Testing measurement invariance in a conditional likelihood framework by considering multiple covariates simultaneously

Clemens Draxler (clem@gmx.org)
UMIT TIROL
Andreas Kurz
University of Salzburg

Method Article

Keywords: Mixed logit model, conditional maximum likelihood, item parameter invariance, Rasch model

Posted Date: January 1st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3821799/v1

Additional Declarations: The authors declare potential competing interests as follows: The authors do not have any competing interests.

Testing measurement invariance in a conditional likelihood framework by considering multiple covariates simultaneously

Clemens Draxler[1] and Andreas Kurz[2]

[1]UMIT TIROL - Private University for Health Sciences and Health Technology

[2]University of Salzburg

December 29, 2023

Abstract

This article addresses the problem of measurement invariance in psychometrics. In particular, its focus is on the invariance assumption of item parameters in a class of models known as Rasch models. It suggests a mixed effects or random intercept model for binary data together with a conditional likelihood approach of both estimating and testing the effects of multiple covariates simultaneously. The procedure can also be viewed as a multivariate multiple regression analysis which can be applied in longitudinal designs to investigate effects of covariates over time or different experimental conditions. This work also derives four statistical tests based on asymptotic theory and a parameter-free test suitable in small sample size scenarios. Finally, it outlines generalizations for categorical data in more than two categories. All procedures are illustrated on real-data examples from behavioral research and on a hypothetical data example related to clinical research in a longitudinal design.

Keywords: Mixed logit model, conditional maximum likelihood, item parameter invariance, Rasch model

1 Introduction

This article addresses some basic problems in psychometrics. Its focus is on issues connected with statistical inference on measurement invariance. In general, the term refers to assuming the same measurement principles for different groups of persons or examinees in the population of interest. Specifically, in this work, it refers to testing the hypothesis of invariance of item parameters of the Rasch model (Rasch, 1960, Fischer & Molenaar, 1995) across multiple groups of persons in a conditional maximum likelihood (CML) framework (Andersen, 1970, Pfanzagl, 1993, Skrondal & Rabe-Hesketh, 2022). As such it can also be viewed as a procedure for investigating differential item functioning (e.g., Holland & Wainer, 1993). Statistical tests based on asymptotic theory serving this purpose have been discussed by, e.g., Andersen (1973), Glas and Verhelst (1995), Draxler and Alexandrowicz (2015), Draxler et al. (2022), and Draxler et al. (2023). Tests not based on asymptotic theory and particularly suited for small samples are suggested by Ponocny (2001), Verhelst (2008), Draxler and Zessin (2015), Draxler and Dahm (2020), and Draxler and Kurz (2021). Program packages are readily available, for instance, the R (R core team, 2022) packages eRm (Mair et al., 2023, Mair & Hatzinger, 2007) and tcl (Draxler & Kurz, 2023).

All of these tests consider only one covariate, for example, testing the equality of item parameters between two or more gender groups. In a typical application, when analyzing psychometric data or developing new educational or psychological tests, one is usually interested in more than one covariate, e.g., gender, age, ethnicity, etc. The usual approach then is to carry out a statistical test for each covariate separately. The drawback of it is that error probabilities accumulate. In more sophisticated terms, the so-called probability of the error of the first kind or the type I error probability or the size of the test increases. This means that the probability of falsely rejecting the hypothesis of invariance in at least one of the (multiple) tests is greater than in each of the single tests and is often left uncontrolled (unless no adjustments are made). In other words, one obtains an uncontrolled probability of wrongly stating that at least one covariate (one or more) does have an effect or does violate the invariance assumption when none of the covariates do. For example, when five covariates are considered and five tests each with a predetermined size of 0.05 are carried out, the probability that at least one of them results in an error of the first kind amounts to $1-(1-0.05)^5=0.226$, provided the tests are independent (i.e., conditional on the result of any of the tests the type I prob. of any other test does not change). When covariates are correlated (when their true correlations are not 0) this number will be smaller depending on the sizes of correlations. Simple adjustments of the type I error probabilities of the individual tests are based on the assumption of independence of the covariates. Hence, such procedures are too conservative when covariates correlate, but the true correlations are usually unknown which prevents researchers from making appropriate and exact adjustments in practice.

This article suggests a solution. It discusses an approach and a model that considers any desirable number of covariates (as long as parameters are identified) and allows both estimating and testing their effects simultaneously in a conditional likelihood framework. Thus, only one hypothesis test is needed whose size or probability of the error of the first kind can be predetermined and controlled at any level. The model is a generalization of the Rasch model and can be viewed as a mixed model with logit link function, i.e., a mixed logit model, since it considers random (persons or examinees) and fixed (items and covariates) effects. It can also be viewed as a multivariate multiple regression model for binary (or other categorical) data (Cox, 1958). It is multivariate since it considers multiple items (to which persons respond) and it considers multiple (more than one) covariates or predictors or explanatory variables that linearly affect the logits of item response probabilities. The model not only allows for binary or nominal covariates. It also considers linear effects of real-valued covariates. This model has already been discussed by Gürer and Draxler (2022) in the context of machine learning and penalizing techniques of conditional likelihood functions but without providing a respective hypothesis test.

The remainder of this text is organized as follows. Sec. 2 introduces the model and discusses the theoretical foundation and technical issues of the approach. Sec. 3 derives four different test statistics based on asymptotic theory serving the present purpose. It also presents a parameter-free test that can be used in small sample scenarios. Sec. 4 gives an outline on generalizations. Sec. 5 provides real data examples and notes further applications in longitudinal designs. Sec. 6 gives a discussion and final remarks.

2 Theoretical foundation

Consider a parametric family of probability distributions specified by a psychometric model and indexed by parameters taking values in parameter space Θ being an open subset of Euclidean space. Assume that the true unknown probability distribution generating the observations in a sample space (from which the data are sampled) is a member of that family. The observations are obtained by the binary responses, e.g., correct, or incorrect, of a number (sample) of persons or examinees to a number of items. Additionally, data on a number of covariates are collected. The psychometric model is given by

$$P(Y_{ij}=1)=\frac{\exp\left(\tau_i+\alpha_j+\sum_p \delta_{jp}x_{ip}\right)}{1+\exp(\tau_i+\alpha_j+\sum_p \delta_{jp}x_{ip})},\ i=1,\dots,n,\ j=1,\dots,k,\ p=1,\dots,q, \quad (1)$$

where $Y_{ij}\in\{0,1\}$ is the binary response of person i to item j and $\tau_i\in\mathbb{R}$ is a person parameter usually interpreted as an ability, proficiency, or attitude. The parameters $\alpha_j\in\mathbb{R}$ and $\delta_{jp}\in\mathbb{R}$ characterize item effects and are usually interpreted as easiness or attractiveness. The former represents a baseline parameter of the respective item or a general level of easiness of the item (i.e., when all covariate values are 0) and the latter a conditional effect of item j given covariate p (i.e., a slope parameter). The x quantities are the observed covariate values, i.e., x_{ip} is the value observed for person i in respect of covariate p. Setting all the δ parameters equal to 0 (no covariate has an effect) yields the Rasch model as a special case with the α parameters as the item parameters. Assume that the k responses of every single person are independent, i.e., local independence, and the persons are drawn independently from a population of interest. Let the binary responses of all n persons in the sample to all k items be arranged in an $n\times k$ matrix denoted by $\boldsymbol{Y}$, i.e., the response matrix. Then, by using matrix multiplication, the joint distribution of all these responses is obtained by

$$
\begin{aligned}
P(\boldsymbol{Y}=\boldsymbol{y}) &= \prod_{i=1}^{n}\prod_{j=1}^{k} P(Y_{ij}=y_{ij}) \\
&= \underbrace{\frac{1}{\prod_i \prod_j \left(1+\exp(\tau_i+\alpha_j+\sum_p \delta_{jp}x_{ip}\right)}}_{C(\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{\delta})} \exp\left(\sum_i\sum_j\sum_p \left(y_{ij}(\tau_i+\alpha_j+\delta_{jp}x_{ip})\right)\right) \\
&= C(\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{\delta}) \exp\left(\boldsymbol{\tau}^\top \boldsymbol{r} + \boldsymbol{\alpha}^\top \boldsymbol{s} + \boldsymbol{\delta}^\top \boldsymbol{t}\right),
\end{aligned}
$$

where $\boldsymbol{\tau}^\top = (\tau_1,\ldots,\tau_n)$ denotes an $n \times 1$ matrix (i.e., a column vector of length n) containing as elements all person parameters, $\boldsymbol{\alpha}^\top = (\alpha_1,\ldots,\alpha_k)$ a $k \times 1$ matrix of baseline parameters, and $\boldsymbol{\delta}^\top = (\boldsymbol{\delta}_1^\top,\ldots,\boldsymbol{\delta}_q^\top)$ with $\boldsymbol{\delta}_p^\top = (\delta_{1p},\ldots,\delta_{kp})$ a $kq \times 1$ matrix of the effects of all q covariates on all k items. Both τ and α parameters are nuisances in the present problem. The only parameters of interest are the δ parameters. For identifiability let $\alpha_1 = 0, \delta_{1p} = 0\ \forall p$, i.e., one element of $\boldsymbol{\alpha}$ and one of each $\boldsymbol{\delta}_p$ is not free. Note that this is not a sufficient condition for the parameters to be identified. It can immediately be seen from the factorization criterion that the statistics $\boldsymbol{R}$ (with realization $\boldsymbol{r}$), $\boldsymbol{S}$ (with realization $\boldsymbol{s}$), and $\boldsymbol{T}$ (with realization $\boldsymbol{t}$) are sufficient for the class of distributions. It is a member of a multiparameter exponential family with $(\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{\delta}) \in \Theta \subseteq \mathbb{R}^{n+k(q+1)}$ as its natural parameter space, i.e., the first factor is a normalizing constant (that does not depend on $\boldsymbol{Y}=\boldsymbol{y}$) and the second factor depends on the observations only through the sufficient statistics

$$\boldsymbol{R}^\top = (R_1,\ldots,R_n),\ R_i = \sum_{j=1}^{k} Y_{ij},$$

$$\boldsymbol{S}^\top = (S_1,\ldots,S_k),\ S_j = \sum_{i=1}^{n} Y_{ij},$$

$$\boldsymbol{T}^\top = (\boldsymbol{T}_1^\top,\ldots,\boldsymbol{T}_q^\top),\ \boldsymbol{T}_p^\top = (T_{1p},\ldots,T_{kp}),\ T_{jp} = \sum_{i=1}^{n} Y_{ij}x_{ip}.$$

