The Mokken Scale: A Critical Discussion

Edward E. Roskam and Arnold L. van den Wollenberg
University of Nijmegen

Paul G. W. Jansen
The Netherlands Postal and Telecommunications Services

The Mokken scale is critically discussed. It is argued that Loevinger's H, adapted by Mokken and advocated as a coefficient of scalability, is sensitive to properties of the item set which are extraneous to Mokken's requirement of holomorphy of item response curves. Therefore, when defined in terms of H, the Mokken scale is ambiguous. It is furthermore argued that item-selection free statistical inferences concerning the latent person order appear to be insufficiently based on double monotony alone, and that the Rasch model is the only item response model fulfilling this requirement. Finally, it is contended that the Mokken scale is an unfruitful compromise between the requirements of a Guttman scale and the requirements of classical test theory.

The so-called "Mokken scale analysis," first developed by Mokken (1971), has recently become more widely used, especially in the Netherlands (e.g., Niemoeller & van Schuur, 1983; Stokman & van Schuur, 1980) and in Germany (e.g., Henning, 1976). Mokken and Lewis (1982) gave a brief account of the Mokken scale and some related issues. Although they referred to a number of recent publications, such as Henning (1976) and Molenaar (1982a,b), they ignored the fact that Henning had reported rather disappointing results, and that Jansen (1981, 1982a,b) and Jansen, Roskam, and van den Wollenberg (1982) had raised a number of critical points with respect to the use of Loevinger's H (as adapted by Mokken) to indicate "scalability."

This paper reiterates the points raised by Jansen et al., and in addition, some related issues are discussed in a fairly non-technical way; for technical details, the reader is referred to Jansen (1982b, 1983) and Jansen, Roskam, and van den Wollenberg (1984). First, the Mokken scale is briefly described. In later sections, Mokken's scalability concept is contrasted with the authors' own position on this issue, and the scalability coefficient H is critically examined.

Monotone homogeneity defines a class of item response models with certain general and desirable properties. Any parametric model belonging to this class adds additional specific and restrictive properties, which may be more than is desirable. Hence, a scalability coefficient and a corresponding procedure of scale analysis which is directed toward nonparametric properties should be considered desirable objectives, and can be put to use to investigate the extent to which data satisfy the properties of monotone homogeneity (or holomorphy, respectively) without involving parameter estimation. The critical question of whether or not this objective is met by Mokken's scalability coefficient and procedure for scale analysis is discussed below.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 3, September 1986, pp. 265–277

0146-6216/86/030265-13$1.90

The Mokken Scale

Double Monotony and Coefficient *H*

A set of dichotomous items having monotone item characteristic curves (ICCs) is called *monotone homogeneous;* the one-, two-, and three-parameter logistic models all have this property, as does the normal ogive model. When the ICCs of a monotone homogeneous model are parallel in terms of horizontal translation, the model is called *holomorphic*. The one-parameter logistic (Rasch) model has this property; the two- and three-parameter logistic models and the normal ogive model do not. A holomorphic model is very simple, as only one parameter for person location and one parameter for item location are needed.

A more general, *nonparametric* definition of holomorphy only requires that the probability of a positive response is an increasing function of a person's latent trait position, and a decreasing function of the item's difficulty. Consequently, no two ICCs intersect, and this is called *double monotony*. The term nonparametric in this context means that no assumptions are made concerning the particular functional form of the ICCs.

For doubly monotone item response functions, Mokken (1971) derived a number of nonparametric properties of the marginal response distribution for a population of persons. For instance, holomorphy implies a unique ordering of the items: The ordering of the response probabilities of the items is independent of a person's latent trait position; the ordering is the same for all conceivable samples. The role of persons and items is completely symmetrical, so the statement may be reversed.

In addition, Mokken (1971, p. 152) developed a "coefficient of scalability", based on (but different from) Loevinger's *H*. For pairs of items, i and j, H_{ij} is defined as

$$H_{ij} = \frac{p_{ij} - p_i p_j}{p_i(1 - p_j)} \qquad (p_i < p_j) \quad , \tag{1}$$

where p_i refers to the proportion of persons with positive responses to item i, and p_{ij} refers to the proportion of persons with positive responses to both items i and j ($i,j = 1, \ldots, k$). H_{ij} is equal to ϕ/ϕ_{max}, where ϕ is the binary inter-item correlation, and ϕ_{max} is its maximum value for given p_i and p_j. In a perfect Guttman scale, or scalogram, $p_{ij} = p_i$ for $p_i < p_j$, and so $\phi = \phi_{max}$, and $H_{ij} = 1$.

