DOCUMENT RESUME

ED 310 126 TM 013 699

AUTHOR Engelen, R. J. H.
TITLE A Review of Different Estimation Procedures in the Rasch Model. Research Report 87-6.
INSTITUTIO Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE Sep 87
NOTE 39p.; Also cited as Project Psychometric Aspects of Item Banking No. 21.
AVAILABLE FROM Mediatheek, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Bayesian Statistics; *Chi Square; Comparative Analysis; *Estimation (Mathematics); *Latent Trait Theory; Literature Reviews; Mathematical Models; *Maximum Likelihood Statistics
IDENTIFIERS *Paired Comparisons; *Rasch Model

ABSTRACT

A short review of the different estimation procedures that have been used in association with the Rasch model is provided. These procedures include joint, conditional, and marginal maximum likelihood methods; Bayesian methods; minimum chi-square methods; and paired comparison estimation. A comparison of the marginal maximum likelihood estimation with all other estimation procedures is then provided. Marginal maximum likelihood estimation is defended as the best procedure, but serious numerical problems exist even when applying this method. These problems are especially evident for distribution-free marginal maximum likelihood estimation. (TJH)

A Review of Different Estimation Procedures in the Rasch Model

Research Report

87-6

R.J.H. Engelen

University of Twente

Division of Educational Measurement and Data Analysis

Project Psychometric Aspects of Item Banking No.21

Colofon:
Typing: Mevr. L.A.M. Padberg
Cover design: Audiovisuele Sectie TOLAB Toegepaste Onderwijskunde
Printed by: Centrale Reproductie-afdeling

A Review of Different Estimation Procedures in the Rasch Model

R.J.H. Engelen

A review of different estimation procedures in the Rasch model / Drs. R.J.H. Engelen – Enschede : University of Twente Department of Education, September, 1987. – 31p.

The Rasch Model and the Marginal Rasch Model

Over the past decades, the Rasch (1960) model has become increasingly popular and proved to be very useful in the theory and analysis of mental tests. The main reasons for the popularity of the Rasch model are the simplicity of the model, as compared with other item response models, and the existence of attractive statistical procedures for estimating its parameters. Therefore, we shall shortly review the different estimation procedures that have previously been used with the Rasch model in the second section. These procedures include maximum likelihood, Bayesian, minimum chi-square and pairwise comparison estimation. The third section consists of a comparison of the marginal maximum likelihood estimation and all other estimation procedures.

Estimation Procedures in the Rasch Model

We start with a test battery consisting of k dichotomously scored items, for which we assume that they all measure the same unidimensional (latent) trait or ability.

Under the Rasch model, the probability that examinee v with ability θ answers item i correctly is given by:

$$(1) \qquad P(X=1|\theta,\epsilon_i) = \theta\epsilon_i/(1+\theta\epsilon_i) \ .$$

where ϵ_i is the item easiness parameter.

On the usual assumption that the items are local (conditional) independent, the probability that examinee v has the response pattern $x=(x_1,\ldots,x_k)$ is given by

$$(2) \qquad P(X=x|\theta,\epsilon) = \prod_{i=1}^{k} \theta\epsilon_i^{x_i}/(1+\theta\epsilon_i) \ .$$

where $x_i=1$ if item i is answered correctly and $x_i=0$ otherwise, and $\epsilon=(\epsilon_1,\ldots,\epsilon_k)$.

If independence applies at person level, i.e., if all examinees answer the items independently of each other, the joint probability of the response patterns for all N examinees can be written down as:

$$(3) \qquad P(X_1=x_1,\ldots,X_N|\theta,\epsilon) = \prod_{v=1}^{N} \prod_{i=1}^{k} \theta_v\epsilon_i^{x_{vi}}/(1+\theta_v\epsilon_i) \ .$$

where θ_v is the ability of examinee v, $\theta=(\theta_1,\ldots,\theta_N)$ and $x_{vi}=1$ if person v answers item i correctly and 0 otherwise.

At this place it is important to mention that it is not necessary that all examinees are administered the same set of items. It may happen that examinee one answers item 2 and 4, whilst examinee two has the items 2, 3 and 5 to solve. If this is the case, one speaks of an incomplete design. If the mechanism by which the items are administered is ignorable with respect to likelihood inference (Rubin, 1976), as is the case if, for example, items are administered randomly to

persons, all estimation procedures that are treated in the following sections will be applicable for this special case with the necessary adjustments. The main adjustment concerns the introduction of a random matrix D, with $d_{ij}=1$ if person i has been administered item j, and $d_{ij}=0$ otherwise. According to the value of d_{ij}, the likelihood function (3) is changed appropriately. To this point, it is not clear yet if adaptive and customized testing do influence the likelihood. Although the estimation procedures are identical for both designs, it is important to stress that the error of estimation in the item parameters can be larger in the incomplete design, since fewer examinees answer to any particular item. Since incomplete designs are for the rest very comparable with complete designs, we will confine ourselves to complete designs.

