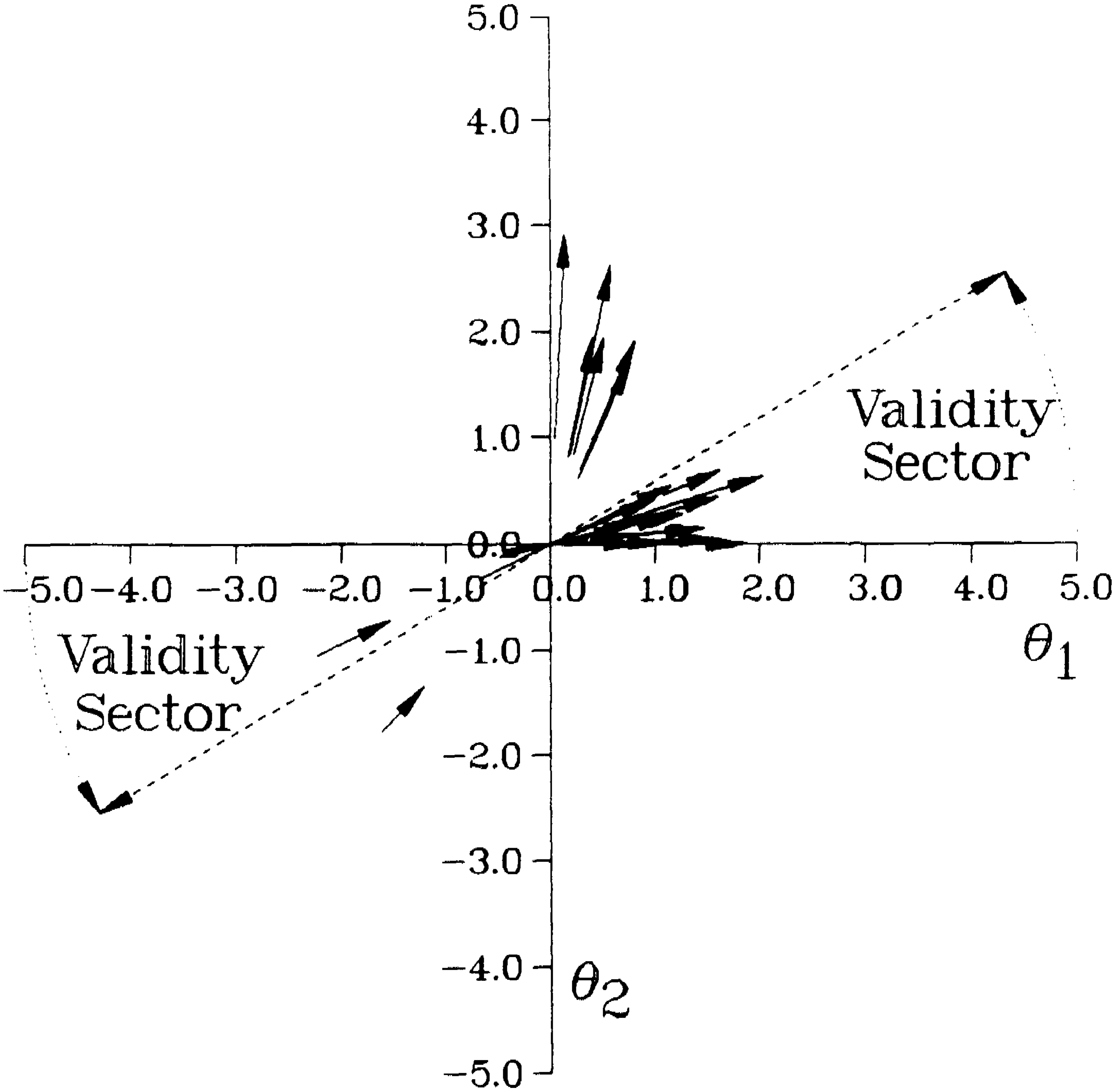

5.0
4.0
3.0
2.0
1.0
0.0
−1.0
−2.0
−3.0
−4.0
−5.0
−5.0
−4.0
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
4.0
5.0
Validity Sector
Validity Sector
θ_1
θ_2