The *item coefficient of scalability*, H_i, is defined by the number of "errors" in response patterns involving item i divided by the number of errors which is expected under the hypothesis of marginal independence (where an error is defined in terms of the perfect scalogram; see Mokken, 1971, p. 151ff., for details):

$$H_i = \frac{\sum_{j=1}^{k} (p_{ij} - p_i p_j)}{\sum_{j=1}^{i-1} p_j(1-p_i) + \sum_{j=i+1}^{k} p_i(1-p_j)} \qquad (p_1 < p_2 < \ldots < p_k) \quad . \tag{2}$$

Finally, the coefficient of scalability, H, is defined as a weighted sum of H_is:

$$H = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} (p_{ij} - p_i p_j)}{\sum_{i=1}^{k} \left[\sum_{j=1}^{i-1} p_j(1-p_i) + \sum_{j=i+1}^{k} p_i(1-p_j) \right]} \qquad (p_1 < p_2 < \ldots < p_k) \quad . \tag{3}$$

Since H is essentially based on counting order reversals, it is also a type of average weighted inter-item Kendall's tau.[1]

The Concept of Scalability and Mokken Analysis

Mokken (1971) defined a scale as any set of positively correlated items, for which all values of H_i exceed some positive constant c. A lower bound of .30 is advocated as a suitable value of c. The term "scale," in this context, appears to refer to a set of items satisfying *any* model. The value of c as an acceptable lower bound for H is in a sense arbitrary, just as setting a probability level is arbitrary in accepting or rejecting a null hypothesis.

The problem with Mokken scale analysis is that the literature must be searched in order to gain a more specific understanding of Mokken's concept of scalability. In particular, the questions of where and in what way double monotony is implied or required can be a matter of debate. The reason for this is that Mokken (1971) was not very explicit in specifying the item response model with respect to which his concept of scalability is defined. Mokken (1971) states: "We prefer to use the coefficient of homogeneity, H, as a criterion of scalability in the sense of monotone homogeneity. . . . Our coefficient of scalability as such will be our sole criterion of scalability" (p. 182).

Ignoring the fact that this quotation implies a circular definition of scalability, the position can be adopted that a Mokken scale is any set of items which have monotone ICCs over a unidimensional latent trait continuum. This definition is evidently implied by Mokken and Lewis (1982, pp. 418, 421). The Rasch model, the two- and three-parameter logistic models and the normal ogive model would all be special cases of the Mokken scale defined in this manner. The Guttman scalogram can be considered a limiting case, where the ICCs are step functions.

However, the concept of scalability becomes more complicated because in both Mokken (1971) and Mokken and Lewis (1982) it also seems that *double* monotony is additionally implied in the concept of the Mokken scale. Mokken (1971, p. 149), introducing H, wrote:

> We may add that in the definition of the coefficient of scalability H, the *order* of the population difficulties . . . is essential. Therefore the property of monotone homogeneity may not be sufficient if we relax the deterministic requirements and we admit a probabilistic model. We may then need holomorphic or doubly monotone sets of items . . .

In their abstract, Mokken and Lewis (1982) considered the Mokken model to be a natural generalization of Guttman scaling. The Guttman scale, as is generally known, represents a weak ordering of both persons and items, and as such it represents a case of double monotony. Mokken and Lewis also emphasized that monotone homogeneity may not be the ultimate goal of scaling, and that double monotony may be additionally required for test administration.

The confusion about the definition of the Mokken scale is of a more than casual nature, as will be seen below. This confusion is related to the fact that although Mokken (1971), as well as Mokken and Lewis (1982), introduced the concept of a scaling model (a nonparametric model of doubly monotone items), they defined a scale by means of a scalability coefficient. Apparently, Mokken and Lewis do not use the word "scale" to refer to a set of items satisfying *any* given model.

---

[1]Cliff, 1983, points out that "in view of . . . the purely descriptive origins of this index, an underlying trait model is apparently not necessary for it" (p. 291).

Scalability

The Scaling Model

Mokken and Lewis (1982, p. 422, footnote 3) pointed out that they do not use the term "scale" in the strict sense in which it is used in axiomatic measurement theory. They use the term in an open way, meaning "a set of items which have certain properties". In measurement theory, a scale is defined as the numerical representation of some quality, obtained from the theoretical structure of the responses to a set of items. Thus, scalability always refers to some item response model, for example, Guttman scalability or Rasch scalability. A coefficient such as H carries some information concerning the structure in the data, but in order to interpret and evaluate that information, it must be interpretable in terms of the properties of a model.

In order to specify the response model that acts as the (implicit) reference for Mokken scalability, it is necessary to follow more closely the reasoning which led Mokken to H and to the procedure of scale analysis. Like Guttman, Mokken concentrated on the prediction of single item responses from the raw score and the item ordering. Related to this is the fact that the scalogram, with a sufficiently large and well-spread number of items, can discriminate very adequately among persons with different latent positions. The objective of scaling persons is not only to reproduce their response patterns from their raw scores, but also to discriminate among them. However, Mokken (1971, pp. 41–42, 70–71) considered the deterministic Guttman scale to be too rigid and obviously not realistic. Furthermore, he found (1971, p. 64) that none of the existing criteria of scalability was satisfactory as a measure of Guttman scalability. His answer to these problems was an adaptation of Loevinger's H as a coefficient of scalability.

