Open access • Journal Article • DOI:10.1177/014662168200600404

A nonparametric approach to the analysis of dichotomous item responses

Robert J. Mokken, Charles Lewis

Institutions: University of Groningen

Published on: 01 Sep 1982 - Applied Psychological Measurement (SAGE Publications Inc.)

Topics: Item response theory, Classical test theory, Differential item functioning, Equating and Mokken scale

A Nonparametric Approach to the Analysis of Dichotomous Item Responses

Robert J. Mokken
Centraal Bureau voor de Statistiek, The Netherlands

Charles Lewis
Rijksuniversiteit Groningen, The Netherlands

An item response theory is discussed which is based on purely ordinal assumptions about the probabilities that people respond positively to items. It is considered as a natural generalization of both Guttman scaling and classical test theory. A distinction is drawn between construction and evaluation of a test (or scale) on the one hand and the use of a test to measure and make decisions about persons' abilities on the other. Techniques to deal with each of these aspects are described and illustrated with examples.

Modern item response theory can be viewed in the historical perspective provided by preceding test theory, e.g., Guttman scaling theory and classical test theory. This paper presents a nonparametric elaboration of a type of model implied by most of the current parametric latent trait models. The analysis is restricted to responses to dichotomous items of the "pass-fail" type, in which one alternative is designated as "positive" with respect to the latent ability of interest. Only ordinal assumptions are used about the item response functions, however, without any further specifications to a particular parametric family of curves.

In many situations—frequent in areas such as attitude scaling, the analysis of voting in legislative bodies, and market research—either the items are difficult to obtain or the level of information concerning item quality is low. Researchers in such situations are therefore more comfortable discussing abilities and item difficulties at the ordinal level than in the interval or ratio terms provided by more pretentious (if not overpretentious) parametric approaches. More than a decade of experience with applications of the scale construction and evaluation procedures to be described here has suggested that this nonparametric approach can be quite useful and satisfactory in such research situations.

The main reference to this nonparametric approach to problems in this domain is Mokken (1971). A recent introduction is provided by Stokman and Van Schuur (1980), and relevant computer programs are described in the *STAP User's Manual* (Technisch Centrum FSW, 1980; see also Van de Wijgaart, 1981; Henning, 1976; and Lippert, Schneider, & Wakenhut, 1978).

The first section of this article presents basic assumptions and corresponding definitions of terminology; the second describes scaling procedures available for constructing and evaluating tests; the

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 4, Fall 1982, pp. 417–430

0416-6216/82/040417-14$1.70

final section describes procedures recently developed (Lewis, 1981) to aid in making inferences about the abilities of individuals.

Assumptions and Definitions

Both in *test construction* on a calibrating sample of people and in *test administration* (ability estimation) where single individuals are confronted with a test, the people can be considered to be selected from a certain population through a process itself governed by some probability distribution.

Monotone Homogeneity and Local Independence

A first property and corresponding assumption is dictated by the primary requirement of *unidimensional* measurement. The positive response probability π_{ia} of person a may be considered as a quantity which, in theory, can be observed or estimated for item i directly through repeated applications to a person a. Consequently, all persons (a) could be ordered according to the magnitude of π_{ia} for item i, if these probabilities were known. If the attribute the items constituting a test are measuring is one dimensional, a necessary requirement is that the items order *all* persons similarly.

In *test construction*, then, only items which order all persons similarly should be selected: The detection of such a set of items supports unidimensionality of the attribute and at the same time provides a set of items to measure (i.e., order) persons along the established dimension. However, in that case, it is not difficult to see that for each item—and uniformly for all items—the persons can be ordered along a continuum in such a way that the probabilities π_{ia} increase monotonely along that continuum. This is exactly the form of monotone increasing item response function which abounds in the parametric models of item response theory and provides an explicit argument for their importance as measuring devices.

As a consequence, the probabilities π_{ia} can be represented as functions $P_i(\theta)$ of the underlying (latent) attribute or ability θ. The property that for all items i the functions $P_i(\theta)$ are monotone increasing in θ can be designated as *monotone homogeneity* (Meredith, 1965, p. 430–432). This gives the monotone increasing item operating characteristic curves familiar in latent trait models. The scaling or test construction procedures to be described have been designed primarily to detect or construct one-dimensional scales or tests consisting of sets of monotonely homogeneous items. However, additional considerations may make it desirable to restrict test construction to a subclass of monotonely homogeneous items, characterized by a second type of monotony.

Prior to discussing this topic, however, another assumption, *local independence*, should be made explicit. This assumption states that for any person with a given value of θ the responses to a given set of items are statistically independent. This assumption is basic to most item response theories.

Doubly Monotone Sets of Items

The $P_i(\theta)$ may be considered as *local* difficulties, measuring the difficulty of item i for a person located at point θ along the ability continuum. In general, for sets of monotonely homogeneous items the tracelines can intersect. This implies that for persons located at different sides of such a point of intersection the *order* of these local difficulties will be reversed for the two items concerned.

