# Rejoinder to "The Mokken Scale: A Critical Discussion"

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

**Charles Lewis**
**Rijksuniversiteit Groningen, The Netherlands**

**Klaas Sijtsma**
**Vrije Universiteit, The Netherlands**

The nonparametric approach to constructing and evaluating tests based on binary items proposed by Mokken has been criticized by Roskam, van den Wollenberg, and Jansen. It is contended that their arguments misrepresent the objectives of this approach, that their criticisms of the role of the *H* coefficient in the procedures are irrelevant or erroneous, and that they fail to distinguish the inherent requirements (and limitations) of general nonparametric models and procedures from those of parametric ones. It is concluded that Mokken's procedures provide a useful tool for researchers in the social sciences who wish to construct and evaluate tests for measuring theoretically meaningful latent traits while avoiding the strong parametric assumptions of traditional item response theory.

In their article, Roskam, van den Wollenberg, and Jansen (1986, henceforth referred to as RWJ) criticize the nonparametric approach to constructing and evaluating tests based on binary items proposed by Mokken (1971, henceforth M) and more recently described by Mokken and Lewis (1982, henceforth ML). Their criticism and our response are the most recent additions to a discussion that has been carried out in the Netherlands over the last few years: Jansen (1982a, 1982b, 1983), Jansen, Roskam, and van den Wollenberg (1982, 1984), Molenaar (1982a, 1982b, 1982c), Sijtsma (1984, in press-a), and Sijtsma and Prins (in press).

The main points in our rejoinder to RWJ's critique may be summarized as follows:

1. Their arguments seem to be based on a selective reading and perception of the background and conceptual development of the theory and procedures in M, leading to a misrepresentation of the relevant objectives;
2. Their criticisms of the role of the *H* coefficient in the implementation of these procedures appear to be irrelevant or erroneous;
3. Their failure to distinguish the inherent requirements (and limitations) of general nonparametric models and procedures from those of parametric models leads to confusion in comparing Mokken's procedures with those appropriate for the Rasch (one-parameter logistic) model for binary items.

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 10, No. 3, September 1986, pp. 279–285*

*0146-6216/86/030279-07$1.60*

## Theory and Procedures

In their discussion of "scalability" and "the Mokken scale," RWJ seem to confuse two issues treated separately in M and ML: the question of the item response models under consideration, and the problem of selecting and evaluating sets of items relative to these models. (For ease of accessibility, references which follow will be largely limited to the more recent ML, although essentially the same treatment may be found in much greater detail in M.)

Two nonparametric item response models are discussed in ML (pp. 418–419), neither of which is referred to as "the Mokken scale." In the first of these, item response functions are assumed to have the property of monotone homogeneity (i.e., monotonically increasing in ability); in the second, these functions are assumed to be doubly monotone (i.e., additionally, non-intersecting). Contrary to the implications in RWJ, there is no confusion, at least not on the part of the present authors, regarding these two models. The former is seen as the more basic of the two, incorporating a minimal requirement for unidimensional measurement of persons. The latter possesses an extra feature, namely a unique difficulty ordering of items for all persons, which may be desirable for certain purposes. One of these purposes, namely obtaining a posterior distribution for the "ability class" to which a person belongs, is outlined in ML.

The second general issue, that of selecting and evaluating items, is treated separately in ML (pp. 422–425). When a "scale" (or "test") is defined (ML, p. 422) as a set of items which are positively correlated, with every item having an $H_i$ coefficient greater than or equal to a given positive constant, this should not be seen as identifying a particular model, but rather as describing, apart from sampling errors, the result of the scaling procedures referred to in ML. In this sense, it may be thought of as an operational definition. Incidentally, the problem of sampling errors in this context is given extensive attention in M, a point not noted in RWJ.

As always when models are used, not every item set which satisfies these operational requirements must conform to the model of monotone homogeneity, nor must every item set for which monotone homogeneity holds satisfy these requirements for every population distribution of abilities. This does nothing to alter the fact that the procedures proposed by Mokken and summarized in his definition of a scale are directed toward the selection or evaluation of a set of monotonely homogeneous items. Specifically, the requirement of positive inter-item correlations provides a necessary condition for monotone homogeneity, and "sufficiently high" values of the $H_i$ coefficient serve the objective of obtaining "sufficiently steep" item response functions, relative to the distribution of ability in the population under study. This latter requirement will be discussed below.

From the results reported in Part II of M, and in a wide variety of studies carried out since then (one of which is summarized in ML, p. 425), it can be seen that these procedures have allowed researchers to produce tests which have proved valuable from both practical and substantive theoretical points of view. They have also allowed critical evaluation of existing tests, including the identification of nonmonotonic items, as in the example reported by M (pp. 234–237).

