KM 28(1988)
pag 115-126

Mokken's approach to reliability estimation extended to multicategory items

I.W. Molenaar[1]
K. Sijtsma[2]

Abstract

In this paper, the method of reliability estimation in the doubly monotone model is extended to multicategory items. This extension is based on the theory for the case of dichotomous items as presented by Mokken (1971) and further developed by Sijtsma and Molenaar (1987).

---

1 Vakgroep Statistiek en Meettheorie FPPSW, Rijksuniversiteit Groningen Oude Boteringestr. 23, 9712 GC Groningen, tel. 050-636185

2 Vakgroep Arbeids- en Organisatiepsychologie, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, tel. 020-5485517.

Introduction

In this paper we deal with the problem of estimating the reliability of testscores, which are the unweighted sums of multicategory item scores. The number of items is denoted by k, and the number of answer categories of an item by m + 1. It is assumed that the number of answer categories is identical for all items in a given test.

The reliability method to be discussed is based on the nonparametric model of double monotonicity (Mokken, 1971) as defined for the case where items consist of at least three ordered answer categories (Molenaar, 1986). The method is an extension of a method proposed by Sijtsma and Molenaar (1987; also see Molenaar & Sijtsma, 1984; Sijtsma, 1987) for reliability estimation based on the model of double monotonicity for dichotomously scored items. One important property of this latter method is that it estimates the reliability without systematic bias. Classical methods, like e.g. Cronbach's alpha and Guttman's lambda-2, are systematic lower bounds to the reliability when considered in the same circumstances as the Sijtsma and Molenaar method, and when applied to a sample almost always underestimate the population reliability. The sampling variability of the three methods has about the same size. Unless a lower bound is the prime goal, it seems recommendable to use the method proposed by Sijtsma and Molenaar when the reliability within the model of double monotonicity is assessed by the test constructor.

It thus seems worth while to extend the reliability method to multicategory items. In this paper the theory of this extension is given. The extended method has already been implemented in the computer program MSP (Debets & Brouwer, 1986); it was applied to empirical test data by Debets, Sijtsma and Molenaar (1987). As the proposed method is based on the double monotonicity, it should not be applied when serious violations of this property are suspected. For checks of this property see Debets, Sijtsma and Molenaar (1987), and in more detail Molenaar (1986).

The doubly monotone model for multicategory items.

The model of double monotonicity is a nonparametric Item Response Theory (IRT), in which the Item Characteristic Curves (ICC's) are monotonely nondecreasing functions of the latent attribute and, moreover, they are not allowed to intersect. These two properties together determine the potential shapes of the ICC's. Since the ICC's are not defined by means of a parametric function, the model of double monotonicity is called nonparametric.

Central in the doubly monotone model for multicategory items is the notion of an item step. An item with m + 1 ordered answer categories is viewed as a sequence of m imaginary dichotomous item steps, which are ordered along the latent measurement scale. We denote the score on item i by X_i, and, furthermore, the score on step g by Y_{gi}, which indicates whether the item step is passed by the examinee. The item score X_i takes integer values ranging from 0 to m, while Y_{gi} can only take the values zero (failed) and one (passed). The relation between X_i and Y_{gi} is

$$X_i = \sum_{g=1}^{m} Y_{gi}. \qquad (1)$$

It may be noted that several item steps within the same item are dependent. If $Y_{gi} = 1$, then all preceding item steps are necessarily passed. If $Y_{gi} = 0$, then the subsequent steps are necessarily failed. It follows that the doubly monotone model for k multicategory items can not be reduced to the model of double monotonicity for km dichotomous item steps, because the model assumes that item (step) responses are locally independent.

Assume next that each person p (p = 1, ..., n) has a value ξ_p on the latent measurement scale, where the cumulative distribution function of ξ is denoted by $G(\xi)$. Furthermore, all item scores are assumed to be distributed independently given ξ. If the test measures only one psychological attribute, then local independence and unidimensionality of measurement coincide (Hambleton & Swaminathan, 1985).

The Item Step Characteristic Curve (ISCC) may now be defined as follows, denoting probabilities by π :

$$\pi_{gi}(\xi) = \text{Prob}(Y_{gi}=1|\xi) =$$

$$= \text{Prob}(X_i \geq g|\xi). \qquad (2)$$

The ISCC thus gives the probability that, given ξ, the item step score equals one, or that the item score equals or exceeds the value g. In the model of double monotonicity for multicategory items the ISCC's are monotonely nondecreasing, and do not intersect.