Sometimes in the *administration* of tests, this may well be seen as undesirable, as it might be desirable for the difficulty order of the items in a test to be the same for all persons to which the test is administered. Other considerations in test administration, such as classifying persons in terms of

their performance on the test, may imply the requirement that for all persons the (local) difficulty order of the items be the same. This latter requirement is fulfilled whenever the item response functions of a set of test items do not intersect, implying a second monotony property for sets of monotonely homogeneous items.

The idea can be formalized by introducing a parameter to characterize the difficulty of an item and by using this to specify a second monotony property. Assume that for each item i of interest, there is a unique value of θ such that $P_i(\theta) = .5$. Define the *difficulty* b_i of item i as this value of θ, giving $P_i(b_i) = .5$.

A set of item response functions with the property of monotone homogeneity (increasing in θ) and the additional property of decreasing monotony in b_i is necessary for double monotony. More precisely, for a set of items, if

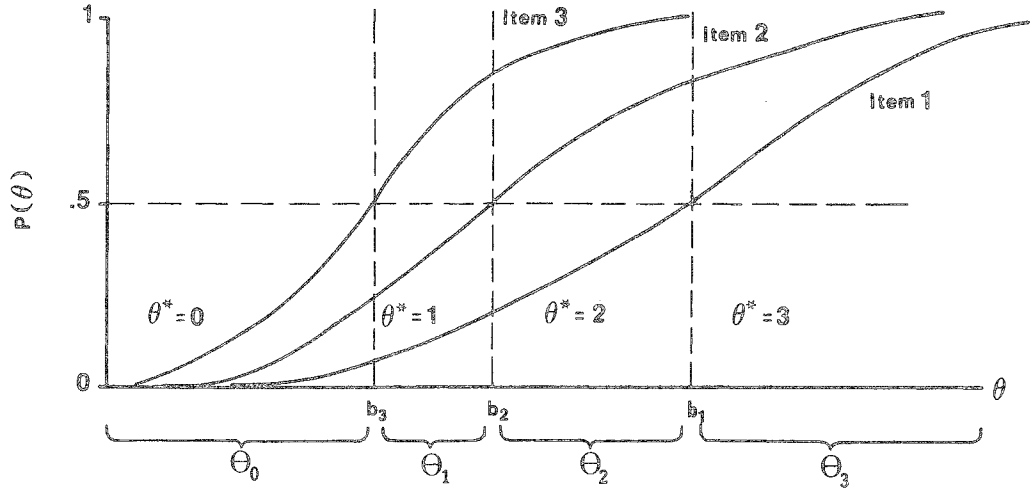
1. $P_i(\theta)$ is monotone increasing in θ and
2. $b_i < b_j$, then $P_i(\theta) > P_j(\theta)$ for all θ,

then that set of items shall be called *doubly monotone*.

Figure 1 illustrates double monotony for three items. Their item response functions do not intersect, so for each θ position along the axis the local difficulty ordering is the same. In Figure 1 a convention of item numbering is used which shall be maintained in formulas throughout this article: The items are numbered according to decreasing difficulty order, the most difficult item (largest value of b_i) numbered 1 and so forth. Consequently, if difficulty $b_i > b_j$, then item number $i < j$.

The discussion can be summarized as follows. If n items are selected for a test to measure (i.e., order) persons unidimensionally according to the attribute or ability θ, response functions are heeded which are monotone increasing in θ. This property of monotone homogeneity is *sufficient* for the purpose of *test construction*. For the purpose of *test administration* it may be *necessary* to require an additional selective restriction in the form of the requirement of decreasing monotony in b_i. There should then be restriction to sets of doubly monotone items, a subclass of that of monotonely homogeneous items.

Figure 1
Double Monotony: Classes and Class Scores



This simple model encompasses as special cases virtually all the parametric continuous or discrete latent trait or latent structure models which have been suggested in the literature, of which the normal ogive and the related logistic functions with three, two, or just one item parameter, as well as ordered latent class models, may be mentioned here. The use of monotone items is obvious in all these cases, but the additional aspects and properties of doubly monotone sets, as introduced by Mokken (1971), seem to have been unnoticed in the literature.

Note that double monotony (as defined here) is a property of a *set* of items. The reader may verify that a set of two- or three-parameter logistic items with varying discrimination (slope) parameters would not be doubly monotone, since the item response functions would intersect.

Some Properties and Scalability

With the simple item response model based on monotone homogeneity, occasionally strengthened to double monotony, it is possible to derive a number of simple nonparametric properties of the marginal response distribution for a population of persons. (They are nonparametric in the sense that they do not depend on the form of the item response functions or the form of the population distribution of θ.) Only a few of these properties are discussed here. The interested reader is referred to Mokken (1971, chap. 4) for proofs, additional results, and details. The results permit a definition of a scale and form the basis for the test construction and evaluation procedures to be described in the next section.