Thus, the $n \times 1$ matrix $\boldsymbol{R}$ contains the row sums or the person scores of the response matrix $\boldsymbol{Y}$, the $k \times 1$ matrix $\boldsymbol{S}$ the column sums or item scores, and the $kq \times 1$ matrix $\boldsymbol{T}$ weighted column sums, where the weights are given by the respective covariate values. Hence, further considerations can be restricted to the distributions of the sufficient statistics. The joint distribution of $\boldsymbol{R}$, $\boldsymbol{S}$, and $\boldsymbol{T}$, the marginal distribution of $\boldsymbol{R}$, and the conditional distribution of $\boldsymbol{S}, \boldsymbol{T}$ given $\boldsymbol{R} = \boldsymbol{r}$ are obtained by

$$P(\boldsymbol{R} = \boldsymbol{r}, \boldsymbol{S} = \boldsymbol{s}, \boldsymbol{T} = \boldsymbol{t}) = C(\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \exp(\boldsymbol{\tau}^\top \boldsymbol{r} + \boldsymbol{\alpha}^\top \boldsymbol{s} + \boldsymbol{\delta}^\top \boldsymbol{t}) h(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{t}),$$

$$P(\boldsymbol{R} = \boldsymbol{r}) = C(\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \exp(\boldsymbol{\tau}^\top \boldsymbol{r}) \prod_{i=1}^{n} \gamma_{r_i}(\boldsymbol{\alpha}, \boldsymbol{\delta}, x_{i1}, \ldots, x_{iq}),$$

and

$$\begin{aligned} P(\boldsymbol{S} = \boldsymbol{s}, \boldsymbol{T} = \boldsymbol{t} \mid \boldsymbol{R} = \boldsymbol{r}) &= \frac{P(\boldsymbol{R} = \boldsymbol{r}, \boldsymbol{S} = \boldsymbol{s}, \boldsymbol{T} = \boldsymbol{t})}{P(\boldsymbol{R} = \boldsymbol{r})} \\ &= \frac{\exp(\boldsymbol{\alpha}^\top \boldsymbol{s} + \boldsymbol{\delta}^\top \boldsymbol{t})}{\prod_{i=1}^{n} \gamma_{r_i}(\boldsymbol{\alpha}, \boldsymbol{\delta}, x_{i1}, \ldots, x_{iq})} h(\boldsymbol{s}, \boldsymbol{t} \mid \boldsymbol{r}), \end{aligned}$$

where $h(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{t}) = h(\boldsymbol{s}, \boldsymbol{t} \mid \boldsymbol{r})$ is a combinatorial function denoting the number of potential $n \times k$ response matrices that yield $\boldsymbol{R} = \boldsymbol{r}$, $\boldsymbol{S} = \boldsymbol{s}$, and $\boldsymbol{T} = \boldsymbol{t}$. It can be ignored since it does not depend on any of the parameters. The function γ_{r_i} denotes an elementary symmetric function of order r_i, where $r_i \in \{0, \ldots, k\}$ denotes the score of person i which can be an integer from 0 to k. In the present problem, it is not only a function of all the item parameters (i.e., baseline and effect parameters) but also all the covariates. For persons with different covariate values one yields different γ functions (provided that the δ parameters are not 0). Only persons with exactly the same covariate values (in respect of every covariate) yield the same γ functions. They are given by

$$
\begin{aligned}
&\gamma_0(\boldsymbol{\alpha}, \boldsymbol{\delta}, x_{i1}, \dots, x_{iq}) = 1\\
&\gamma_1(\boldsymbol{\alpha}, \boldsymbol{\delta}, x_{i1}, \dots, x_{iq}) = \exp(\alpha_1 + \sum_p \delta_{1p} x_{ip}) + \cdots + \exp(\alpha_k + \sum_p \delta_{kp} x_{ip})\\
&\gamma_2(\boldsymbol{\alpha}, \boldsymbol{\delta}, x_{i1}, \dots, x_{iq}) = \exp\big(\alpha_1 + \alpha_2 + \sum_p (\delta_{1p} + \delta_{2p}) x_{ip}\big)\\
&\quad + \cdots + \exp\big(\alpha_{k-1} + \alpha_k + \sum_p (\delta_{k-1,p} + \delta_{kp}) x_{ip}\big)\\
&\gamma_3(\boldsymbol{\alpha}, \boldsymbol{\delta}, x_{i1}, \dots, x_{iq}) = \exp\big(\alpha_1 + \alpha_2 + \alpha_3 + \sum_p (\delta_{1p} + \delta_{2p} + \delta_{3p}) x_{ip}\big)\\
&\quad + \cdots + \exp\big(\alpha_{k-2} + \alpha_{k-1} + \alpha_k + \sum_p (\delta_{k-2,p} + \delta_{k-1,p} + \delta_{kp}) x_{ip}\big)\\
&\vdots\\
&\gamma_{k-1}(\boldsymbol{\alpha}, \boldsymbol{\delta}, x_{i1}, \dots, x_{iq}) = \exp\big(\alpha_1 + \cdots + \alpha_{k\text{-}1} + \sum_p (\delta_{1p} + \cdots + \delta_{k-1,p}) x_{ip}\big)\\
&\quad + \cdots + \exp\big(\alpha_2 + \cdots + \alpha_k + \sum_p (\delta_{2p} + \cdots + \delta_{kp}) x_{ip}\big)\\
&\gamma_k(\boldsymbol{\alpha}, \boldsymbol{\delta}, x_{i1}, \dots, x_{iq}) = \exp\big(\alpha_1 + \cdots + \alpha_k + \sum_p (\delta_{1p} + \cdots + \delta_{kp}) x_{ip}\big).
\end{aligned}
$$

Thus, γ_1 is composed of a sum over all items, each summand being a function of the baseline and the q covariate effect parameters associated with the respective item (i.e., one α par. and one δ par. for each covariate), γ_2 is composed of a sum over all (potential) pairs of items, each summand being a function of the two baseline and the $2q$ covariate effect parameters associated with the respective pair (i.e., one α par. for each item and one δ par. for each item and each covariate), γ_3 is composed of a sum if all (potential) triples of items, each summand being a function of the three baseline and the $3q$ covariate effect parameters associated with the respective triple, etc.

The conditional distribution $P(\boldsymbol{S} = \boldsymbol{s}, \boldsymbol{T} = \boldsymbol{t} \mid \boldsymbol{R} = \boldsymbol{r})$ does not depend on the person parameters. Treating it as function of the remaining parameters and taking the logarithm yields the conditional log likelihood

$$
\ell(\boldsymbol{\alpha}, \boldsymbol{\delta}) = \boldsymbol{\alpha}^\top \boldsymbol{s} + \boldsymbol{\delta}^\top \boldsymbol{t} - \sum_{i=1}^{n} \log \gamma_{r_i}(\boldsymbol{\alpha}, \boldsymbol{\delta}, x_{i1}, \dots, x_{iq}), \tag{2}
$$

where the additive constant $\log h(\boldsymbol{s}, \boldsymbol{t} \mid \boldsymbol{r})$ is omitted.

From general likelihood theory and exponential families as well as particular results for the conditional likelihood case (Andersen, 1970, Pfanzagl, 1993) one readily obtains the score function, the Fisher information matrix, estimates of the parameters, and their properties. The vector-valued score function denoted by $\boldsymbol{D}$ is given by the first order partial derivatives of $\ell(\boldsymbol{\alpha}, \boldsymbol{\delta})$ with respect to all free α and δ parameters. Note that the first α parameter and the first δ parameter for each covariate has been set to 0 for identifiability. Thus, it has only length $(k-1)(q+1)$. It is a function of the last $k-1$ elements of $\boldsymbol{s}$ and every $\boldsymbol{t}_p$ as well as the free α and δ parameters. It is simply given by the differences of the observed and expected values of the sufficient statistics $\boldsymbol{S}$ and $\boldsymbol{T}$ conditional on $\boldsymbol{R} = \boldsymbol{r}$ which holds generally for exponential families. The Fisher information matrix denoted by $\boldsymbol{F}(\boldsymbol{\alpha}, \boldsymbol{\delta})$ can be obtained from the second order partial derivatives of $\ell(\boldsymbol{\alpha}, \boldsymbol{\delta})$ with respect to all free parameters. Details and computational issues on information matrix and score function are given in Appendix A.