Note that the Rasch model is unidentifiable; if the item parameters are all multiplied with a constant c, and if all person parameters are divided by that same number, the probability statement in (3) does not change. For this reason, a constraint has to be imposed on the set of parameters. The choice of the appropriate constraint depends on the particular problem at hand and will therefore be imposed on the place needed.

In this paper it is assumed that both sets of parameters, i.e., item and person parameters, are unknown, and that they all have to be estimated from the data. For this purpose, the following estimation procedures are

available: maximum likelihood (unconditional or joint, conditional and marginal), Bayesian (hierarchical and marginal hierarchical), pairwise comparison and minimum chi-square estimation

Since maximum likelihood and Bayesian estimation procedures have a general nature and can be applied in many different settings, these two estimation procedures will be discussed for the general case in more detail now.

To this end, assume that the random variables $X_1,\ldots,X_n$, are independent and identically distributed with density $f(x|\alpha)$, where α is unknown and possibly vector valued. Maximum likelihood estimation is based on the principle that one should pick that value of α that makes the observed data most probable. To achieve this, the likelihood function is maximized with respect to the unknown parameter α. If the likelihood function is sufficiently smooth, as will often be the case, this can be done by differentiating the likelihood function, or equivalently, the loglikelihood, with respect to α, equating the derivative to zero, and finally solving the equation(s). Under the (mild) assumption that the density function $f(x|\alpha)$ satisfies certain regularity conditions, maximum likelihood estimates have very nice large sample features. First, maximum likelihood estimates are consistent, i.e., the estimates converge to the true parameter. Secondly, the maximum likelihood estimator of α is asymptotically normally distributed with mean α and with a variance-covariance matrix equal to the reciprocal of the Fischer

information matrix. Thirdly, the maximum likelihood estimator is efficient, i.e., the data is used in an optimal way. Furthermore, if the density function f belongs to an exponential family, maximum likelihood estimators are asymptotically equivalent to uniform minimum variance unbiased estimators (UMVU).

In Bayesian estimation, it is additionally assumed that the parameter α itself is random. Note that Bayesian techniques can also be interpreted from a frequentist point of view (Box & Tiao, 1973). The distribution of the parameter α, then, expresses the belief of the researcher in the possible values of α. This distribution of α is choosen prior to the observation of the data, and is therefore termed 'a priori' distribution. After having observed the data, one can compute, with the help of Bayes rule, the posterior distribution. This posterior distribution is proportional to the product of the prior distribution and the likelihood function, and incorporates all information that is available for the unknown parameter α. Assuming that the prior distribution is characterized by a parameter Ω, the objective is now to estimate this parameter Ω. In order to estimate this unknown parameter Ω, one now could use a maximum likelihood approach, i.e., use that estimator for Ω that maximizes the posterior distribution, or equivalently, the mode of the posterior distribution. Since the posterior distribution is a distribution function however, one could also use the median or the mean of the posterior to estimate

the unknown parameter Ω. It depends on the nature of the problem which of these methods should be applied, although it seems that modal Bayes estimation seems to do the best job, in most cases (O'Hagan, 1976).

It is important to note that maximum likelihood and Bayes estimation are equivalent if the sample size is large, since in that case the prior information as used in the Bayesian estimation plays an insignificant role. These two procedures are also equivalent if the prior distribution is non-informative with respect to the unknown parameter, i.e., in the case of a flat prior (Lehmann, 1983).

Joint Maximumum Likelihood (JML)

In this method, also called unconditional maximum likelihood, both sets of parameters are estimated simultaneously. This is done by maximizing the joint likelihood function (3) over all the parameters. An estimate can be found by differentiating (3) with respect to that parameter, equate this derivative to zero and solve the resulting equation. For the Rasch model the resulting set of equations is given by:

$$\begin{aligned} \sum_{i=1}^{k} x_{vi} &= \sum_{i=1}^{k} \theta_v \epsilon_i / (1+\theta_v \epsilon_i) \text{ for all } v=1,\ldots,N \\ \sum_{v=1}^{N} x_{vi} &= \sum_{v=1}^{N} \theta_v \epsilon_i / (1+\theta_v \epsilon_i) \text{ for all } i=1,\ldots,k \end{aligned} \tag{4}$$

Note that (4) consists of implicit equations, and hence that iterative procedures to solve (4) have to be used. Furthermore, the equations in (4) have the well known form: "observed" = "expected".

Since the same set of items may be administered to different populations, the constraint that is most appealing in this setting is one on the item parameters. Two constraints have been used: $\sum_{i=1}^{k} \epsilon_i = 1$ and $\epsilon_1 = 1$. The latter constraint has the disadvantage that for item parameter estimates on the samethe standard errors for the estimates will be larger than in the first constraint (de Gruijter, personal communication). Therefore, the constraint $\sum_{i=1}^{k} \epsilon_i = 1$ will be used.