Although it is possible to agree with Mokken that none of the existing coefficients is satisfactory, H is not a satisfactory alternative, as will be shown. In order to have a satisfactory coefficient of scalability, it is desirable to have a model or theory about the way responses may differ from what is expected when perfect scalability is assumed. Such a theory is patently absent in the Guttman scalogram, for the simple reason that the scalogram is a deterministic model with no stochastic elements in it.

Consequently, a probabilistic "Guttman model" should serve as the basis for the construction of the coefficient of scalability. The ideal is to order and select a set of items with respect to the degree to which they satisfy the requirements of some sort of probabilistic analogue of the scalogram model. There appear to be good reasons (cf. Jansen, 1983; Roskam, 1983; Roskam & Jansen, 1984) to argue that the Rasch model is the only probabilistic generalization of the property of composite transitivity, which is the core axiom of the scalogram model (Ducamp & Falmagne, 1969).

However, a researcher may not be interested in a specific parametric model, but rather in a class of models defined by certain broad properties. These properties could, for instance, be monotony of item response curves, or the more restrictive property of holomorphy, which excludes both the two-parameter logistic and the normal ogive models. But in such a case, in order to be useful, a coefficient of scalability must give information about the deviations from the expected structure; the degree to which the properties of the class of models are satisfied should be reflected in the scalability coefficient.

Requirements for a Coefficient of Scalability

In the previous section, the almost trivial conclusion was drawn that a coefficient of scalability must give information with respect to the hypothesis that the items constitute a scale with clearly specified properties. In particular, H should indicate the extent to which the items satisfy an explicit though nonparametric item response model. This does not imply that rigid criteria are required for rejecting that hypothesis, nor that the sampling distribution of the statistic should be known under various deviations

from the model, but it should at least behave properly. This means at least that the coefficient increases when items known to deviate from the model requirements are eliminated, and that it decreases when an item known to satisfy those requirements is replaced by one which is known to violate them. Note that this implies that any test of goodness of fit is by itself a scalability coefficient. Conversely, a coefficient which is not explicitly related to a scale model is not a scalability coefficient.

With regard to the behavior of goodness-of-fit statistics and scalability coefficients, some aspects can be distinguished:

1. The sensitivity of a statistic to violation of essential properties of the model;
2. The sensitivity of a statistic to irrelevant properties of the data;
3. The differential sensitivity of a statistic to violation of essential properties of the model, given differences in irrelevant aspects.

A statistic should be sensitive to violations of the properties (axioms) of the model; however, it need not be sensitive to violations of all axioms. For example, van den Wollenberg (1982) showed for the case of the Rasch model that his statistic Q_1 is sensitive to violations of monotony and sufficiency, but not to violations of local independence and unidimensionality; for Q_2 the situation is reversed, so together these statistics constitute a complete test of model fit. A statistic should not be sensitive to properties of the data which are immaterial in the scale model. As will be shown, this requirement is not met by the scalability coefficient H.

It is quite possible that fit statistics react differentially to model violations dependent upon intrinsically irrelevant properties of the data. A quite common example is that of statistical information. Information is irrelevant to model violation; but, when the model is violated, a fit statistic will covary with the amount of statistical information in the data.

To conclude, a scalability coefficient may be insensitive to some types of model deviations, and in this case more than one coefficient is needed for a full characterization; but a scalability coefficient may never react to irrelevant properties of the data when the scaling model holds. For the case of the Mokken scale, this implies that a pertinent coefficient of scalability should be a test of monotone homogeneity and/or double monotony of ICCs, as these are defining characteristics. The question of whether the coefficient proposed by Mokken, H, satisfies the above requirements will be investigated below.

What Does H Measure?

The meaning of H will be revealed by scrutiny of the relation between H and specified properties of item response structures. It is implied by Mokken that H expresses the degree of monotone homogeneity of a set of items. However, the quotations show that holomorphy was additionally implied (cf. Mokken, 1971, pp. 148–149). Since both the Guttman scalogram and the Rasch model are special cases of the Mokken scale, they should be homogeneous in the sense of Mokken's H.

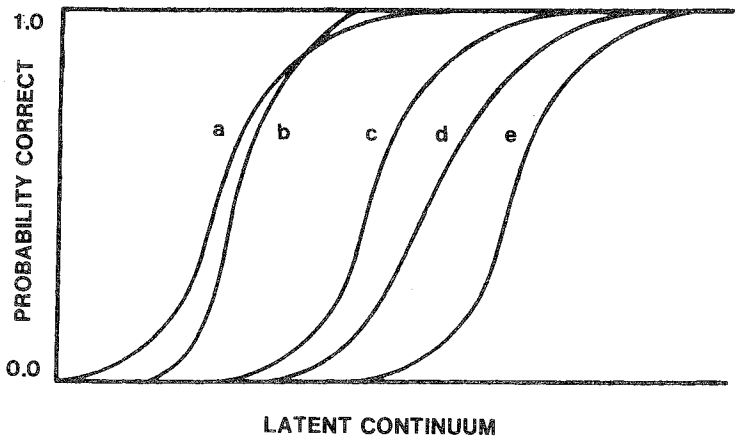
For pairs of items, i and j, H_{ij} is simply equal to ϕ/ϕ_{max}. Therefore, it follows immediately that H_{ij}, H_i, and H are equal to unity when a set of items forms a perfect Guttman scale. Indeed, H operates satisfactorily for the scalogram. In this section it is shown that the same does not hold for probabilistic response models, including the Rasch model.