The *population difficulty* π_i of an item i is defined as the probability of a positive response to that item in the population of persons under study. It is obtained by integration of the item response function $P_i(\theta)$ over the population distribution of θ.[1] If double monotony is assumed for a set of items, then the ordering of the population difficulties is related to that of the latent difficulties, b_i, by

$$b_i > b_j \text{ if and only if } \pi_i < \pi_j \, . \qquad [1]$$

For items whose response functions intersect, the ordering of the π_i will depend on the population distribution of θ because the ordering of the $P_i(\theta)$ is different for different values of θ. Thus, within the family of monotonely homogeneous items, inferring the order of the latent difficulties from that of the population difficulties (or sample estimates of these) is only unambiguously justified for sets of doubly monotone items.

Turning to the bivariate response distribution for pairs of items, let $\pi_{ij}(1, 1)$ denote the probability of a positive response to both items i and j in the population of persons. For a person with a given θ, based on the assumption of local independence, this probability is the product $P_i(\theta)P_j(\theta)$. Integrating this product over the population distribution of θ gives $\pi_{ij}(1, 1)$.

If (0, 1) scoring is used for the two items and the scores are called u_i and u_j, then the population covariance between two items i and j which are monotonely homogeneous is equal to

$$\mathrm{Cov}(u_i, u_j) = \pi_{ij}(1,1) - \pi_i\pi_j \geq 0 \, . \qquad [2]$$

If it is further insisted that P_i and P_j are strictly increasing in θ, then the covariance between u_i and u_j will always be positive. For monotonely homogeneous items i and j the joint probability of positive response is therefore larger than expected under conditions of marginal independence in the population (namely, $\pi_i\pi_j$).

---

[1]This integration is of a more general type than the one usually used in calculus, to provide valid results for discrete as well as continuous θ.

The above results provide a framework in terms of which a coefficient of scalability may be selected. For a discussion and criticism of the coefficients used in Guttman scale analysis, the interested reader is referred to Mokken (1971, chap. 2.5). On the basis of this study, the coefficient of homogeneity H, originally suggested by Loevinger (1947, 1948), was seen to be particularly appropriate for use with the model presented in this article. Other authors, including Green (1954, p. 357), Torgerson (1958, p. 326) and, more recently, Van Naerssen (1972), have also discussed the comparative advantage of this coefficient over others. Of its general properties it may first be mentioned that H is based on the property of monotone homogeneity. Secondly, it can be derived as a variance ratio involving the simple test score (number correct) in a form familiar to users of Guttman scaling methods. Finally, it can be written in terms of item coefficients H_i evaluating the scalability of a particular item with respect to the other items as a scale.

The item coefficient of scalability of item i is defined by

$$H_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} [\pi_{ij}(1,1) - \pi_i\pi_j] \Big/ \sum_{\substack{j=1 \\ j\neq i}}^{n} \pi_{ij}^{(0)} \qquad [3]$$

where

$$\pi_{ij}^{(0)} = \begin{cases} \pi_i(1-\pi_j) & \text{if } i < j \ ; \\ (1-\pi_i)\pi_j & \text{if } i > j \ . \end{cases} \qquad [4]$$

From this definition, it can be shown that H_i can never exceed unity. The coefficient of scalability for a set of n monotonely homogeneous items can then be given by

$$H = \sum_{j\neq i}\sum [\pi_{ij}(1,1) - \pi_i\pi_j] \Big/ \sum_{j\neq i}\sum \pi_{ij}^{(0)}$$

$$= \sum_{j\neq i}\sum \pi_{ij}^{(0)} H_i \Big/ \sum_{j\neq i}\sum \pi_{ij}^{(0)} \qquad [5]$$

The coefficient of scalability H for a set of n items (i.e., a test or scale) therefore is a weighted sum of the item scalability coefficients H_i of its constituting items.[2]

From these expressions and from Equation 2 it follows that for monotonely homogeneous tests of n items, all H_i will be non-negative and, for all practical purposes, positive. Combining this with the general upper bound given above gives $0 < H_i \leqslant 1$ for sets of monotonely homogeneous items. Moreover, it is clear that if the values of the constituting item coefficients H_i are all greater than a given positive constant c, then the same is true for H:

$$\text{If for } i=1(1)n, c < H_i, \ (0 < c < 1),$$

$$\text{then } 0 < c < H \leq 1. \qquad [6]$$

H is always at least as large as the smallest item's coefficient of homogeneity in the test.

[2] A similar formula can be given in terms of H_{ij}, defined on the 2×2 table for any pair of items i and j. H_{ij} is just the value of H for a set of two items. The reader may verify easily that $H_{ij} = \text{phi}/\text{phi}_{max}$, a familiar coefficient in the analysis of 2×2 tables.

Sample estimates $\hat{H}$ and $\hat{H}_i$ can be obtained by inserting the sample relative frequencies corresponding to π_i and π_{ij} (1, 1) in Equations 3 and 5. Asymptotic sampling theory for these estimates is completely developed by Mokken (1971, chap. 4. 3), and includes the following results:

1. One-sided tests for a scale ($H = 0$ vs. $H > 0$) and for individual items ($H_i = 0$ vs. $H_i > 0$);
2. Confidence intervals for H (and H_i); and
3. Tests of equality of H (and H_i) for different populations.

