Another note on the usefulness of Mokken scaling

Sijtsma, K.

Published in:
Psychologische Beiträge

Publication date:
1986

Psychologische Beiträge, Band 28, 1986, S. 425–432

Another note on the usefulness of MOKKEN scaling[1]

K. SIJTSMA[2]

Summary, Zusammenfassung, Résumé

In this paper two topics are treated which are inspired mainly by a recent paper in this journal by JANSEN, ROSKAM and VAN DEN WOLLENBERG (1984). First, the interpretation of LOEVINGER's coefficient H as a measure of reproducibility of subjects' item responses is discussed. Second, the combination of coefficient H and double monotonicity into MOKKEN's scaling procedure is treated. Finally, it is concluced that coefficient H is a useful tool in test construction.

Ein weiterer Hinweis auf die Brauchbarkeit des Skalierungsverfahrens von MOKKEN

In diesem Beitrag werden zwei Gegenstände besprochen, die besonders von einem Artikel von JANSEN, ROSKAM und VAN DEN WOLLENBERG (1984) in dieser Zeitschrift inspiriert sind. Erstens wird die Interpretation des Homogenitätskoeffizienten H von LOEVINGER als Maß für die Reproduzierbarkeit von Antwortmustern diskutiert. Zweitens wird die Kombination des Koeffizienten H und die doppelte Monotonie im Skalierungsverfahren von MOKKEN besprochen. Es stellt sich heraus, daß der Koeffizient H ein brauchbares Verfahren für die Konstruktion psychologischer Tests darstellt.

Remarques complémentaires sur l'efficacité de la technique scalométrique de MOKKEN

Dans cet article on traite deux sujets qui sont inspirés par un article de JANSEN, ROSKAM et VAN DEN WOLLENBERG (1984) recemment publié dans ce même journal. D'abord l'interpretation du coefficient H de LOEVINGER comme une mesure de la reproduction des réponses est discutée. En suite la combinaison du coéfficient H et la monotonie double dans de MOKKEN sont traites. On finit par concluire que le coéfficient H est un instrument utile pour la construction des tests la technique scalomêtrique. (W. Lohr et S. Mönikheim)

1 The author is grateful to Charles Lewis and Ivo W. Molenaar for their critical comments on an earlier draft of this paper.

2 Klaas Sijtsma, Vakgroep Arbeids- en Organisatiepsychologie, Vrije Universiteit, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands.

MOKKEN尺度の有効性についての追加資料

本論文は，JANSEN,ROSKAM and VAN DEN WOLLENBERG(1984)による本誌最近の論文によって主として提起されたふたつの論点を扱う．第一に，被験者の項目反応の再現可能性の測度としてのLOEVINGER のH係数（Homogenitätskoeffizient）の解釈について論ずる．第二にH係数と二重単調性を組合せてMOKKEN尺度の方法に用いることについて扱う．最後に係数Hがテストの作成に有効な道具であるという結論が出される．

（山下清美　K. Yamashita ）

Introduction

The present paper contains a reaction to a paper by JANSEN, ROSKAM and VAN DEN WOLLENBERG (1984) in this journal, which criticizes the role of coefficient H (LOEVINGER, 1948) in the context of the MOKKEN (1971) model. The MOKKEN model has e.g. been discussed briefly by HENNING (1976), STOKMAN and VAN SCHUUR (1980) and MOKKEN and LEWIS (1982), and the interested reader is referred to these papers. Coefficient H takes an important place in MOKKEN's item response model, and its role has become the subject of an extensive debate in several journals. We restrict our attention mainly to the discussion in Psychologische Beiträge (JANSEN, 1982; JANSEN et al., 1984; SIJTSMA, 1984).

In order to prevent a tedious repetition of arguments, this paper is confined to two topics which are central in the recent paper by JANSEN et al. (1984). They are, in order of treatment:

1. Reproducibility of item responses from a subject's test score.
2. Combining coefficient H and double monotonicity into one scaling procedure.

Reproducibility of a subject's item scores

The well-known GUTTMAN (1944, 1950) model is characterized by the property of reproducibility of the item scores given a subject's raw score on the test. Reproducibility means that given a raw score x on a test consisting of k dichotomously scored items, the x easiest items have been answered positively, whilst the k − x hardest items have been answered

negatively. Consequently there are only k + 1 item score patterns in accordance with the GUTTMAN model.

In empirical applications the deterministic GUTTMAN model usually does not hold. The degree to which empirical data deviate from the GUTTMAN model can be expressed by means of e.g. LOEVINGER's (1948) coefficient of scalability H. The definition and properties of H have been discussed elsewhere (e.g. LOEVINGER, 1948; MOKKEN, 1971; JANSEN, 1982; SIJTSMA, 1984), and they will therefore not be repeated here. MOKKEN (1971, p. 59) has concluded that H possesses the best properties of all the coefficients of GUTTMAN scalability which have been proposed.