Coefficient H and Monotone ICCs

In the following, sets of (doubly) monotonous item response curves are discussed; these sets can be characterized by two properties:

1. *Steepness,* the slope of the ICC;

Figure 1
Various ICC Patterns Which Cause Different Degrees of Violation of Guttman Scalability



2. *Closeness,* the distances between the item location parameters.

Figure 1 shows ICCs for several items which may cause different degrees of violations of Guttman scalability. ICCs *a* and *b* cause many violations, but *a* and *e* cause almost no violations at all, and *b* and *e* even less; other pairs of items, e.g., *c* and *e*, cause a fairly high number of violations. It can be observed that the ICCs in Figure 1 are very steep as compared to their closeness. As a consequence, the probability of a non-scalogram response pattern is low. Stated more precisely: If the slope of ICCs is steep, H will be close to unity. The lower the slope, the lower H will be, other things being equal. Similarly, if ICCs are close together, non-scalogram response patterns are likely to occur.

For a given sample or population of persons, therefore, H will vary in two ways, even for doubly monotone items:

1. The steeper the ICCs, the higher the value of H;
2. The closer the ICCs, the lower the value of H.[2]

Now consider a set of monotone homogeneous items and two samples of persons, one with a small variance and one with a large variance. For the sake of argument, let the mean of the two distributions coincide, and assume that the location of the items is symmetric around this mean.

In the sample with small variance, a large proportion of persons will produce violations of Guttman scalability and hence a small H. If the person variance is extremely small and therefore degenerates to a single latent trait position, the response probabilities become perfectly predictable from the marginal item p values, and H becomes zero. This means, of course, that the items cannot discriminate among the persons, since there is nothing to discriminate among.

In the sample with large variance, relatively few violations will occur and hence H will be large. However, if the variance of the persons becomes extremely large, so that the relative spacing of the items

[2]A proof of the relation between H and closeness is given in Jansen (1982b, 1983); with respect to steepness, the authors have no formal proof, but it should be quite clear that the effect of flatness on H is analogous to the effect of closeness on H. A simulation study by Wierda (1984) also showed the effects of the person distribution on H.

is small, and if their slope is relatively flat, the items act almost as a single item and H will be small again.

Alternatively, it may be useful to examine the effect of the spacing of the items, and how it interacts with the variance of the persons. As an extreme case, a set of items with coinciding ICCs may be considered. When the ICC is steep, H will be high: In the limiting case of the ICC being a step function, H equals 1, indicating a perfect scalogram. The discrimination by means of this scalogram is poor, as only the response patterns (0,0) and (1,1) can occur. If, on the other hand, the ICCs are close together but flat, H will be low, although the raw score may discriminate reliably among the persons if their variance on the latent continuum is fairly large. This is a case where the persons differ considerably in terms of their probabilities of correct responses and a large number of (almost) equivalent items gives a reliable estimate of those probabilities.

Thus it is seen that properties of the items and of the sample of persons, which in themselves are not related to monotone homogeneity, affect the value of H. Two data sets that are equally homogeneous in the sense of holomorphy may show a high value of H in one data set and a low value in the other. This is due to the combined effect of the variance of the persons, the spacing of the items, and the steepness of the ICCs.

From the above it follows that H should not be considered as a coefficient of scalability in the sense of monotone homogeneity, and certainly not as a measure of holomorphy (see also Molenaar, 1982a, p. 29); this is contrary to the suggestions of Mokken (1971, pp. 148–149) which were accepted at the time by, among others, Henning (1976). The effort of Molenaar (1982c) to construct alternative test procedures for the characteristics of the Mokken scale should also be seen in this light. It is a well-established fact that H is sample dependent, and is thus sensitive to properties of the data which have nothing to do with the assumed scale model.

H is based on counting the *errors* in response patterns, where error is defined in the sense of the perfect Guttman scalogram. It is obvious that both lack of monotone homogeneity and lack of holomorphy give rise to errors, but at the same time, probabilistic deviations from the scalogram pattern, which are quite acceptable, are also treated as errors. The very fact that H is based on errors (i.e., deviations from the perfect, deterministic scalogram) introduces the ambiguity of its meaning and interpretation with respect to the probabilistic concept of (double) monotony.