It may therefore be concluded that H seems to satisfy fully four prerequisite criteria for a coefficient of scalability proposed originally by White and Saltz (1957, p. 82), together with a fifth one extending its usefulness:

1. Its theoretical maximum is 1 and hence invariant over scales.
2. Its theoretical minimum is 0, assuming monotone homogeneity and hence invariant over scales.
3. It is possible to evaluate scales as a whole with H and also to evaluate the scalability of individual items with the item coefficients H_i.
4. It is possible to test theoretically interesting hypotheses about H and H_i.
5. It is possible to construct approximate confidence intervals for H and H_i.

Approximate tests of the equality of the H values of a set of items for different populations can also be implemented. From these considerations H (and H_i) will be used as criterion of scalability for all procedures to be described.

A scale (or scalable set) can be defined in terms of H, which as such will be the sole criterion of scalability.[3] This makes it possible to discard the many additional and cumbersome criteria of scalability of the more orthodox methods of Guttman scaling.

A scale is defined in simple terms:

> A scale is a set of items which are all positively correlated and with the property that every item coefficient of scalability (H_i) is greater than or equal to a given positive constant c ($0 < c < 1$).

Equation 6 implies that the value of H, the coefficient testing the scalability of the set of items as a whole, will then also be greater than c, which can be designated as the scale-defining constant. Some degrees of scalability may be distinguished in terms of the overall coefficient H. Although an empirical basis for this distinction is still lacking, for practical use the following classification of scales was suggested:

$.50 \leqslant H$: a strong scale;
$.40 \leqslant H < .50$: a medium scale; and
$.30 \leqslant H < .40$: a weak scale.

The concept of a strong scale corresponds to the original strong requirements for a Guttman scale, values near unity indicating nearly perfect scales. Experience based on a variety of applications has shown that the usual practice, based on a lower bound $c = .30$, performs quite satisfactorily, delivering long and useful scales. Recent studies by Molenaar (1982a) based on the simulation of Rasch (1980) items for different population distributions may invite the investigation of if and to what extent other values (e.g., $c = .15$) can be admissible.

Description of Scaling Procedures

The procedures presuppose dichotomous items. In the case of multicategory items the researcher has to define the "correct" or "positive" alternative beforehand.

[3]The term "scale" (or, for that matter, "test") is used here because it was, and probably still is, a familiar concept in the practice of social research. Obviously, it has no immediate bearing on the basic concept of a scale in axiomatic theories of measurement, where it denotes the abstract triple of a numerical system, an empirical system, and a set of mapping rules.

The scalability of a set of items may be investigated from many angles and from different levels of analysis that may require different approaches and correspondingly different procedures. Most of these possibilities have been incorporated into a package of programs (Technisch Centrum FSW, 1980) as options. The most important alternatives are the following:

1. The evaluation of a set of items as one scale;
2. The construction of a scale from a given pool of items;
3. Multiple scaling, the construction of a number of scales from a given pool of items;
4. The extension of an existing scale by means of a larger pool of items; and
5. The investigation of the double monotony of a set of items.

These procedures will be briefly indicated below. For further details the reader is referred to Mokken (1971) or the corresponding manual (Technisch Centrum FSW, 1980).

The Evaluation of a Set of Items as One Scale

In most cases where the traditional techniques of Guttman scaling were applied, the researcher was to have selected beforehand a set of n items that could be considered homogeneous with respect to some variable. With the aid of these techniques, the whole set of items is then evaluated as just one scale, and as a result some defective items are eliminated in the end. The procedures are based on the estimates of the item coefficients H_i, the scale coefficient H, and an analysis of the 2×2 tables for all item pairs, which should not be negatively correlated.

Such an evaluation can take place at two levels of statistical analysis. The first level implies a test of the criterion of random response (marginal independence: H or $H_i = 0$ for some i). The corresponding estimates $\hat{H}$ and $\hat{H}_i$ are used to test these null hypotheses. The hypothesis of scalability may be rejected for the whole set of items when its H value proves not to be significantly greater than zero according to the test.

The second level aims at the estimation of corresponding population values. In this case confidence intervals are sought for the population coefficients H and H_i as estimated with the sample coefficients $\hat{H}$ and $\hat{H}_i$ at a given level of confidence $(1 - \alpha)$.

Constructing a Scale from a Pool of Items

Here primary interest is in the exploration of the homogeneity of a pool of items which, on the basis of their content alone, are thought to be more or less homogeneous. From this pool must be selected a set of items, as large as possible, which satisfies the scale criteria and which may be used for the ultimate measurement of the variable. Apart from the statistical criteria used, the general structure of this problem is a familiar one in statistical methods aiming at the optimal selection of n variables from a larger pool.

The method summarily sketched here aims at a rather straightforward "maximization" of $\hat{H}$ in terms of the definition of a scale given in the previous section. It consists of a stepwise and recursive technique of constructing a scale from a given set of items. A value for the defining constant c is chosen (e.g., .30 in most applications). The analysis then proceeds from the cross-tabulations (2×2 tables) for all item pairs starting with the best (or a given) pair of items and progressively adding items to the scalable set, ending with a test of the completed scale.