Finally, it should be noted that the procedures summarized in Mokken's definition are not directed toward the second item response model mentioned above, which assumes doubly monotone response functions. As discussed in ML (p. 424), if this model is of interest to the researcher, the procedures may be expanded to include checks of necessary conditions for double monotony. For additional possibilities, see Molenaar (1982b, 1982c) and Schriever (1985).

## Coefficient *H*

Given that RWJ's critique of Mokken's procedures is largely directed toward the role played therein by Loevinger's coefficient $H$ and Mokken's variation $H_i$, it seems appropriate to respond to their contentions on this issue in some detail. In their discussion of the $H$ coefficient, RWJ rightly emphasize with their

quotation from M (p. 149) its dependence on the order of the *global* population difficulties of the items, but they wrongly infer from this quote that the assumption of double monotony is implied in the use of $H$. When double monotony holds, the order of the population difficulties will be the same for every distribution of abilities, and will then reflect the uniform ordering of the *local* difficulties for persons. The intention of the quoted passage was to indicate that this may be a desirable state of affairs in some applications. With monotone homogeneity alone, however, it would be possible for two different population distributions of ability to show differences in the ordering of global item difficulties. Nonetheless, monotone homogeneity would be indicated by a nonnegative value of $H$ in both cases. As already stated, $H$ and $H_i$ cannot identify items that are not doubly monotone with respect to the other items. Additional criteria are available for that purpose. Consequently, only criticisms of $H$ with respect to the monotone homogeneity model will be further considered.

RWJ repeatedly emphasize that "scalability coefficients" should not be sensitive to properties of an item set or of the population distribution of ability which are not related to the item response model of interest, in the present case that of monotone homogeneity. In specific cases the value of this normative statement should be assessed by considering whether or not this sensitivity is undesirable.

Three factors are discussed by RWJ with respect to their influence on $H$. These are the slopes of the item response functions, the distances between the item location parameters, and the population variance of ability: $H$ is said to be an increasing function of each, when the other two are held constant. It may be noted that none of these three is well-defined in a nonparametric framework where arbitrary monotonic transformations of the ability scale are allowed. Even in a parametric framework such as that of Birnbaum's two-parameter logistic model, the item slopes, inter-item distances, and ability variance are determined only up to a scalar factor: Multiplying all abilities by a positive constant $c$ has the effect of multiplying the ability variance by $c^2$, the inter-item distances by $c$, and the slopes by $1/c$.

Restricting attention to this parametric framework, and keeping in mind the relationship among the three factors, what are the more specific results? When the population variance is zero, so is the value of $H$. RWJ see this as a shortcoming of $H$ when viewed as a scalability coefficient. This is exactly what should happen under the model assumption of local independence. Furthermore, for this case there is no relevant information available regarding nontrivial monotone homogeneity of the item set, and this is accurately reflected by the value of $H$. The items may be seriously nonmonotonic but, with no variability in the population of persons under study, it will be impossible to detect this state of affairs. More generally, as population variance decreases with fixed item characteristics, there is less information available which might reveal possible violations of monotone homogeneity, and smaller values of $H$ may be thought of as warning the researcher of this.

Inter-item distances play a somewhat different role. Decreasing distances while holding slopes and variance fixed will generally reduce $H$, but the value achieved when inter-item distances are zero is a function of the other two factors. Items which are further apart have the potential of providing relevant information regarding monotone homogeneity over a greater range of the ability continuum than items with similar locations. The extent to which this potential is realized depends on the size of the population variance in a given situation, and this is reflected in the combined sensitivity of $H$ to both inter-item distance and population variance (with item slopes fixed).

Finally, given the relationship among the three factors, it is possible to describe the effect of the third—item slopes—in terms of the other two. Reducing slopes while fixing variance and distances is equivalent to holding the slopes fixed and reducing the other two factors. Thus, the sensitivity of $H$ to slope may be thought of as sensitivity to variance and distance, which has been interpreted above as providing information regarding the degree to which the setup (consisting of item and population properties) allows an evaluation of the monotone model. Item response functions with steeper slopes provide more complete positive evidence for monotone homogeneity, and this is reflected in a higher value of $H$.

RWJ, having concluded that $H$ is not a satisfactory coefficient of scalability, go on to consider its

relationship to the reliability coefficient for the simple sum score in the context of classical test theory. Citing Molenaar and Sijtsma (1984), they warn that $H$ is not a measure of reliability, a conclusion also found in M (p. 150) and in Sijtsma (1984). In particular, RWJ are concerned that a monotone homogeneous item set with a high test reliability may not be judged suitable, due to a low value of $H$. Considering the simple case of equivalent items (i.e., items having identical item response functions), it may be shown (Molenaar & Sijtsma, 1984) that $H$ equals the inter-item correlation (which is constant for item pairs and equals the reliability of each item). Test reliability, on the other hand, is an increasing function of this correlation and of the number of items in the test, given by the Spearman-Brown formula. Thus low inter-item correlation (due, with fixed item slope, to a small population variance) and a large enough number of items will result in low $H$ and high reliability.