Integrating (2) across ξ yields the unconditional probability

$$\pi_{gi} = \int \pi_{gi}(\xi)dG(\xi) =$$

$$= \text{Prob}(Y_{gi}=1) = \text{Prob}(X_i \geq g), \qquad (3)$$

which is the proportion of persons in the population having an item step score $Y_{gi} = 1$. This means that these persons have items scores equal to $X_i = g, g+1, \ldots, m$. In the sample the proportion in (3) is estimated by means of

$$\hat{\pi}_{gi} = \sum_{h=g}^{m} n_{hi}/n, \qquad (4)$$

where n_{hi} denotes the number of persons having an item score $X_i = h$.

For items i and j, and answer categories g and h, we define the bivariate proportion

$$\pi_{gi,hj} = \int \pi_{gi}(\xi)\pi_{hj}(\xi)dG(\xi) =$$

$$= \text{Prob}(Y_{gi}=1, Y_{hj}=1) =$$

$$= \text{Prob}(X_i \geq g, X_j \geq h). \qquad (5)$$

Three cases are distinguished. First, the situation where $i \neq j$, implying that $\pi_{gi,hj}$ is an observable quantity. This proportion is estimated by means of

$$\hat{\pi}_{gi,hj} = \sum_{e=g}^{m} \sum_{f=h}^{m} n_{ei,fj}/n, \qquad (6)$$

where e and f denote category numbers, and $n_{ei,fj}$ denotes the number of persons who have item scores $X_i = e$ and $X_j = f$, respectively.

Second, we consider the case where $i = j$ and $g \neq h$, meaning that the same item is of interest in two independent administrations, but different item steps are considered:

$$\pi_{gi,hi} = \int \pi_{gi}(\xi)\pi_{hi}(\xi)dG(\xi). \tag{7}$$

In practice, independent replications are usually not available. Consequently, (7) must be approximated on the basis of a single administration of the test. The same is true for the third case, where $i = j$ and $g = h$:

$$\pi_{gi,gi} = \int \pi_{gi}^2(\xi)dG(\xi). \tag{8}$$

The univariate proportion in (3), as well as the observable bivariate proportions in (5) and the unobservable bivariate proportions in (7) and (8), are needed to estimate the reliability of the test score. First, we shall give a definition of the reliability of test scores based on an unweighted sum of polychotomous item scores. Then we go on by proposing methods to approximate the unobservable bivariate proportions in (7) and (8).

Reliability of test scores

The test score is defined as

$$X = \sum_{i=1}^{k} X_i = \sum_{i=1}^{k} \sum_{g=1}^{m} Y_{gi}. \tag{9}$$

We assume two stochastic processes to underlie the behaviour of a person taking the test. First, a person is randomly selected from the population characterized by $G(\xi)$. Second, given this person, a test score is randomly generated from his/her distribution of test scores across independent replications of the test. These assumptions lead to the well known variance decomposition

$$\sigma^2(X) = \sigma_\xi^2[E(X|\xi)] + E_\xi[\sigma^2(X|\xi)]. \tag{10}$$

The first term on the right can be interpreted as the true score variance, and the second term as the average error variance. Using this decomposition, the reliability $\rho_{XX'}$ is defined as

$$\rho_{XX'} = \sigma_{\xi}^2[E(X|\xi)]/\sigma^2(X). \tag{11}$$

The denominator can be estimated directly from empirical test data. In order to estimate the numerator, we rewrite it in terms of the item steps. This can be accomplished as follows. Using (2), the term within brackets in the numerator of (11) can be written as

$$E(X|\xi) = \sum_{i=1}^{k} \sum_{g=1}^{m} \pi_{gi}(\xi). \tag{12}$$

Substitution of (12) in the numerator yields

$$\sigma_{\xi}^2[E(X|\xi)] = \sigma_{\xi}^2[\sum_{i=1}^{k} \sum_{g=1}^{m} \pi_{gi}(\xi)] \quad . \tag{13}$$