The CML estimate of $(\boldsymbol{\alpha}, \boldsymbol{\delta})$ is defined by

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\delta}}) := \underset{(\boldsymbol{\alpha},\boldsymbol{\delta})\in\mathbb{R}^{k(q+1)}}{\arg\max} \; \ell(\boldsymbol{\alpha}, \boldsymbol{\delta})$$

which is obtained by solving $\boldsymbol{D} = \mathbf{0}_{(k-1)(q+1)}$ for the α and δ parameters. An R code is provided as supplementary material in an online repository which uses a numerical procedure known as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Broyden, 1970, Fletcher, 1970, Goldfarb, 1970, Shanno, 1970). In order to obtain the usual asymptotic properties of maximum likelihood estimates a mild regularity condition has to be considered in the CML case (Andersen, 1970, Pfanzagl, 1993), i.e., the values of the person parameters (τ parameters) must not be too extreme. Then, it holds that

$$(\widehat{\boldsymbol{\alpha}}_*, \widehat{\boldsymbol{\delta}}_*) \xrightarrow{P} (\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*)$$

and

$$\sqrt{n}\Big((\widehat{\boldsymbol{\alpha}}_*, \widehat{\boldsymbol{\delta}}_*) - (\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*)\Big) \xrightarrow{D} N\big(\mathbf{0}_{(k-1)(q+1)}, \boldsymbol{F}^{-1}(\boldsymbol{\alpha}, \boldsymbol{\delta})\big)$$

when the number of persons $n \to \infty$, where $\boldsymbol{F}^{-1}(\boldsymbol{\alpha}, \boldsymbol{\delta})$ is the asymptotic covariance matrix of the CML estimate. Note that the notation $\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*$ is used to indicate that the respective vector contains only the free parameters.

Finally, a brief note on identifiability of parameters. A parameter is obviously not identified or estimable when the conditional distribution given the observed value of its sufficient statistic is only 0 and 1, i.e., when exactly one response pattern is associated with the respective value of the sufficient statistic. Such data are completely uninformative. Necessary and sufficient conditions for all the parameters to be identified have only been given for the Rasch model thus far (Fischer, 1981).

3 Statistical tests

Let the following two subclasses of distributions or, equivalently, parameter spaces be defined by

$$\Theta_1 = \{(\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \mid (\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \in \Theta, \boldsymbol{\delta} = \mathbf{0}\}$$

and

$$\Theta_2 = \{(\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \mid (\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \in \Theta, \boldsymbol{\delta} \neq \mathbf{0}\},$$

$\Theta_1 \cup \Theta_2 = \Theta$. Of interest is the hypothesis that the true unknown distribution (or the true parameters) satisfies

$$(\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \in \Theta_1$$

(as assumed by the Rasch model) against the alternative

$$(\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \in \Theta_2.$$

The latter represents the scenario that at least one of the q covariates has an effect on at least one item.

Four different test statistics derived from the properties of the CML estimates based on asymptotic theory may be used. Let $\widehat{\boldsymbol{\alpha}}_0$ denote the restricted CML estimate of $\boldsymbol{\alpha}$, i.e., the argument of the maximum of $\ell(\boldsymbol{\alpha}, \boldsymbol{\delta})$ given $\boldsymbol{\delta} = \mathbf{0}$. Thus, the vector $\widehat{\boldsymbol{\alpha}}_0$ contains the estimates of the item parameters of the Rasch model. Then, one obtains a likelihood ratio test statistic (Neyman and Pearson, 1928, Wilks, 1938) by evaluating $\ell(\boldsymbol{\alpha}, \boldsymbol{\delta})$ at both the restricted and the unrestricted estimates:

$$LR = -2\Big(\ell\big(\widehat{\boldsymbol{\alpha}}_0, \boldsymbol{\delta} = \mathbf{0}\big) - \ell\big(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\delta}}\big)\Big).$$

This is simply a generalization of the well-known Andersen likelihood ratio test (Andersen, 1973) for the case of more than one covariate and non-binary covariates. A Rao score test statistic (Rao, 1948) can be obtained by

$$RS = \boldsymbol{D}^\top(\widehat{\boldsymbol{\alpha}}_0, \boldsymbol{\delta} = \mathbf{0})\boldsymbol{F}^{-1}(\widehat{\boldsymbol{\alpha}}_0, \boldsymbol{\delta} = \mathbf{0})\boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_0, \boldsymbol{\delta} = \mathbf{0}),$$

where both the score function $\boldsymbol{D}$ and the information matrix $\boldsymbol{F}$ are evaluated only at the restricted estimates. Thus, $\boldsymbol{\delta}$ need not be estimated at all. This is quite a remarkable feature of the score test which sets it apart from others. Note that the score test is also called Lagrange multiplier test (mainly in econometrics) since the test statistic can be expressed in terms of Lagrange multipliers (Silvey, 1959). A Wald test statistic (Wald, 1943) is given by

$$W = \widehat{\boldsymbol{\delta}}_*^\top \boldsymbol{\Sigma}\big(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\delta}}\big)\widehat{\boldsymbol{\delta}}_*,$$

where the notation $\widehat{\boldsymbol{\delta}}_*$ (again) is used to indicate that the respective vector is assumed to contain only the estimates of the free δ parameters. The matrix $\boldsymbol{\Sigma}$ denotes the covariance matrix of $\widehat{\boldsymbol{\delta}}_*$. It is obtained by omitting the first $k-1$ rows and columns (that refer to $\boldsymbol{\alpha}_*$ which is not of interest) of the complete covariance matrix $\boldsymbol{F}^{-1}$ evaluated at the unrestricted estimates $\big(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\delta}}\big)$. Finally, a gradient test statistic (Terrel, 2002, Lemonte, 2016) is obtained by

$$G = \boldsymbol{D}_*^\top(\widehat{\boldsymbol{\alpha}}_0, \boldsymbol{\delta} = \mathbf{0})\widehat{\boldsymbol{\delta}}_*,$$

where $\boldsymbol{D}_*^\top = (\boldsymbol{D}_1^\top, \ldots, \boldsymbol{D}_q^\top)$ is evaluated at the restricted estimates. All four test statistics have a common limiting distribution when $(\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \in \Theta_1$ holds true and $n \to \infty$. It is the central χ^2 distribution with $df = (k-1)q$, i.e., $k-1$ free δ parameters per covariate.

Note that the gradient test is a relatively recent development in the theory of statistics. It is, thus, still little known in psychological and educational communities. It has only been discussed in the context of psychometric problems in three articles. Draxler et al. (2022, 2023) discuss it in a conditional likelihood and Zimmer et al. (2023) in a marginal likelihood framework. The test statistic can be derived from a combination of the Rao score and Wald test statistics. An obvious computational advantage of it is that it does not depend on an information matrix.

3.1 A parameter-free test

Test statistics whose (exact) distribution is known (since it does not depend on unknown nuisance parameters) or can practically well approximated in case of small sample sizes may also be easily derived. This approach considers conditioning on the observed values of sufficient statistics for all nuisance parameters ($\boldsymbol{\tau}$ and $\boldsymbol{\alpha}$) which are given by the row and column sums of the response matrix. Thus, it considers the conditional distribution of the sufficient statistics for the δ parameters which are the only parameters of interest (in the present problem). The marginal distribution of the vectors of sufficient statistics for all nuisance parameters is given by

$$P(\boldsymbol{R}=\boldsymbol{r},\boldsymbol{S}=\boldsymbol{s}) = C(\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{\delta})\exp(\boldsymbol{\tau}^\top\boldsymbol{r}+\boldsymbol{\alpha}^\top\boldsymbol{s})\sum_{\Omega}\exp(\boldsymbol{\delta}^\top\boldsymbol{t}).$$

The conditional distribution is then obtained by

$$P(\boldsymbol{T}=\boldsymbol{t}\mid\boldsymbol{R}=\boldsymbol{r},\boldsymbol{S}=\boldsymbol{s}) = \frac{P(\boldsymbol{R}=\boldsymbol{r},\boldsymbol{S}=\boldsymbol{s},\boldsymbol{T}=\boldsymbol{t})}{P(\boldsymbol{R}=\boldsymbol{r},\boldsymbol{S}=\boldsymbol{s})} = \frac{\exp(\boldsymbol{\delta}^\top\boldsymbol{t})}{\sum_{\Omega}\exp(\boldsymbol{\delta}^\top\boldsymbol{t})},$$

where Ω denotes the restricted sample space (that follows from the conditioning) consisting of all potential $n\times k$ response matrices yielding row and column sums $\boldsymbol{R}=\boldsymbol{r}$ and $\boldsymbol{S}=\boldsymbol{s}$ (i.e., yielding the same row and column sums as the observed response matrix). The summations on the right sides of the two equations have accordingly to be taken over all elements of Ω. Note that this is a summation over all potential values of the vector-valued statistic $\boldsymbol{T}$, i.e., matrices in the sample space can yield different values of $\boldsymbol{T}$, where the range of possible values for each element of the vector $\boldsymbol{T}$ is determined by the condition $\boldsymbol{R}=\boldsymbol{r}$ and $\boldsymbol{S}=\boldsymbol{s}$. Thus, it is a normalizing constant ensuring that the respective probabilities sum up to 1. Again, treating this conditional distribution as a function of the only remaining parameter vector and taking the logarithm yields a conditional log likelihood function

$$\ell(\boldsymbol{\delta}) = \boldsymbol{\delta}^\top \boldsymbol{t} - \log \sum_{\Omega} \exp(\boldsymbol{\delta}^\top \boldsymbol{t}).$$

Since this conditional distribution is also a multiparameter exponential family (as can immediately be seen) all well-known results of likelihood and asymptotic theory in respect of properties of the estimates and the distribution of the four χ^2 test statistics hold true and are principally applicable in this case too. Technical details are given in Appendix B.