Figure 3

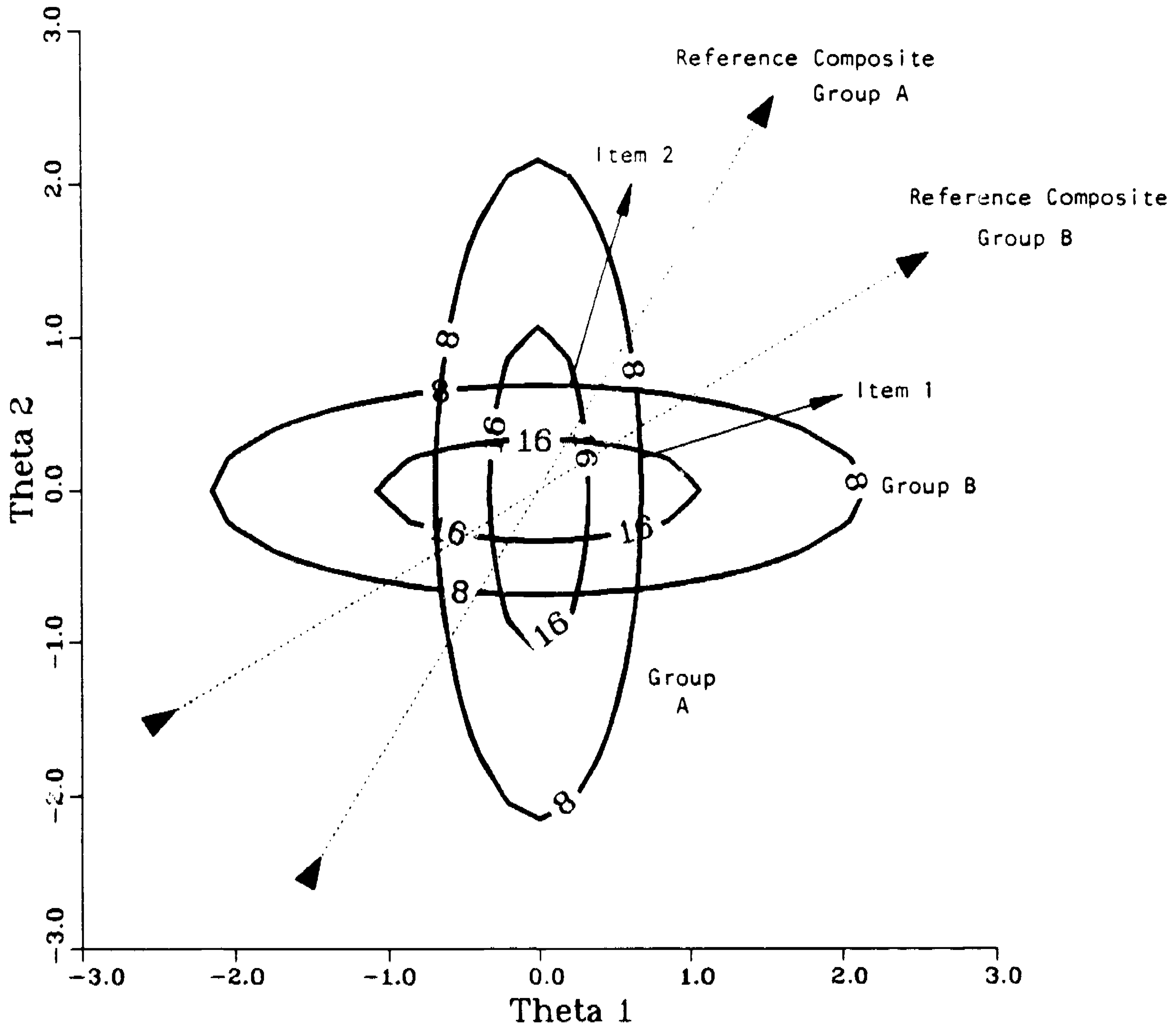

Reference Composite
Group A
Item 2
Reference Composite
Group B
Item 1
Group B
Group A
8
16
Theta 2
Theta 1
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0