The evidence for monotone homogeneity of the item set in this example remains limited, regardless of the number of items, due to the limited range of abilities. This condition is reflected by $H$, but not by the test reliability. More generally, high reliability may be attained with seriously nonmonotonic items. It should be clear that the procedures proposed by Mokken are not directed merely at producing reliable tests, but rather tests for which relatively complete information is available which is consistent with monotone homogeneity of the component items.

It may be noted that M (pp. 142–147) also proposed methods for estimating test reliability based on the assumption of doubly monotone items. Since $H$ and test reliability express different properties of a given setup, these reliability estimates are intended to be used in addition to $H$. For further theoretical developments with respect to reliability estimation based on Mokken's approach, see Sijtsma and Molenaar (in press).

RWJ further comment on the result of item selection based on $H$, claiming that only well-spaced items with steep slopes will survive. Molenaar's (1982a) numerical examples suggest that, if the population variance of abilities is large enough, this will not be the case. Sijtsma and Prins (in press) have extensively studied this selection process by means of simulated data. They considered two questions: (1) In what order are items selected from a monotonely homogeneous or doubly monotone set of items? and (2) How does $H$ behave during the selection process?

In the simulations by Sijtsma and Prins, Rasch items were used, with locations uniformly spaced between $-4$ and 4 on the conventional Rasch scale, and the mean of the (normal) ability distribution fixed at zero. Population standard deviation and the number of items available in the range were allowed to vary. Some relevant results of this study with respect to doubly monotone item sets may be summarized as follows:

1. Usually, the most extreme items are the first two to be selected, while the third item selected lies halfway between the extremes, and is thus of intermediate difficulty. This last result was also mathematically shown to be true (i.e., without the help of simulated data).
2. During the selection of the first five or six items, $H$ decreases relatively quickly; but when additional items are selected, $H$ tends to stabilize at a value dependent on the population standard deviation. A standard deviation of 1.5, for instance, resulted in an "asymptotic" $H$ value of .43. Under these conditions, where $H$ is relatively unaffected by test length, long tests can be constructed (see also Molenaar, 1982a, and Wierda, 1984).

This sort of behavior is exactly what would be expected from a procedure aimed at selecting one-dimensional items.

The example of equivalent items, used earlier to illustrate the distinction between $H$ and reliability, is also relevant in this context. If the population variance is sufficiently large to produce an inter-item correlation which exceeds the criterion value for $H$ chosen by the researcher, then the value of $H$ for a test consisting of these items is independent of test length. Additional results of Sijtsma and Prins' (in press) study regarding sets of monotone homogeneous items with varying slopes are not directly relevant to RWJ's critique and thus are not discussed here.

Finally, the work of Molenaar (1982a, 1982b, 1982c, in press) with respect to the role of $H$ in Mokken's procedures should be mentioned, as well as that of Sijtsma (in press-b), who studied a person coefficient based on $H$, intended to detect persons having aberrant item response patterns. An exact significance test for $H$ was derived in Molenaar (1982c), where a form of $H$ was proposed for multicategory items (see also Molenaar, in press).

## Parametric Versus Nonparametric Models

Nonparametric models, due to their generality of application, offer fewer possibilities of estimation and inference than specific parametric models. In place of parametric model estimation, more summary methods and statistics like coefficient $H$ must be relied upon. RWJ seem to ignore this and tend to approach the whole problem of selecting, evaluating, and using sets of items from an inherently parametric viewpoint, the more so because they manifest themselves as rather committed to the Rasch model.

This state of affairs, which has been noted earlier at various points, is especially clear in RWJ's discussion of ability estimation, where they state: "In the absence of a specified probabilistic response model connecting latent trait and manifest behavior, there is no obvious way of estimating the person parameter" (p. 273). Here "obvious" presumably refers to maximum likelihood estimation as it is applied in the Rasch and other parametric latent trait models. The fact that ML (pp. 426–429) described a nonparametric Bayesian approach to estimating latent ability classes for the special case of doubly monotone items is completely ignored (see also Lewis, 1983). Instead, RWJ discuss in some detail estimation properties which could only conceivably be achieved in the context of a parametric model, and contrast these with the more limited results available for the simple sum score in a nonparametric setting. Even here, their treatment is selective. Thus, for clarity, the results obtained in M (pp. 128–129, 138–142) and summarized in ML (p. 426) are repeated here. For any monotonely homogeneous set of items, and any population distribution of abilities having positive variance,

1. The latent ordering of the persons is, apart from sampling variation, given by that of their observed simple scores. The precision of this estimated order depends on the number of items (test length) and the item sampling model (item selection procedure).
2. The simple score is positively correlated with the value of the latent ability.

