This paper reviews the literature about item response models for the subject level and aggregated level (group level). Group-level item response models (IRMs) are used in the United States in large-scale assessment programs such as the National Assessment of Educational Progress and the California Assessment Program. In the Netherlands, these models are useful to the National Institute for Educational Measurement, especially for the Dutch National Assessment Program of Educational Achievement. After a short introduction on IRMs on the subject level, a comprehensive treatment is given of the following estimation methods for subject-level parameters: joint maximum likelihood, conditional maximum likelihood, marginal maximum likelihood, logit based parameter estimation, the Bayesian approach, and other estimation procedures. A group-level IRM describes the probability of a correct response from an examinee selected at random from a specific group. The following group-level models are described: the group fixed-effects model, the two-parameter and three-parameter normal-normal model, the normal-logistic model, and the California Assessment Program model. Analogies and differences between group-level and subject-level IRMs are discussed. Group-level IRMs may be justified as aggregate descriptions of IRMs on subject-level, and they may be interpreted analogously. Group-level IRMs are implied by subject-level IRMs only when within-group ability distributions are identical except for location. For the subject-level, the addition of an examinee increases the number of incidental (ability) parameters; however, for the group-level, the number of ability parameters does not increase. (RLC)
Item Response Theory at Subject- and Group-Level

Hilde Tobi

Division of Educational Measurement and Data Analysis

University of Twente
Item Response Theory

at

Subject- and Group-Level

Hilde Tobi
Abstract

This paper contains a review of the literature about item response models for the subject and aggregated-level (group-level).

After a short introduction on item response models on subject-level, a comprehensive treatment is given of the following estimation methods for subject-level parameters; joint maximum likelihood, conditional maximum likelihood, marginal maximum likelihood, logit based parameter estimation, the Bayesian approach, and some less familiar procedures.

A group-level item response model describes the probability of a correct response from an examinee selected at random from a specific group. The following group-level models are described; the group fixed-effects model, the two- and three-parameter normal-normal model, the normal-logistic model and the Californian Assessment Program (CAP) model.

Finally, the analogies and differences between group-level and subject-level item response models are discussed.
Introduction

Item response theory (IRT) can be seen as a reaction to the well-documented shortcomings of classical test theory (Fischer, 1974; Hambleton & Swaminathan, 1985).

An item response model specifies a relationship between the observable item performance of an examinee and the latent trait or ability assumed to underlie the performance on that item. Item characteristic curve (ICC) is a central construct in item response theory. Generally, an ICC is a monotonically increasing mathematical function ranging from zero to one that gives the probability of an examinee with a given ability level answering the item correctly. In the one-parameter model, also called the Rasch model, sufficient statistics are available: the relative item difficulty can be estimated independent of the sample of examinees used, and estimators of the relative examinee ability are independent of the particular subset of items from a certain item domain. This feature makes item response models particularly useful in comparative studies, where performance of (groups of) examinees are compared.

There has been an increasing interest among assessment and evaluation researchers for models to analyse data at an aggregated level. This interest has initiated the formulation of item response models for groups of subjects, such as
Subject- and Group IRT

schools or sex (Bock, Mislevy & Woodson, 1982). These group-level item response models are used in the United States of America in large scale assessment programs like the National Assessment of Educational Progress (NAEP) and the Californian Assessment Program (CAP). In the Netherlands these models are useful to the National Institute for Educational Measurement (CITO), especially for the Dutch National Assessment Program of Educational Achievement (PPON).

Item response models at subject-level

As mentioned in the introduction, the development of item response theory started with models formulated at the level of an individual subject. In this paragraph these item response models and their estimation procedures will be discussed.