Practically, applying asymptotic theory does not make much sense in this case since the enumeration of all potential matrices, i.e., all elements contained in Ω, is computationally infeasible for cases with usual numbers of persons and items. Miller and Harrison (2013) solved the complicated combinatorial problem of determining the exact number of matrices, i.e., the cardinality of Ω, by deriving a recursive algorithm based on graph theory but computing it is nevertheless very intensive and this number itself does not suffice for the purpose of determining the exact conditional distribution of the statistic $\boldsymbol{T}$. One needs to enumerate all matrices. Fortunately, algorithms designed to sample each element of Ω with approximately the same probability (i.e., to obtain a simple random sample) can be applied to approximate the conditional distribution of $\boldsymbol{T}$. Verhelst (2008), for instance, suggests a Markov chain Monte Carlo algorithm whose stationary distribution is given by a discrete uniform distribution of the elements of Ω. Miller and Harrison (2013) suggest an exact sampling approach, i.e., each element is selected with exactly the same probability. Given $\boldsymbol{\delta} = \boldsymbol{0}$ (i.e., the hypothesis of interest) and having obtained a simple random sample of matrices (by applying one of these algorithms) the exact conditional distribution of $\boldsymbol{T}$ can be arbitrarily well approximated by simply considering the distribution of relative frequencies with which the different values of every element of $\boldsymbol{T}$ are observed in the random sample (of all matrices drawn). Verhelst's algorithm seems to be the most efficient choice in respect of computing time so far (e.g., Draxler & Nolte, 2018) and it is readily available as an R package called RaschSampler (Verhelst et al., 2007).

An obvious choice of a test statistic is the Rao score since score function and information matrix (given in Appendix B) have to be evaluated only at $\boldsymbol{\delta} = \boldsymbol{0}$. As such it can be viewed as a parameter-free test. Since the score function is simply given by the difference of observed and expected value of $\boldsymbol{T}$ and since the expected value is a constant the exact distribution of the score function is the same as the conditional distribution of $\boldsymbol{T}$ itself. Accordingly, it can be arbitrarily well approximated from a simple random sample of matrices drawn. One simply computes the mean for each element of $\boldsymbol{T}$ of all matrices drawn as an approximation or estimate of the respective expected value and considers the difference to the observed value in every one of the matrices drawn. Similarly, one obtains an approximation of the information matrix by simply computing the sample covariances from all the matrices drawn. Henceforth, one obtains a value of the Rao score test statistic for each matrix drawn and the p-value for the observed response matrix is obtained from that distribution.

This test can of course be recommended in scenarios of small sample sizes when a poor approximation (of the exact distribution) by χ^2 is to be expected.

3.2 Power function of the tests

The power of a test is a function of the unknown parameters given its size (type I error probability) and the ample size. In the multiparameter case it seems to be convenient and practical to use a function of all the unknown parameters. It is typically called an effect measure. In the present problem, it is easily obtained by dividing the respective χ^2 test statistic by the informative sample size (e.g., Cohen, 1988, Draxler, 2010, Draxler & Alexandrowicz, 2015). This yields a sort of pseudo R^2 which can be interpreted as the proportion of variance explained by the covariates considered. With the term informative sample the following is meant. Persons with a score (i.e., row sum in the response matrix $\boldsymbol{Y}$) of 0 or k have to be excluded from the total sample since their responses are completely uninformative. They do not contribute to the test statistics' values.

The Rasch model does not allow any differences in the logits of response probabilities between persons with different covariate values (only between persons with different person parameters). Thus, the effect (of all the covariates) is 0. In this case, the power of the tests equals their given size (i.e., type I error prob.). Figure 1 shows examples of power curves for different informative sample sizes given a type I error probability of 0.05 and given the number of degrees of freedom of the test is 20. Note that the case $df = 20$ can be obtained in different scenarios regarding the numbers of covariates and items, for instance, when $q = 1\,, k = 21$ or $q = 2\,, k = 11$ or $q = 4\,, k = 6$. In case of an informative sample of size 300, for example, the

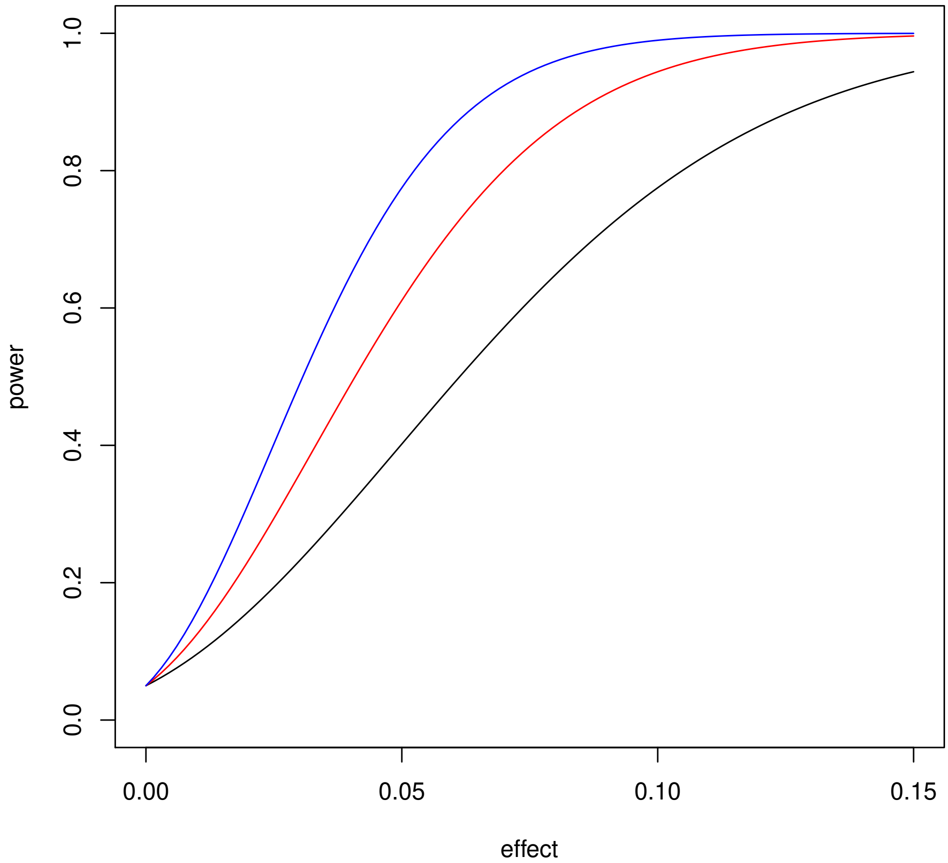

Figure 1. Power curves given a type I error probability of 0.05 and $df = 20$. The black line represents the case of an informative sample size of 200, the red line of 300, and the blue line of 400.

power yields 0.61 given an effect of 0.05 (one twentieth of explained variance by the covariates), and it yields 0.94 given an effect of 0.1 (one tenth of explained variance). Thus, when the true effect is 0.05 or greater the Rasch model is rejected with a probability of at least 0.61. When the true effect is 0.1 or greater the Rasch model is rejected with a probability of at least 0.94. Note that the validity (and accuracy) of all these considerations depends on the χ^2 approximation of the distribution of the respective test statistics, i.e., the non-central χ^2 with $df = (k-1)q$ and non-centrality parameter given by the product of the effect and informative sample size (in case of an effect of 0 it reduces to the central χ^2, of course) (Draxler & Alexandrowicz, 2015). If it is poor the power function may also be inaccurate.

4 Outline of generalizations

Given the theoretical foundation of the approach of testing invariance of item parameters discussed in this work, a generalization to models that consider item responses in more than two nominal or ordinal categories is obvious and straightforward. It is also of great practical interest. For instance, the partial credit model (Masters, 1982) that considers ordinal responses in potentially more than two categories is one of the most popular and most frequently applied models in psychometric problems. In general, the following multiparameter exponential family may principally be considered:

$$P(\boldsymbol{Y} = \boldsymbol{y}) = C(\boldsymbol{\vartheta}, \boldsymbol{\theta}) \exp\left(\boldsymbol{\vartheta}^\top \boldsymbol{u}_1 + \boldsymbol{\theta}^\top \boldsymbol{u}_2\right) h(\boldsymbol{y}),$$

where $(\boldsymbol{\vartheta}, \boldsymbol{\theta}) \in \Theta$ with $\boldsymbol{\vartheta}$ as a vector of nuisance parameters and $\boldsymbol{\theta}$ a vector of parameters of interest. The former typically represents characteristics of persons and the latter is a vector of parameters that represent characteristics of items and their response categories as well as the effects of covariates. The vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are the observed values of their respective sufficient statistics which are functions of the data $\boldsymbol{y}$ (being nonnegative integers representing the responses of the persons to the items) and the covariate values.

The derivation of the conditional distribution given the observed values of the sufficient statistics for the nuisance parameters $P(\boldsymbol{U}_2 = \boldsymbol{u}_2 \mid \boldsymbol{U}_1 = \boldsymbol{u}_1)$, the respective conditional likelihood function, CML estimates, and their properties is straightforward for the given class of models. Statistical tests are obtained along the same lines as discussed in Sec. 3.

5 Data Examples

This sec. is aimed at illustrating the application of the estimation and testing procedures on a number of real-data and hypothetical examples and the interpretation of the results.

5.1 Example 1

The first example of data can be found in the R package sirt (Robitzsch, 2022). The files are called data.pisaMath.rda and data.pisaRead.rda. The former contains binary responses of 565 students to 11 mathematics items and the latter binary responses of 623 students to 12 reading items. Both consider three covariates: gender which is binary, a real-valued index of the socio-economic status (of the students' families) called hisei, and migration background abbreviated by migra which is again a binary covariate. Results of estimates and respective standard errors are shown in Table 1. As can be seen standard errors are smaller for the estimates referring to the real-valued covariate hisei. The largest standard errors are obtained for the parameters referring to covariate migra, i.e., this covariate provides little information since the proportion of students with migration background in the sample is quite low. Appendix C provides additional tables with Z test statistics for each single parameter, i.e., testing the hypothesis that the true value of the respective parameter is 0 against the alternative of $\neq 0$. These are simply obtained by dividing the respective estimate by its standard error, i.e., the square of it yields the Wald test statistic with $df = 1$.