A serious problem with joint maximum likelihood estimation is the fact that the item parameters are not estimated consistently This is due to the fact that we have a problem with structural and incidental parameters (Neyman and Scott, 1948). These problems are the result of the fact that with the introduction of another examinee, we also introduce a new person parameter. This has the effect that the number of parameters increases indefinitely, so that standard maximum likelihood estimation does not apply in this case. A heuristic interpretation for this phenomenon is the following: although with each new person we get additional information about the item parameters, we also introduce bias since the person parameter is not known. Solutions to the general problem have been given by Kiefer and Wolfowitz

(1956), Lehmann (1959), Rasch (1960) and Andersen (1970, 1973).

Kiefer and Wolfowitz (1956) showed that one can consistently estimate the structural parameter, if one assumes that the incidental parameters are independent and identically distributed. Furthermore, this distribution can be also estimated consistently. Engelen (1987) used this for the special case of the Rasch model, and called this semiparametric estimation. This will be discussed more extensively in the section marginal maximum likelihood.

Using Lehmanns (1959) notion of conditional estimation Andersen (1970, 1973), proved that a solution for the problem of structural and incidental parameters can be given, if there exists 'sufficient' statistics for the incidental parameters that do not depend on the structural parameters. Note that this was the most important assumption that led to the Rasch model (Rasch, 1960). This solution has been termed 'conditional maximum likelihood estimation' by Andersen and will be discussed in greater detail below.

The most famous example of structural and incidental parameters has been given by Neyman and Scott (1948). They considered a sequence of independent normally distributed random variables X_{ij}, $i=1,\ldots,n$, $j=1,\ldots,k$ such that $X_{i1},\ldots,X_{ik}$ have mean μ_i and variance σ^2. They showed that the (inconsistent) maximum likelihood estimate of σ^2, can be adjusted with a factor $(k-1)/k$ to yield a consistent estimate. It was long believed that this factor could also

been used in the Rasch model, i.e., if the estimates of ϵ_i where multiplied by (k-1)/k that consistent estimate would be the result (Wright & Douglas, 1977; Andersen, 1980). However, the proof for this fact has never been given since Andersen's proof only applies to the special case k=2; a generalization of this proof for larger k has never been given up to now. In a simulation study by van den Wollenberg (1986) it was shown that the factor (k-1)/k does not apply for k>2. Even stronger, van den Wollenberg showed that there does not exist a universal factor to adjust the estimate of ϵ_i in order to get a consistent estimate. This factor would have to depend on the distribution of the item difficulties and on the ability distribution.

Conditional Maximum Likelihood (CML)

This method is based on the fact that in the Rasch model a 'sufficient' statistic for the incidental parameter θ_i, namely the number of correctly answered items by person i, exists.

The concept of sufficiency, as introduced by Fisher (1950), was based on the fact, that some part of the data carries no information about the unknown distribution and that therefore X can be replaced by some statistic T=T(X) without loss of information. Many nice features of sufficiency can now be derived; all of these are based on the

fact that for making inference one can confine oneself to a sufficient statistic.

Sufficiency as defined by Andersen (1970, 1973) however, does not necessarily have the same features. Basicly, Andersen's definition of sufficiency is an extension of the earlier definition of Fisher's concept of sufficiency. Since these two definitions of sufficiency are not equivalent, all results that are derived from Andersen's new definition should be carefully checked. This has been done by Andersen (1973) in most cases, only in the case of the principle "information" there are some discrepancies. For instance, in the Rasch model, if one conditions on the total score of person v, one can show that no information about that person's ability is lost, but there seem reasons to believe that this is not true for the information about the items, i.e., by proceeding in this way one discards information (Engelen, forthcoming 1988).

For the special case of the Rasch model, Andersen's notion of sufficiency means that the total score of person v is a sufficient statistic for the ability θ of that person in the presence of the item parameter ϵ. Note the contrast with the ordinary principle of sufficiency, where the total score of person v and the number correct on item i are (jointly) sufficient statistics for the person parameter θ and the item parameter ϵ. Denote the total score statistic for person v as T_v. The conditional probability for the score pattern x_v, given $T_v=t_v$ can now be derived:

$$(5) \qquad P(X_v=x_v|T_v=t_v) = P(X_v=x_v, T_v=t_v)/P(T_v=t_v)$$

Noting that $P(T_v=t_v) = \sum_{\Sigma x_v=t_v} P(X_v=x_v)$ and that $P(X_v=x_v, T_v=t_v) = P(X_v=x_v)$, the probability statement in (4) can be rewritten into

$$(6) \qquad P(X_v=x_v|T_v=t_v) = \frac{\prod_{i=1}^{k} \epsilon_i^{x_{vi}}}{\sum_{x_{vi}=t_{vi}} \prod_{i=1}^{k} \epsilon_i^{\lambda_{vi}}}$$

In this form, the likelihood function (6) contains no item parameters anymore, and estimates of item parameters can be evaluated by ordinary maximum likelihood. Andersen (1970,1973) showed that these estimates are consistent and have asymptotically a normal distribution. Starting from (6), i.e., regarding (6) as a model on itself, no problems would be encountered with maximum likelihood estimation, for example, the item parameters would be estimated correctly. A rationale for (6) as a model can, however, not be given.