The probability of a correct response $X_{vi}=1$ from an examinee $v$ selected at random from a certain population to item $i$, can be written as a function of the examinees ability $\theta_v$ and a vector of item parameters $\theta_i$:

$$P_{vi} = P(X_{vi}=1) = H_1(\theta_v, \theta_i)$$

where $H_1(\theta_v, \theta_i)$ is a continuously differentiable function of $\theta_v$. Usually, $H_1(\theta_v, \theta_i)$ is either the normal-ogive or the logistic curve. For the two-parameter normal-ogive model (Lord & Novick, 1968) the probability that an examinee $v$ with ability level $\theta_v$ passes item $i$ is given by
\[ P_{v_i} = P(X_{v_i}=1) = \int_{-\infty}^{a_i(\theta_v-b_i)} \Phi(t) \, dt, \]

where \( b_i \) is the item difficulty, \( a_i \) is the discrimination parameter and \( \Phi(t) \) is the normal density function. The normal-ogive has a point of inflexion at \( \theta = b_i \); at this point the probability of a correct answer is 0.5, and the slope of the curve is \((2\pi)^{-1/2}a_i\).

For the logistic function model the probability is:

\[ P_{v_i} = P(X_{v_i}=1) = \left[ 1 + \exp\left( -D \frac{a_i(\theta_v-b_i)}{b_i} \right) \right]^{-1}, \]

\( D \) is an arbitrary constant. When \( D=1.7 \), the normal-ogive and the logistic item response functions are almost equal. The logistic model is often preferred because of its mathematical convenience.

The two-parameter models can be modified to take guessing into account. If \( c_i \) denotes the guessing parameter, i.e. the lower asymptote, the three-parameter logistic model becomes

\[ P_{v_i} = c_i + (1-c_i) \left[ 1 + \exp\left( -D \frac{a_i(\theta_v-b_i)}{b_i} \right) \right]^{-1} \]

Much attention has been paid to the one-parameter logistic model, also called the Rasch model. In this model all items have the same discriminating power, i.e. \( a_i \) is a constant for
all items in the test. The ICC's only differ in their location, indicated by the item difficulty parameter $b_i$. The Rasch model is given by

$$P_{vi} = \frac{\exp(\theta_v - b_i)}{1 + \exp(\theta_v - b_i)}$$

The advantage of the Rasch model is that the total test score is a sufficient statistic for the examinee's ability (Fischer, 1974).

**Estimation**

In this paragraph a short review of the available estimation procedures for the item response theory models will be given. Some of the advantages of these procedures will also be discussed. First the most often used procedure, the joint maximum likelihood estimation, will be described.

**Joint Maximum Likelihood Estimation (JML)**

Let the $(N \times n)$ matrix $U$ contain the responses of $N$ examinees on $n$ items, in such a way that

$$U = [x_{11} x_{21} \ldots x_{N1}].$$

where $x_{iv}$ is a column vector which contains the responses $x_{iv}$ of examinee $v$ to all $n$ items. Under local independence the likelihood function is
Subject- and Group IRT

\[ L(U | \theta_1, \theta_2, \ldots, \theta_N, \xi_1, \xi_2, \ldots, \xi_n) = \prod_{v=1}^{N} L(x_v | \theta_v) = \prod_{v=1}^{N} \prod_{i=1}^{n} P_{vi}^{x_{vi}} (1-P_{vi})^{(1-x_{vi})}, \]

where \( \boldsymbol{\xi}_i \) is a vector containing the item parameters of item \( i \). To calculate the maximum likelihood estimates of \( \theta=(\theta_1, \theta_2, \ldots, \theta_N) \) and \( \boldsymbol{\xi}_i \) (for \( i=1, \ldots, n \)), the following likelihood equations have to be solved

\[ \delta \ln L / \delta m_k = 0, \]

where \( m_k \) is the \( k \)-th element of the vector \( m=[\theta, \xi_1, \xi_2, \ldots, \xi_n] \). So for the three parameter model \( m \) contains \( N+3n \) elements and because of the indeterminacy in the model, \( N+3n-2 \) parameters have to be estimated.

There are some problems limiting the use of JML; see Hambleton & Swaminathan (1985, p. 135) for a discussion of these. The main problems are that solving a system of so many nonlinear equations takes a lot of computing time and that the parameter estimates may take on values outside the accepted range. A more fundamental problem with JML estimation is that the item parameters are not estimated consistently. When simultaneous estimation of item and ability parameters is attempted, the number of ability parameters increases with the addition of each examinee. Therefore the estimators of the (structural) item parameters...
will not always converge to their true value since the number of (incidental) ability parameters increases too.