The results of the four χ^2 tests and their observed effects (observed test statistic divided by informative sample size) as well as the power obtained for the respective observed effect (post hoc power) are shown in Table 2.

The parameter-free Rao score test based on an approximate simple random sample of matrices (i.e., each one drawn with approx. the same prob.) of size $2^{13} - 1 = 8191$ using the R package RaschSampler (the max. number allowed by the package) yields for the math data $RS = 87.581$, p-value < 0.001 and for the reading data $RS = 51.795$, p-value $= 0.019$. The procedure of drawing random samples of matrices and computing RS has been replicated a number of times (i.e., about 20 times) to check the accuracy and reliability of the approximation. The results do not differ substantially and are, thus, quite stable. Furthermore, it can be observed that practically relevant percentiles like the 90th, 95th, and 99th of the distribution of RS obtained from the random samples drawn do barely deviate from the respective percentiles of the χ^2 distribution with $df = (k-1)q$. Thus, the parameter-free Rao score test yields reliable results.

One may also make the following additional comparisons of models by using information criteria like AIC (Akaike, 1974) and BIC (Schwarz, 1978) as well as statistical tests. Table 3 shows AIC and BIC values computed for four models and both data examples: model 1 is the Rasch model that considers no covariate, model 2 considers only gender as a covariate, model 3 considers gender and hisei as covariates, and model 4 considers all three covariates available

Table 1. Conditional maximum likelihood estimates with standard errors in parentheses for both data examples.

				math				reading
item	baseline	gender	hisei	migra	baseline	gender	hisei	migra
2	−0.134	0.142	−0.179	−0.089	−1.096	−0.482	−0.157	0.333
	(0.205)	(0.278)	(0.137)	(0.538)	(0.252)	(0.368)	(0.194)	(0.550)
3	−0.881	−0.141	−0.276	−0.764	−5.428	−0.752	−0.221	0.112
	(0.209)	(0.295)	(0.146)	(0.739)	(0.356)	(0.476)	(0.241)	(0.789)
4	1.348	0.368	−0.289	0.267	2.194	0.244	0.047	0.445
	(0.222)	(0.294)	(0.145)	(0.498)	(0.452)	(0.761)	(0.377)	(0.927)
5	0.275	0.445	−0.090	0.151	−0.199	−0.710	−0.103	0.879
	(0.206)	(0.277)	(0.137)	(0.498)	(0.271)	(0.390)	(0.205)	(0.597)
6	1.208	0.286	0.041	0.688	0.107	−0.478	0.388	0.144
	(0.221)	(0.292)	(0.147)	(0.505)	(0.282)	(0.405)	(0.218)	(0.573)
7	0.005	0.502	−0.339	0.733	−1.693	−0.641	−0.324	0.002
	(0.203)	(0.274)	(0.135)	(0.492)	(0.246)	(0.360)	(0.189)	(0.543)
8	−0.006	0.612	−0.108	0.616	−0.016	−0.368	−0.210	−0.087
	(0.205)	(0.275)	(0.136)	(0.492)	(0.275)	(0.403)	(0.212)	(0.577)
9	−0.265	1.375	0.153	0.068	−1.846	−0.847	−0.152	0.329
	(0.207)	(0.280)	(0.140)	(0.502)	(0.246)	(0.359)	(0.189)	(0.542)
10	0.092	0.755	−0.137	0.433	−2.361	−0.940	−0.359	−0.016
	(0.205)	(0.275)	(0.137)	(0.491)	(0.246)	(0.359)	(0.188)	(0.559)
11	−0.625	1.229	0.039	0.702	−3.847	−1.198	−0.529	−0.132
	(0.207)	(0.277)	(0.138)	(0.496)	(0.270)	(0.394)	(0.204)	(0.659)
12	–	–	–	–	−0.412	−1.174	0.015	0.264
	–	–	–	–	(0.265)	(0.376)	(0.198)	(0.556)

Note. Item 1 omitted. Its parameters are set to 0 for identifiability.

Table 2. Results of four χ^2 tests for both data examples.

math data	*LR*	*RS*	*W*	*G*
test statistic	89.971	87.662	85.572	91.857
df	30	30	30	30
p-value	< 0.001	< 0.001	< 0.001	< 0.001
observed effect	0.170	0.165	0.162	0.173
power	> 0.999	> 0.999	> 0.999	> 0.999
reading data	*LR*	*RS*	*W*	*G*
test statistic	53.318	52.255	51.273	54.150
df	33	33	33	33
p-value	0.014	0.018	0.022	0.012
observed effect	0.088	0.086	0.084	0.089
power	0.996	0.996	0.995	0.997

Note. Power values shown refer to the power of the tests for the observed effect given a type I error probability of 0.05.

gender, hisei, and migra. According to *AIC* model 3 is most appropriate (in relation to the other choices) for both math and reading data, i.e., gender and hisei seem to have a considerable effect on the item responses, whereas migra does not. *BIC* prefers model 2 (considering only gender) for both math and reading data. *BIC* is known to be conservative. It prefers simple models with less parameters since it penalizes much more for additional parameters than *AIC*.

Likelihood ratio tests are also readily applicable for comparing models. For the math data one obtains for comparing models 3 and 4, i.e., testing the hypothesis that the slope parameters

Table 3. Information criteria for four models and both data examples.

	math data		reading data	
model	*AIC*	*BIC*	*AIC*	*BIC*
1	4853.5	4896.2	3534.6	3583.2
2	4799.5	4842.3	3514.8	3563.3
3	4795.1	4880.6	3508.2	3605.2
4	4803.5	4931.7	3525.3	3670.9

of covariate migra are all 0, $LR = 11.636$, $df = 10$, p-value $= 0.31$. The comparison of models 2 and 3, i.e., testing the hypothesis that the slope parameters of covariate hisei are all 0, yields $LR = 24.39$, $df = 10$, p-value $= 0.007$, and comparing models 1 (Rasch model) and 3, i.e., testing the hypothesis that the slope parameters of both gender and hisei are all 0, yields $LR = 78.33$, $df = 20$, p-value < 0.001. Hence, given sizes of the tests typically used in practice, in all three tests model 3 has to be accepted or chosen which is in accordance with the results of the AIC. Similar results are obtained for the reading data.

5.2 Example 2

The second example refers to the admission procedure for secondary level general education in Austria (i.e., Verbund Mitte Austria). Pfaffel and Ecker (2023) recently presented the main elements of the procedure as well as initial results on predictive validity in the first year of study. Data stem from examinees or participants that were accepted into the teacher training program at a university in Austria. This example presents only an analysis of one part of the data which refers to measuring social understanding. The data contain 426 binary responses to 10 items and consider four covariates: gender which is binary, age, and the mean grade (1 is the best and 5 the worst) of the participants' first year of study.

Table 4. Conditional maximum likelihood estimates with standard errors in parentheses for Example 2.

item	base	gender	age	mean grade
2	−1.694 (3.731)	0.512 (1.203)	0.060 (0.136)	−0.430 (0.805)
3	0.775 (3.405)	0.496 (1.185)	−0.037 (0.115)	−0.624 (0.789)
4	−0.569 (3.207)	0.163 (1.093)	−0.040 (0.109)	−0.517 (0.750)
5	−2.920 (3.172)	1.261 (1.075)	0.011 (0.109)	−0.651 (0.739)
6	−2.454 (3.274)	1.253 (1.109)	0.035 (0.114)	−0.546 (0.753)
7	−1.664 (3.163)	0.655 (1.072)	−0.048 (0.108)	−0.589 (0.739)
8	0.260 (3.443)	0.185 (1.168)	0.000 (0.119)	−0.733 (0.778)
9	−4.098 (3.686)	0.143 (1.145)	0.127 (0.138)	−0.152 (0.787)
10	−2.177 (3.165)	0.827 (1.074)	−0.046 (0.108)	−0.340 (0.741)

Note. Item 1 omitted. Its parameters are set to 0 for identifiability.

Table 5. Results of four χ^2 tests for Example 2.

	LR	RS	W	G
test statistic	29.966	28.110	26.843	31.734
df	27	27	27	27
p-value	0.316	0.405	0.472	0.242
observed effect	0.108	0.101	0.097	0.114
power	0.910	0.885	0.865	0.929

Note. Power values shown refer to the power of the tests for the observed effect given a type I error probability of 0.05.

Tables 4, 5, and C3 in Appendix C show results. Results of estimates and respective standard errors are shown in Table 4. Standard errors are generally quite large since the informative sample size is only 278, i.e., 148 persons responded correctly to all items and are, thus, completely uninformative. The negative estimates (of the slope parameters) of covariate mean grade imply that the items get harder compared to item 1 with increasing mean grades (i.e., worse grades) of the persons (but standard errors are all larger than the negative deviation of the estimates from 0). Table 5 shows the results of the χ^2 tests, i.e., testing the hypothesis that the slope parameters of all covariates are 0.

5.3 Example 3

An application that goes beyond typical psychometric problems referring to a wider empirical and experimental context may be the following. Since the approach discussed in this work uses a random intercept model or mixed effects model (for binary data) with the person parameters as random effects it is perfectly suited for modeling responses of persons in longitudinal designs and, thus, investigating effects of covariates over time (i.e., a number of discrete time points). In clinical research, for instance, one may observe diseases and symptoms of patients repeatedly at particular points in time and is typically interested in covariates like gender, age, treatment conditions, drug dosages, etc. The theoretical groundwork of such a regression analysis of binary sequences (together with conditional distributions and tests) dates back to the work of Cox (1958).