An important drawback of the CML estimation procedure might be that examinees with all items correct or all items wrong, have to be eliminated from the sample, since in that case no conditional item estimates can be obtained. That no estimates exist for these persons, can be easily seen from (6), since in that case, both sides are equal to one. The only information that we can draw now is that for examinees

with all items correct (wrong), the items were all too easy (difficult).

The denumerator in (6) is termed elementary symmetric function. The evaluation of these functions is a tedious task, and was for a long time possible only for a small number of items (Hambleton & Swaminathan, 1985). In a paper by Verhelst et all (1984), it is shown that no serious problems are encountered anymore, and that one can handle as many as 1000 items now.

After estimates of the item parameters have been obtained, one can estimate the person parameters by considering these estimates as the true values, substituting these values in the likelihood (3), and obtaining maximum likelihood estimates of the person parameters in the usual way. Since the number of persons is usually large, so that the item parameter estimates have a very small standard error, the effect of treating estimated values as known, seems appropriate. The precise effect of this procedure is however not known, yet. The effects of the replacement of true item parameter values by estimated values will be analyzed with the help of a simulation study simulation (Engelen, forthcoming 1988)

Marginal Maximum Likelihood (MML)

The person parameters are now regarded as independent and identically distributed random variables. In other words, it is assumed that there exists a distribution function of ability F and that persons are exchangable, i.e., the ability of a randomly drawn person is an outcome of this distribution. For the Rasch model, we can evaluate the probability for a score pattern x, given the population of interest, by integrating the probability (2) over the population density dF(θ):

$$(7) \qquad P(X=x|F,\epsilon) = \int_0^\infty P(X=x|z,\epsilon)dF(z)$$

The integral in (7) is evaluated as a Stieltjes-integral; if there exist a derivative of F, then dF(z) can be replaced by f(z)dz and we have an ordinary Riemann integral.

In this marginal likelihood function, no person parameters are present anymore, since they have been integrated out. Hence, (7) is a function of the item parameters $\epsilon_1,\ldots,\epsilon_k$ and the ability distribution function F only.

Substituting (3) into the marginal probability function and rearranging leads to:

$$(8) \qquad P(X=x|F,\epsilon) = \prod_{i=1}^{k} \epsilon_i^{x_{vi}} \int_0^{\infty} z^{\sum x_{vi}} / \left(\prod_{i=1}^{k} (1+z\epsilon_i) \right) dF(z)$$

Substituting $b_x = \prod_{i=1}^{k} \epsilon_i^{x_i}$, and $D_x(z,\epsilon) = z^{\sum x_i} / \left(\prod_{i=1}^{k} (1+z\epsilon_i) \right)$, the marginal probability function for the responses of all N examinees is given by:

$$(9) \qquad P(X_1=x_1,\ldots,X_1=x_1|F,\epsilon) = \prod_{v=1}^{k} P(X_v=x_v|F,\epsilon)$$

$$= \prod_{x} \left(b_x \int_0^{\infty} D_x(z,\epsilon) dF(z) \right)^{M_x},$$

where M_x is the number of examinees with response pattern x.

From this starting point, a few different routes have been followed. First, one can assume that F belongs to a special parametric family, indexed by a parameter Φ. Then, one can estimate Φ along with the item parameters $\epsilon_1,\ldots,\epsilon_k$. A common choice for F has been a lognormal distribution with mean $\exp(\mu+\frac{1}{2}\sigma^2)$ and variance $(\exp(\sigma^2)-1)\exp(2\mu+\sigma^2)$ (so $\Phi=(\mu,\sigma^2)$). Recall that a random variable Y is lognormally distributed with mean $\exp(\mu+\frac{1}{2}\sigma^2)$ and variance $(\exp(\sigma^2)-1)\exp(2\mu+\sigma^2)$ if log Y is normally distributed with mean μ and variance σ^2.

Good results with this ability distribution where obtained by Thissen (1982), Andersen & Madsen (1977), Mislevy

(1984) and Sanathanan & Blumenthal (1978). Most of these authors assumed that the item parameters where known beforehand, so that only μ and σ^2 had to be estimated.

Note, however, that the form of the ability distribution need not to be known beforehand. Hence, this method lacks a basic common sense interpretation.

