Comparison of the Nonparametric Mokken Model and Parametric IRT Models Using Latent Class Analysis

Dato N. M. de Gruijter
Leiden University

A nonparametric Mokken analysis of test data generally results in the rejection of different items as misfitting than an analysis with a parametric item response theory model. This is due to differences between the methods of analysis employed. Croon (1991) demonstrated that the assumption of "double monotony" of the nonparametric Mokken model can be tested with a latent class analysis using the EM procedure. This allows a comparison of the Mokken model of double monotony and parametric item response models within the same framework. The Mokken model was compared with parametric models using simulated data. It was demonstrated that latent class analysis provides a consistent comparison of item response models. *Index terms: EM algorithm, item fit, item response theory, latent class analysis, model comparisons, Mokken model, Rasch model.*

The Rasch model and the three-parameter logistic model are examples of unidimensional item response theory (IRT) models for dichotomous items with monotonic item response functions (IRFs). In these models, the IRF—the probability of a correct response to an item as a function of a latent trait θ, $P_i(X = 1 | \theta)$—has a specific parametric form. Mokken (1971; Mokken & Lewis, 1982) proposed procedures for nonparametric IRT analysis. A nonparametric analysis is useful if parametric models are too restrictive in a particular application. A disadvantage of a nonparametric scale is that it does not allow strong inferences with respect to θ, for example.

The first model discussed by Mokken (1971) had monotonic IRFs. The second Mokken model, which sometimes has been referred to in the literature as the Mokken scale (Roskam, Van den Wollenberg, & Jansen, 1986), also specifies that the IRFs of any two items do not intersect. This property is called *double monotony.* The model can be viewed as a stochastic version of the Guttman scale. Rosenbaum (1987), Sijtsma (1988), and Sijtsma & Meijer (1992) discussed procedures for testing the property of double monotony.

Item selection within the context of Mokken scale analysis has been based on a coefficient of scalability that was criticized by Roskam et al. (1986; for a reaction see Mokken, Lewis, & Sijtsma, 1986). In a study using a Mokken analysis and a Rasch analysis (Meijer, Sijtsma, & Smid, 1990), some items fit the Rasch model but did not fulfill the requirements of the double monotony model. According to the authors, this result was theoretically impossible. They argued that the only explanation for these results was that the methods they used for evaluating item fit had different properties.

The advantage of comparing the Mokken model with parametric models within the same analysis framework is clear. Several authors have done preliminary work to provide this framework. For example, Mooijaart (1980) used a least squares procedure for the estimation of double monotonic response probabilities in a latent class analysis (LCA). Sijtsma (1988) imposed inequality constraints on the item test regressions, using total test score as a proxy for θ. Croon (1991) applied the EM algorithm to the latent class problem associated with the property of double monotony. Because other item

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 1, March 1994, pp. 27-34

0146-6216/94/010027-08$1.65

response models also can be treated as latent class models, it has become possible to compare the Mokken double monotony model with specific parametric models.

In this paper, latent trait analysis and LCA are introduced using the Rasch model. Then several item response models and the EM algorithm for estimating parameters of these models are discussed. Finally, the possibility of comparing these models within LCA is illustrated using a small simulation study.

Latent Trait and Latent Class Analysis Using the Rasch model

In the Rasch model, the probability of a correct response to item i as a function of θ is

$$P_i(\theta) = P_i(X = 1 | \theta) = \exp(\theta - b_i)/[1 + \exp(\theta - b_i)] \ , \qquad (1)$$

where b is the difficulty parameter for item i, and X is the person's scored response to an item.

Item parameters can be estimated using maximum likelihood methods. Assume that an n-item test is given to N examinees. In unconditional maximum likelihood estimation, b_i ($i = 1, \ldots, n$) and θ_r ($r = 1, \ldots, N$) are estimated simultaneously by maximizing

$$L = \prod_{r=1}^{N} \prod_{i=1}^{n} P_i(\theta_r)^{x_{ir}}[1 - P_i(\theta_r)]^{1-x_{ir}} \qquad (2)$$

with respect to the parameters, where $x_{ir} = 1$ when examinee r has answered item i correctly, and $x_{ir} = 0$ otherwise. The simultaneous estimation of item and θ parameters is statistically unsatisfactory. Andersen (1973) demonstrated the biasedness of this estimation method (see also de Gruijter, 1990; Divgi, 1986). In the Rasch model, the likelihood can be simplified: Total scores for examinees and items are sufficient for the estimation of the parameters. The sufficiency of total scores implies that at most $n - 1$ different θ values are estimated (examinees with 0 or perfect scores are eliminated from the analysis).