As an example consider investigating seasonal affective disorder or seasonal depression.

Table 6. Conditional maximum likelihood estimates with standard errors (se) for Example 3.

	spring	summer	se	autumn	se	winter	se
baseline	0	-1.268	0.886	0.383	0.932	2.231	1.134
gender	0	-0.396	0.256	-0.540	0.273	-0.406	0.324
age	0	0.023	0.020	0.028	0.020	0.009	0.025
treatment	0	0.070	0.255	-0.779	0.273	-0.703	0.328

Note. Spring season is used as a baseline. The respective parameters are set to 0.

Participants or patients respond in each of the four seasons of the year. In the simplest case it is only a binary response, i.e., depression is present or not. Thus, one obtains four binary responses from every patient, i.e., one per season. The covariates considered are: gender (binary), age in years, and treatment condition (binary), i.e., half of the sample of patients receives a treatment, the other half does not. Hypothetical data of $n = 600$ patients are generated assuming the following: The baseline parameters (α parameters) are chosen as to characterize a realistic scenario that the prevalence of seasonal depression is generally lower in spring and summer seasons and higher in autumn and winter. The slope parameters (δ parameters) of the covariates gender and age are chosen to be 0 (i.e., no effect of gender and age on seasonal depression), whereas the choice of the slope or effect parameters of the treatment condition reflects a reduction of the probabilities of observing depression in the autumn and winter seasons in the treatment group. The spring season is selected to be time point 1 and is, thus, considered as a baseline, i.e., α and δ parameters referring to time point 1 are set to 0 (for identifiability). Thus, the parameters referring to the other time points or seasons express their effects relative to time point 1 (spring). Tables 6, 7, and 8 show results. It can immediately be seen that covariate treatment has an effect in the autumn and winter seasons. The respective estimates are distinctly smaller than 0 indicating a decline of prevalence of depression in the treatment group (as an effect of the treatment). From the estimates of the baseline parameters (α parameters) it can also be seen that the prevalence rate generally drops in summer (compared to spring) and increases in autumn and winter.

When comparing different models using information criteria one yields the following results. According to both AIC and BIC the data support the model considering the treatment group as the only covariate (relative to the other choices or models), i.e., it is the only covariate that

Table 7. Z test statistics with respective two-sided p-values for each single free parameter for Example 3, i.e., parameter referring to spring omitted.

	summer	p-value	autumn	p-value	winter	p-value
baseline	−1.431	0.153	0.411	0.681	1.967	0.049
gender	−1.545	0.122	−1.983	0.047	−1.253	0.210
age	1.160	0.246	1.345	0.179	0.366	0.715
treatment	0.274	0.784	−2.852	0.004	−2.142	0.032

Table 8. Results of four χ^2 tests for Example 3.

	test statistic	df	p-value	obs. effect	power
LR	19.173	9	0.024	0.041	0.887
RS	18.979	9	0.025	0.041	0.887
W	18.583	9	0.029	0.040	0.878
G	19.375	9	0.022	0.041	0.887

Note. Power values shown refer to the power of the tests for the observed effect given a type I error probability of 0.05.

has an effect (as expected).

Data and the complete list of results are provided in an online repository.

6 Final remarks

A commented R code for all the analyses discussed in this article can be found in an online repository (link to the url: `https://anonymous.4open.science/r/MixedLogit-DB75/`). It contains a function called estimation which provides conditional maximum likelihood estimates (of baseline and slope parameters), standard errors, statistical tests, and information criteria. It depends on the package psychotools (Zeileis et al., 2023). It requires only two arguments, i.e., response matrix and covariate matrix (both must be numerical, no data frames or other R objects). In respect of the covariate matrix each column must contain the respective covariate values of the persons. Thus, it must have as many columns as covariates are considered. In case of one covariate only it must also be a matrix, i.e., a one column matrix. Additionally, a file is provided with an R script for an analysis using the parameter-free Rao score test that depends on the package RaschSampler (Verhelst et al., 2007). At the moment both can only be used with complete data matrices, i.e., no missing values are allowed. Persons with missing values have to be excluded from the analysis. The authors currently work on an extension of the code in two respects: the consideration of missing values and a generalization of the approach to mixed effects logit models that consider responses in potentially more than two categories. This would, for instance, allow testing item parameter invariance in the partial credit model (Masters, 1982) which is a model for ordinal responses. Once this additional work is completed the extended code will be included in the R package tcl (Draxler & Kurz, 2023) for the next update of the package.

The conditional likelihood approach involves computational issues that are particularly noteworthy. In the present context, the γ functions depend on all the covariate values of a person. Thus, for persons with different covariate values one yields different γ functions (provided the δ parameters are not 0). The more covariates are considered in an application, and even more so when real-valued covariates are used, the less likely is it for two or more persons to obtain exactly the same covariate values. The R code, therefore, computes the γ functions for every single person in the sample separately which is, of course, computationally intensive. Nevertheless, computation times are certainly acceptable for typical sample sizes, i.e., up to a few thousand persons. For the examples presented in this article it is only a matter of seconds. Thus, the approach does not seem to be of any substantial practical limitation. When considering only one or very few covariates, in particular, binary ones this code will be (rather) inefficient (but not necessarily unacceptable or unusable). The more persons are contained in the sample whose covariate values are exactly the same the more inefficient is it to compute their γ functions separately, of course.

The main objective of this article relates to a typical psychometric problem. It discusses a mixed effects logit model and an inferential approach of measurement or item parameter invariance for multiple (potentially real-valued) covariates simultaneously. Thus, it avoids carrying out multiple statistical tests and the accumulation of respective error probabilities. An additional aim of this article is to provide an incentive for researchers to apply such a mixed effects model in longitudinal designs and to investigate effects of covariates or predictors or explanatory variables over a period of time (or different experimental conditions) as illustrated in Example 3. Such applications of mixed effects models for binary data, even though dating back to the work of Cox (1958), are not quite well-known and, thus, rather seldom in behavioral research.

At the core of this work is the conditional maximum likelihood approach. It eliminates random effects (i.e., person parameters) and, in case of conditioning on the observed values of both row and column sums of the response matrix, also effects of other nuisance parameters (i.e., α parameters). This approach has a long tradition, in particular, in the Rasch modeling framework. It implies that the item parameters are estimated independently of the person parameters. Furthermore, by eliminating random effects or the person parameters one also yields a solution of a technical problem related to the properties of the estimates discussed by Neyman and Scott (1948). Another solution that is also widely used in psychometric problems is the marginal maximum likelihood approach. Roughly speaking, it eliminates the effects of individual person parameters (i.e., random effects) by averaging over an assumed population. Thus, it comes at the cost of an assumption on the (unknown) distribution of the person parameters but, principally, it is straightforwardly applicable in the present problem too.

The last remark on the concept of conditioning is of a deeper theoretical and philosophical nature and involves the following argument. Different schools of statistical inference (in particular, frequentist and Bayesian) only agree on the problem of conditioning when the statistic is ancillary (a notion that goes back to R.A. Fisher), i.e., when its probability distribution does not depend on the parameters of interest. This is, of course, not the case in the present problem. The distributions of the row and column sums of the response matrix $\boldsymbol{R}$ and $\boldsymbol{S}$ do depend on all parameters of the model. A further extensive discussion on the conditional approach, particularly, in reference to psychometric problems has been given by Skrondal and Raabe-Hesketh (2022).

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

[2] Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, *32*(2), 283–301. https://doi.org/10.1111/j.2517-6161.1970.tb00842.x

[3] Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140. https://doi.org/10.1007/BF02291180

[4] Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1), 76–90.

[5] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* . Lawrence Erlbaum Associates, Hillsdale,NJ, 2 edition.

[6] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, *20*(2), 215–242. http://www.jstor.org/stable/2983890

[7] Draxler, C. (2010). Sample size determination for Rasch model tests. *Psychometrika*, 75(4), 708–724.

[8] Draxler, C., & Alexandrowicz, R. W. (2015). Sample size determination within the scope of conditional maximum likelihood estimation with special focus on testing the Rasch model. *Psychometrika*, *80*(4), 897-919. https://doi.org/10.1007/s11336-015-9472-y

[9] Draxler, C., & Dahm, S. (2020). Conditional or pseudo exact tests with an application in the context of modeling response times. *Psych*, *2*(4), 198–208. https://doi.org/10.3390/psych2040017

[10] Draxler, C., & Kurz, A. (2021). Conditional inference in small sample scenarios using a resampling approach. *Stats*, *4*(4), 837–849. https://doi.org/10.3390/stats4040049

[11] Draxler, C., & Kurz, A. (2023). tcl: Testing in conditional likelihood context (0.2.0) [Computer software]. https://CRAN.R-project.org/package=tcl

[12] Draxler, C., Kurz, A., Gürer, C., and Nolte, J. P. (2023). An Improved Inferential Procedure to Evaluate Item Discriminations in a Conditional Maximum Likelihood Framework. *Journal of Educational and Behavioral Statistics 0*(0). https://doi.org/10.3102/10769986231183335

[13] Draxler, C., Kurz, A., & Lemonte, A. J. (2022). The gradient test and its finite sample size properties in a conditional maximum likelihood and psychometric modeling context. *Communications in Statistics – Simulation and Computation*, *51*(6), 3185–3203. https://doi.org/10.1080/03610918.2019.1710193

[14] Draxler, C., & Nolte, J. P. (2018). Computational precision of the power function for conditional tests of assumptions of the Rasch model. *Open Journal of Statistics*, *8*(6), 873–884. https://doi.org/10.4236/ojs.2018.86058

[15] Draxler, C., & Zessin, J. (2015). The power function of conditional tests of the Rasch model. *AStA Advances in Statistical Analysis*, *99*(3), 367-378. https://doi.org/10.1007/s10182-015-0249-5

[16] Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *46*(1), 59–77. https://doi.org/10.1007/BF02293919

[17] Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, Recent Developments, and Applications*. Springer. https://doi.org/10.1007/978-1-4612-4230-7

[18] Fletcher, R. (1970). A New Approach to Variable Metric Algorithms. *The computer journal*, 13(3), 317–322.