Therefore, one can try to estimate the ability distribution jointly with the estimation of the item parameters. For this purpose, Bock & Aitkin (1981) used a discrete distribution over a finite number of points and called this histogram the empirical distribution. Although they claim that they now freed the marginal maximum likelihood procedure from arbitrary assumptions about the ability distribution, this is not true. Since they use preassigned values for the nodes of the ability distribution function, and since these nodes are not changed during the iteration process used to estimate the ability distribution function and the item parameters, Bock and Aitkin are actually working in the parametric setting again. De Leeuw & Verhelst (1986) and Engelen (1987) showed that one can in fact estimate the ability distribution function jointly with the item parameters. Furthermore, both authors showed that this can be done consistently, under certain suitable regularity conditions. The ability distribution function turns out to be a step function, where the number of steps is a function of the number of items only. All this will be discussed in more detail in the third section.

Bayesian Estimation

In the Bayesian framework, one starts with imposing reasonable prior distributions for the parameters of interest. Reasonable in this context should be understood as ease of computation for or believe in the particular prior chosen. Then using Bayes rule, one can, having observed the data, compute the a posteriori distribution. This a posteriori distribution now, will be used as the base of further inference.

Bayesian estimation always improves on maximum likelihood estimation, if it is reasonable to assume that one or more subsets of parameters can be considered as exchangeable members of corresponding populations. If no prior information is available, and one uses a non-informative prior for a parameter, i.e., a flat (uniform) one, than Bayesian estimation is equivalent to maximum likelihood estimation.

Historically, Bayesian estimation started with the specification of a parametric prior distribution. Later on, this changed into the specification of empirical priors, i.e., priors that are estimated from the data, and hierarchical Bayesian estimation, where a prior is specified for the parameters in the prior distribution. The latter has

one clear advantage: hierarchical Bayes is far more flexible than ordinary Bayes estimation.

In principle, one can distinguish three different Bayesian estimation procedures in the Rasch model: (i) both item and person parameters are subject to prior information; i.e., prior distributions for item as well as person parameters are assumed; (ii) only a prior distribution for the person parameter is specified; and (iii) only a prior distribution function for the item parameters is specified.

Procedure (iii) has never been used in the Rasch model before, since in most applications the item parameters are known beforehand or are believed to be estimated reasonable by one of the maximum likelihood procedures. This restricts the discussion to the first two procedures.

First, we will discuss the first procedure, since the second one can be seen as a special case. We will do this for hierarchical Bayes estimation. The starting point for the analysis is likelihood function (3). Using Bayes rule, the posterior distribution f of the observed data and all the parameters is proportional to the product of this likelihood and the prior distribution g of the parameters:

$$(10) \qquad f(\mathbf{X},\theta,\epsilon) \propto L(\mathbf{X}|\theta,\epsilon)g(\theta,\epsilon).$$

Now one has to chose a prior distribution for the item and person parameters. Swaminathan and Gifford (1982) show that the analysis can be effectively reduced if one makes the

reasonable assumption that the item and person parameters are independently distributed. They also assume that the distributions for item and person parameters have the same form (both distributions multivariate lognormal), a standard approach, but nevertheless not free of criticism. So, we have

(11) $\log \theta_v \sim N(\mu_\theta, \Phi_\theta)$; $\log \epsilon_i \sim N(\mu_\epsilon, \Phi_\epsilon)$.

To complete the hierarchical Bayes structure, prior distributions for the so-called hyperparameters $\mu_\theta, \Phi_\theta, \mu_\epsilon, \Phi_\epsilon$ have to be specified. For the means μ_θ and μ_ϵ, a flat uniform prior is chosen, and since μ_θ and Φ_ϵ are variances, inverse χ^2 distributions with parameters τ and β seem appropriate. Note that these are conjugate priors. Finally, Swaminathan and Gifford showed that reasonable values for the hyperparameters are between 5 and 15 for τ and about 10 for β. Working all this out, they find the likelihood of the posterior distribution, which they use as a base for further inference. For more specific details, see Swaminathan and Gifford (1982).

Note that the classical objection against Bayesian procedures applies in this case also: no empirical evidence for the choice of the priors is given. On the other hand, considering the flexibility of hierarchical Bayes estimation, this need not to be a serious problem.

An other approach is given by Mislevy (1986), who uses the same structure for the item parameters, but changes the

prior of the person parameters. For the person parameters, Mislevy offers a choice between a nonparametric prior in the form of a histogram and a mixture of normal components. Again, natural conjugate priors are chosen for the hyperparameters. Note that the term nonparametric is misplaced; the nodes of the histogram are fixed in advance and are not estimated from the data. See also Engelen (1987) for a discussion of this in the marginal model.

The results of Mislevy (1986) and Swaminathan and Gifford (1982, 1986) show that hierarchical Bayesian estimation yields good results. This is especially true for the case of the three parameter logistic item response models, where maximum likelihood estimation performs rather badly, even for a very large number of examinees.