**Conditional Maximum Likelihood Estimation (CML)**

Conditional Maximum Likelihood estimation is based on the availability of a sufficient statistic for the ability parameters. In the Rasch model the total test score is a sufficient statistic for the ability parameter. Since the Rasch model is a member of the exponential family (Fischer, 1974) the conditional probability of observing the response vector $X_v$ does not depend on the ability parameter $\theta_v$.

$$P(X_v=x_v | T_v=t_v) = \frac{P(X_v=x_v, T_v=t_v)}{P(T_v=t_v)} = \frac{P(x_v-A_v)}{P(T_v-t_v)}$$

where $\{x | \Sigma x_i = t_v\}$ is the set of all possible response patterns $x=[x_1,x_2,..,x_n]$ with total sum score $t_v$. It can easily be seen that examinees with all items wrong or all items correct have to be eliminated from the sample since in that case there is only one $\{x | \Sigma x_i = t_v\}$.

The above obtained estimators of the item parameters are consistent and have an asymptotic normal distribution (Andersen, 1970).

After the estimation of the item parameters, the ability parameters are commonly estimated by substitution of the item
parameter estimates in the Rasch model. The precise effect of using the estimated parameters instead of the parameter values is not known (Engelen, 1989).

**Marginal Maximum Likelihood Estimation (MML)**

In the marginal approach, it is assumed that there exists an ability distribution \( F \) and that the ability of a randomly selected examinee is a realization of this distribution \( F \). The probability of observing any response pattern \( \mathbf{x} \), given the population, can be evaluated by integrating over the population density. So

\[
P(\mathbf{x} = \mathbf{x} | F, \mathbf{1}) = \int_{0}^{\infty} P(\mathbf{x} = \mathbf{x} | \mathbf{1}, \theta) \, dF(\theta) = \pi_x
\]

where \( \mathbf{1} = [\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_n] \). Note that \( \mathbf{x} \) and \( \theta \) are vectors with random variables now.

There are \( 2^n \) response patterns and if \( N_{\mathbf{x}} \) is the number of examinees with response pattern \( \mathbf{x} \), then the loglikelihood function is given by (Hambleton & Swaminathan, 1985)

\[
\ln(L) = n_x \sum_{x=1}^{2^n} \ln \pi_x + \text{constant.}
\]

There is no ability parameter anymore in this likelihood function, so the maximum likelihood estimators are obtained.
by differentiating \( \ln(L) \) with respect to the item parameters, setting them equal to zero and solving the equations.

Numerical issues can be a problem in solving these equations. Another problem is the ability distribution \( F \). However, Engelen (1987a) showed for the Rasch model that one can estimate the ability function jointly with the item parameters, without making any assumption on the ability distribution.

The advantages of MML over CML estimation are that no examinees have to be eliminated from the data and that it is also applicable to the two- or three-parameter logistic model. A disadvantage of MML estimation is that no estimators of the individual ability parameters are available, but only information about the distribution of the ability is obtained.

**Logit-based parameter estimation**

An important reason for investigating the possibilities of logit-based parameter estimation is the expected low computer costs of the procedures. Logit-based parameter estimation has been explored by Verhelst and Molenaar (1988) for the Rasch model and by Baker (1987) for the two-parameter logistic model.

Verhelst and Molenaar (1988) transform an initial \( \sqrt{n} \)-consistent estimator into an asymptotically efficient one. Let \( L_N(\theta) \) be the log-likelihood function of parameter \( \theta \) and let

\[
\Delta_N(\theta) = N^{-1/2} \frac{\delta L_N(\theta)}{\delta \theta}.
\]
Assume that $\Delta_N(\theta)$ is asymptotically normal distributed $N(0,I_\theta)$, with $I_\theta$ the Fisher-information matrix per single observation. If $\theta_N^{(0)}$ (the starting value) is any $\sqrt{N}$-consistent estimator then

$$\theta_N^{(1)} = \theta_N^{(0)} + N^{-1/2} [I_{\theta_N^{(0)}}]^{-1} \Delta_N(\theta_N^{(0)})$$

is asymptotically normal with

$$[\sqrt{n}(\theta_N^{(1)} - \theta)] \rightarrow N(0,I_\theta^{-1}).$$

Since all persons with the same raw score will end up with the same $\theta$-estimate, they can be treated as having the same ability value. This notion is used by Verhelst and Molenaar (1988) and by Baker (1987) to introduce least-squares logit estimation.