[19] Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, Recent Developments, and Applications* (pp. 69–95). Springer. https://doi.org/10.1007/978-1-4612-4230-7_5

[20] Goldfarb, D. (1970). A Family of Variable-Metric Methods Derived by Variational Means. *Mathematics of Computation*, *24*(109), 23–26. https://doi.org/10.2307/2004873

[21] Gürer, C., & Draxler, C. (2022). Penalization approaches in the conditional maximum likelihood and Rasch modelling context. *British Journal of Mathematical and Statistical Psychology, 00*, 1–38. https://doi.org/10.1111/bmsp.12287

[22] Holland, P. W. and Wainer, H. (1993). *Differential Item Functioning*. Lawrence Erlbaum Associate.

[23] Lemonte, A. J. (2016). The gradient test. Another likelihood-based test. Academic Press. https://doi.org/10.1016/C2015-0-00195-2

[24] Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–20. https://doi.org/10.18637/jss.v020.i09

[25] Mair, P., Rusch, T., Hatzinger, R., Maier, M. J., and Debelak, R. (2023b). *eRm: Extended Rasch Modeling*. R package Version 1.0-4.

[26] Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

[27] Miller, J. W., & Harrison, M. T. (2013). Exact sampling and counting for fixed-margin matrices. *The Annals of Statistics*, *41*(3), 1569–1592. https://doi.org/10.1214/13-AOS1131

[28] Neyman, J. and Pearson, E. S. (1928). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part II. *Biometrika*, 20A(3/4), 263–294.

[29] Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*(1), 1–32. https://doi.org/10.2307/1914288

[30] Pfanzagl, J. (1993). On the consistency of conditional maximum likelihood estimators. *Annals of the Institute of Statistical Mathematics*, *45*(4), 703–719. https://doi.org/10.1007/BF00774782

[31] Pfaffel, A. & Ecker, B. (2023). *Evaluation der Aufnahmeverfahren für Lehramtsstudien der Primarstufe und Sekundarstufe Allgemeinbildung an den Pädagogischen Hochschulen und Universitäten in Österreich*. Be+Be- Verlag, Heiligenkreuz.

[32] Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika*, 66(3), 437–459.

[33] R core team (2022). R: A language and environment for statistical computing (4.2.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

[34] Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *44*(1), 50–57. https://doi.org/10.1017/S0305004100023987

[35] Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

[36] Robitzsch, A. (2022). sirt: Supplementary item response theory models (3.12-66) [Computer software]. https://CRAN.R-project.org/package=sirt

[37] Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111), 647–656.

[38] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

[39] Silvey, S. D. (1959). The Lagrangian multiplier test. *The Annals of Mathematical Statistics*, *30*(2), 389–407.

[40] Skrondal, A., & Rabe-Hesketh, S. (2022). The role of conditional likelihoods in latent variable modeling. *Psychometrika*, *87*(3), 799–834. https://doi.org/10.1007/s11336-021-09816-8

[41] Terrell, G. R. (2002). The gradient statistic. *Computing Science and Statistics*, *34*(34), 206–215.

[42] Verhelst, N. D. (2008). An efficient MCMC algorithm to sample binary matrices with fixed marginals. *Psychometrika*, *73*(4), 705–728. https://doi.org/10.1007/s11336-008-9062-3

[43] Verhelst, N. D., Hatzinger, R. & Mair, P. (2007). The Rasch sampler. Journal of Statistical Software, *20*(4), 1–14. https://doi.org/10.18637/jss.v020.i04

[44] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1), 60–62.

[45] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(3), 426–482. https://doi.org/10.2307/1990256

[46] Zeileis, A., Strobl, C., Wickelmaier, F., Komboz, B., Kopf, J., Schneider, L., & Debelak, R. (2023). *psychotools: Infrastructure for Psychometric Modeling*. R package version 3.6.0.

[47] Zimmer, F., Draxler, C., & Debelak, R. (2023). Power analysis for the Wald, LR, score, and gradient tests in a marginal maximum likelihood framework: applications in IRT. *Psychometrika*, *88*(4), 1249–1298. https://doi.org/10.1007/s11336-022-09883-5

A Appendix A: Technical details on score function and information matrix

The vector-valued score function is given by

$$\boldsymbol{A}(s_2,\dots,s_k,\boldsymbol{\alpha},\boldsymbol{\delta}) = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\alpha},\boldsymbol{\delta})}{\partial \alpha_2} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\alpha},\boldsymbol{\delta})}{\partial \alpha_k} \end{pmatrix} = \begin{pmatrix} s_2 - E(S_2) \\ \vdots \\ s_k - E(S_k) \end{pmatrix} = \begin{pmatrix} s_2 - \sum_i \gamma_r^{-1} \frac{\partial \gamma_r}{\partial \alpha_2} \\ \vdots \\ s_k - \sum_i \gamma_r^{-1} \frac{\partial \gamma_r}{\partial \alpha_k} \end{pmatrix},$$

$$\boldsymbol{D}_p(t_{2p},\dots,t_{kp},\boldsymbol{\alpha},\boldsymbol{\delta}) = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\alpha},\boldsymbol{\delta})}{\partial \delta_{2p}} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\alpha},\boldsymbol{\delta})}{\partial \delta_{kp}} \end{pmatrix} = \begin{pmatrix} t_{2p} - E(T_{2p}) \\ \vdots \\ t_{kp} - E(T_{kp}) \end{pmatrix} = \begin{pmatrix} t_{2p} - \sum_i \gamma_r^{-1} \frac{\partial \gamma_r}{\partial \delta_{2p}} \\ \vdots \\ t_{kp} - \sum_i \gamma_r^{-1} \frac{\partial \gamma_r}{\partial \delta_{kp}} \end{pmatrix} \forall p,$$

$$\boldsymbol{D}(s_2,\dots,s_k,t_{21},\dots,t_{k1},\dots,t_{2q},\dots,t_{kq},\boldsymbol{\alpha},\boldsymbol{\delta}) = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{D}_1 \\ \vdots \\ \boldsymbol{D}_q \end{pmatrix}.$$

Note that all expected values are conditional on $\boldsymbol{R} = \boldsymbol{r}$.

The Fisher information matrix is the covariance of the score, i.e.,

$$\boldsymbol{F}(\boldsymbol{\alpha},\boldsymbol{\delta}) = Cov\big(\boldsymbol{D}\big) = E\big(\boldsymbol{D}\boldsymbol{D}^\top\big).$$

It is a positive-definite square matrix of order $(k-1)(q+1)$ (i.e., the number of free parameters) and can be obtained by

$$\boldsymbol{F}(\boldsymbol{\alpha},\boldsymbol{\delta}) = -E\left(\frac{\partial^2 \ell(\boldsymbol{\alpha},\boldsymbol{\delta})}{\partial(\boldsymbol{\alpha}_*^\top,\boldsymbol{\delta}_*^\top)^\top \partial(\boldsymbol{\alpha}_*^\top,\boldsymbol{\delta}_*^\top)}\right).$$

The R code provided in an online repository uses the following reparameterization to compute the Fisher information matrix. Let $\boldsymbol{\beta}_i : \mathbb{R}^{k(q+1)} \to \mathbb{R}^k$,

$$\boldsymbol{\beta}_i(\boldsymbol{\alpha},\boldsymbol{\delta},x_{i1},\dots,x_{iq}) = \begin{pmatrix} \beta_{i1} \\ \vdots \\ \beta_{ik} \end{pmatrix} = \begin{pmatrix} \alpha_1 + \sum_{p=1}^q \delta_{1p} x_{ip} \\ \vdots \\ \alpha_k + \sum_{p=1}^q \delta_{kp} x_{ip} \end{pmatrix} \forall i$$

denote a k-vector of person specific item parameters. Note that for two arbitrary persons i and l one yields $\boldsymbol{\beta}_i = \boldsymbol{\beta}_l$ only if $x_{ip} = x_{lp}\,\forall p$ and provided the δ parameters are not 0. Otherwise, $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}_l$. Thus, potentially all persons can have different item parameters (depending on their covariate values). Let $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_n^\top)$ and

$$\boldsymbol{J} = \frac{\partial \boldsymbol{\beta}}{\partial(\boldsymbol{\alpha}_*^\top, \boldsymbol{\delta}_*^\top)}$$

be its Jacobian, i.e., a $nk \times (k-1)(q+1)$ matrix of first order partial derivatives with respect to all free parameters $\boldsymbol{\alpha}_*^\top = (\alpha_2, \ldots, \alpha_k)$ and $\boldsymbol{\delta}_*^\top = (\boldsymbol{\delta}_{1*}^\top, \ldots, \boldsymbol{\delta}_{q*}^\top)$, $\boldsymbol{\delta}_{p*}^\top = (\delta_{2p}, \ldots, \delta_{kp})$. Then, the Fisher information matrix is obtained by

$$\boldsymbol{F}(\boldsymbol{\alpha}, \boldsymbol{\delta}) = \boldsymbol{J}^\top \boldsymbol{F}(\boldsymbol{\beta}) \boldsymbol{J},$$