Minimum Chi-Square Estimation

Another estimation procedure has been proposed by Fischer and Scheiblechner (1970) and Fischer (1974): the minimum chi-square estimation. This procedure starts with the observation that

(12) $n_{ij}/n_{ji} = \epsilon_i/\epsilon_j$.

where n_{ij} stands for the number of examinees that respond correctly to item i but incorrectly to item j. With the easier notation

(13) $\delta_i = 1/\epsilon_i$.

Fischer uses

(14) $$\sum_{i<j} \frac{(n_{ij}\delta_i - n_{ij}\delta_i)^2}{\delta_i\delta_j(n_{ij}+n_{ji})} .$$

as a chi-square criterion. Now, (15) is minimized with respect to the item parameters δ_i, which yields estimates of these parameters. Subsequently, the person parameters can now be estimated, in the same way as with conditional maximum likelihood, by using the estimated values of the item parameters as the true ones, and maximizing the resulting likelihood expression.

An advantage of this method is its fastness. Furthermore, although the n_{ij} are dependent, Fischer and Scheiblechner (1970) claim, as a result of their simulation studies, that the distribution of (15) is approximately distributed as chi square. That this is true in the general case, has however never be shown, neither has the contrary.

Paired Comparison Estimation

In this method, the Rasch model is rewritten as a model for paired comparison with ties (Bradley,1976). In the latter method one compares the responses of a subject responding to a pair of items. Therefore, the Rasch model is rewritten in the following way:

$$(15) \qquad P(X_{vi}=x_{vi}|\theta_v,\delta_i) = \theta_v(\theta_v+\delta_i)^{-1}.$$

This is done by substituting $\epsilon_i=\epsilon_i^{-1}$ and a simple rewriting of (1). For a pair of items (i,j), one can now consider the four possible patterns of an examinee v. These patterns then, give informa.'on about the relative difficulties of the two items for that examinee. In other words, one considers these patterns as the outcomes of a paired comparison experiment. In that case, three basic different outcomes can be distinguished:

$(X_{vi} > X_{vj})$	item i correct and j not
$(X_{vi} < X_{vj})$	item j correct and i not
$(X_{vi} = X_{vj})$	both items correct or both items incorrect.

The first outcome can now be interpreted as a comparison showing that, for examinee v, item i is likely to be more

easy than item j. The other outcomes are interpreted analogously.

The probabilities for the possible outcomes can now be evaluated for the Rasch model:

$$
\begin{aligned}
&P(X_{vi} > X_{vj}) = \theta_v\delta_j[(\theta_v+\delta_i)(\theta_v+\delta_j)]^{-1} \\
(16)\quad &P(X_{vi} < X_{vj}) = \theta_v\delta_i[(\theta_v+\delta_i)(\theta_v+\delta_j)]^{-1} \\
&P(X_{vi} = X_{vj}) = (\theta_v{}^2+\delta_i\delta_j)[(\theta_v+\delta_i)(\theta_v+\delta_j)]^{-1}.
\end{aligned}
$$

If one now conditions on the event on a non-tie, or equivalently on the event that the total test score for the two items is one, the result is the Bradley-Terry model from the paired comparison literature:

$$
(17)\quad P(X_{vi} > X_{vj} \mid X_{vi} = X_{vj}) = \delta_j(\delta_i+\delta_j)^{-1} = \tau_{ij}.
$$

Note that in equation (12) the person parameter θ_v has disappeared; for any examinees the probability described in (12) is independent of that examinees ability. This means that the likelihood for a comparison of two items takes the form

$$
(18)\quad L(\delta_i,\delta_j \mid a_{ij},) = \prod_{v=1}^{N} \delta_j(\delta_i+\delta_j)^{-aij}\delta_i(\delta_i+\delta_j)^{-aji},
$$

where a_{ij} is the number of times $(X_{vi} > X_{vj})$ is observed. For n items however, the outcomes of the comparisons $(X_{vi} > X_{vj})$ and $(X_{vk} > X_{vi})$ are not independent. It is shown by van der

Linden and Eggen (forthcoming, september 1987), that the number of independent comparisons for an examinee with total test score t on n items, is given by min(t,n-t). Denoting the set of independent comparisons for a set of n items by J, the likelihood for this set is given by

$$(19) \qquad L(\delta_1,\ldots,\delta_n|a_{ij.}) = \prod_J \delta_j(\delta_i+\delta_j)^{-a_{ij}}\delta_i(\delta_i+\delta_j)^{-a_{ji}} .$$

An iterative algorithm for obtaining maximum likelihood estimates has already been given in the general paired comparisons setting independently by Zermelo (1929) and Ford (1957). Furthermore, they showed that these maximum likelihood estimates exist and are unique if the following necessary and sufficient condition is satisfied: For every partition of the set in two non-empty subsets, for some item in the first set and some item in the second one, the outcome $(X_{vi}>X_{vj})$ has occurred at least for one value of v. That this is a weak, almost always satisfied condition has been showed by Fischer (1981), who found the same condition for the existence and uniqueness of conditional maximum likelihood estimates. In contrast with conditional maximum likelihood estimation, this method is not limited to a small number of items, since elementary symmetric functions of order greater than two do not have to be calculated in this approach. Note that since the item parameters are estimated by maximum likelihood, they are estimated consistently.