Consequently, for monotonely homogeneous sets of items, the observed simple score can be used as a nonparametric measurement of the latent trait. The fact that these scores are highly correlated with Rasch estimates of ability, for which they serve as sufficient statistics, may be seen as an additional advantage of this choice if the Rasch model should happen to hold for the item set.

## Conclusions

RWJ's critique of the nonparametric procedures proposed by Mokken is apparently inspired by their commitment to the Rasch model and associated procedures for selecting and evaluating unidimensional sets of items. It may, therefore, be useful to conclude by giving some indication of the relation of applications of these two approaches. In particular, it will be noted that the Rasch model is not as population-independent as is claimed on the basis of the technical separability in the likelihood of its parameters and corresponding sufficient statistics. Rasch scaling may be thought of, in the context of the two-parameter logistic model, as selecting a subset of items having identical slopes and rescaling these slopes to unity; the choice of items then determines both the variance of the population and the distance between items.

To illustrate this point, the reader is asked to imagine a small academic world, for instance the Low Countries, investigating for its population a unidimensional ability scale for which (unknown to anyone)

only two-parameter logistic items can be found. Three academic teams, let us say from Nijmegen, Groningen, and Amsterdam, are busy selecting items and constructing tests to measure this important ability, all using procedures associated with the Rasch model. As always and everywhere, science starts with judgment, which in this case requires a prior selection of items to present to respondents. It so happens that residents of Groningen are rather strict, so that the Groningen team's prior selection is inspired by an implicit predilection for steep items. On the other hand, the temperament of Nijmegen induces its team to be biased towards flat items. The unprincipled Amsterdammers prove to be inclined to something in between. After a while, all three teams happily report splendid fits to the Rasch model. Comparison of their results, however, shows differences in the assessment of the dispersion of this ability in the Dutch population, Groningen showing it to be rather dispersed, Nijmegen insisting on a homogeneous population, and Amsterdam, of course, taking an intermediate position.

An application of the nonparametric procedures described in M to the combined set of all three centers leads to the following results. The Nijmegen set is rejected in all cases. With a high value of the scale-defining constant for $H$ (say .50), only the Groningen set is selected, and a subsequent check for double monotony leaves that set intact. With a somewhat lower value of the constant (say .30), both the sets of Groningen and Amsterdam are selected, but the subsequent analysis of double monotony indicates violations of this property. Inspection results in a partition into two separate doubly monotone sets, corresponding to the original sets, each with a different constant slope. This leads to the final question of which set to use for future Rasch scaling.

## References

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*. The Hague: Rijks Psychologische Dienst.

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

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.

Lewis, C. (1983). Bayesian inference for latent abilities. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 224–251). San Francisco: Jossey-Bass.

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. New York-Berlin: Walter 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 [Another weighing of the Mokken scaling procedure]. *Tijdschrift voor Onderwijsresearch, 7,* 172–181.

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

Molenaar, I. W. (in press). Een vingeroefening in item response theorie voor drie geordende antwoordcategorieën [An exercise in item response theory for three ordered answer categories]. In G. F. Pikkemaat & J. J. A. Moors (Eds.), *Liber Amicorum Jaap Muilwijk*. Groningen: Econometrisch Instituut.

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

Roskam, E. E., van den Wollenberg, A. L., & Jansen, P. G. W. (1986). The Mokken scale: A critical discussion. *Applied Psychological Measurement, 10,* 265–277.

Schriever, B. F. (1985). *Order dependence*. Amsterdam: Centrum voor Wiskunde en Informatica.

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

Sijtsma, K. (in press-a). Another note on the usefulness

of Mokken scaling. *Psychologische Beiträge*.
Sijtsma, K. (in press-b). A coefficient of deviance of response patterns. *Kwantitatieve Methoden*.
Sijtsma, K., & Molenaar, I. W. (in press). Reliability of test scores in nonparametric item response theory. *Psychometrika*.
Sijtsma, K., & Prins, P. M. (in press). Itemselectie in het Mokken model [Item selection in the Mokken model]. *Tijdschrift voor Onderwijsresearch*.
Wierda, F. W. (1984). *Mokkenschaalanalyse: Bijdrage aan een discussie* [Mokken scale analysis: Contribution to a discussion]. Unpublished master's thesis, University of Groningen.

## Acknowledgments

*The authors express their thanks to Ivo Molenaar for his useful comments on an earlier version.*

## Authors' Addresses

Send requests for reprints or further information to Robert J. Mokken, Centraal Bureau voor de Statistiek, Postbus 959, 2270 AZ Voorburg, The Netherlands; Charles Lewis, Educational Testing Service, Princeton NJ 08541, U.S.A.; or Klaas Sijtsma, Vakgroep Arbeids en Organisatiepsychologie, Vrije Universiteit, 1081 HV Amsterdam, The Netherlands.