In the case of the Rasch model the logit model is

$$\text{logit } p_{1|s} = \theta_s - b_i,$$

in which $p_{1|s}$ denotes the probability of a person with score $s$ answering item $i$ correctly and $\theta_s$ the ability of persons with $s$ item answers correctly. Verhelst and Molenaar (1988) note that this model is not the same as the Rasch model, because in regression models the observed variables are functionally independent of the dependent variable while in
the Rasch model they are completely dependent. Verhelst and Molenaar (p. 288-292) compared weighted-least squares (WLS) estimators with CML estimators in some data settings. The WLS ability estimates sometimes failed to increase with total score. However, for simulated (perfect) data the WLS estimates multiplied by a constant came very close to the CML estimates.

Baker (1987), organized the data for each item in a $s \times 2$ contingency table. Here $s$ denotes the number of ability groups with midpoint $\theta_j$, containing $f_j$ examinees. For the two-parameter logistic model the logit ($P_{ij}$) is given by

$$\text{logit}(P_{ij}) = a_i(b_i - \theta_j).$$

Baker used a two stage iterative procedure for the joint estimation of item and ability parameters. In the first stage the ability parameters are substituted by their estimates and

$$x^2 = \sum_{j=1}^{s} f_j P_{ij} (1-P_{ij}) \{ \log(p_{ij}/(1-p_{ij})) - [a_i(b_i - \theta_j)] \}^2$$

is minimized to estimate the item parameters. In the second stage

$$x^2 = \sum_{i=1}^{n} f_i P_{ij} (1-P_{ij}) \{ \log(p_{ij}/(1-p_{ij})) - [a_i(b_i - \theta_j)] \}^2$$
is minimized to estimate $\theta_j$ for each ability group separately, while the item parameters are substituted by their estimates. Stage 1 and 2 are repeated until a convergence criterion is reached. Baker performed a simulation study in which the results improved as test length and sample size increased as well as when the test difficulty and the group ability were matched. Surprisingly, although the item parameters were sometimes poorly estimated, the ability estimates correlated high with the underlying ability parameters.

In conclusion, though logit based parameter estimation in item response theory is less expensive than ML estimation, the precision of the estimates is also less.

**Bayesian approach**

In the Bayesian approach prior distributions are imposed on the parameters of interest. Then, after the data is obtained Bayes' theorem is used to compute the posterior distribution.

Bayesian estimation starts with the specification of a certain parametric prior distribution or with the specification of empirical priors estimated from the data. Hierarchical Bayesian estimation arises if a distribution is specified for the parameters in the prior distribution.

The hierarchical Bayesian estimation procedure will be discussed in further detail because of its flexibility. However, the objection against Bayes’ procedures that no empirical evidence for the choice of the priors is given still applies to some extent. Here hierarchical Bayesian
estimation will be illustrated considering the three parameter model (as in Hambleton & Swaminathan, 1985)

\[ P_{v_i} = c_i \times (1-c_i)(1+\exp[-D a_i(\theta_v - b_i)])^{-1} \]

Let \( f(\theta_v) \) be the prior believe about the ability of examinee \( v (v=1,2,..,N) \) and let \( f(a_i), f(b_i) \) and \( f(c_i) \) be the prior believes about the parameters of item \( i (i=1,2,..,n) \). The joint posterior density of the parameters \( \theta, a, b, \) and \( c \) is

\[ f(\theta, a, b, c|x) \propto L(x|\theta, a, b, c) \prod_{i=1}^{n} f(a_i) f(b_i) f(c_i) \prod_{v=1}^{N} f(\theta_v) \]

It is necessary to take into account the restrictions of the parameter considered when specifying the prior. For example, since \( a_i \) is generally positive, an appropriate prior for \( a_i \) would be the chi-square distribution. The next stage is to specify the distributions of the parameters of the prior distribution. Once these distributions are specified, the values of the parameters \( \theta, a, b, \) and \( c \) that maximize the joint posterior distribution can be obtained.