The item parameters of the Rasch model also can be estimated by conditioning on total scores. In the conditional maximum likelihood method, θs are not estimated. The method gives statistically satisfactory estimates, but it becomes computationally difficult with longer tests (Gustafsson, 1980).

The third estimation method is marginal maximum likelihood (MML). In MML estimation, θs are eliminated by integration over the θ distribution. The observed frequencies have a multinomial distribution with a likelihood proportional to

$$L = \prod \phi(\mathbf{x})^{N_x} = \prod_{r=1}^{N} \int \left\{ \prod_{i=1}^{n} P_i(\theta)^{x_{ir}}[1 - P_i(\theta)]^{1-x_{ir}} \right\} g(\theta) d\theta \ . \qquad (3)$$

In Equation 3, **x** is a response vector with item responses $x_i = 1$ and $x_i = 0$, and the product is over response vectors. The likelihood in Equation 3 is maximized with respect to the item parameters and parameters of the distribution $g(\theta)$, which belongs to a particular family of distributions. For example, $g(\theta)$ might be a normal distribution that has two parameters: the mean and the variance (Bock & Aitkin, 1981). For estimation purposes, the continuous θ distribution can be approximated by a discrete distribution with k θ values, $\theta_1 < \theta_2 < \ldots \theta_k$, and relative frequencies v_j ($j = 1, \ldots, k$).

It is also possible to consider the discrete distribution as a distribution in its own right (see also Langeheine & Rost, 1988) and to maximize the likelihood with respect to this distribution. De Leeuw & Verhelst (1986) showed that there is an important relationship between conditional maximum likelihood and MML with a special kind of discrete distribution in which k is approximately $n/2$. An analysis of response data with a discrete distribution is equivalent to a LCA (e.g., Hagenaars, 1990) in which the latent item probabilities satisfy the Rasch model restrictions.

In unconditional maximum likelihood estimation, θs are estimated simultaneously with item parameters. The estimated θs are given by a monotonic transformation of the total score scale. When item parameters are estimated with MML there are various θ estimators, such as the expected a posteriori estimator and the Bayes modal estimator (Bock & Aitkin, 1981). Macready & Dayton (1992) suggested incorporating costs of classification errors in a classification procedure on the basis of a LCA. In situations in which differentiation between examinees is needed on the basis of one test, total scores may be better to use: Generally the number of total scores considerably exceeds the number of latent classes in an analysis. In the Rasch model, total score is a sufficient statistic for θ. In other models, total score is not sufficient, but when expected total score is a monotonically increasing function of θ it is a useful indicator of θ level.

Estimating Item Parameters of Some Latent Class Models by the EM Algorithm

Let the θ continuum be approximated by k ordered latent classes $\theta_1 < \theta_2 < \ldots < \theta_k$ with relative frequencies v_j ($j = 1, \ldots, k$). Let the probability of a correct response to item i given latent class j be written as

$$p_{ij} = P_i(X = 1 | \theta_j) \ . \tag{4}$$

The property of double monotony can be defined in terms of a latent class model with k ordered categories ($j = 1, \ldots, k$). Let the n items ($i = 1, \ldots, n$) be ordered from difficult to easy. Then the Mokken double monotony model is defined as

$$\theta_j \leq \theta_{j'} \rightarrow p_{ij} \leq p_{ij'} \ , \tag{5}$$

and for items h and i

$$h \leq i \rightarrow p_{hj} \leq p_{ij} \ . \tag{6}$$

The Mokken double monotony model in Equations 5 and 6 is referred to as Model M.

The response probabilities for the Rasch model, which is a submodel of the Mokken double monotony model, have the parametric form

$$p_{ij} = \exp(\theta_j - b_i)/[1 + \exp(\theta_j - b_i)] \ . \tag{7}$$

The model in Equation 7 is referred to as Model R.

Another parametric submodel of Model M is

$$p_{ij} = c + (1 - c)\exp(\theta_j - b_i)/[1 + \exp(\theta_j - b_i)] \ , \tag{8}$$

where c is a common guessing parameter (i.e., a guessing parameter that is equal for all items). The model in Equation 8 is referred to as Model G. Model R can be viewed as a special case of Model G with $c = 0$. The b_is and the latent class parameters in Equations 7 and 8 are defined on a difference scale. In order to define the scale, one parameter should be set to a particular value, for example $\theta_k = 0$.