The right hand side of (13) can be expanded noting that it is the variance of an unweighted linear combination. The expanded version of (13) equals

$$\sigma_{\xi}^2[E(X|\xi)] = \sum_{i=1}^{k} \sum_{g=1}^{m} \sum_{j=1}^{k} \sum_{h=1}^{m} \sigma_{\xi}[\pi_{gi}(\xi), \pi_{hj}(\xi)] =$$

$$= \sum_{i} \sum_{g} \sum_{j} \sum_{h} \{E_{\xi}[\pi_{gi}(\xi)\pi_{hj}(\xi)] - E_{\xi}[\pi_{gi}(\xi)]E_{\xi}[\pi_{hj}(\xi)]\}. \tag{14}$$

Using results in (3) and (5) yields

$$\sigma_{\xi}^2[E(X|\xi)] = \sum_{i} \sum_{g} \sum_{j} \sum_{h} (\pi_{gi,hj} - \pi_{gi}\pi_{hj}). \tag{15}$$

Substitution of this result in the reliability formula (11) yields

$$\rho_{XX'} = \sum_{i} \sum_{g} \sum_{j} \sum_{h} (\pi_{gi,hj} - \pi_{gi}\pi_{hj}) \; / \; \sigma^2(X). \tag{16}$$

As pointed out before, bivariate proportions where i = j should be approximated, since no repeated administrations of the same item in identical circumstances are available. In the next section it is explained how these approximations are obtained.

Approximations to bivariate proportions

Based on Mokken's (1971, p. 147) Method One of approximating unobservable bivariate proportions in the dichotomous case, Sijtsma and Molenaar (1987) have proposed an estimation method using the average of four distinct approximations. This method has led to a reliability estimate which is for practical purposes unbiased when the model of double monotonicity holds. In this section, this method is generalized to the multicategory case.

As a starting point, the bivariate proportions in the numerator of (16) are arranged in a km * km matrix, denoted by Π. Along the marginals of this matrix, the proportions π_{gi} from (3) are ordered according to increasing magnitude. Within the matrix, the position of the bivariate proportions $\pi_{gi,hj}$ corresponds to the marginals π_{gi} and π_{hj}. It can easily be shown that, given the ordering of the marginals π_{gi}, the rows and columns are monotonely nondecreasing when the model of double monotonicity holds.

First, we discuss the approximation of $\pi_{gi,hi}$, with $g \neq h$. Second, the case where $g = h$ is considered. We use the definition of $\pi_{gi,hi}$ in (7). Following the rationale presented by Mokken (1971) for the dichotomous case, one of the probabilities in the integrand of (7) is replaced by an approximation of this probability.

Except for ties, in the model of double monotonicity the ordering of the unconditional and conditional probabilities is identical. We use the adjacent probabilities of $\pi_{gi}(\xi)$ or $\pi_{hi}(\xi)$ in this order to approximate these probabilities. Approximations to proportions are denoted by $\tilde{\pi}$. Now, except when it belongs to the first or the last row or the first or last column of Π, $\pi_{gi,hi}$ can be approximated in four different ways. First, we consider the case where $\pi_{ej}(\xi)$ is the larger neighbour of $\pi_{gi}(\xi)$ or $\pi_{hi}(\xi)$, respectively. According to the logic of Mokken's Method One, both $\pi_{gi}(\xi)$ and $\pi_{hi}(\xi)$ can be approximated by means of their larger neighbour, after it is multiplied by a conveniently chosen constant. Insertion of this approximation for one of the probabilities in (7), and then integrating, yields a quantity which can be estimated from a single test administration. Applying Mokken's Method One, $\tilde{\pi}_{gi,hi}$ equals

$$\tilde{\pi}_{gi,hi} = \int \pi_{ej}(\xi)\pi_{gi}/\pi_{ej}\ \pi_{hi}(\xi)dG(\xi) =$$

$$= \pi_{ej,hi}\ \pi_{gi}/\pi_{ej}, \qquad \text{for } \pi_{gi} < \pi_{ej}; \qquad (17a)$$

$$\tilde{\pi}_{gi,hi} = \int \pi_{gi}(\xi)\ \pi_{ej}(\xi)\pi_{hi}/\pi_{ej}dG(\xi) =$$

$$= \pi_{gi,ej}\ \pi_{hi}/\pi_{ej}, \qquad \text{for } \pi_{hi} < \pi_{ej}. \tag{17b}$$

Second, we consider the case where $\pi_{fl}(\xi)$ is the smaller neighbour of either $\pi_{gi}(\xi)$ or $\pi_{hi}(\xi)$. This smaller neighbour can also be used to approximate the other ISCC's. The final results of Method One are

$$\tilde{\pi}_{gi,hi} = \pi_{fl,hi}\ \pi_{gi}/\pi_{fl}, \qquad \text{for } \pi_{gi} > \pi_{fl}; \tag{17c}$$

$$\tilde{\pi}_{gi,hi} = \pi_{gi,fl}\ \pi_{hi}/\pi_{fl}, \qquad \text{for } \pi_{hi} > \pi_{fl}. \tag{17d}$$

It can easily be checked that if one of the probabilities in (7) is either the smallest or the largest in the order of marginal probabilities of Π, there are only three approximations to $\pi_{gi,hi}$. If both are extreme probabilities or marginals, only two approximations are possible.