Figure 4

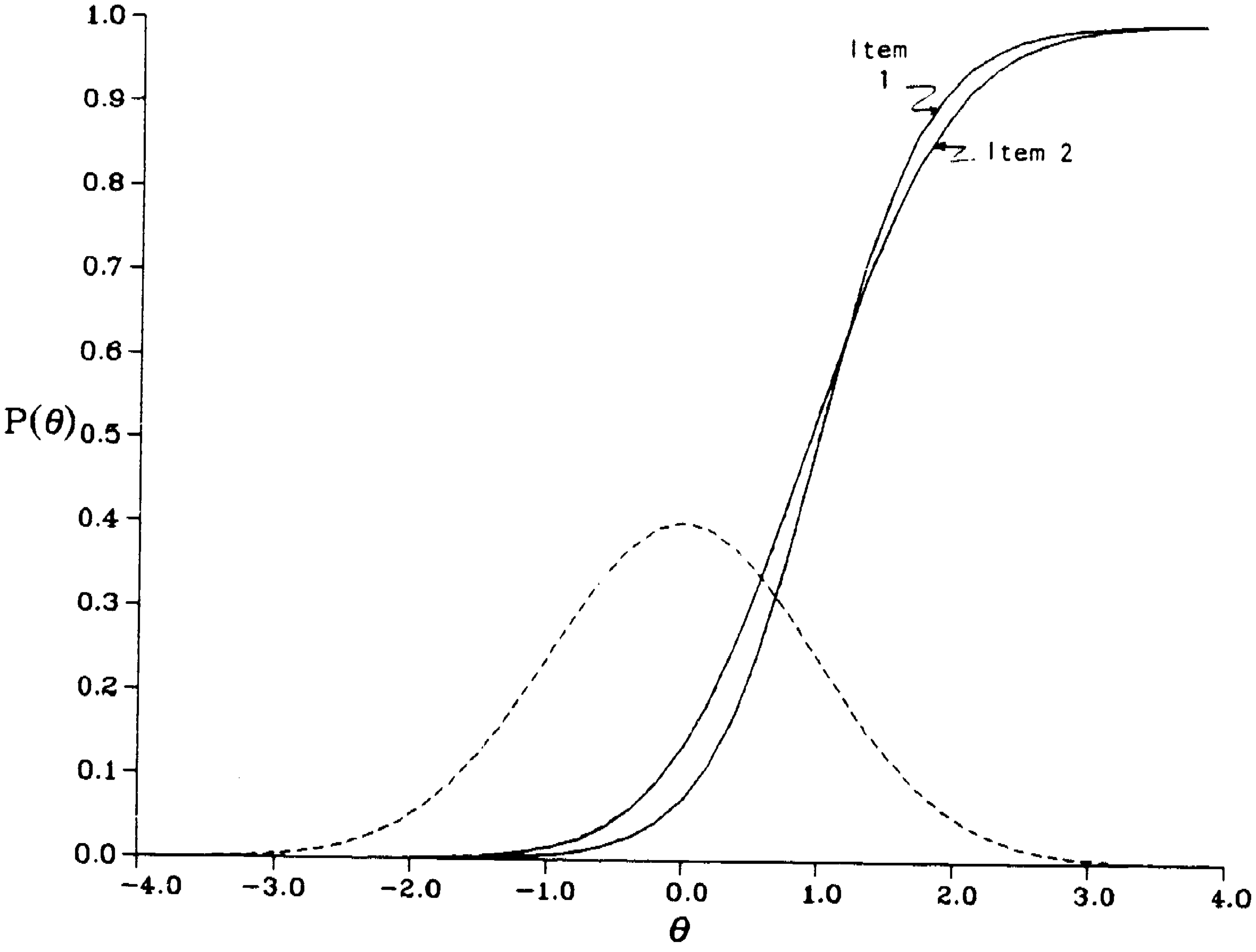
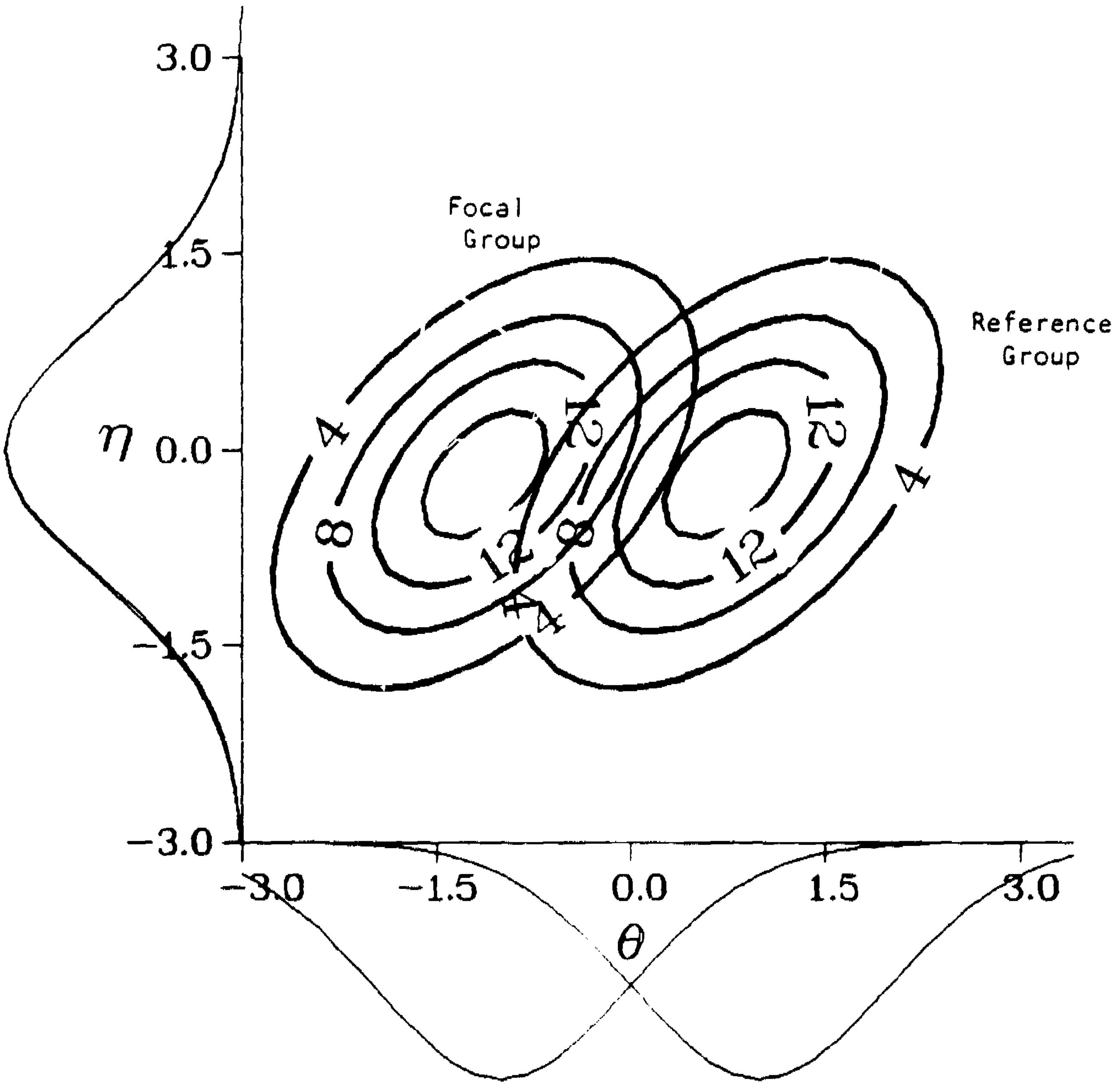

1.0
0.9
0.8
0.7
0.6
P(θ) 0.5
0.4
0.3
0.2
0.1
0.0
−4.0
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
4.0
θ
Item 1
Item 2

Figure 5



Focal Group
Reference Group
3.0
1.5
0.0
-1.5
-3.0
η
-3.0
-1.5
0.0
1.5
3.0
θ
4
8
12
12
8
4
12
12
4
8
12
12
4