After item parameters have been obtained, the person parameters can be obtained by maximum likelihood estimation where the real values of the item parameters are replaced by their estimates.

More on the Marginal Rasch Model

In this section we will compare the different estimation procedures in greater detail and mention some other features of the marginal Rasch model that have not been discussed before.

First, we have to explain why one wants to use the marginal Rasch model instead of the Rasch model itself. As explained before, one can not use the Rasch model in combination with joint maximum likelihood estimation, since the resulting item parameter estimates are not consistent. Remains the possibility of the conditional model. The main reason not to use the conditional model is the loss of information (mentioned earlier).

Secondly, the marginal Rasch model can be seen as a model on itself, just like the unconditional Rasch model, and was introduced as such by Cressie and Holland (1985). In doing so, Cressie and Holland used the notion of manifest probabilities (Lazersfeld & Henry, 1968), i.e., the proportion of examinees in a certain given population who obtain a particular pattern of right and wrong responses.

These manifest probabilities can be explained by an -unobservable- latent trait model if that model correctly predicts the data. Then they show that the Rasch model is a model that can predict these manifest probabilities correctly, given that the data satisfies certain conditions. These conditions will be discussed later. Note that there is no rational explanation for the conditional Rasch model.

Furthermore, it is not clear at all which of the properties derived for the conditional maximum likelihood estimation by Andersen (1970, 1973), are really true. How well do the two different conceptions of sufficiency as given by Fisher and Andersen match ? Another reason is that the conditional model is only applicable to the one-parameter logistic model and not with the two- or three- parameter logistic models. The reason for the latter is that in the more parameter logistic models, no simple 'sufficient' statistics for ability exist.

Next, we shall discuss some advantages and disadvantages of marginal maximum likelihood estimation in the Rasch model in comparison with the other maximum likelihood and the minimum chi-square and pairwise comparison approaches. First, no persons have to be eliminated from the data to be able to obtain estimates for the item parameters. In the other maximum likelihood approaches, persons with all items correct or wrong as well as items that have been answered correctly by all examinees have to be eliminated from the initial data-set. Also, in the minimum chi-square and in the pairwise

comparison approach, one can r . use the complete dataset for estimation. This has the effect that not all the available information in the data-set is used. Secondly, marginal maximum likelihood estimation is also applicable in the two- and three- parameter logistic models, while the others are not. Unconditional maximum likelihood estimation does not work in more parameter logistic models, since the estimates of the guessing parameter drift out of their bands (Mislevy, 1986). The unconditional maximum likelihood, the minimum chi square and the pairwise comparison approaches do not apply in more parameter logistic item response models, since the notion of sufficiency, which is the uniform base for all these estimation procedures, is violated in these models (Fischer, 1974). The main disadvantage of the marginal approach is that no estimates of the person parameters are obtained, only information about the distribution of ability is achieved. Note that the main purpose of a test is often to get information about the ability of the examinees taking the test. With marginal maximum likelihood estimation, this information is not available; only the ability distribution function can be estimated. However, this ability distribution estimate could be used, for example, as an instrument to measure the differences between different schools or different curricula.

Important to note is further that marginal maximum likelihood estimation yields consistent item parameter estimates and that together with the estimated distribution

function of ability, one achieves a reasonable fit in most cases.

Compared with modal Bayes estimation, the marginal maximum likelihood approach yields the same results and is hence equivalent. This subject to the constraint that no prior distribution is put on the item parameters.

The conditions on the manifest probabilities as given by Cressie and Holland (1983), are exactly the same as the conditions that de Leeuw and Verhelst needed to be able to estimate the (empirical) ability distribution function. The conditions of Engelen (1987) only show that is possible to estimate the ability distribution consistently: it is not proven that these estimates exist. To be able to do this, one needs additional constraints like the ones given in de Leeuw and Verhelst.

It is important to stress the fact that one should use empirical marginal maximum likelihood, or equivalently, empirical Bayes estimation, instead of the parametric approach. This is necessary since one never has an exact indication of the true form of the ability distribution function; therefore this function should be estimated empirically.

Although marginal maximum likelihood estimation seems the most appropriate one, serious numerical problems exist, especially for distribution free marginal maximum likelihood estimation.

References

Andersen, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimators. Journal of the Royal Statistical Society, 32, 283–301.

Andersen, E.B. (1973). Conditional inference and models for measuring. Copenhagen: Mentalhygiejnisk Verlag.

Andersen, E.B. (1980). Discrete statistical models with social sciences applications. Amsterdam: North–Holland.

Andersen, E.B. & Madsen, M. (1977). Estimating the parameters of the latent population distribution. Psychometrika, 42, 357–374.