Besides the four approximations in (17a) up to (17d), four additional approximations are possible when the direction of the measurement scale is reversed, see Sijtsma and Molenaar (1987) for the dichotomous case. On the level of items this means that the items are scored in the opposite direction. One case will be given in detail.

In (17a), $\pi_{gi}(\xi)$ was approximated by the nearest ISCC $\pi_{ej}(\xi)$, where $\pi_{ej}(\xi) \geq \pi_{gi}(\xi)$ for all ξ because of double monotonicity:

$$\tilde{\pi}_{gi}(\xi) = \pi_{ej}(\xi)\ \pi_{gi}/\pi_{ej}. \tag{18}$$

Reversal of the scale direction yields

$$1 - \tilde{\pi}_{gi}(\xi) = [1 - \pi_{ej}(\xi)](1 - \pi_{gi})/(1 - \pi_{ej}), \tag{19}$$

which can be written as

$$\tilde{\pi}_{gi}(\xi) = \pi_{ej}(\xi)[1 - \pi_{gi}]/(1 - \pi_{ej}) + (\pi_{gi} - \pi_{ej})/(1 - \pi_{ej}). \tag{20}$$

Substitution for $\pi_{gi}(\xi)$ in (7) and integrating yields

$$\tilde{\pi}_{gi,hi} = \pi_{ej,hi}(1 - \pi_{gi})/(1 - \pi_{ej}) - \pi_{hi}(\pi_{ej} - \pi_{gi})/(1 - \pi_{ej}). \tag{21a}$$

Results for the other three cases are given without further derivations. These 'reversed' cases match (17b), (17c) and (17d) for the original scale:

$$\tilde{\pi}_{gi,hi} = \pi_{gi,ej}(1 - \pi_{hi})/(1 - \pi_{ej}) - \pi_{gi}(\pi_{ej} - \pi_{hi})/(1 - \pi_{ej}); \quad (21b)$$

$$\tilde{\pi}_{gi,hi} = \pi_{fl,hi}(1 - \pi_{gi})/(1 - \pi_{fl}) + \pi_{hi}(\pi_{gi} - \pi_{fl})/(1 - \pi_{fl}); \quad (21c)$$

$$\tilde{\pi}_{gi,hi} = \pi_{gi,fl}(1 - \pi_{hi})/(1 - \pi_{fl}) + \pi_{gi}(\pi_{hi} - \pi_{fl})/(1 - \pi_{fl}). \quad (21d)$$

We have in total eight approximations. Following Sijtsma and Molenaar (1987), the mean of these approximations is taken to be the final approximation which is inserted in the reliability formula (16). This mean is denoted by $\bar{\pi}_{gi,hi}$. If in a first or last row or column not all eight are available, then the final result is obtained by taking the mean of the available proportions only. Molenaar and Sijtsma (1984) showed that for the dichotomous case the mean across several approximations is only biased to a negligible extent.

It may be noted that the approximation to $\pi_{gi,gi}$ now is straightforward. Since the integrand of (8) only contains $\pi_{gi}(\xi)$, there are two approximations concerning the original scale, and two when the scale direction is reversed. Final approximations are obtained by taking the mean of the individual approximations.

Special Cases

For the dichotomous case, Sijtsma and Molenaar (1987) discuss several situations in which approximations to dichotomous bivariate proportions are problematic. Solutions to these problems are proposed.

In the computer program MSP (Debets & Brouwer, 1986) provisions have been made to meet similar difficulties in the polychotomous case. Following Sijtsma and Molenaar (1987), each approximation should lie in the interval

$$\pi_{gi}\pi_{hi} \leq \tilde{\pi}_{gi,hi} \leq \min(\pi_{gi}, \pi_{hi}). \quad (22)$$

Since the mean of all approximations is used to estimate the reliability, the program only checks whether this mean lies between the bounds in (22). If not, it is replaced by the appropriate bound. Furthermore, alternative approximation methods are used when π_{ji} or π_{hi}, or both, belong to a string

of identical proportions. In such cases, the choice of adjacent elements becomes problematic. Since the discussion of the solutions to this problem would take much space, we prefer to give only a brief ouline.