Coefficient H and Item Selection

What will happen if H_i is used to select or eliminate items from a pool of items? The foregoing shows that items with flat ICCs will be rejected, as well as items with ICCs close to others. Only items which are well spaced and have steep ICCs will survive. This, in itself, is not undesirable, but it might be obtained at the expense of items which might be perfectly doubly monotonic with the surviving items, and which would contribute substantially to the reliability of the raw score or the estimated latent parameter.

When items are rejected by the scalability coefficient, even though they are perfectly in accordance with the scaling model, it is feasible to lower the critical value of H for admission to a scale. Some simulation studies by Molenaar (1982a) appear to invite, according to Mokken and Lewis (1982), investigation of whether and to what extent other values (e.g., $c = .15$) can be admissible. However, there appears to be no way to derive a criterion of "admissibility" from the theory of the Mokken scale. Furthermore, Mokken and Lewis (1982, p. 422) stated that a certain value of c performs "quite satisfactorily" without giving any explicit criterion for what they consider satisfactory. Stating that this is a matter of "practicality" is likewise gratuitous when no criteria are specified.

What makes a scale "work"? Reliability and validity may, of course, be conceivable and useful criteria. Only recently, Molenaar and Sijtsma (1984) have begun to study relationships between H and

classical reliability. One conclusion is that Loevinger's H and Cronbach's α measure different things, and it may therefore not be wise to compare values of H and α. For instance, if the ICCs of a set of items are close and/or relatively flat, H will be low, but the reliability of the raw score can be high in a sample of persons with a fairly large variance on the latent continuum, and will increase with increasing number of items. Given the lack of specification of model properties or objectives, lowering the critical value of H_i is gratuitous, and was in fact refuted by some simulation studies by Jansen (1982b).

The present authors are willing to concede that H is not only sensitive to properties of the ICCs in a perfectly monotone set of items, but also to lack of monotone homogeneity, as it was meant to be. This implies that lowering the critical value of c, in order to accept items which are closely spaced and/or have a flat ICC, will also cause acceptance of nonhomogeneous items, or items which would not even fit any unidimensional item response model. A high H may be indicative of a "near-scalogram" with well-spaced items, though at the expense of items which would have survived, for example, a test of Rasch homogeneity, or items which would increase the reliability of the raw score.

It might be added that tests of Rasch homogeneity (e.g., Q_1 and Q_2; van den Wollenberg, 1982) are independent of the person parameters and as such are sample free. This does not mean, of course, that the sensitivity or power of such tests is sample free. Such tests will also not lead to rejection of the model in those cases where the model is trivially satisfied (as was demonstrated by Wood, 1978); that, however, can be judged independently.

What Is a Mokken Scale As Defined by H?

Loevinger's H was originally intended to be a measure of Guttman scalability (cf. Torgerson, 1958, p. 325). In Loevinger's terminology, a homogeneous test corresponds to a perfect Guttman scale. Mokken appears to have chosen Loevinger's coefficient—with a slight modification—as a useful index of scalability for a set of doubly monotone items (i.e., in the context of probabilistic latent trait theory, without parametric assumptions).

H can be viewed as a measure that expresses how well a set of items sorts a set of persons into piles, where each item can be seen as a probe which cuts the group into two consecutive parts on the latent continuum (e.g., those who pass and those who fail, or those who agree and those who disagree). It means that a set of items partitions the set of persons into a unidimensional quasi-order.[3]

Granting that each cut should be as neat as possible, with as few persons falling at the "wrong" side of the cut, it does make sense to use H as a coefficient of scalability, ignoring whatever assumptions might be made about the probabilistic character of the response process. It then follows that items which cut close to each other cannot unambiguously order people by their manifest response pattern. In the same vein, an item with a flat ICC is a poorly discriminating item, and one with a steep ICC will provide fine discriminations and clear person ordering.

Using H in this way as a means for item selection may lead to selection of only a few items, which will permit differentiation of the people into only a few score groups. Conversely, a test with a large number of items of approximately equal difficulty may return a low value of H, but may yield fairly reliable scores by virtue of the large number of items.

These considerations simply mean that the Mokken scale is an approximate Guttman scale, and H is nothing but what it was originally intended for, namely a measure of Guttman scalability. In other words, a Mokken scale as defined by H is an imperfect Guttman scale, with the degree of imperfection specified by the value of H and dependent on the sample of persons.

---

[3]This notion of partitioning the persons is also employed by Guttman and Lingoes (cf. Lingoes, 1968) for multidimensional models.

Guttman Scaling, Mokken Scaling, Rasch Scaling, and Classical Test Theory

In the absence of a specified probabilistic response model connecting the latent trait and manifest behavior, there is no obvious way of estimating the person parameter. This section will consider the question of what can be said about the relation of manifest and latent score in the nonparametric Mokken scale.