Multiple Scaling

Sometimes allowance should be made for the existence of more than one dimension or latent variable and, hence, for the existence of more than one corresponding scale. The procedure provides for

this contingency by the availability of criteria to look for various scales. The procedure starts with the selection of items for the first scale. After the selection of this scale, the remaining pool of items, including the items that were rejected, are subjected to the same procedure once more for the purpose of selecting a second scale; this process is repeated until no more scales can be found.

Extension of an Existing Scale

Once a scale has been found and has demonstrated its usefulness in research, it may be worthwhile to develop and extend it further in subsequent research. In the exploratory phases referred to above, scales often consist of only a few, say four to six, items. The old scale may then be extended by trying out and adding new items in order to get a larger scalable set of items for future use. A modification of the former procedure serves that purpose. Starting at once with the original set of scalable items, new items may be added along the lines described.

Investigating Double Monotony

The coefficient of scalability, H and H_i, and the definition of scalability are related to sets of monotonely homogeneous items. Therefore, the procedures of test construction based on them may at best be expected to result in the selection of sets of monotone items. Consequently, if it is desirable to restrict selection to sets of doubly monotone items, the items will have to satisfy that stronger condition. Additional criteria and procedures will then be necessary in order to weed out further a monotonely homogeneous set of items in search of "defective" items, trimming it down to a doubly monotone set.

There are a few rough and ready but correspondingly simple means to perform such an inspection. An obvious method is to split a calibrating sample of persons in various groups and to investigate whether the difficulty order of the items is invariant across groups (Molenaar, 1982b). One way is to divide the sample into quantiles of the distribution of the test score (number correct), and to inspect the invariance of item difficulty order. If the sample is split according to $n - 1$ observed score classes (excluding the extremes for lack of information) and plotting the difficulty proportions for each score class, the empirical item-test regression functions are checked.

Another check arises from a result of double monotony involving the following symmetric matrices of probabilities:

$$\Pi = [\pi_{ij}(1,1)]; \quad n \times n \qquad [7]$$

$$\Pi^{(0)} = [\pi_{ij}(0,0)]; \quad n \times n \qquad [8]$$

where π_{ij} (0, 0) is the probability of a {0, 0} response to items i and j (responses incorrect or not positive) and where the diagonal elements π_{jj} (1, 1) and π_{jj} (0, 0) are not specified. Assume that the rows (and columns) are numbered according to the item ordering, i.e., $i < j$, $\pi_i < \pi_j$. Then, double monotony implies that the elements of row i of Π will *increase* monotonically with column index j. Similarly, the elements of row i of $\Pi^{(0)}$ will *decrease* monotonically with increasing column index j. By symmetry, the same is true for the rows. (Of course, significance margins can be taken into account in actual analysis.)

Example of an Analysis

Detailed illustrations of all the scaling procedures described above (applied to political science problems) are found, for instance, in Mokken (1971, chaps. 6–9) and Stokman (1977). The following example uses only the simplest of the scaling procedures described above: evaluation of a set of items as one scale. The example is taken from a recent study of subjective sleep quality, described by Mulder-Hayonides van der Meulen, Wijnberg, Hollanders, De Diana, & Van den Hoofdakker (1980).

The General Sleep Quality Scale consists of 14 dichotomous items (agree/disagree). These are chosen to measure sleep complaints ("positive" response) ranging from mild to very severe (see Table 1). The original study was carried out with a group of 80 depressive patients using 27 items, from which the 14 presented here were selected. An inspection of the matrix of sample item intercorrelations (not reproduced here) revealed that no pair of items was negatively correlated.

In Table 1 the sample proportions of positive response ($\hat{\pi}_i$) and sample $\hat{H}_i$ values are given, together with $\hat{H}$ for the total scale. All H_i were significantly greater than zero ($p < 6 \times 10^{-17}$ for all 14 tests simultaneously).

More interestingly, the smallest $\hat{H}_i$ was .37 for item 12, and the total $\hat{H}$ of .52 would suggest that this is a "strong" scale by the criteria given earlier.

A further inspection, which for reasons of space is omitted here, showed that there was no reason to doubt the double monotony of the scale. Of course, no test or scale developed for general use should be constructed on the basis of a small sample from a very specific population. The actual scale was developed and cross-validated extensively.

Table 1
General Sleep Quality Scale

Item	Content	$\hat{\pi}_i$	$\hat{H}_i$
1	Often don't sleep at all	.16	.75
2	Get up often	.34	.57
3	Mostly toss and turn a lot	.41	.53
4	Wake up several times	.43	.62
5	Mostly sleep very badly	.43	.44
6	Feeling of sleeping just a few hours	.49	.56
7	Don't sleep more than 5 hours	.50	.49
8	Mostly sleep well*	.53	.61
9	Mostly fall asleep easily*	.54	.52
10	Feeling don't sleep long enough	.58	.53
11	Often lie awake more than half hour	.61	.46
12	Hard to fall asleep again, after waking up	.65	.37
13	Tired feeling after getting up	.66	.44
14	Feel well rested after getting up*	.70	.51
Total ($\hat{H}$)			.52

*Scored negatively.