Suppose the n-item test has been administered to N examinees. The N response vectors, $\mathbf{x}_r$ ($r = 1, \ldots, N$), are obtained with item responses x_{ir}. For any of the above three models the log likelihood equals (apart from a constant):

$$\log L = \sum_{r=1}^{N} \log\left(\sum_{j=1}^{k} P_{jr} v_j \right) \tag{9}$$

with

$$P_{jr} = P(\mathbf{x}_r | \theta_j) = \prod_{i=1}^{n} p_{ij}^{x_{ir}}(1 - p_{ij})^{1-x_{ir}} \ . \tag{10}$$

The log likelihood is maximized with respect to the parameters. For Model M, these parameters are the latent class response probabilities with the restrictions given by Equations 5 and 6. For Model R, the n b_i parameters and $k - 1$ θ_j parameters must be estimated. For Model G, the c parameter must be estimated as well. The latent class population proportions v_j are parameters to be estimated under the restrictions $0 \leq v_j \leq 1$ and $\Sigma v_j = 1$, or they can be fixed. Bock & Aitkin (1981) also used a fixed latent distribution, with fixed θ_j as well as fixed v_j. A fixed distribution is, however, too restrictive when different models are compared because the fixed distribution might fit one model better than another model.

The EM algorithm can be used to estimate the parameters. First, assume that class membership of all N examinees is known. In that case the number of examinees in each latent class, N_j, and the number of examinees in each latent class that answered item i ($i = 1, \ldots, n$) correctly, N_{ij}, are known. N_j and N_{ij} are sufficient for the latent class probabilities, and log L can be rewritten as

$$\log L = \sum_{j=1}^{k} \sum_{i=1}^{n} [N_{ij} \log p_{ij} + (N_j - N_{ij}) \log (1 - p_{ij})] + C \ . \tag{11}$$

Maximization of log L with respect to the parameters is called the M step of the EM algorithm. Maximization with respect to the parameters—b_i, θ_j, and possibly c in the case of a parametric model—is relatively straightforward. For Model M, the probabilities p_{ij} are obtained that deviate from N_{ij}/N_j in a weighted least squares sense under the monotony constraints (Croon, 1991).

Each iteration in the EM algorithm consists of two steps: the M step and the E (expectation) step. Using Bayes' rule, the missing class membership is estimated in the E step on the basis of the parameter estimates from the M step. In other words, new values of N_j and N_{ij} are obtained as

$$N_j = \sum_{r=1}^{N} P_{jr} v_j \bigg/ \left(\sum_{l=1}^{k} P_{lr} v_l \right) \tag{12}$$

and

$$N_{ij} = \sum_{r=1}^{N} x_{ir} P_{jr} v_j \bigg/ \left(\sum_{l=1}^{k} P_{lr} v_l \right) \ . \tag{13}$$

When the latent class probabilities v_j are free, new values of these probabilities can be obtained as

$$v_j = N_j / N \ . \tag{14}$$

The iterations continue until convergence has been reached. Croon (1991) warned that the EM solution for Model M might give a local maximum. Good starting values are important.

A specific model, M_1, with likelihood L_1 can be evaluated with the likelihood ratio statistic

$$G^2(\mathrm{M}_1 | \mathrm{M}_0) = -2 \log(L_1/L_0) \ , \tag{15}$$

where M_0 is a model without restrictions, with likelihood L_0:

$$L_0 = \prod \phi(\mathbf{x})^{N_\mathbf{x}} \ , \tag{16}$$

where $\phi(\mathrm{x})$ is given by N_x/N for all observed response patterns. Given the correctness of the models, the statistic—under certain conditions—is approximately distributed as χ^2 with degrees of freedom equal to the difference between the number of free parameters in M_0 (the number of observed response patterns s with a maximum equal to $n^2 - 1$) and that in M_1; for Model M the number of free parameters is set equal to $n \times k - e$, where e is the number of active equality constraints.

In applications like the one discussed here, the number of cells, or response patterns, may be very large. As a result, there generally will be many empty cells or cells with an expected value less than 5. In practice, cells are combined and the degrees of freedom for M_0 are $s^* < s$. Agresti (1990) discussed some of the problems encountered with testing models under these conditions. It is perhaps best to use the outcomes heuristically when a choice between competing models is to be made.

Two hierarchical models M_1 and M_2 can be compared using

$$G^2(\mathrm{M}_2 | \mathrm{M}_1) = G^2(\mathrm{M}_2 | \mathrm{M}_0) - G^2(\mathrm{M}_1 | \mathrm{M}_0) \ . \tag{17}$$

When a model is rejected, a less restrictive model might be tested; one possibility is to test whether an increase in the number of latent classes is useful. The parametric models (Models R and G) used here are submodels of Model M because they are more restrictive models. They are not, however, hierarchically related to Model M: their sets of free parameters are not subsets of the set of free parameters of the Mokken model.