According to Sijtsma (1984, p. 430), a large value ($.30 \leqslant H \leqslant 1$) of H means that subjects having equal test scores often answered the same items correctly. JANSEN et al. (1984, p. 726, 727) have responded to this statement by presenting an example in which a large value of H can go together with a modest degree of reproducibility, thus claiming to show that H is not a proper measure of reproducibility. Their example entails two subjects denoted by v and w, having the same position on a latent trait ξ, which implies that their attribute parameters are equal: $\xi_v = \xi_w$. Their response behaviour is described by means of the RASCH model. In their example, JANSEN et al. calculate the probability that v and w have the same score on a binary item i, and conclude that the minimum probability equals .50. In the RASCH model this is true when $\xi_v = \xi_w = \delta_i$, where δ_i is the item difficulty measured on the same scale as the latent attribute. As no assumptions with respect to the distributions of subjects and items have been made, JANSEN et al. claim that one can construct examples in which H is large but reproducibility according to their definition is modest.

With respect to the example of JANSEN et al. (1984), the following remarks seem to be justified.

First, when the GUTTMAN model holds, each score pattern across the items in the test can be reproduced. Consequently, the individual item scores of two persons v and w with $\xi_v = \xi_w$ are equal with probability one. This probability is defined across hypothetical replications of presenting the same items to the same persons. In practical applications the GUTTMAN model is too restrictive, so that some score patterns can be reproduced, but others can not. In that situation the average probability across pairs of persons (with $\xi_v = \xi_w$) of identical scores on a single item, is smaller than one. Occasionally, the probability for a single pair of persons may be quite small as JANSEN et al. have shown in their example. At the same time coefficient H may be quite large, however, e.g. $H > .30$ (MOKKEN, 1971, p. 185).

The explanation for this seemingly paradoxical situation is that JANSEN et al. use a definition of reproducibility which is not suited for coefficient H. They express the degree of reproducibility as the probability that two persons with identical attribute parameters have the same item score pattern. Coefficient H, however, does not express a probability, nor is it based on probabilities as defined by JANSEN et al. (1984). Given the item popularities, it is a descending linear function of the number of error patterns in a sample of data (e.g. MOLENAAR and SIJTSMA, 1984). Furthermore, it is based on all k items in the test and the total group of persons tested. When evaluating coefficient H by means of a definition of reproducibility on which it is not based, it should not be too surprising to find seemingly contradictory results.

Second, on further studying the definition of reproducibility as presented by JANSEN et al., one may reach the conclusion that it is rather restrictive in considering only identical response patterns. Carrying their line of reasoning a bit further, it seems to be reasonable that two subjects with equal attribute parameters may differ with respect to one or more item responses. This is fully in accordance with the probabilistic nature of the response process in e.g. the RASCH model. It implies that two subjects with equal attribute parameters need not necessarily have the same item response patterns.

In Table 1 the probabilities are given that two subjects ($\xi_v = \xi_w$) have response patterns which differ with respect to m items ($0 \leqslant m \leqslant k$). The column denoted by m = 0 shows a few probabilities of reproducibility as defined by JANSEN et al.

Table 1

Rasch probability that two subjects v and w ($\xi_v = \xi_w = 0$) have response patterns on k items that differ in m positions

Note: $k = 2$: $\delta_1 = -0.5; \delta_2 = 0.5$
$k = 3$: $\delta_1 = -1.0; \delta_2 = 0.0; \delta_3 = 1.0$
$k = 4$: $\delta_1 = -1.5; \delta_2 = -0.5; \delta_3 = 0.5; \delta_4 = 1.5$

k \ m	0	1	2	3	4
2	.28	.50	.22	–	–
3	.18	.42	.32	.08	–
4	.14	.37	.34	.15	.02

Furthermore, it is possible to derive from the table the probability that v and w have at most m different responses, or phrased positively, at least k − m identical responses. When e.g. k = 4 the probability that v and w have at least two identical responses equals .14 + .37 + .34 = .85 for the parameters chosen. It is to be expected on theoretical grounds that the larger the differences between item difficulties and the larger the item discriminations, the smaller the probability that response patterns differ on many positions. This means that reproducibility, according to both the usual and alternative definitions, increases under such circumstances. A problem remains where to put a borderline for acceptably versus unacceptably large differences between two response patterns, given the same subject parameter for both. This problem requires a deeper study, which is not the purpose of the present paper.

Alternative approaches to evaluating score patterns. Within the context of probabilistic item response theory, approaches to the problem of evaluating response patterns of individual persons are presented by e.g. LEVINE and DRASGOW (1982) and TATSUOKA (1984). Deviations of individual subjects' response patterns with respect to the perfect GUTTMAN patterns can be expressed by many measures, see e.g. HARNISCH and LINN (1981). It should be noted, however, that coefficient H is an overall measure, that does not allow the detection of individual subjects' aberrant response patterns. One may thus define coefficients based on a comparable definition of an error pattern as is used with respect to H, which are the subject counterparts (see e.g. CLIFF, 1977) of the H-coefficients for item pairs and for individual items with respect to the other items (MOKKEN, 1971, p. 148).