Sample Independence and Simple Sum Scores

For monotone homogeneous item sets, the expectation of the simple sum score is a monotone function of the latent trait score (here and in the following it should be understood that statements about persons and items can be interchanged in the case of doubly monotone items). Mokken and Lewis (1982) correctly state: "The person order . . . always is given by the expected proportion correct . . . and that person order is item selection free" (p. 426). However, there is more to it. The person order on the latent level is free of the effects of item selection by definition. Expected proportions cannot be observed, so inferences about person order are to be based on the order of observed proportions. What matters is the question of whether such statistical inferences about the person order are item-selection free. Secondly, though the rank order of the expected proportions correct is item-selection free in any monotone model, the statistical properties of orderings of observed proportions correct may not be item-selection free. (In this discussion, "item-selection free" can only refer to inferences about the positions of persons or items on the latent trait, and the inferences to be considered must be specified.)

The crucial point is that the probability that the observed proportion correct of one person exceeds that of another person is in general *not independent* of the selection of items. Consider the simplest case of two persons and one item. In this case, any requirement which is necessary for sample independence to hold is also necessary for sample independence for any sample of persons and items. For two persons and one item, there are two possible raw scores: *0* or *1*. If one person responds *1* and the other responds *0*, it is inferred that the first person's latent trait position is above the other's. This is the obvious and certain inference if the items satisfy the scalogram structure perfectly. In a probabilistic context, the primary concern is the *probability* of the event that one person responds *1* to the item, and the other responds *0*, given that only one of them responds *1*. If this (conditional) probability is to be independent of the item selected, the Rasch model follows (cf. Fischer, 1974, p. 197; Roskam, 1983, p. 82; Roskam & Jansen, 1984).

In the Rasch model, the conditional likelihoods of the persons' marginals given the sufficient statistics for the parameters of the items (i.e., the item marginals) are independent of the selection of the items. They are functions only of the person parameters (Fischer, 1974, Equation 13.5.7). Conversely, the person marginals are sufficient statistics for their true latent scores, and the rank order of these marginals is identical to the rank order of their estimated latent score. Furthermore, these statistics provide consistent estimators.

Mokken and Lewis appear to confuse the rank order of expected proportions (which cannot even be observed) and the probability of a rank order of observed proportions. Because of that, they seem to contend that persons' simple sum scores allow item-independent inference of their true rank order. The maximum likelihood latent rank order is, of course, equal to the observed rank order, but it is by itself not necessarily item-selection free. If that were true, then, for example, the normal ogive model with equal slopes would share this property of sample independent ordering with the Rasch model. In the Rasch model, the person parameters can be estimated independently of the item selection, and therefore, the latent person ordering is also estimated independently of the item selection. It can furthermore be argued that any latent trait model can only identify the latent parameters up to an ordinal transformation

(stretching or shrinking of the latent continuum can simply be cancelled by compensatory deformation of the ICCs, without affecting any statistical property of the model). Therefore, a set of items satisfies the Rasch model if and only if there exists a monotone transformation of the latent continuum which makes all ICCs one-parameter logistic. If, in this way, a set of normal ogive ICCs could be transformed into a holomorphic set of logistic ICCs, the normal ogive model would have sufficient statistics for its parameters, which it has not. The existence of sufficient statistics, and hence of sample independent parameter estimates, holds if and only if the model belongs to the exponential family with separable parameters. This leads to the Rasch model for dichotomous items (Andersen, 1980, pp. 241–242; Fischer, 1974, pp. 420–421), and excludes models such as the normal ogive model.

Though sample-free estimation of the persons' latent parameters implies sample-free estimation of the persons' latent rank order, the reverse need not be true. As far as is known, sample-free inference of the persons' latent rank order has not been shown to hold for any model except the Rasch model. In view of the considerations presented here, it can be conjectured that sample-free inferences about the persons' latent rank order require at least double monotony, and, even stronger, that it imposes the Rasch model. Considering the admissibility of a monotone transformation of the latent continuum, any model which permits sample-free inferences about the persons' latent rank order should be equivalent to the Rasch model.

The statistical properties of the Rasch model are more powerful properties of statistical inference than the property that expected marginals reflect the true latent score rank order. Note that Mokken's (1971) approach was:

> To formulate a scaling model . . . and to investigate what general properties would be preserved in the manifest data *irrespective* of the specific parametric form of the latent structure and the resulting *distribution of subjects (population distribution)* [all italics added]. (p. 174)

In summary:

1. For any holomorphic item set, the rank order of the persons' expected proportions correct is independent of the selection of items. Obviously, with increasing number of items the rank order of the observed proportions correct is asymptotically equal to the true rank order. (Little appears to be known about the asymptotic behavior of the observed rank order.)
2. The likelihood of observed proportions correct, conditional upon sufficient statistics for the instrumental, "incidental" item parameters, is independent of item sampling if and only if the Rasch model holds, and it is a function only of the structural person parameters. In this context, the parameters of interest (e.g., the person parameters) are called structural parameters, whereas the instrumental parameters (item parameters) are called "incidental" or "nuisance" parameters (cf. Fischer, 1971).