Description of Ability Estimation

When a set of items has been shown to form a satisfactory test or scale—for instance, by the methods treated in the previous section—the next step is test administration: using the scale to obtain estimates of the abilities of individuals. Common practice favors use of the simple score x (the number of positive responses), which is adopted by fiat in classical test theory, formally justified in the deterministic Guttman model, and—among stochastic item response models—uniquely associated with the Rasch model (1980) as a minimal sufficient statistic for θ.

Assuming only monotonely homogeneous items, it has been noted by Lord and Novick, (1968, p. 386) and by Mokken (1971, p. 139) that the simple score has a monotone increasing regression on the ability θ. Moreover, the variates x and the underlying ability θ (assuming an arbitrary distribution for θ) are positively correlated, which warrants the use of x in linear structural models such as LISREL (Jöreskog, 1981). Moreover, it can be shown that for monotonely homogeneous items the "local" person order (i.e., $\theta_a < \theta_b$) always is given by the expected proportion correct (or expected score), and that person order is item selection free, i.e., not dependent upon the particular selection of items. Since in the nonparametric approach only ordinal properties of θ are of interest, this result provides a justification (by, e.g., Mokken, 1971) for the use of the simple score in the present context as well.

The possibility, especially when working with shorter tests, of going beyond the simple score as a point estimate of θ is the subject to which the rest of this section is devoted. The two greatest technical problems associated with making inferences about individual abilities based on a relatively short test are (1) the relatively limited amount of information provided by the responses themselves and (2) the general inapplicability of results which are only asymptotically valid.

Working in the context of the nonparametric model developed here for the case of double monotony, the method to be described addresses the first of these problems by allowing the introduction of prior knowledge into the analysis (via Bayes' theorem). It avoids the second by using exact, small sample results. In addition, working directly with posterior distributions enables (1) obtaining interval estimates for θ, (2) making classification decisions based on utilities associated with different mastery levels, and (3) analyzing responses obtained in tailored adaptive testing situations.

Definitions and Basic Results

Begin by defining for a set of n doubly monotonous items $n + 1$ ordered ability classes Θ_i, according to the subdivision of the ability scale given by the item difficulties b_i, where (as usual) the items have been ordered so that $b_1 > b_2 > \ldots > b_n$.
Specifically, let

$$\begin{aligned} \Theta_0 &= \{\theta \mid \theta < b_n\} \\ \Theta_1 &= \{\theta \mid b_n \leq \theta < b_{n-1}\} \\ &\vdots \\ \Theta_n &= \{\theta \mid b_1 < \theta\} \ . \end{aligned} \qquad [9]$$

Remembering that $P_i(b_i) = .5$ for each item, a given ability θ will be a member of class Θ_i if

$P_j(\theta) \geq .5$ for $j = n, n-1, \ldots, n-i+1$ (the i "easiest" items) and
$P_j(\theta) < .5$ for $j = n-i, n-i-1, \ldots, 1$ (the $n-i$ "most difficult" items).

Based on these ordered classes, there may also be assigned, for a person with ability θ, a corresponding "class score" θ^* as follows:

$$\theta^* = i \quad \text{if } \theta \in \Theta_i \quad \text{for } i = 0,\ldots,n \ . \qquad [10]$$

Note that the class score of a person may be directly interpreted as the *number* of items which the person *dominates*, in the sense of having a positive response probability of at least .5 for each of these items. Note also that "number of items dominated" (the class score) in the nonparametric context plays the same role as number of positive responses (the simple score) does for the Guttman model: "reproducing" the location of a person's ability relative to the item difficulties. The situation described above is illustrated in Figure 1.

Specification of Prior Knowledge

The main goal of the approach to ability estimation adopted here is to obtain posterior distributions for class scores θ^*, based on prior knowledge combined with the information from item responses. To do this necessitates beginning with a specification of prior knowledge regarding the unknown probabilities of positive response for person a, denoted by π_{ia} rather than $P_i(\theta_a)$, to emphasize that the probabilities, and not the ability θ_a, are the parameters being discussed. From double monotony and the order of the b_i is derived

$$\pi_{1a} < \pi_{2a} < \ldots < \pi_{na} \quad \text{for all } a \ . \qquad [11]$$

In addition to using this critical piece of prior information, a natural conjugate prior density is selected for the π_{ia}, which takes the form of a product of beta densities, restricted to conform with Equation 11. For information regarding natural conjugate priors in general and beta densities in particular, the reader is referred to Novick and Jackson (1974, especially chaps. 5 and 6). Here, it will have to suffice to say that this choice provides a flexible family within which a wide variety of prior knowledge may be specified. Moreover, when such a prior is combined with information from the item responses, the posterior density for the π_{ia} belongs to the same family: a product of betas with the restriction of Equation 11.