If, e.g., π_{gi} belongs to a string of three identical proportions, the other two proportions are regarded as the smaller and the larger neighbour, respectively, and the approximation procedure is applied straightforwardly. If there are, e.g., four identical proportions, π_{gi} has three neighbours. This means that $\pi_{gi}(\xi)$ is three times approximated in the 'original' case, as well as three times in the 'reversed' case. Assuming that π_{hi} does not belong to a string of identical proportions, the average $\bar{\pi}_{gi,hi}$ is based on a total of ten approximations. If a string consists of just two proportions, only one neighbour is considered, which is the proportion identical to π_{gi}.

Discussion

Dichotomization of multicategory item scores has at least two disadvantages. The first is that if the scores on a multicategory item have a skew distribution, dichotomization will often lead to one category in which most scores accumulate, while the other category is almost empty. The item thus has an extreme difficulty, and, consequently, it often correlates weakly with the total test score. The result is a badly discriminating item.

The second disadvantage of dichotomization is that the reliability of dichotomous items is usually less than that of multicategory items (Nunnally, 1978, p.595, 596). A short test consisting of multicategory items may have the same reliability as a longer test consisting of dichotomous items.

On the other hand, some items may violate the requirements of double monotonicity and a sufficiently high H coefficient in their polychotomous form, while showing more model conformity after dichotomization. It will then depend on validity and reliability considerations whether a shorter scale of polychotomous items is preferred to a longer version with dichotomous scoring. The availability of a more finely graded sum score may also influence the choice.

Some researchers may feel tempted to try different dichotomizations for each item, until one is found that is favorable as regards model conformity and/or avoids extreme difficulty values. The present authors prefer using the same dichotomization, based on substantive considerations, for all

items. The availability of a model for multicategory items, moreover, implies that there is frequently no need to dichotomize at all.

The formulation of the reliability problem for multicategory items further complements the polychotomous Mokken model as proposed by Molenaar (1982; 1986). The approach to reliability estimation as discussed in this paper, has been implemented in the computer program MSP (Debets & Brouwer, 1986) which also contains the other proposals by Molenaar. The dichotomous model simply is a special case of the polychotomous model, implying that both dichotomous and polychotomous date can be analyzed by MSP. Because of this feature the Mokken model is a flexible model which is easily applied to several kinds of data.

References

Debets, P. & Brouwer, E. (1986). User's Manual MSP. Publication nr. TC130. Amsterdam: Technisch Centrum FSW, Universiteit van Amsterdam.

Debets, P., Sijtsma, K. & Molenaar, I.W. (1987). Mokken-schaalanalyse voor meercategoriële items: het programma MSP. In L.Th. van der Weele (ed.), SSS87, Proceedings of the Third Symposium Statistical Software, Rekencentrum RUG, Groningen , p. 203-220.

Hambleton, R.K., & Swaminathan, H. (1985). Item Response Theory. Boston: Kluwer Nijhoff Publishing.

Mokken, R.J. (1971). A Theory and Procedure of Scale Analysis. The Hague: Mouton.

Molenaar, I.W. (1982). Mokken scaling revisited. Kwantitatieve Methoden, 8, 145-164.

Molenaar, I.W. (1986). Een vingeroefening in item response theorie voor drie geordende antwoordcategorieën. In: G.F. Pikkemaat & J.J.A. Moors (eds.), Liber Amicorum Jaap Muilwijk. Groningen: Econometrisch Instituut, p. 39-57.

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

Nunnally, J.C. (1978). Psychometric Theory. New York: McGraw-Hill.

Sijtsma, K. (1987). Reliability estimation in Mokken's nonparametric item response model. In: W.E. Saris & I.N. Gallhofer (eds.), Sociometric Research, Vol. 1: Data Collection and Scaling. London: MacMillan Press Ltd.

Sijtsma, K. & Molenaar, I.W. (1987). Reliability of test scores in nonparametric item response theory. Psychometrika, 52, 79-97.

Ontvangen: 03-11-1987
Geaccepteerd: 19-01-1988