Compromising Between Classical Test Theory and Guttman Scalability

When only simple properties of the manifest responses can be used to order or score persons, only two methods come to mind:

1. Taking the raw score; but this is tantamount to assuming the Rasch model, or to ignoring latent trait theory altogether and resorting to classical test theory.
2. Scoring the person by the rank of the most difficult item which was answered positively; this is tantamount to requiring a perfect scalogram pattern.

In the absence of a specified probabilistic response model, the only testable property of the responses is their closeness to a perfect scalogram pattern. This is measured by H. However, H is just one of several possible ways of assessing Guttman scalability. H implies a particular weighing of deviations from perfect scalability, and is sample dependent. Recently, Raju (1982) and Cliff (1983) discussed various other

coefficients of homogeneity or consistency of response patterns, such as Horst's (1953) adaptation of the KR–20 coefficient.

Guttman (1950) has pointed out that scalability and discrimination are different matters. "Item analysis may find that items discriminate (or not) regardless of scalability" (Guttman, 1950, p. 185). Scalability implies that each item's score is a function of the total (or latent) score, whereas discriminability refers to the correlation between item score and total score.

In classical test theory, scoring (discriminating) people reliably is a primary concern. This is sometimes confusingly called "scaling the subjects." In item response theory, scaling or scalability of both persons and items is a central issue, and item response characteristics are taken into consideration. The Mokken scale appears to compromise between these two approaches in that scalability in the Mokken sense appears to refer primarily to persons, but a nonparametric probabilistic item response model is assumed. Hence, a set of items is sought by which the persons are neatly partitioned into cumulative score groups, exhibiting the scalogram pattern as closely as possible. It follows that certain items will spoil this pattern more than others, in particular those which have their ICCs relatively close to others, and/or have relatively flat ICCs. Mokken scale analysis will eliminate such items—even though they might very well fit a specified ICC model, such as Rasch's or Birnbaum's—for no other reason than that they spoil the scalogram pattern. It therefore appears that the technique of Mokken scale analysis runs counter to its objective, namely to identify or scale nonparametrically a set of monotone items. Conversely, if its objective were formulated not in terms of scalability but in terms of discriminability of persons, it also runs counter to that objective: Eliminating items with low H_i might decrease the reliability of the simple sum score, because the test is shortened. Also, a smaller number of items means a smaller range of scores and hence less differentiation among the persons.

The above does not imply that a coefficient for the efficiency of person discrimination by means of an item set is not relevant. For many practical purposes this is of utmost importance. However important a coefficient of person discrimination may be, discriminability and scalability are not the same thing,[4] and cannot be pursued simultaneously in a single analysis. Mokken and Lewis (and others) might argue that H is intended to have something to do with discriminability (which it appears to have), but then discriminability acts as a criterion of scalability, and contributes to the confusion about what a Mokken scale is.

Recently, Sijtsma (1984), reacting to Jansen (1982b), described Mokken scale analysis as a two-stage process, consisting of a first phase in which items are selected using H (indicating Guttman scalability), and a second phase in which the double monotony of these selected items is investigated. In a rejoinder, Jansen et al. (1984) questioned this "union" into one scaling procedure of two techniques that (1) have theoretically different bases, and (2) can be conflicting in practice, as illustrated above.

Conclusion

It appears that two issues are involved in discussing the merits of the Mokken scale. The first concerns monotone homogeneity, holomorphy, and sample independence, and the other concerns the meaning and usefulness of the H coefficient. It should be clear that H is not a measure of monotone homogeneity or of holomorphy, and that it is not sample independent. At this moment, reliable tests for holomorphy do not appear to exist, except in the Rasch model.

The Mokken scale (with Mokken's refinement of Loevinger's H) appears to be a revival of the

[4]An interesting case has been demonstrated by Wood (1978): A set of random data cannot, under the Rasch model, be distinguished from perfectly Rasch-homogeneous data obtained from persons with identical person parameters.

Guttman scale. In addition, it claims to have a probabilistic flavor which would make it work for the broad class of unspecified monotone response models. However, this claim cannot be substantiated, as the procedure can do no more than judge how response patterns deviate from the perfect scalogram. Some of the drawbacks of the Mokken scale procedure were aptly summarized by Niemoeller and van Schuur (1983): "Scales consisting of different items as indicators of the same latent trait may give rise to different scale values for the subjects, dependent upon the items used; the value of the coefficient of scalability, *H*, depends upon the homogeneity of the group of subjects in the analysis" (p. 147).

References

Andersen, E. (1980). *Discrete statistical models with social science applications*. Amsterdam: North-Holland.

Cliff, N. (1983). Evaluating Guttman scales: Some old and new thoughts. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement*. Hillsdale NJ: Lawrence Erlbaum.

Ducamp, P., & Falmagne, J. C. (1969). Composite measurement. *Journal of Mathematical Psychology, 6,* 359–390.