The hierarchical Bayes' procedure yields good results, even in cases where maximum likelihood estimation performs rather badly (Hambleton & Swaminathan, 1985; Engelen, 1987b).
Other estimation procedures

Under the assumption that the two-parameter model fits the data and the ability is $N(0,1)$ distributed, one may consider the procedure described by Lord & Novick (1968, ch. 16.10) using point biserial correlation coefficients.

For the Rasch model other procedures are available. Minimum chi-square estimation is such a procedure, proposed by Fischer (1974). Let $N_{ij}$ denote the number of examinees that answers item $i$ correct and item $j$ wrong and let $N_{ji}$ be the number of examinees that answers item $j$ correct but item $i$ wrong. If the Rasch model fits the data

$$\frac{N_{ij}}{N_{ji}} = \frac{\exp(-b_i)}{\exp(-b_j)} = \exp(b_j - b_i).$$

Let $\delta_i = \exp(b_i)$, then

$$\Sigma_{i<j} \frac{(n_{ij}\delta_i - n_{ji}\delta_j)}{b_i b_j (n_{ij} + n_{ji})}$$

is the quantity to be minimized.

The Rasch model rewritten as a model for paired comparisons with ties, resulted in estimation by paired comparison. Here the responses of an examinee $v$ to a pair of items are compared. These response patterns give information about the relative difficulty of the two items for examinee $v$. For more details, see Engelen (1987b).
Item response theory for aggregated data

Introduction

According to Mislevy and Reiser (1983), there are two dimensions along which an application of IRT to large-assessment settings can vary: (1) the level at which an item response model is defined and (2) the level at which ability estimates are produced. The marginal maximum likelihood estimation procedure maintains the subject-level definition of an item response model, but just gives information about the ability distribution in the sample. In this chapter the focus will be on the group-level definition of item response models and their relationship to the more familiar subject-level models.

In contrast to item response models for the subject-level, a group-level item response model does not describe the probability of a response to an item from a specific examinee, but describes the probability of a response from an examinee selected at random from a specific group. By groups are meant salient groups, segments of a population (subpopulations) that can easily be identified such as sex, race, social economical class and urbanity. Salient groups make it possible to decide on curriculum issues concerning certain subpopulations. Furthermore, the items are classified in narrowly defined skill domains.
Item response models for groups

The probability of a correct response $x_{vgi}$ to item $i$ by an examinee $v$, selected at random from a subpopulation $g$ can be written as a function of $\theta_g$, the "ability" level of that subpopulation and the item parameters $\beta_i$:

$$P_{gj} = P(x_{vgi}=1) = H_i(\theta_g, \beta_i).$$

$H_i(\theta_g, \beta_i)$ is a (with respect to $\theta_g$) continuously differential and generally monotonically increasing function ranging from 0 to 1. Furthermore, $N_{gi}$ is the frequency of attempts to answer item $i$ by members from group $g$, out of which $R_{gi}$ were correct responses. The probability of observing the vector $R_g=(R_{g1}, R_{g2}, \ldots, R_{gn})$ correct responses among $N_g=(N_{g1}, N_{g2}, \ldots, N_{gn})$ attempts can be written as

$$\prod_{i=1}^{n} P_{gj}^{R_{gi}} (1-P_{gj})^{N_{gi}-R_{gi}}.$$ 

It is assumed that the responses of different examinees given the attainment level of the subpopulation $g$, are independent.

The following part is heavily based on Mislevy (1983), who shows under what conditions group-level item response models with $H_i(\theta_g, \beta_i)$ are implied by subject-level item response models with $H_i(\theta_{vg}, \beta_i)$.