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

Bock, R.D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179–197.

Box, G.E.P. & Tiao, E.G. (1973). Bayesian inference and Statistical analysis. Reading, Mass.: Addison–Wesley.

Cressie, N. & Holland, P.W. (1983). Characterizing the manifest probabilities of latent trait models. Psychometrika, 48, 129–141.

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

Eggen, T.J.H.M. & van der Linden, W.J. (1986). The use of paired comparison with ties. Research Report 86-8, University of Twente, Department of Education. Enschede, The Netherlands.

Engelen, R.J.H. (1987). Semiparametric estimation in the Rasch model. Research Report 87-1, University of Twente, Department of Education. Enschede, The Netherlands.

Fischer, G.H. (1974). Einführung in die Theorie psychologischer Tests: Grundlagen und Anwendungen. Bern: Verlag Hans Huber.

Fischer, G.H. (1981). On the existence and uniqueness of maximum likelihood estimates in the Rasch model. Psychometrika, 46, 59-76.

Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. Monthly Notices Royal Astronomical Society, 80, 758-770.

Ford, L.R.J. (1957). Solution of a ranking problem from binary comparisons. American Mathematical Monthly, 64, 241-252.

Hambleton, R.K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Annals of Mathematical Statistics, 27, 887-903.

Lazersfeld, P.F. & Henry, N.W. (1968). Latent structure analysis. Boston: Houghton Mifflin.

Lehmann, E.L. (1959). Testing statistical hypotheses. New York: Wiley.

Lehmann, E.L. (1983). Theory of point estimation. New York: Wiley.

Mislevy, R.J. (1984). Estimating latent distributions. Psychometrika, 49, 359-381.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. Psychometrika, 51, 177-195.

Neyman, J. & Scott, E.L. (1948). Consistent estimates based on partially consistent observations. Econometrika, 16, 1-32.

O'Hagan, A. (1976). On posterior joint and marginal modes. Biometrika, 63, 329-333.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago:The University of Chicago Press.

Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 801-827.

Sanathanan, L. & Blumenthal, S. (1978). The logistic model and estimation of latent structure. Journal of the American Statistical Association, 73, 794-799.

Swaminathan, H. & Gifford, J.A. (1982). Bayesian estimation in the Rasch model. Journal of Educational Statistics, 7, 175-191.

Swaminathan, H. & Gifford, J.A. (1985). Bayesian estimation in the two parameter logistic model. Psychometrika, 50, 349–364.

Thissen, D.B. (1982). Marginal maximum likelihood estimation for the one–parameter logistic model. Psychometrika, 47, 175–186.

Wright, B.D. & Douglas, G.A. (1977). Conditional versus unconditional procedures for sample free item analysis. Educational and psychological measurement, 37, 47–60.

Zermelo, E. (1929). Die Berechnung der Turnier–Ergebnisse als ein Maximum–problem der Wahrscheinlichkeitsrechnung. Mathematisch Zeitschrift, 29, 436–460.

Titles of Recent Research Reports

RR-86-1 W.J. van der Linden, The use of test scores for classification decisions with threshold utility

RR-86-2 H. Kelderman, Item bias detection using the loglinear Rasch model: Observed and unobserved subgroups

RR-86-3 E. Boekkooi-Timminga, Simultaneous test construction by zero-one programming

RR-86-4 W.J. van der Linden, & E. Boekkooi-Timminga, A zero-one programming approach to Gulliksen's matched random subtests method

RR-86-5 E. van der Burg, J. de Leeuw, & R. Verdegaal, Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features

RR-86-6 W.J. van der Linden, & T.J.H.M. Eggen, An empirical Bayes approach to item banking

RR-86-7 E. Boekkooi-Timminga, Algorithms for the construction of parallel tests by zero-one programming

RR-86-8 T.J.H.M. Eggen, & W.J. van der Linden, The use of models for paired comparisons with ties

RR-86-9 H. Kelderman, Common item equation using the loglinear Rasch model

RR-86-10 W.J. van der Linden, & M.A. Zwarts, Some procedures for computerized ability testing

RR-87-1 R. Engelen, Semiparametric estimation in the Rasch model

RR-87-2 W.J. van der Linden (Ed.), IRT-based test construction

RR-87-3 R. Engelen, P. Thommassen, & W. Vervaat, Ignatov's theorem: A new and short proof

RR-87-4 E. van der Burg, & J. de Leeuw, Use of the multinomial jackknife and bootstrap in generalized nonlinear canonical correlation analysis

RR-87-5 H. Kelderman, Estimating a quasi-loglinear models for the Rasch table if the number of items is large

RR-87-6 R. Engelen, A review of different estimation procedures in the Rasch model

Research Reports can be obtained at costs from Mediatheek, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

[redacted]tment of

UCATION

A publication by
the Department of Education
of the University of Twente

[redacted]ede
[redacted]nds