Figure 6

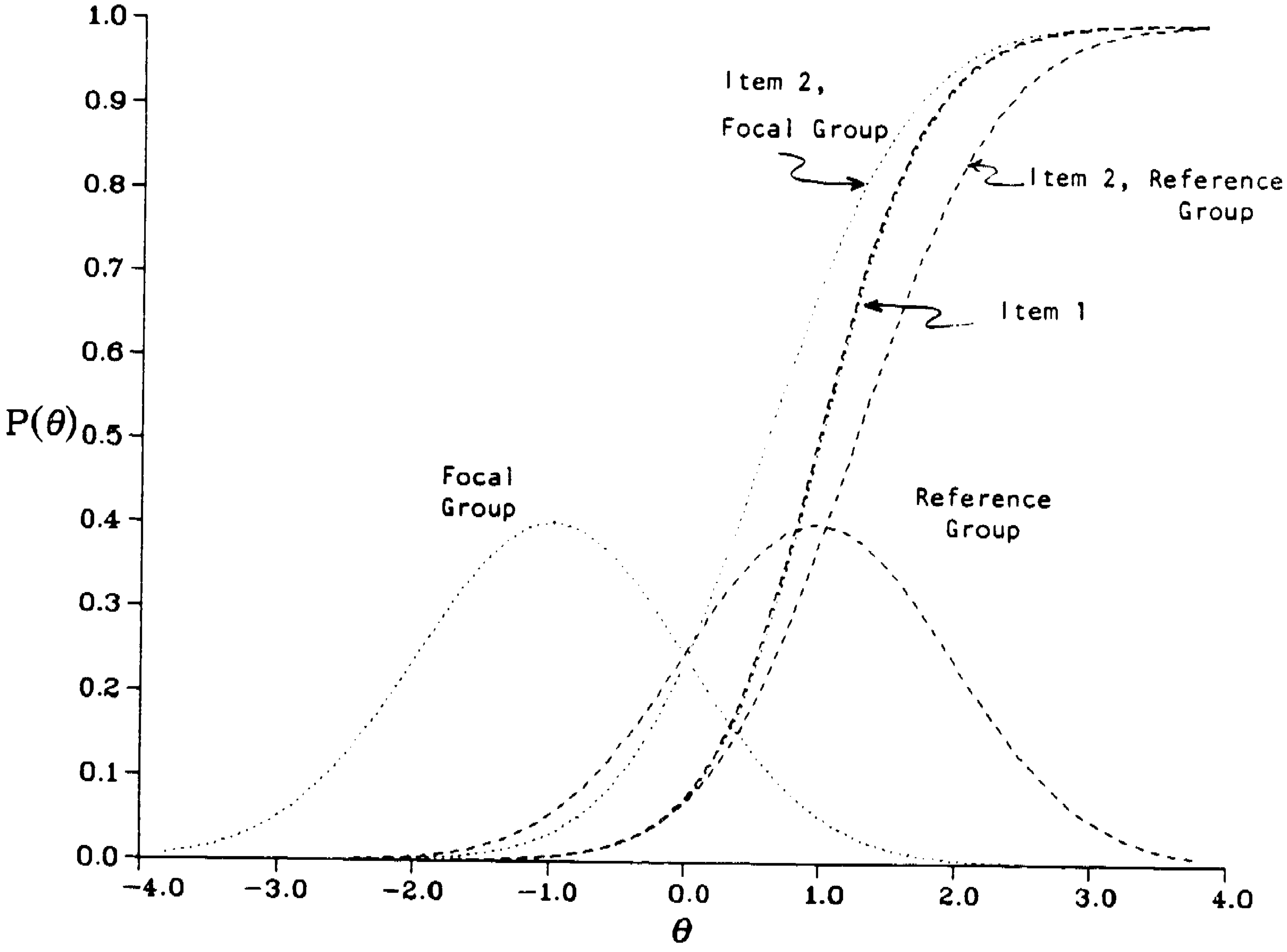

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
P(θ)
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
θ
Item 2,
Focal Group
Item 2, Reference
Group
Item 1
Focal
Group
Reference
Group