Let $H_i(\theta_{vg}, \beta_i)$ be the subject-level item response curve of item $i$. Let $E_i$ be a continuous random nuisance variable
with mean zero and density function \( f_i \). The value of the response of a randomly selected member \( v \) of group \( g \), \( x_{vgi} \), is assumed to depend on the fixed item threshold value \( \beta_i \) and the person’s ability \( \theta_{vg} \). The possible values of this response are defined as follows

\[
\begin{cases}
  x_{vgi} = 1 & \text{if } \theta_{vg} + E_i > \beta_i \text{ or equivalently } \\
  \text{if } h_i = E_i + (\theta_{vg} - \theta_g) > \beta_i - \theta_g \\
  x_{vgi} = 0 & \text{otherwise.}
\end{cases}
\]

Let \( d_i \) be the density function of \( h_i \). The probability of a correct answer to item \( i \) by a random member \( v \) from group \( g \) is then given as

\[
P(X_{vgi} = 1 \mid \theta_g, \beta_i) = \int_{-\beta_i - \theta_g}^{\infty} d_i(h) \, dh = H_i(\theta_g - \beta_i) = H_i(\theta_g, \xi_1),
\]

where \( \xi_1 \), again, is the vector containing the item parameter of item \( i \).

Since ability only appears in the form of the mean group ability, it is assumed that all populations have the same ability distribution except for location. This assumption of homoscedasticity is a strong one and needs to be tested.

To test the assumption of homoscedasticity, the item parameters of the subject-level item response model need to be known or estimated. This means that at least two responses have to be elicited from each examinee. All the within-group
ability distributions should belong to the same known parametric distribution. The group ability parameters too, should follow a known parametric distribution. For more details about procedures and tests see Mislevy (1984).

So, recapitulating, \( H_i(\theta_g, I_i) \) is a group-level item response curve for item \( i \) under the assumption of the subject-level item response curve \( H_i(\theta_{vg}, I_i) \) and equal ability distributions within groups except for location.

Except in some special cases no simple closed form expression may exist for \( d_i \) and \( H_i(\theta_g, I_i) \). These exceptions are: (1) the group fixed effects model, (2) the two and three parameter normal-normal model, (3) normal logistic models and (4) the CAP model. These models will be discussed in the following paragraphs.

The group fixed-effects model

Reiser (1980, 1983) suggests the group fixed effects model, where it is assumed that grouping accounts for all variation among examinees. So, \( \theta_{vg} = \theta_g \) for \( v=1,2,...,N \) and \( g=1,2,...,m \). Because it is assumed that each examinee responds to only one item in the item domain, variability at the subject-level is considered as independent within-group error. The model is formulated as a logit model where

\[
z_{gi} = \log \frac{P(X_{vgi}=1)}{P(X_{vgi}=0)} = \log \frac{P(X_{vgi}=1)}{1-P(X_{vgi}=1)}
\]

and
\[ z_{gi} = b_{i} + k_{g}' \Theta a_{i}. \]

Here \( b_{i} \) and \( a_{i} \) are the item parameters, \( k_{g}' \) is a 1xm row vector from a design matrix \( K \) and \( \Theta \) is a mx1 vector of contrasts to be estimated among the sampled groups. The product \( k_{g}'\Theta \) specifies a weighted combination of effects from \( \Theta \) to produce the relative scale position of group \( g \) \((g=1,2,\ldots,m)\).

The log likelihood for the given data is

\[
\log L = \sum_{g=1}^{m} \sum_{i=1}^{n} (R_{gi} \log P(X_{vgi}=1|\Theta,a_{i},b_{i}) + (N_{gi} - R_{gi}) \log (1-P(X_{vgi}=1|\Theta,a_{i},b_{i}))) + \text{const.,}
\]

where \( R_{gi} \) is the number of correct responses in group \( g \) on item \( i \) and \( N_{gi} \) is the number of examinees in group \( g \) who respond to item \( i \). Parameters are estimated by an iterative procedure using Fisher's efficient scoring method, i.e.:

\[
\begin{bmatrix} h \\ a \\ \theta \end{bmatrix}^{t+1} = \begin{bmatrix} h \\ a \\ \theta \end{bmatrix}^{t} + (I_{t}(h,a,\Theta))^{-1} \begin{bmatrix} \delta h/\delta a \\ \delta h/\delta a \\ \delta h/\delta \theta \end{bmatrix}
\]

where \( t \) is the iteration step. If \( I_{t}(h,a,\Theta) \) is not of full rank, the method does not converge. Asymptotic standard
errors of the estimators are available as functions of the diagonal elements of \([I_t(h, a, \theta)]^{-1}\).