Individual IRFs also could be inspected. Model IRFs p_{ij} can be compared with their expectations on the basis of the model N_{ij}/N_j (Hoijtink, 1991) by computing residuals. The investigation of the residuals might lead to the detection of the least fitting items and possible causes for misfit of a model.

Comparison of Models Using LCA: An Example With Simulated Data

Method

Dataset 1 was generated with the following characteristics: (1) 5 latent classes with θs = –1.50, –.75, 0.0, .75, 1.50; (2) 2,000 examinees, 400 in each latent class; (3) 7 items with equally spaced location parameters b (1.20, . . . , –1.20); and (4) $c = .25$. Equation 8 (Model G) was used for generating the item scores.

The resulting 0-1 scores were analyzed with the three models in which the property of double monotony holds: Model M (Equations 5 and 6), Model R (Equation 7), and Model G (Equation 8). In each analysis, five latent classes with νs equal to .20 were used; therefore, the true latent distribution was used. For Model M, the empirical rank order of item difficulties was used to order the items from difficult to easy. For Model G, the starting value for c was set equal to .15, in order not to favor the true model with $c = .25$ over the Rasch model with $c = 0.0$. The estimation process was stopped when the maximum change in p_{ij} between iterations was less than .001.

Results

Table 1 shows the obtained values of G^2 for Dataset 1. Both Model M and Model R have high G^2 values in relation to the degrees of freedom for these models. Model G is a specific submodel of Model M and is acceptable. When Model G is the correct model, the more general Model M should be valid too. However, in the present situation the parametric model is the one to be preferred. Model G needs fewer parameters than the nonparametric Model M and it allows stronger inferences to be made, because it is parametric. In this particular case, Model R must be rejected; when Model R is compared to Model G, of which it is a submodel, it becomes obvious that the addition of a guessing parameter is needed.

Table 1
G^2 for Models M, G, and R, the Number of Free Parameters of the Model (n_f), and the degrees of freedom (df) for Datasets 1 and 2

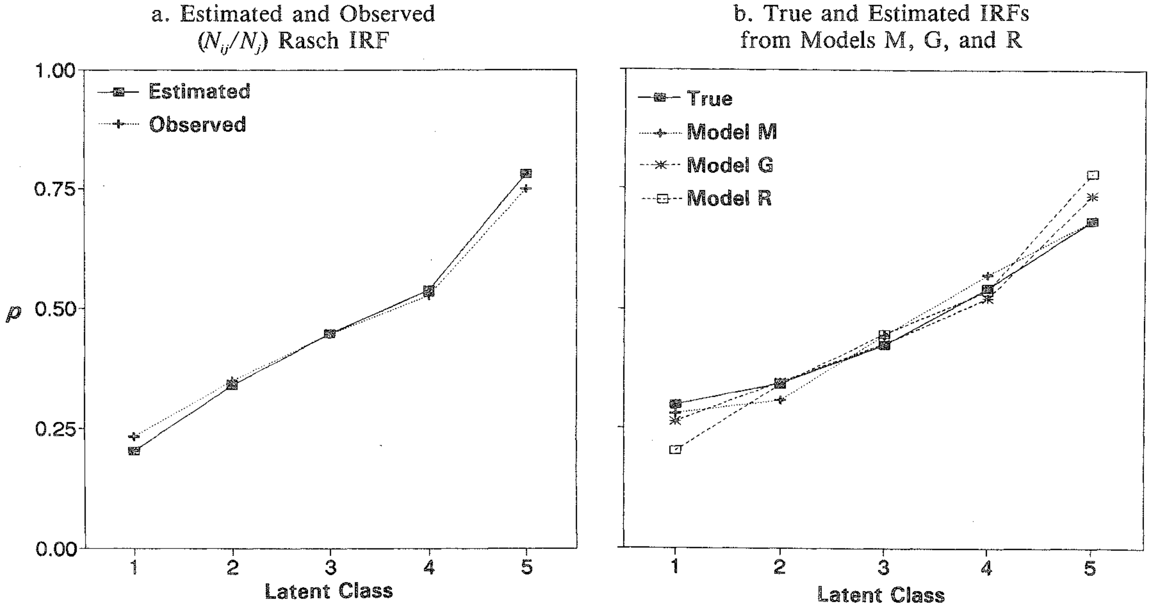
Model	G^2	n_f	df
Dataset 1			
M	48.842	29	32
G	59.355	12	49
R	69.497	11	50
Dataset 2			
M	52.356	23	42
G	62.984	12	53
R	73.720	11	54

Results from a Mokken LCA might be useful in selecting a parametric model if the assumption of a Mokken scale is not rejected. Inspection of the item probabilities might suggest a specific parametric model. It is easier, however, to estimate and compare competing parametric models directly.