where $\boldsymbol{F}(\boldsymbol{\beta})$ is the information matrix for the reparameterized case of person-specific item parameters. It is a block-diagonal matrix with n blocks each one of order k. Thus, a block represents the contribution (of information) of a single person. The entries of the ith block on the diagonal are given by

$$\gamma_{r_i}^{-1} \frac{\partial \gamma_{r_i}}{\partial \beta_{ij}} \left(1 - \gamma_{r_i}^{-1} \frac{\partial \gamma_{r_i}}{\partial \beta_{ij}}\right) \quad \forall j$$

and off-diagonal by

$$\gamma_{r_i}^{-1} \frac{\partial^2 \gamma_{r_i}}{\partial \beta_{ij} \partial \beta_{il}} - \gamma_{r_i}^{-2} \frac{\partial \gamma_{r_i}}{\partial \beta_{ij}} \frac{\partial \gamma_{r_i}}{\partial \beta_{il}} \quad \forall j \neq l \quad l = 1, \ldots, k,$$

where the γ functions have also to be reparameterized accordingly, i.e.,

$$
\begin{aligned}
\gamma_0(\boldsymbol{\beta}_i) &= 1 \\
\gamma_1(\boldsymbol{\beta}_i) &= \exp(\beta_{i1}) + \cdots + \exp(\beta_{ik}) \\
\gamma_2(\boldsymbol{\beta}_i) &= \exp(\beta_{i1} + \beta_{i2}) + \exp(\beta_{i1} + \beta_{i3}) + \cdots + \exp(\beta_{i,k-1} + \beta_{ik}) \\
\gamma_3(\boldsymbol{\beta}_i) &= \exp(\beta_{i1} + \beta_{i2} + \beta_{i3}) + \exp(\beta_{i1} + \beta_{i2} + \beta_{i4}) + \cdots + \exp(\beta_{i,k-2} + \beta_{i.k-1} + \beta_{ik}) \\
&\vdots \\
\gamma_{k-1}(\boldsymbol{\beta}_i) &= \exp(\beta_{i1} + \cdots + \beta_{i,k-1}) + \cdots + \exp(\beta_{i2} + \cdots + \beta_{ik}) \\
\gamma_k(\boldsymbol{\beta}_i) &= \exp(\beta_{i1} + \cdots + \beta_{ik})
\end{aligned}
$$

$\forall i$ and with $\beta_{ij} = \alpha_j + \sum_p \delta_{jp} x_{ip}$.

B Appendix B: The case of conditioning on both $\boldsymbol{R} = \boldsymbol{r}$ and $\boldsymbol{S} = \boldsymbol{s}$

In case of conditioning on both $\boldsymbol{R} = \boldsymbol{r}$ and $\boldsymbol{S} = \boldsymbol{s}$ the CML estimate of $\boldsymbol{\delta}$ is obtained by

$$\widehat{\boldsymbol{\delta}} := \underset{(\boldsymbol{\delta}) \in \mathbb{R}^{kq}}{\arg\max}\ \ell(\boldsymbol{\delta}).$$

Given mild conditions from general likelihood theory and the conditional case (Andersen, 1970, Pfanzagl, 1993) it holds that

$$\widehat{\boldsymbol{\delta}}_* \xrightarrow{P} \boldsymbol{\delta}_*$$

and

$$\sqrt{n}\big(\widehat{\boldsymbol{\delta}}_* - \boldsymbol{\delta}_*\big) \xrightarrow{D} N\big(\mathbf{0}_{(k-1)q}, \boldsymbol{F}^{-1}(\boldsymbol{\delta})\big),$$

when $n \to \infty$, where $\boldsymbol{F}^{-1}(\boldsymbol{\delta})$ denotes the asymptotic covariance matrix of $\boldsymbol{\delta}_*^\top = (\boldsymbol{\delta}_{1*}^\top, \dots, \boldsymbol{\delta}_{q*}^\top),\ \boldsymbol{\delta}_{p*}^\top = (\delta_{2p}, \dots, \delta_{kp})$.

The vector-valued score function, i.e., a $(k-1)q \times 1$ matrix, is given by

$$\boldsymbol{D}_p(t_{2p}, \dots, t_{kp}, \boldsymbol{\delta}) = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\delta})}{\partial \delta_{2p}} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\delta})}{\partial \delta_{kp}} \end{pmatrix} \ \forall p,$$

$$\boldsymbol{D}(s_2, \dots, s_k, t_{21}, \dots, t_{k1}, \dots, t_{2q}, \dots, t_{kq}, \boldsymbol{\delta}) = \begin{pmatrix} \boldsymbol{D}_1 \\ \vdots \\ \boldsymbol{D}_q \end{pmatrix}.$$

and the Fisher information matrix by

$$\boldsymbol{F}(\boldsymbol{\delta}) = -E\left(\frac{\partial^2 \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}_* \partial \boldsymbol{\delta}_*^\top}\right).$$

The four test statistics based on asymptotic theory are then obtained by

$$LR = -2\big(\ell(\boldsymbol{\delta} = \mathbf{0}) - \ell(\widehat{\boldsymbol{\delta}})\big),$$

$$RS = \boldsymbol{D}^\top(\boldsymbol{\delta} = \mathbf{0})\boldsymbol{F}^{-1}(\boldsymbol{\delta} = \mathbf{0})\boldsymbol{D}(\boldsymbol{\delta} = \mathbf{0}),$$

$$W = \widehat{\boldsymbol{\delta}}_*^\top \boldsymbol{F}(\widehat{\boldsymbol{\delta}})\widehat{\boldsymbol{\delta}}_*,$$

$$GR = \boldsymbol{D}^\top(\boldsymbol{\delta} = \mathbf{0})\widehat{\boldsymbol{\delta}}_*.$$

When $(\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \in \Theta_1$ is true and $n \to \infty$ their common limiting distribution is the central χ^2 with $df = (k-1)q$.

C Appendix C: Additional results from Examples 1 and 2, Sec. 5.1. and 5.2

Table C1. Z test statistics referring to parameters of Table 1.

		math				reading		
item	baseline	gender	hisei	migra	baseline	gender	hisei	migra
2	−0.654	0.511	−1.300	−0.165	−4.349	−1.311	−0.809	0.606
3	−4.215	−0.478	−1.895	−1.034	−15.250	−1.579	−0.920	0.142
4	6.076	1.252	−1.995	0.536	4.849	0.321	0.125	0.481
5	1.332	1.610	−0.655	0.304	−0.736	−1.820	−0.502	1.472
6	5.454	0.979	0.279	1.362	0.378	−1.181	1.778	0.251
7	0.026	1.835	−2.505	1.490	−6.875	−1.783	−1.717	0.004
8	−0.031	2.225	−0.791	1.253	−0.058	−0.913	−0.991	−0.150
9	−1.281	4.909	1.094	0.135	−7.501	−2.360	−0.803	0.607
10	0.448	2.739	−1.007	0.881	−9.602	−2.617	−1.910	−0.029
11	−3.027	4.442	0.285	1.417	−14.275	−3.042	−2.599	−0.200
12	–	–	–	–	−1.556	−3.122	0.077	0.474

Note. Item 1 omitted. Its parameters are set to 0 for identifiability.

Table C2. The corresponding p-values of Table C1.

		math				reading		
item	baseline	gender	hisei	migra	baseline	gender	hisei	migra
2	0.513	0.609	0.194	0.869	< 0.001	0.190	0.418	0.544
3	< 0.001	0.633	0.058	0.301	< 0.001	0.114	0.358	0.887
4	< 0.001	0.211	0.046	0.592	< 0.001	0.748	0.900	0.631
5	0.183	0.107	0.512	0.761	0.461	0.069	0.615	0.141
6	< 0.001	0.327	0.780	0.173	0.705	0.237	0.075	0.802
7	0.980	0.066	0.012	0.136	< 0.001	0.075	0.086	0.997
8	0.975	0.026	0.429	0.210	0.954	0.361	0.322	0.880
9	0.200	< 0.001	0.274	0.893	< 0.001	0.018	0.422	0.544
10	0.654	0.006	0.314	0.378	< 0.001	0.009	0.056	0.977
11	0.002	< 0.001	0.776	0.156	< 0.001	0.002	0.009	0.841
12	–	–	–	–	0.120	0.002	0.939	0.635

Note. Item 1 omitted. Its parameters are set to 0 for identifiability.

Table C3. Z test statistics referring to parameters of Table 4 with corresponding p-values in parentheses.

item	base	gender	age	mean grade
2	−0.454 (0.650)	0.425 (0.671)	0.437 (0.662)	−0.534 (0.593)
3	0.228 (0.820)	0.419 (0.675)	−0.319 (0.750)	−0.791 (0.429)
4	−0.177 (0.859)	0.149 (0.881)	−0.371 (0.711)	−0.689 (0.491)
5	−0.920 (0.357)	1.173 (0.241)	0.100 (0.920)	−0.881 (0.378)
6	−0.750 (0.453)	1.130 (0.258)	0.309 (0.757)	−0.725 (0.468)
7	−0.526 (0.599)	0.611 (0.541)	−0.444 (0.657)	−0.797 (0.426)
8	0.076 (0.940)	0.158 (0.874)	0.000 (1.000)	−0.942 (0.346)
9	−1.112 (0.266)	0.125 (0.901)	0.919 (0.358)	−0.193 (0.847)
10	−0.688 (0.491)	0.770 (0.441)	−0.426 (0.670)	−0.458 (0.647)

Note. Item 1 omitted. Its parameters are set to 0 for identifiability.