Figure 7

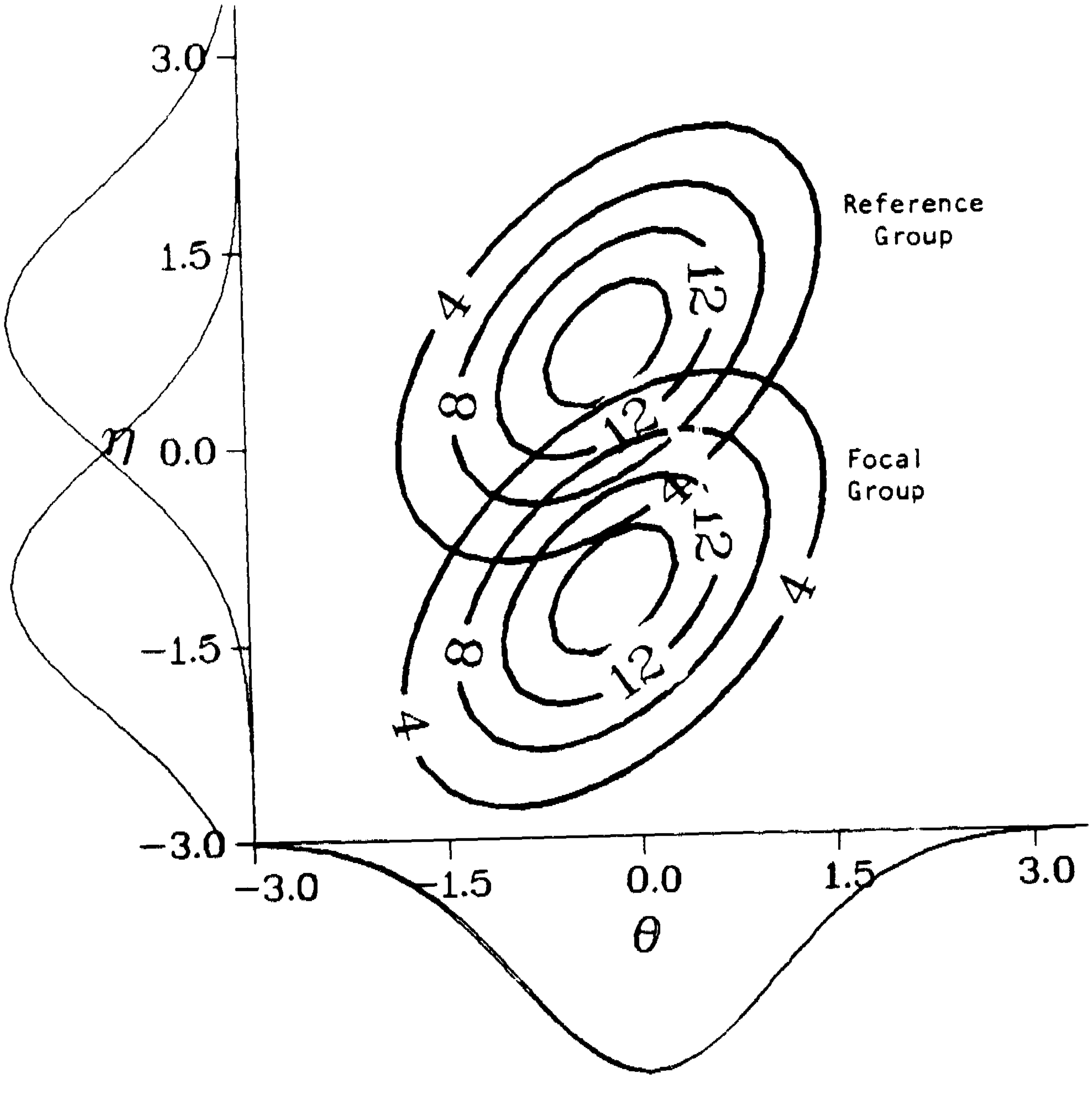

3.0
1.5
η 0.0
−1.5
−3.0
−3.0
−1.5
0.0
1.5
3.0
θ
4
8
12
12
4
12
4
8
12
4
Reference Group
Focal Group

Figure 8

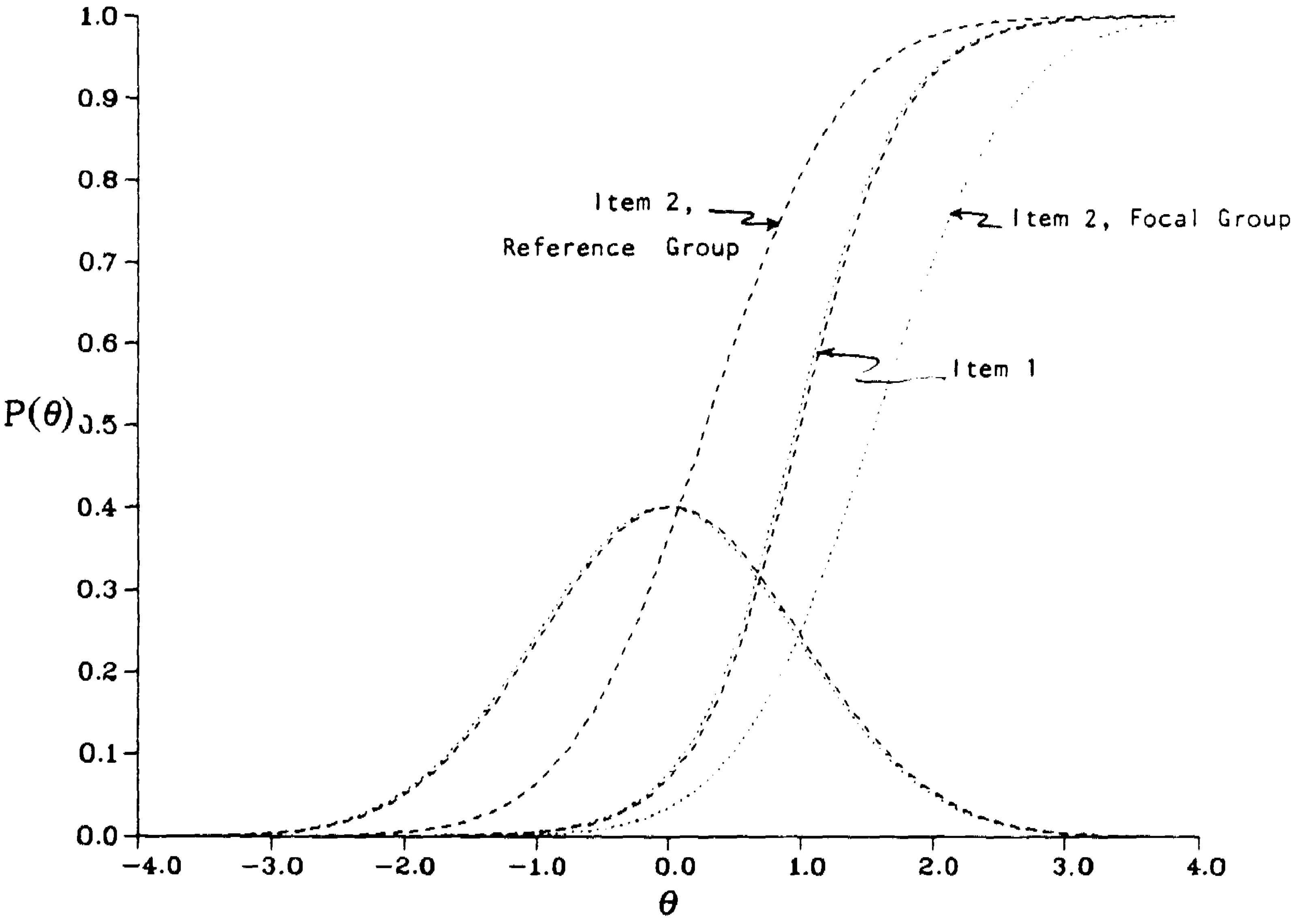

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
P(θ)
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
θ
Item 2, Reference Group
Item 2, Focal Group
Item 1