Reproducibility and double monotonicity

JANSEN et al. (1984, p. 732–734) object to a scaling procedure which combines two different objectives of measurement. In the case of MOKKEN scaling, these are a high degree of reproducibility and ordering persons and items according to the doubly monotonic model (or only ordering persons according to the monotonely homogeneous model). The authors correctly state that the two objectives can be contradictory, and for that reason reject MOKKEN scaling in its present (e.g. MOKKEN and LEWIS, 1982) or reformulated (SIJTSMA, 1984) condition.

One reason for disagreeing with this point of view is that reproducibility and double monotonicity can be checked independently. When finding that just one of these objectives is reached, the researcher nevertheless has obtained information on the usefulness of his/her test. Conse-

quently, measures can be taken in order to improve the test with respect to the other objective.

Another reason for objecting to the conclusion of JANSEN et al. is that in constructing a test, occasionally contradictory measurement objectives are often pursued. A test may e.g. comply perfectly with the RASCH model, but its information function values (e.g. LORD, 1980; also JANSEN et al., 1984, p. 732) may be low in the region of ξ where the population of interest is located. This means that measurement may be considered specifically objective (e.g. FISCHER, 1974), but at the same time it is hardly possible to discriminate among persons by means of their estimated parameters, $\hat{\xi}$. This situation may occur when e.g. the test is relatively hard or easy in the population under investigation, or when the test consists of a small number of items. Although RASCH-homogeneity and a high information function are not always both obtained, there seems to be no apriori reason why one should not pursue them when they are both considered desirable. In the same vein, pursuing reproducibility and double monotonicity as desirable measurement objectives within one scaling procedure should be understood.

Finally, JANSEN et al. (1984, p. 732; see also JANSEN, 1982, p. 104) seem to object to the fact that MOKKEN scaling combines procedures from classical psychometrics (coefficient H), and item response theory (double monotonicity). It is not necessary to distinguish in principle between the two approaches. Scaling persons and items according to e.g. the doubly monotonic model or the RASCH model neglects the possibly desirable objectives of reproducibility and reliability (e.g. WOOD, 1978; MOLENAAR and SIJTSMA, 1984; SIJTSMA, 1984; 1987). Consequently, in the context of item response theory response patterns across items are also studied for their atypicality, and the use of the information function is advocated in order to discriminate persons reliably. In the nonparametric monotonely homogeneous and doubly monotonic models these approaches are hardly applicable since the latent parameters ξ and δ cannot be estimated (see, however, MOKKEN and LEWIS, 1982). LOEVINGER's H and the classical reliability concept (MOKKEN, 1971) are useful tools in the context of these models, however.

Discussion

Many of the issues raised by JANSEN et al. (1984) have not been discussed in the present paper. Instead, two main topics have been treated. The first one is the definition of reproducibility as used by JANSEN et al. in connection to coefficient H. The second one is the combination into one scaling procedure of occasionally contradictory measurement

objectives which come from different approaches to psychological measurement.

There exist several measurement objectives, which often are not realized all at the same time in one set of data. It often depends on the application of the specific test which measurement ideals are stressed. By no means can a formal model, e.g. the RASCH model or the MOKKEN model, tell the scientist which measurement ideals he/she should pursue.

References

Cliff, N.: A theory of consistency of ordering generalizable to tailored testing. Psychometrika, 1977, 42, 375–401.

Fischer, G. H.: Einführung in die Theorie psychologischer Tests. Bern: Huber, 1974.

Guttman, L.: A basis for scaling qualitative data. American Sociological Review, 1944, 9, 139–150.

Guttman, L.: The basis for scalogram analysis. In: S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), Measurement and prediction. Princeton: Princeton University Press, 1950.

Harnisch, D. L., & Linn, R. L.: Analysis of item response patterns: Questionnable test data and dissimilar curriculum practices. Journal of Educational Measurement, 1981, 18, 133–146.

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

Jansen, P. G. W.: Homogenitätsmessung mit Hilfe des Koeffizienten H von Loevinger: Eine kritische Diskussion. Psychologische Beiträge, 1982, 24, 96–105.

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

Levine, M. V., & Drasgow, F.: Appropriateness measurement: Review, critique and validating studies. British Journal of Mathematical and Statistical Psychology, 1982, 35, 42–56.

Loevinger, J.: The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. Psychological Bulletin, 1948, 45, 507–530.

Lord, F. M.: Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Lawrence Erlbaum, 1980.

Mokken, R. J.: A theory and procedure of scale analysis. The Hague: Mouton, 1971.

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

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

Sijtsma, K.: Useful nonparametric scaling: A reply to Jansen. Psychologische Beiträge, 1984, 49, 95–110.

Sijtsma, K.: Reliability estimation in Mokken's nonparametric item response model. In: W. E. Saris, & I. N. Gallhofer (Eds.), Sociometric research Volume 1: Data collection and scaling. London: MacMillan, 1987.

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

Tatsuoka, K. K.: Caution indices based on item response theory. Psychometrika, 984, 49, 95–110.

Wood, R.: Fitting the Rasch model – A heady tale. British Journal of Mathematical and Statistical Psychology, 1978, 31, 27–32.