Goodness of fit of the model can be assessed using the Pearson’s chi-square or the likelihood ratio statistic.

**Two- and three-parameter normal-normal model**

Normal-normal indicates that both the subject-level item response density function and the subpopulation ability density function \(g_g\) are normal, in which case the group-level item response density functions are normal as well (Mislevy, 1983).

Let \(c_i\) be the guessing parameter, \(\beta_i\) the item threshold and \(\sigma_i\) the standard deviation of item \(i\) in a subject-level normal-ogive three parameter model. The probability of observing a correct response to item \(i\) by an examinee with ability \(\theta_{vg}\) is given as

\[
P(X_{vgi}) = c_i + (1-c_i) \Phi[(\theta_{vg} - \beta_i)/\sigma_i]
\]

Within the groups, \(\theta_{vg}\) is normally distributed with mean \(\theta_g\) and variance \(\sigma_g^2\). The probability of observing a correct answer of a randomly selected person from group \(g\) is then equal to
Results for the two-parameter normal ogive model at group-level follow as a special case of the three-parameter model in which \( c_i = 0 \).

**Normal-logistic model**

Mislevy (1983) shows how homoscedastic normal groups and a subject-level two- or three-parameter normal ogive item response model imply the existence of a corresponding group-level item response model. There is no similar result for logistic item response models, because the convolution of a logistic density with another logistic or normal distribution does not result in either a logistic or a normal density. There is, however, a possibility of approximating the logistic density with a normal one by \( \Phi(z) = \Psi(1.7z) \). In that case a logistic subject-level item response model is assumed to fit with item parameters \( \beta_i, \sigma_i \) and \( c_i \). This subject-level item response model is approximated by a normal subject-level item response model with item parameters \( \beta_i, 1.7\sigma_i \) and \( c_i \). If ability is assumed to be normally distributed in the subpopulations, then the procedure in the previous paragraph can be followed, resulting in an approximate group-level item response model.
Finally, the basic model in the Californian Assessment Program (Mislevy & Bock, 1984). This model is formulated at the level of detail necessary for diagnosing curricular effects: school level and skill element. Again the design permits every examinee to answer only one item on each skill domain. The probability of a random examinee $v$ from school $g$ answering item $i$ correctly is equal to

$$P(X_{vgi}=1) = \frac{\exp[(\theta_g - \beta_i)/\sigma_i]}{1 + \exp[(\theta_g - \beta_i)/\sigma_i]}$$

$$= \Phi((\theta_g - \beta_i)/\sigma_i).$$

Here $\theta_g$ is the average ability level of examinees in school $g$ for the skill element of interest. Item parameters $\beta_i$ and $\sigma_i$ are the item threshold and dispersion, respectively. The probability of a school pattern of numbers correct attempts $R_g = [R_{g1}, R_{g2}, \ldots, R_{gn}]$, given the total numbers of attempts $N_g = [N_{g1}, N_{g2}, \ldots, N_{gn}]$ is

$$P(R_g|N_g, \theta_g, \beta_i, \sigma_i) = \prod_{i=1}^{n} \frac{N_{gi}}{R_{gi}} P_{gi} R_{gi} (1-P_{gi})^{N_{gi}-R_{gi}}.$$

This equation is essential in the parameter estimation procedure. If this equation is employed in a design, where an examinee might sometimes respond to more than one item,
the school and item estimates are consistent but the resulting standard error of estimation would tend to be a little too small (Mislevy & Bock, 1984, p. 7).

The estimation procedure needs the assumption that the distribution of school scores in the sample is approximately normal, but it need not be assumed that the distribution is approximately normal in the population itself. Furthermore, the estimation procedure is based on the assumption that the model holds and uses the marginal maximum likelihood approach. After the calibration of the items a goodness-of-fit test is applied to evaluate this assumption.