Figure 1a shows Rasch data on Item 1, the most difficult item. In Figure 1b, data resulting from different models are compared to each other and to the proportions correct according to the true model. The various IRFs are easy to compare, because the latent class frequencies were equal for all analyses.

The figures show that the Model R IRF for Item 1 deviates from the other IRFs at the lowest θ level; the IRF fails to reproduce the effect of the guessing parameter. Because the Rasch IRF tries to fit the data at low θ levels, it also fails at the highest θ level. The absolute difference (D) between

Figure 1
IRFs for the Most Difficult Item


a. Estimated and Observed (N_{ij}/N_j) Rasch IRF

b. True and Estimated IRFs from Models M, G, and R the model p_{ij} and the values N_{ij}/N_j, averaged over latent classes, was .0167. For Model G the value of D was .0058. The lowest value was obtained in Model M: $D = .0031$.

For the other items, D was lower in the case of Model R. For all items, the lowest D was obtained with Model M, the least restrictive model. The more parsimonious Model G was also adequate, which makes the diagnostic value of the Mokken analysis small in this particular case.

Dataset 2 was based on 11 latent classes with equally spaced θs, ranging from $\theta = -2.50$ to $\theta = 2.50$, and class sizes equal to 20, 60, 140, 240, 340, 400, 340, 240, 140, 60, 20. This distribution is a rough approximation to the normal distribution. The other generating parameters were the same as those used for Dataset 1. The same analyses were conducted—the imposed latent class structure (with 5 classes) differed from the true distribution (with 11 classes). The G^2 values also are given in Table 1. For this dataset, Model M gave satisfactory results. The results for the other two models were essentially similar to those for Dataset 1.

Discussion

Within the framework of LCA it is possible to compare the adequacy of several parametric IRT models and that of the nonparametric model with the property of double monotony. The adequacy of the nonparametric model and the other models can be evaluated using a unified approach.

When a LCA is done, an analysis of item fit for a nonparametric model can be based on the same method as the item fit analysis for other models. Therefore, with respect to item fit, the difference in the approach between nonparametric and parametric approaches decreases.

In the study by Meijer et al. (1990), items for the Mokken scale were selected using the scalability coefficient. For the Rasch model, items were removed on the basis of item fit. This resulted in a theoretically impossible discrepancy between the Mokken and Rasch scales. "Impossible" results can be avoided by a unified approach, as suggested here.

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Andersen, E. B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31–44.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology, 44,* 315–331.

de Gruijter, D. N. M. (1990). A note on the bias of UCON item parameter estimation in the Rasch model. *Journal of Educational Measurement, 27,* 285–288.

de Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11,* 183–196.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23,* 283–298.

Gustafsson, J.-E. (1980). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement, 40,* 377–383.

Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis.* Newbury Park CA: Sage.

Hoijtink, H. (1991). The measurement of latent traits by proximity items. *Applied Psychological Measurement, 15,* 153–169.

Langeheine, R., & Rost, J. (Eds.). (1988). *Latent trait and latent class models.* New York: Plenum Press.

Macready, G. B., & Dayton, C. M. (1992). The application of latent class models in adaptive testing. *Psychometrika, 57,* 71–88.

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and Rasch approach to IRT. *Applied Psychological Measurement, 14,* 283–298.

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

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

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken scale: A critical discussion." *Applied Psychological Measurement, 10,* 279–285.

Mooijaart, A. (1980). Latent class analysis (LCA) with order restrictions on the latent parameters. *MDN (Methoden en Data Nieuwsbrief van de sociaal wetenschappelijke sectie van de VVS), 5*(2), 22–37.

Rosenbaum, P. (1987). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology, 40,* 157–168.

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.

Sijtsma, K. (1988). *Contributions to Mokken's nonparametric item response theory.* Amsterdam: Free University Press.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16,* 149–157.

Author's Address

Send requests for reprints or further information to Dato N. M. de Gruijter, Educational Research Center, Leiden University, Boerhaavelaan 2, 2334 EN Leiden, The Netherlands. Internet: buooon@rulmvs.leidenuniv.nl.