Figure 9

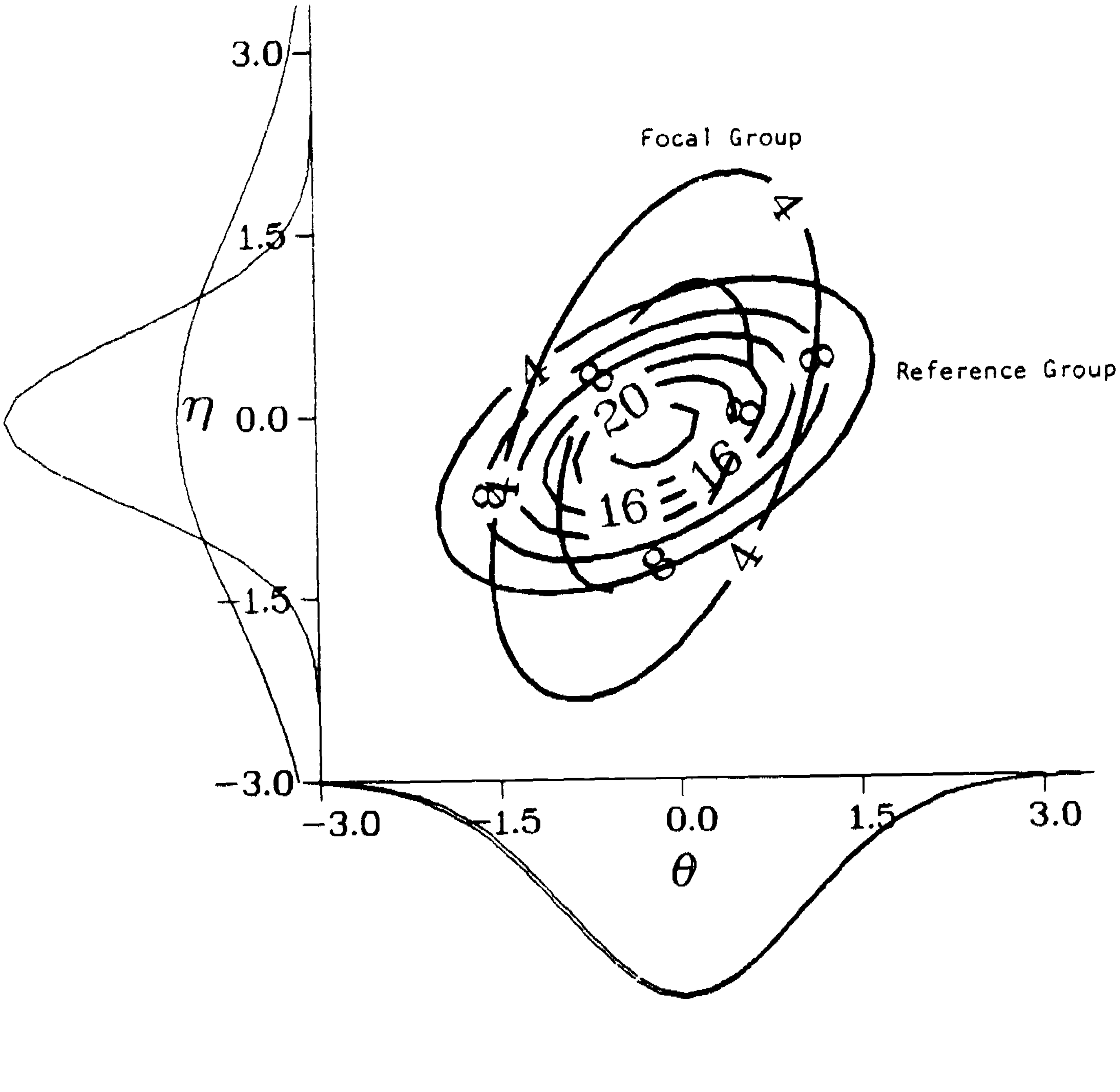

Focal Group
Reference Group
3.0
1.5
η 0.0
-1.5
-3.0
-3.0
-1.5
0.0
1.5
3.0
θ
4
8
16
20