The relation between group-level and subject-level item response models

Group-level item response models may be justified as aggregate descriptions of item response models on subject-level and interpreted analogously. Group-level item response models are implied by subject-level item response models only when within-group ability distributions are identical except for location (Mislevy, 1983).

In the context associated with the previous described models, every examinee answers only one item of each skill domain; hence individual ability levels can not be estimated. Even if some distinguished skill domains can be considered as one latent trait, there are still too few observations of each examinee, and ability estimates will have a considerable
measurement error. A more complex design with each examinee taking a few items per skill domain would provide more reliable estimates. In this design only one observation of each examinee will be used to estimate the group ability parameter.

Both at subject and group level, parameters are undetermined in their scale and in order to eliminate these indeterminacies some parameters (the number depends on the item response model) could be fixed arbitrarily. However, there is an important difference too. For the subject-level the addition of an examinee increases the number of incidental (ability) parameters. For the group-level, however, the number of ability parameters does not increase.

If a test indicates that the homoscedasticity assumption is not realistic, the detection of aberrant response patterns will become very interesting. But if only one observation is available, procedures used on subject-level as described by Kogut (1987a,1987b,1988) are not applicable.

So future research should try and find closed form expressions for a group-level item response model with less severe restrictions on the ability distribution within groups. Homoscedasticity tests and methods for detecting aberrant response patterns should be refined and adapted.
References


Titles of recent Research Reports from the Division of Educational Measurement and Data Analysis, University of Twente, Enschede, The Netherlands.

RR-90-2 H. Tobi, Item Response Theory at subject- and group-level
RR-90-1 P. Westers & H. Kelderman, Differential item functioning in multiple choice items
RR-89-6 J.J. Adema, Implementations of the Branch-and-Bound method for test construction problems
RR-89-5 H.J. Vos, A simultaneous approach to optimizing treatment assignments with mastery scores
RR-89-4 M.P.F. Berger, On the efficiency of IRT models when applied to different sampling designs
RR-89-3 D.L. Knol, Stepwise item selection procedures for Rasch scales using quasi-loglinear models
RR-89-2 E. Boekkooi-Timmenga, The construction of parallel tests from IRT-based item banks
RR-89-1 R.J.H. Engelen & R.J. Jannarone, A connection between item/subtest regression and the Rasch model
RR-88-18 H.J. Vos, Applications of decision theory to computer based adaptive instructional systems
RR-88-17 H. Kelderman, Loglinear multidimensional IRT models for polytomously scored items
RR-88-16 H. Kelderman, An IRT model for item responses that are subject to omission and/or intrusion errors
RR-88-15 H.J. Vos, Simultaneous optimization of decisions using a linear utility function
RR-88-14 J.J. Adema, The construction of two-stage tests
RR-88-13 J. Kogut, Asymptotic distribution of an IRT person fit index
RR-88-12 E. van der Burg & G. Dijksterhuis, Nonlinear canonical correlation analysis of multiway data
RR-88-11 D.L. Knol & M.P.F. Berger, Empirical comparison between factor analysis and item response models
RR-88-10 H. Kelderman & G. Macready, Loglinear-latent-class models for detecting item bias
RR-88-9  W.J. van der Linden & T.J.H.M. Eggen, The Rasch model as a
model for paired comparisons with an individual tie parameter

RR-88-8  R.J.H. Engelen, W.J. van der Linden, & S.J. Oosterloo, Item
information in the Rasch model

RR-88-7  J.H.A.N. Rikers, Towards an authoring system for item
construction

RR-88-6  H.J. Vos, The use of decision theory in the Minnesota
Adaptive Instructional System

RR-88-5  W.J. van der Linden, Optimizing incomplete sample designs for
item response model parameters

RR-88-4  J.J. Adema, A note on solving large-scale zero-one
programming problems

RR-88-3  E. Boekkooi-Tiemenga, A cluster-based method for test
construction

RR-88-2  W.J. van der Linden & J.J. Adema, Algorithmic test design
using classical item parameters

RR-88-1  E. van der Burg & J. de Leeuw, Nonlinear redundancy analysis

Research Reports can be obtained at costs from Bibliotheek,
Department of Education, University of Twente, P.O. Box 217,
7500 AE Enschede, The Netherlands.