Of course, the real need concerns determination of distributions (prior and posterior) for the class score θ_a^*. From Equations 9 and 10, these are directly obtainable in terms of the distributions of the π_{ia}:

$$\begin{aligned}
&\text{Prob } (\theta_a^* = 0) = \text{Prob } (\pi_{na} < .5), \\
&\text{Prob } (\theta_a^* = 1) = \text{Prob } (\pi_{n-1,a} < .5 \text{ and } \pi_{na} \geq .5), \\
&\vdots \\
&\text{Prob } (\theta_a^* = n) = \text{Prob } (\pi_{1a} \geq .5).
\end{aligned} \qquad [12]$$

When working with product beta densities, the probabilities in Equation 12 may be obtained through relatively straightforward numerical integration.

To give an idea of the importance of the prior information contained in the order restriction of Equation 11, consider the following simple example. Suppose a uniform joint prior density is adopted for the π_{ia}, one which gives equal weight to every permissible combination of values. When this density is integrated to obtain the probabilities given in Equation 12, it is found that θ_a^* has not a uniform distribution but a binomial distribution with parameters n and .5. Thus, when $n = 10$, the prior chance assigned to a class score of 0 is only about 1 in 1,000, while a class score of 5 has a prior chance of roughly 1 out of 4. All this is a direct implication of assuming that the items increase monotonically in difficulty for person a.

The above example was not intended to suggest that a uniform prior for the π_{ia} (which is a binomial prior for θ_a^*) will normally be a suitable specification. Instead, a careful examination of alternative priors in the light of available knowledge should typically take place. An interactive computer program, ABILITY, is available to facilitate this process, allowing the investigator to easily specify and obtain information about any chosen series of priors. In the same way, the investigator may combine various priors with hypothetical item response patterns and see the resultant posterior distributions for θ^*. This provides insight into the relative roles played by the prior and by data in any particular case and, as such, offers a guide in the choice of an appropriate prior.

Specifying a prior distribution for the π_{ia} as discussed above has, of course, a basically subjective character. "Subjective" should not, however, be equated with "arbitrary." The specifications should be at least broadly defensible, with regard to both their origins and their consequences.

Example of an Analysis

To illustrate the use of the ability estimation procedure described above, it is convenient to return to the example of the previous section, namely the General Sleep Quality Scale, as administered to Dutch harbor pilots and their wives (De Vries-Griever, De Vries, & Meijman, 1982).

Although there are 14 items in the scale, for the purpose of this analysis, items were grouped together whose observed proportions of positive responses in the original study differed by less than .01. This resulted in the 10 item groups shown in Table 2. These, in turn, defined 11 ordered ability classes, Θ_i, and corresponding possible class scores, θ^* (from 0 to 10).

Remembering that the items are scored so that a positive response refers to an instance of a sleep *complaint*, a person with a class score of 0 would be considered as having a high sleep quality in the sense that his or her chance of a positive response would be less than .5 for each of the 14 items. A person with a class score of 4 would have a chance greater than or equal to .5 of responding positively to the items in only the four "easiest" groups (those measuring the mildest sleep complaints), i.e., Items 10, 11, 12, 13, and 14.

This illustration is restricted to the subpopulation of wives of the youngest group of harbor pilots (ages 29 to 34), and only the measurement of their sleep quality will be considered for the period when their husbands were working. Based on the earlier studies cited in the previous section, a product beta prior distribution was specified for the π_{ia} which assigns the highest prior probability to $\theta_a^* = 3$, with a 96% interval for θ_a^* going from 1 to 5 (inclusive). Thus, it was expected that women in this group would have moderate, but not severe, sleep complaints. The complete prior distribution for θ_a^* used in subsequent analyses is given in Table 2.

Table 3 shows the results of analyzing five selected response records. The first (Person 1) shows the posterior distribution for a woman with no sleep complaints, and the second (Person 2) for a

Table 2
Item Groups, Class Scores, and Prior Distribution for Sleep Quality Scale

Variable	Item Group									
	10	9	8	7	6	5	4	3	2	1
Items in Group	14	13,12	11	10	9,8	7,6	5,4	3	2	1
Class Score (θ^*)	0	1	2	3	4	5	6	7	8	9	10
Prior[a]	.02	.10	.25	.31	.21	.09	.02	–	–	–	–

Table 3
Analysis of Individual Response Records For the Sleep Quality Scale

Person and Variable	Item Group 10	9	8	7	6	5	4	3	2	1
Person 1										
Proportion*	0	0	0	0	0	0	0	0	0	0
Posterior	.38	.45	.14	.03	-	-	-	-	-	-
Person 2										
Proportion	1	1	1	1	1	1	1	1	1	1
Posterior	-	-	-	-	.01	.04	.16	.34	.30	.13	.02
Person 3										
Proportion	1	1	1	1	.5	0	0	0	0	0
Posterior	-	.01	.06	.25	.46	.21	.01	-	-	-	-
Person 4										
Proportion	1	.5	0	1	1	.5	0	0	0	0
Posterior	-	.04	.14	.20	.27	.28	.07	-	-	-	-
Person 5										
Proportion	0	0	1	0	1	0	1	1	0	0
Posterior	.03	.11	.16	.25	.20	.18	.05	.02	-	-	-

*Proportion of positive responses per item group.

woman responding positively to all 14 items. In both cases the posterior represents a compromise between prior and data.

According to well-known results in theoretical statistics, in any analysis the full information contained in the complete response pattern will be necessary to make inferences about a person's ability. (For a criticism of the treatment of response patterns in traditional Guttman scaling procedures, see Mokken, 1971, pp. 153–157). The last three records were chosen to illustrate this effect of pattern on posterior with the number of positive responses held constant (at $x = 6$). A "perfect pattern" results in the most concentrated posterior (Person 3). A less consistent pattern shifts the mode and increases variance and skewness of the posterior (Person 4). Finally, a pattern which shows little consistency with the theoretical ordering of the items leads to a posterior which is actually broader than the prior (Person 5).

Thus, the data in Tables 2 and 3 demonstrate that the method discussed here allows flexible specification of prior knowledge and delivers complete posterior distributions for latent ability, as measured by the class score.

References

De Vries-Griever, A. H. G., De Vries, G. M., & Meijman, T. F. *De Nederlandse Rijksloods. Deel II: Slaap.* Groningen: Subfaculteit Psychologie van de Rijksuniversiteit Groningen, 1982.

Green, B. F. Attitude measurement. In G. Lindzey (Ed.), *Handbook of social psychology (Vol. 1).* Cambridge MA: Addison-Wesley, 1954.

Henning, H. J. Die Technik der Mokken-Skalenanalyse. *Psychologische Beitrage,* 1976, *18,* 410–430.

Jöreskog, K. G. Analysis of covariance structures. *Scandinavian Journal of Statistics,* 1981, *8,* 65–92. (with Discussion)

Lewis, C. Estimating abilities: Inference for random variables. *Kwantitatieve Methoden,* 1981, *2,* 17–34.

Lippert, E., Schneider, P., & Wakenhut, R. Die Verwendung der Skalierungsverfahren von Mokken und Rasch zur Überprüfung und Revision von Einstellungsskalen. *Diagnostica,* 1978, *24,* 252–274.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs,* 1947, *61* (No. 4).

Loevinger, J. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin,* 1948, *45,* 507–530.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading MA: Addison-Wesley, 1968.

Meredith, W. Some results based on a general stochastic model for mental tests. *Psychometrika,* 1965, *30,* 419–440.

Mokken, R. J. *A theory and procedure of scale analysis with applications in political research.* New York: de Gruyter/Berlin: Mouton, 1971.

Molenaar, I. W. De beperkte bruikbaarheid van Jansen's kritiek. *Tijdschrift voor Onderwijsresearch,* 1982, *7,* 25–30. (a)

Molenaar, I. W. Een tweede weging van de Mokkenschaal. *Tijdschrift voor Onderwijsresearch,* 1982, *7,* 172–181. (b)

Mulder-Hayonides van der Meulen, W. R. E. H., Wijnberg, J. R., Hollanders, J. J., De Diana, I. P. F., & Van den Hoofdakker, R. H. *Measurement of subjective sleep quality.* Paper presented at the European Sleep Conference, Amsterdam, 1980.

Novick, M. R., & Jackson, P. H. *Statistical methods for educational and psychological research.* New York: McGraw-Hill, 1974.

Rasch, G. *Probabilistic models for some intelligence and attainment tests* (expanded edition). Chicago: University of Chicago Press, 1980. (Originally published, Copenhagen: Danmarks Paedagogiske Institute, 1960)

Stokman, F. N. *Roll calls and sponsorship: A methodological analysis of third world group formation in the United Nations.* Leyden: Sijthoff, 1977.

Stokman, F. N., & Van Schuur, W. H. Basic scaling. *Quantity and Quality,* 1980, *4,* 5–30.

Technisch Centrum FSW. *STAP user's manual (Vol. 4: Stochastic cumulative scaling, Mokken scale, Mokken test).* Amsterdam: University of Amsterdam, 1980.

Torgerson, W. S. *Theory and methods of scaling.* New York: Wiley, 1958.

Van de Wijgaart, C. Program package STAP: A statistical appendix. *Behavior Research Methods and Instrumentation,* 1981, *13,* 379–380.

Van Naerssen, R. F. Eenvoudige formules voor de optimale spreiding van item-p-waarden. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden,* 1972, *27,* 123–133.

White, B. W., & Saltz, E. Measurement of reproducibility. *Psychological Bulletin,* 1957, *54,* 81–99.

Acknowledgments

The authors express their thanks to a referee and to Ivo Molenaar for their useful comments on an earlier version, and to Geertje Winkel and Margie van der Mark for their excellent typing of the manuscript. The authors are also extremely grateful to Theo Meijman and Adrí de Vries-Griever for their help in providing data, analyses, references, and advice regarding the sleep quality example.

Authors' Addresses

Send requests for reprints or further information to Robert J. Mokken, Netherlands Central Bureau of Statistics, Postbus 959, 2270 AZ Voorburg, The Netherlands, or Charles Lewis, Vakgroep Statistiek and Meettheorie, Rijksuniversiteit Groningen, Oude Boteringestr. 23, 9712 GC Groningen, The Netherlands.