Figure 10

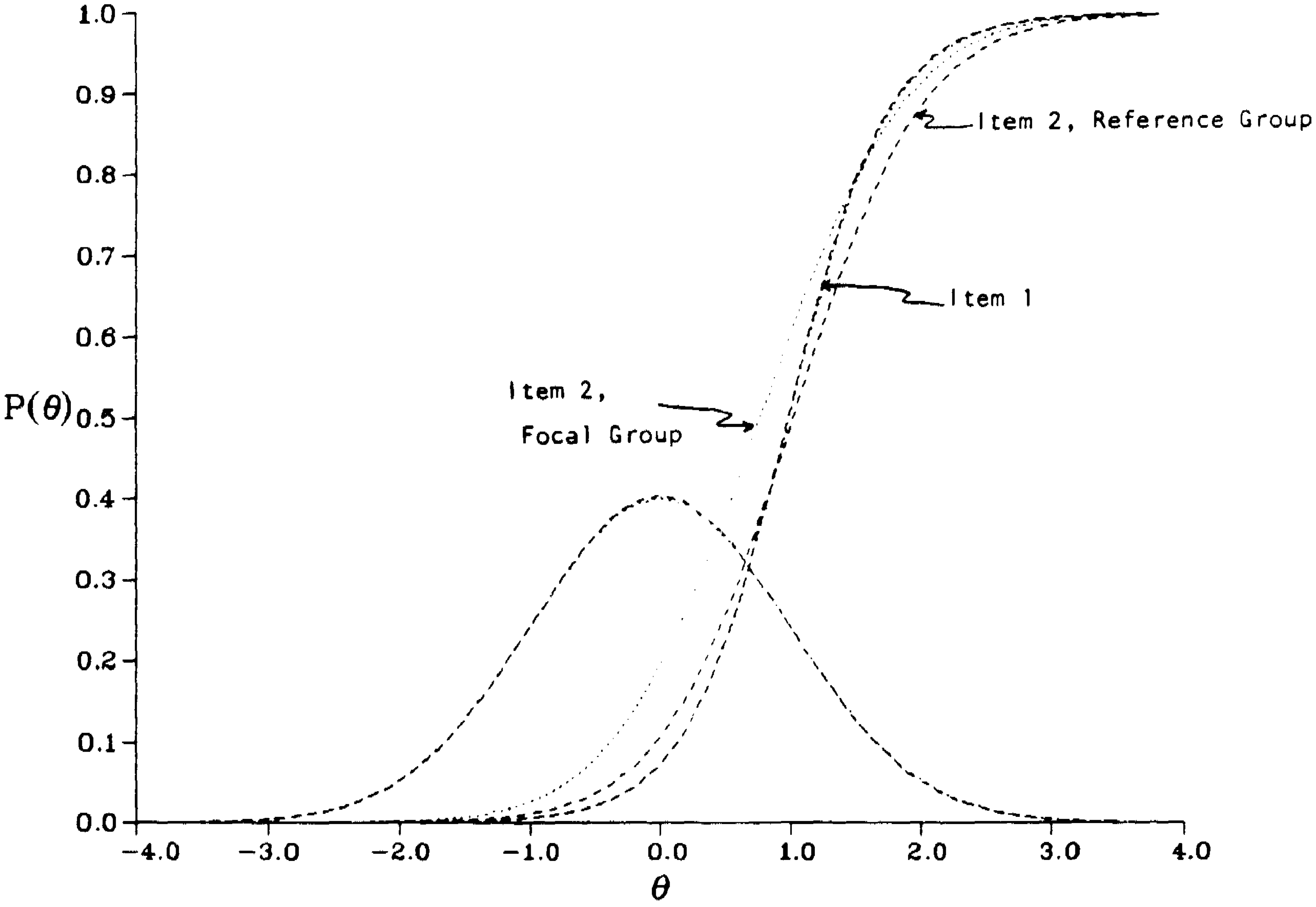
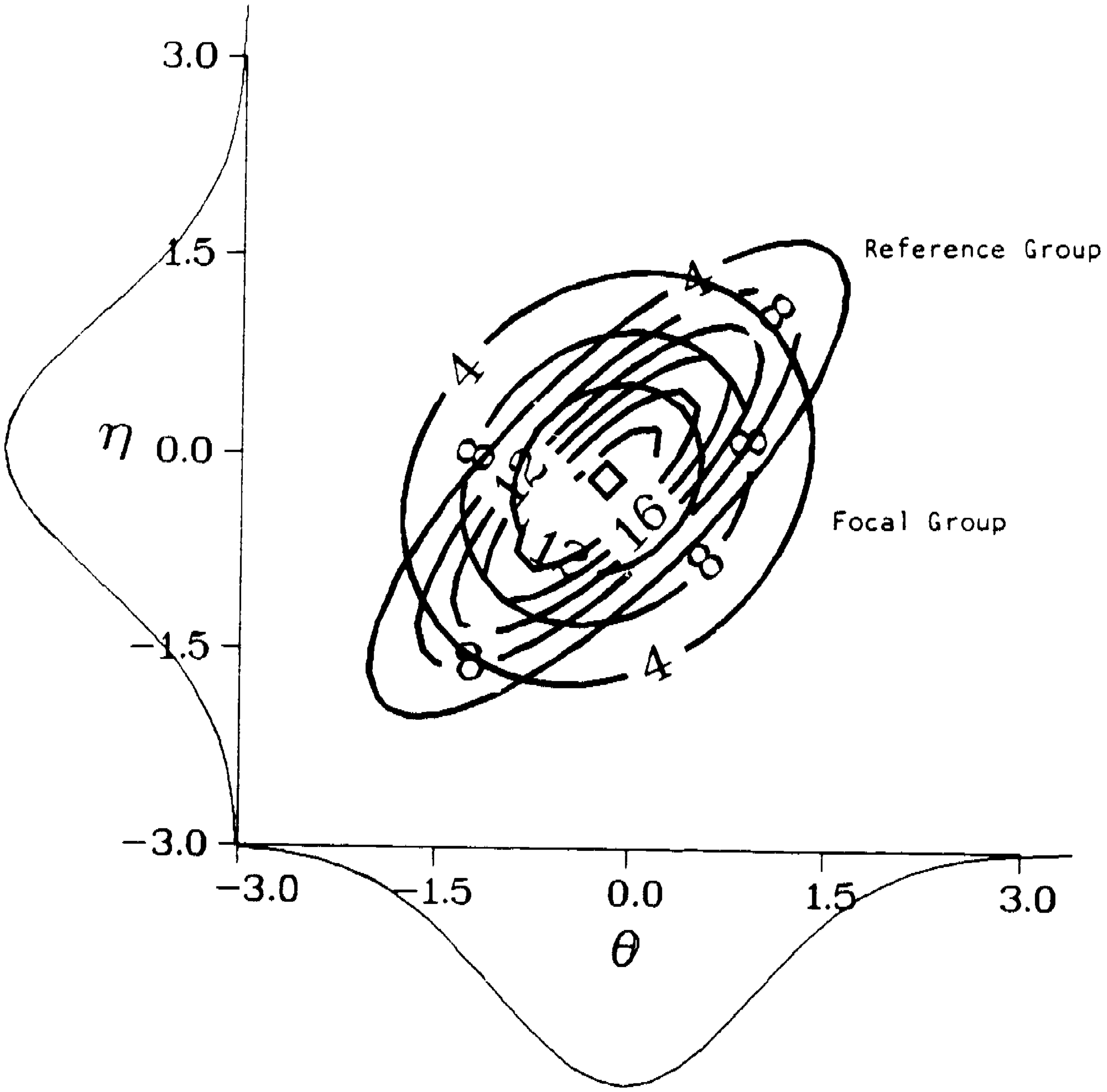
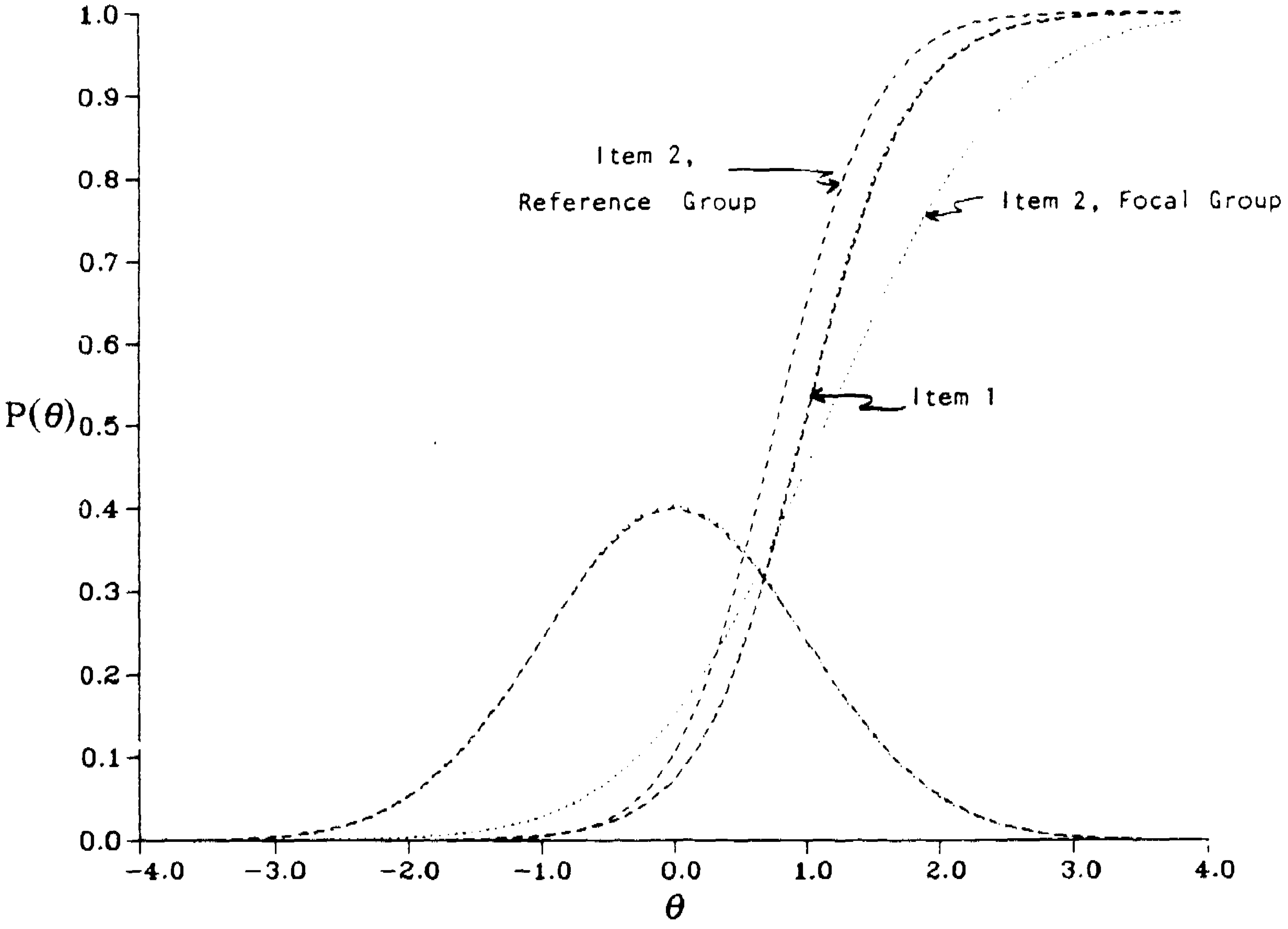

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
P(θ)
−4.0
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
4.0
θ
Item 2, Reference Group
Item 1
Item 2,
Focal Group

Figure 11



Reference Group
Focal Group
4
8
12
16
η
3.0
1.5
0.0
−1.5
−3.0
−3.0
−1.5
0.0
1.5
3.0
θ

Figure 12


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
P(θ)
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
θ
Item 2,
Reference Group
Item 2, Focal Group
Item 1