Guttman, L. (1950). The problem of attitude and opinion measurement. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction*. New York: Wiley.

Fischer, G. H. (1971). Einige Gedanken über formalisierte psychologische Theorieen [Some thoughts on formalized psychological theories]. *Psychologische Beiträge, 13,* 376–383.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to psychological test theory]. Bern: Huber.

Henning, H. J. (1976). Die Technik der Mokkenanalyse [The technique of Mokken scale analysis]. *Psychologische Beiträge, 18,* 410–430.

Horst, P. (1953). Correcting the Kuder-Richardson reliability for dispersion of item difficulties. *Psychological Bulletin, 50,* 371–374.

Jansen, P. G. W. (1981). Spezifisch objektive Messung im Falle monotoner Einstellungsitems [Specifically objective measurement in the case of monotone items]. *Zeitschrift für Sozialpsychologie, 12,* 24–41.

Jansen, P. G. W. (1982a). De onbruikbaarheid van Mokkenschaalanalyse [On the uselessness of Mokken scale analysis]. *Tijdschrift voor Onderwijsresearch, 7,* 11–24.

Jansen, P. G. W. (1982b). Measuring homogeneity by means of Loevinger's *H:* A critical discussion. *Psychologische Beiträge, 24,* 96–105.

Jansen, P. G. W. (1983). *Rasch analysis of attitudinal data*. Doctoral dissertation, University of Nijmegen. The Hague: Rijks Psychologische Dienst.

Jansen, P. G. W., Roskam, E. E., & van den Wollenberg, A.L. (1982). De Mokkenschaal gewogen [The Mokken scale weighed]. *Tijdschrift voor Onderwijsresearch, 7,* 31–42.

Jansen, P. G. W., Roskam, E. E., & van den Wollenberg, A. L. (1984). Discussion on the usefulness of the Mokken procedure for nonparametric scaling. *Psychologische Beiträge, 26,* 722–735.

Lingoes, J. C. (1968). The rationale of the Guttman-Lingoes nonmetric series: A letter to Doctor Philip Runkel. *Multivariate Behavioral Research, 3,* 495–508. (Reprinted in J. C. Lingoes, E. E. Roskam, & I. Borg (Eds.), *Geometric representations of relational data*. Ann Arbor: Mathesis Press, 1979, pp. 249–261.)

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

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6,* 417–430.

Molenaar, I. W. (1982a). De beperkte bruikbaarheid van Jansen's kritiek [The limited usefulness of Jansen's criticism]. *Tijdschrift voor Onderwijsresearch, 7,* 25–30.

Molenaar, I. W. (1982b). Een tweede weging van de Mokkenschaal [A second weighing of the Mokken scale]. *Tijdschrift voor Onderwijsresearch, 7,* 172–181.

Molenaar, I. W. (1982c). Mokken scaling revisited. *Kwantitatieve Methoden, 3,* 145–164.

Molenaar, I. W., & Sijtsma, K. (1984). Internal consistency and reliability in Mokken's nonparametric item response model. *Tijdschrift voor Onderwijsresearch, 9,* 257–268.

Niemoeller, K., & van Schuur, W. H. (1983). Stochastic models for unidimensional scaling: Mokken and Rasch. In D. McKay, N. Schofield, & P. Whitley (Eds.), *Data analysis and the social sciences* (pp. 120–170). London: Frances Pinter.

Raju, N. S. (1982). On tests of homogeneity and maximum KR–20. *Educational and Psychological Measurement, 42,* 145–152.

Roskam, E. E. (1983). Allgemeine Datentheorie [Gen-

eral theory of data]. In H. Feger & J. Bredenkamp (Eds.), *Messen und Testen, Forschungsmethoden der Psychologie, Enzyklopädie der Psychologie, Serie I, Band I* [Measurement and testing, Research methods in psychology, Encyclopedia of psychology, Series I, Vol. I] (pp. 1–135). Göttingen: Hogrefe.

Roskam, E. E., & Jansen, P. G. W. (1984). A new derivation of the Rasch model. In E. Degreef & J. van Buggenhaut (Eds.), *Trends in mathematical psychology* (pp. 293–307). Amsterdam: Elsevier-North Holland.

Sijtsma, K. (1984). Useful nonparametric scaling: A reply to Jansen. *Psychologische Beiträge, 26,* 423–437.

Stokman, F. N., & van Schuur, W. H. (1980). Basic scaling. *Quality and Quantity, 14,* 5–30.

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

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123–140.

Wierda, F. (1984). *Mokkenschaalanalyse: Bijdrage aan een discussie* [Mokken scale analysis: Contribution to a discussion]. Unpublished master's thesis, University of Groningen.

Wood, R. (1978). Fitting the Rasch model—a heady tale. *The British Journal of Mathematical and Statistical Psychology, 31,* 27–32.

Author's Address

Send requests for reprints or further information to Edward E. Roskam, Psychological Laboratory, University of Nijmegen, Montessorilaan 3, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands.