Psychological Test and Assessment Modeling, Volume 56, 2014 (4), 405-432

Multiple group cognitive diagnosis models, with an emphasis on differential item functioning

Ann Cathrice George[1] & Alexander Robitzsch[2]

Abstract

In recent years, cognitive diagnosis models (CDMs) have received a growing attention because of their potential to diagnose achievement on the level of sub-competencies. In the context of that development researchers have introduced relevant tools for the practical application of CDMs, as for example multiple group approaches and differential item functioning (DIF) detection. However, when applying CDMs and these related methods to large scale data, one has to overcome a diversity of obstacles: With a growing number of sub-competencies, the models may, due to a large number of parameters, become often (nearly) non-identifiable and thus extremely hard to estimate. Additionally, significance tests may become significant for the only reason of sample size necessitating adequate effect sizes. The present article aims at two aspects: First, it summarizes existing CDM methods for multiple group models and DIF analyses. Second, it gives hints for their application to large-scale assessment data, amongst others we introduce an adapted estimation routine and an appropriate effect size. Both aspects are illustrated by means of the Austrian educational standards test in mathematics 2012 containing a sample size of 71464 students and 72 items.

Keywords: Cognitive Diagnosis Modelling, Multiple Group Models, Differential Item Functioning, Large-Scale Assessment Data

[1] *Correspondence concerning this article should be addressed to:* Ann Cathrice George, PhD, Federal Institute for Educational Research, Innovation and Development of the Austrian School System; Salzburg, Austria; email: a.george@bifie.at

[2] Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens, Salzburg, Austria

1 Introduction

Over the last decade cognitive diagnosis models (CDMs; DiBello, Roussos, & Stout, 2007; Rupp, Templin, & Henson, 2010) have been actively studied and the number of their applications to educational data has increased. One aim of CDMs is to classify individuals based on their item response patterns with respect to a certain number of so-called sub-competencies, which are assumed to form the gross competency domain to be assessed. The discrete individual values on each such sub-competency establish a multidimensional classification, which is said to provide more diagnostic information compared to a single proficiency score and can therefore be used as empirical basis for the development of targeted feedback and support.

Several specific and general CDMs of various formulations have been proposed in the psychometric literature: Examples of specific CDMs include the deterministic input, noisy “and” gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977) model and the reduced reparameterized unified model (R-RUM; Hartz, 2002; Roussos, Templin, & Henson, 2007); the general approaches divide into the generalized-DINA framework (de la Torre, 2011), the log-linear CDM (Henson, Templin, & Willse, 2009), and the general diagnostic model (GDM; von Davier, 2008).

From a technical point of view, CDMs are restricted latent class models, which demand that students possessing the same combination of sub-competencies exhibit the same item response probabilities (Formann, 2007; Formann & Kohlmann, 1998). Results, which are comparable to those obtained through CDMs, may be achieved through item response theory (IRT; van der Linden & Hambleton, 1997) by joining students with similar trait locations on the latent continuum into a small number of classes (for details cf. Haberman, von Davier, & Lee, 2008). Following these arguments, researchers have argued that CDMs are not an entirely novel class of models, but should rather be seen as specific latent structure models (von Davier, 2009; von Davier & Haberman, 2014). Nonetheless we use the term “CDM”, because it is widespread in the research literature and facilitates readability.

With respect to the practical application of CDMs researchers started to investigate relevant methods for the analysis of educational data within the framework of CDMs: For example, multiple group approaches for CDMs have been suggested (e.g. Johnson et al., 2013; Xu & von Davier, 2008b), which allow the comparison of achievement between different groups of students. In this context, one may also be interested in analyzing differential item functioning (DIF; Penfield & Camilli, 2006) on the level of sub-competencies (Hou, de la Torre, & Nandakumar, 2014). Additional to these methods, a wide range of statistical procedures for checking model validity has been presented, as for example measures of global and local model fit (Chen, de la Torre, & Zhang, 2013), item fit (Kunina-Habenicht, Rupp, & Wilhelm, 2009), or classification accuracy (Cui, Gierl, & Chuang, 2012; DiBello et al., 2007).

Despite these efforts to bring CDMs closer to practical needs, some methodological obstacles in applying these methods still remain, becoming even more severe when dealing with large scale data. Probably the gravest problem is the estimation of CDMs assuming a large number of sub-competencies leading to a high-dimensional model. The number of model parameters grows exponentially with the number of sub-competencies and thus the model soon becomes (almost) non-identifiable. Combining a large number of dimensions (sub-competencies) with a large sample size, as in the case of large scale assessments, memory overflow is likely to appear on current computer systems. Another commonly known problem with large sample sizes is that significance tests are overpowered. In the context of CDMs one is confronted with this overpowerment when applying multiple group variants (see Section 2.2) of the models: Like multiple group models, multiple group CDMs often assume invariant item parameters across groups (Xu & von Davier, 2008a). It is argued that this assumption, which is equivalent to the absence of DIF items, should be ascertained prior to applying a multiple group CDM (Hou et al., 2014). However, due to the aforementioned excessively high statistical power in large samples, significance tests are inadequate for identifying items exhibiting practically relevant DIF.

For the Austrian educational standards test in mathematics 2012 (Breit & Schreiner, 2012) it was decided to reanalyze the data using a multiple group CDM. In the first part of this article, we briefly present the idea of the statistical theory behind CDMs and review the existing methods for performing multiple group analysis and DIF detection within this framework. In each of these points special consideration is given to the application of CDMs to large scale data with a concrete focus on

1) an adaption of the estimation algorithm to prevent memory overflow (Section 2.4) and
2) the development of an effect size measure for DIF (Section 2.5)

In the second part, we illustrate the presented methods using the Austrian educational standards test in mathematics 2012 (Breit & Schreiner, 2012), a large scale assessment involving 71464 students and 72 items measuring 8 sub-competencies. More precisely, we discuss three models, one analyzing the mathematical sub-competencies of the whole student sample, the second comparing the possession of these sub-competencies between boys and girls, and a third model analyzing differences between upper and lower track students. Beyond the pure application of the CDM methods, we also promote some ideas for further handling and analysis of the results, e.g.

3) We conduct analyses of variance to summarize the skill class probabilities and their interactions (Sections 5.2, 5.3 and 5.4).
4) For reporting the differences in skill mastery between the groups we use differences in skill probabilities, which are afterwards transformed to the widely used Cohen's *d* effect size (Sections 4.3 and 5.5).

Our aim is to solve practical obstacles when applying CDMs to large scale data and to present an application, revealing some substantial findings going beyond the results obtained by unidimensional IRT models with continuous skills.

2 Theory

We consider the gross competency domain maths, which is split up into a few sub-competencies. In order to model this *substantial* structure by means of a CDM, we assign each sub-competency a so-called latent categorical *skill* variable, termed α_k (i.e. α_1 corresponds to the first sub-competency and so on). It is assumed that each student possesses a subset out of a total of K skills $\alpha_1, \dots, \alpha_K$.

In this article, we refer to CDMs for dichotomous responses and dichotomous skills, however, more general model variants allow for polytomous items and polytomous skills as well (e.g. Chen & de la Torre, 2013; von Davier, 2008). For specifying which skill is required to solve which item, domain experts have to define a binary $J \times K$ item-by-skill matrix $\mathbf{Q}$, in which the element q_{jk} in the j-th row and the k-th column indicates whether skill k is needed ($q_{jk} = 1$) or not ($q_{jk} = 0$) for correctly responding to item j, $j = 1, \dots, J$. Thus, the so called Q-matrix $\mathbf{Q}$ reflects the substantial theory of how skills contribute to solving each item. Based on $\mathbf{Q}$, a CDM infers the possession of the K skills from the $I \times J$ response matrix of $i = 1, \dots, I$ students. The K skills allow for a total of 2^K different skill patterns, which are termed *skill classes* $\boldsymbol{\alpha}_l$, $l = 1, \dots, 2^K$, in the CDM context.

The results obtained through a CDM analysis are twofold:

1) We obtain the probability that a randomly chosen individual belongs to skill class $\boldsymbol{\alpha}_l$, i.e. 2^K skill class probabilities $P(\boldsymbol{\alpha}_l)$, representing the proportion of students *in the population* possessing a specific combination $\boldsymbol{\alpha}_l = [\alpha_{l1}, \dots, \alpha_{lK}]$ of skills. Because of the assumption $\sum_{l=1}^{2^K} P(\boldsymbol{\alpha}_l) = 1$, the vector of the 2^K skill probabilities $P(\boldsymbol{\alpha}_l)$ is the skill class distribution.
2) *For individual assessment*, each individual i is classified into exactly one of the 2^K skill classes and the dichotomous outcome vector $\boldsymbol{\alpha}_i = [\alpha_{i1}, \dots, \alpha_{iK}]$ is called the i-th student's *skill profile*. As a simple example, consider a CDM with $K = 4$ skills. According to this model, each student is assigned one of the $2^4 = 16$ skill classes $\boldsymbol{\alpha}_l$. For example, the skill class $\boldsymbol{\alpha}_l = [1,1,0,0]$ includes students possessing the skills $\alpha_1$and α_2 but not α_3 and α_4.

2.1 The DINA Model Framework

Because of its simplicity and parsimony in terms of model parameters, the Deterministic Input Noisy-And-Gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977) model is one of the most commonly used CDMs. For the same reason it is used in this section to present the basic statistical concepts underlying CDMs.

The DINA model asserts that students have to possess all skills assigned in $\mathbf{Q}$ to an item for successfully mastering it. To put it differently, the DINA model is non-compensatory, in that a lack in one required skill cannot be compensated for by another skill assumed to be present in this class. The i-th student's probability to master the j-th item involves two components, namely a deterministic one and a probabilistic one. The former states whether the student is expected to master the *j*-th item on the basis of his possessed skills. A student possessing all required skills for item *j* (or even more skills) is expected to master the item, whereas a student lacking at least one required skill is not expected to master the item. This deterministic component is expressed through the dichotomous latent response $\eta_{ij} = 0{,}1$ of student *i* to item *j*, with

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}{}^{q_{jk}}.$$

In case of $\eta_{ij} = 1$, student i is expected to master item j, in case of $\eta_{ij} = 0$ he is not.

The probabilistic component represents the probability of student i to *actually* respond correctly to item j: Students may *slip*, i.e. fail to produce the correct answer, although they are expected to master an item (i.e. $\eta_{ij} = 1$), e.g. due to lack of concentration, distraction, or alike. Analoguously, students who are not expected to master an item (i.e. $\eta_{ij} = 0$) may succeed by luckily *guessing* the correct response. The probabilities δ_{j0} for guessing item j and $(\delta_{j0} + \delta_{j1})$ for *not* slipping item j are modeled as item specific parameters. Including both components, the DINA model is expressed through

$$P\left(X_{ij} = 1 \mid \boldsymbol{\alpha}_i\right) = P\left(X_{ij} = 1 \,\middle|\, \eta_{ij}, \delta_{j0}, \delta_{j1}\right) = (\delta_{j0} + \delta_{j1})^{\eta_{ij}}\, \delta_{j0}{}^{1-\eta_{ij}},$$

denoting the probability of student i to correctly respond item j conditional on a skill profile $\boldsymbol{\alpha}_i$. Note that in DINA models with a simple loading structure (i.e. models in which each item measures exactly one skill) the latent response η_{ij} for an item *j* measuring skill $k' = 1, \dots, K$ reduces to $\eta_{ij} = \alpha_{ik'}{}^{q_{jk'}}$. This corresponds to a stepwise item response function, which can be seen as a multidimensional analogue of the probabilistic Guttman model (Proctor, 1970).

2.2 Multiple group CDMs

Multiple group CDMs (MG-CDMs) are an extension of CDMs for situations, in which more than one (manifest) group of students responds to the same test (e.g. gender groups). The objective of a MG-CDM is to compare the extent to which these groups differ in their skill possession. One simple approach for this purpose could be to estimate one model for all students and to compare the individual classifications of the students given their group membership. However, proceeding that way leads to biased estimation of the group differences (Bock & Zimowski, 1997). Thus, based on pertinent methods for latent trait models, multiple group approaches for CDMs have been suggested (Johnson et al., 2013; Xu & von Davier, 2008a) which incorporate some identification condition for assessing group differences (von Davier & von Davier, 2007). Following Xu and von Davier (2008a) and de la Torre & Lee (2010), the assumption of invariant item parameters across groups is the strongest identification condition. The validity of this assumption, i.e. the assumption that each item works in the same way in each group, should be tested prior to the estimation of a MG-CDM (see Section 2.5).

2.3 Parameter estimation

Parameter estimation of (MG-)CDMs is performed by means of marginal maximum likelihood (MML) estimation. A pertinent way to implement this method is the expectation-maximization (EM; Dempster, Laird, & Rubin, 1977) algorithm (de la Torre, 2009; for one group). The EM algorithm iterates between an E-step and an M-step: In the E-step, expected counts for each item and each group are calculated, which are a prerequisite for the calculation of the required statistics in the M-step. Then, the M-step updates the parameter estimates for the MG-CDM using maximization methods. Finally, the E-step and M-Step alternate until a previously set convergence criterion is attained.

We assume for the following presentation that groups $G_i = g$, $g = 1, \dots, G$, are exhaustive and mutually disjunctive (i.e. each student i belongs to exactly one group g). Furthermore, the assumption of invariant item parameters across groups is posed. For reasons of simplicity, the parameter estimation process is again presented for the example of the DINA model, but it may be extended to more complex CDMs in a straightforward manner (Xu & von Davier, 2008a).

For estimating the DINA model the marginal log-likelihood

$$\log L(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \sum_{i=1}^{I} \log L(\boldsymbol{X}_i, G_i; \boldsymbol{\delta}, \boldsymbol{\gamma}) = \sum_{i=1}^{I} \log[\sum_{l=1}^{L} P(\boldsymbol{X}_i | \boldsymbol{\alpha}_l; \boldsymbol{\delta}) \cdot P(\boldsymbol{\alpha}_l | G_i; \boldsymbol{\gamma})] \quad (1)$$

is maximized. Here, the parameter vector $\boldsymbol{\delta} = [\boldsymbol{\delta}_0, \boldsymbol{\delta}_1] = [\delta_{10}, \dots, \delta_{J0}, \delta_{11}, \dots, \delta_{J1}]$ includes all item parameters and

$$P(\boldsymbol{X}_i | \boldsymbol{\alpha}_l; \boldsymbol{\delta}) = \prod_{j=1}^{J} P\left(X_{ij} = 1 \middle| \boldsymbol{\alpha}_l, \delta_{j0}, \delta_{j1}\right)^{X_{ij}} [1 - P\left(X_{ij} = 1 \middle| \boldsymbol{\alpha}_l, \delta_{j0}, \delta_{j1}\right)]^{1 - X_{ij}}$$

is the probability of a response vector $\boldsymbol{X}_i$ if student i possesses the skills of skill class l, $l = 1, \dots, L$. Note that $P(\boldsymbol{X}_i | \boldsymbol{\alpha}_l, \boldsymbol{\delta})$ is independent of the group-membership since we assume invariant item parameters across groups. Furthermore, the unknown parameter vector $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_G]$ describes all group specific distributions $P(\boldsymbol{\alpha} | g; \boldsymbol{\gamma}_g) = [P(\boldsymbol{\alpha}_1 | g; \boldsymbol{\gamma}_g), \dots, P(\boldsymbol{\alpha}_L | g; \boldsymbol{\gamma}_g)]$. While in case of a full skill space each $\boldsymbol{\gamma}_g = [\gamma_{1g}, \dots, \gamma_{Lg}]$ contains all $L = 2^K$ probabilities $\gamma_{lg} = P(\boldsymbol{\alpha}_l | g)$, in case of a reduced skill space the $\boldsymbol{\gamma}_g = [\gamma_{1g}, \dots, \gamma_{Lg}]$ are resulting vectors of a log-linear smoothed skill space with $L = 1 + K + K \cdot \frac{(K+1)}{2} < 2^K$ parameters. For further details see Section 2.4.

Before the first iteration of the EM algorithm, initial item parameters $\boldsymbol{\delta}$ and skill distribution parameters $\boldsymbol{\gamma}$ have to be chosen. Then, the EM algorithm alternates between the E-step and the M-step described in the following:

E-Step:

a. The individual posterior distribution can be deduced via Bayes' theorem:

$$P\left(\boldsymbol{\alpha}_l \middle| \boldsymbol{X}_i, G_i; \boldsymbol{\delta}, \boldsymbol{\gamma}_g\right) = \frac{P(\boldsymbol{X}_i | \boldsymbol{\alpha}_l; \boldsymbol{\delta}) \, P\left(\boldsymbol{\alpha}_l \middle| G_i; \boldsymbol{\gamma}_g\right)}{\sum_{l=1}^{L} P(\boldsymbol{X}_i | \boldsymbol{\alpha}_l; \boldsymbol{\delta}) \, P\left(\boldsymbol{\alpha}_l \middle| G_i; \boldsymbol{\gamma}_g\right)}, \qquad l = 1, \dots, L$$

b. Two types of expected counts are derived from the posterior: The first count is the expected number

$$I_{lj} = \sum_{i=I}^{I} P(\boldsymbol{\alpha}_l | \boldsymbol{X}_i, G_i; \boldsymbol{\delta}, \boldsymbol{\gamma}_g)$$

of students which are classified into skill class $\boldsymbol{\alpha}_l$ for item $j, j = 1, \dots, J$. Note that in case of no missing data $I_{lj} = I_{lj'}$ for all $j, j' = 1, \dots, J$.

The second count

$$R_{lj} = \sum_{i=I}^{I} X_{ij} \cdot P(\boldsymbol{\alpha}_l | \boldsymbol{X}_i, G_i; \boldsymbol{\delta}, \boldsymbol{\gamma}_g)$$

describes the expected number of students classified in skill class $\boldsymbol{\alpha}_l$ while responding item *j* correctly.

M-Step:

a. The set of item parameters $[\boldsymbol{\delta}_0, \boldsymbol{\delta}_1]$ is updated. The estimating equations are obtained by setting the first derivative of the log-likelihood with respect to the item parameters equal to zero. The derivative only involves the two counts obtained in the E-step. Let

$$I_j^{(0)} = \sum_{l:\, \eta_{lj}=0} I_{jl}$$

be the expected number of students lacking at least one of the skills required for the mastery of item *j* (i.e. $\eta_{lj} = 0$) and

$$R_j^{(0)} = \sum_{l:\, \eta_{lj}=0} X_{ij} \cdot I_{jl}$$

be the expected number of students among $I_j^{(0)}$ who correctly respond to item j. Furthermore let $I_j^{(1)}$ and $R_j^{(1)}$ have the same interpretation except that they belong to students which possess all skills required for item *j* (i.e. $\eta_{lj} = 1$). Based on this definitions the items parameters of item *j* are updated according to

$$\delta_{j0} = \frac{R_j^{(0)}}{I_j^{(0)}} \quad , \quad \delta_{j0} + \delta_{j1} = \frac{R_j^{(1)}}{I_j^{(1)}} \quad .$$

For details see de la Torre (2009).

b. The group-wise skill class distributions $P(\boldsymbol{\alpha}_l | g; \boldsymbol{\gamma})$ are updated. For each group g, the expected number n_{lg} of students in group g and skill class $\boldsymbol{\alpha}_l$ is calculated, namely

$$n_{lg} = \sum_{i \,|\, G_i = g} P(\boldsymbol{\alpha}_l | \boldsymbol{X}_i, G_i; \boldsymbol{\delta}, \boldsymbol{\gamma}_g).$$

Let N_g be the number of students in group g, then the skill class distributions are updated by

$$P(\boldsymbol{\alpha}_l|g;\boldsymbol{\gamma}_g) = \frac{n_{lg}}{N_g}.$$

In an optional step, these skill class distributions may be smoothed by using a log-linear model. For details see the following Section 2.4.

Finally, the E- and M-Step alternate until convergence. Convergence may be achieved if the maximal change between the parameter estimates or the relative change in the deviance is below a specific predefined value or after a maximum number of iterations. Note that the estimation algorithm may also handle sampling weights, which is not presented here.

2.4 Skill space reduction

In cases where models have almost as many parameters as observations, which, consequently, would lead to weakly or non-identifiable skill classes, Xu & von Davier (2008b) proposed to change from the unreduced skill space $P(\boldsymbol{\alpha}_l)$, $l = 1, \ldots, 2^K$, to a log-linear smoothed form of the skill space

$$\log P(\boldsymbol{\alpha}_l|g) = \gamma_{g0} + \sum_{k=1}^{K} \gamma_{kg1}\,\alpha_{lk} + \sum_{k=1}^{K-1}\sum_{m=k+1}^{K} \gamma_{kmg2}\alpha_{lk}\alpha_{lm}\,.$$

Here, the γ_{g0} is an intercept parameter, the parameter γ_{kg1} summarizes the main effects of skill k in group g (allowing for different marginal skill probabilities of the K skills) and γ_{kmg2} captures the interaction of skills k and m in group g. The unknown $\boldsymbol{\gamma}_g$ parameters can be estimated by a generalized least squares estimation within the M-step (Xu & von Davier, 2008b). Within this log-linear parameterization $1 + K + K \cdot \frac{(K+1)}{2}$ parameters are estimated for each group (instead of $2^K - 1$ parameters in the full skill space).

Additionally, forming the log-likelihood over $L = 2^K$ skill classes is computationally demanding for a large number of skills. In applications of large scale assessment data, the representation of the appropriate high-dimensional posterior distributions may lead to problems of memory overflow (e.g. performing an analysis with $K = 16$ skills and $I = 72000$ students leads to two matrices of size $2^{16} \times 71464 = 65536 \times 71464$ for storing the values of the individual likelihood and the posterior). Hence, we propose an alternative approach which uses a much smaller number of skill classes than 2^K: We assume that the multidimensional distribution of dichotomous skills $\boldsymbol{\alpha}$ is obtained by discretizing an underlying multivariate normal distribution $\boldsymbol{\alpha}^*$ at appropriate thresholds (Templin & Henson, 2006). In adopting ideas from item response models with continuous traits, we approximate the continuous distribution $\boldsymbol{\alpha}^*$ by a discrete grid $\boldsymbol{\alpha}_1^*$ with L grid points via quasi Monte Carlo integration (Pan & Thompson, 2007). The discrete grid

$\boldsymbol{\alpha}_1^*$ is finally split into a grid of dichotomous skill classes $\boldsymbol{\alpha}_1$. Thus, choosing a sufficiently high number of grid points L (say $L = 2000$ or $L = 4000$) can adequately represent the log-likelihood (which only depends on the lower dimensional $\boldsymbol{\gamma}$ parameters).

2.5 Differential Item Functioning

As mentioned before, the application of a MG-CDM assumes invariant item parameters across the different groups (Xu & von Davier, 2008a). To assure this assumption one may test each item for parameters differences between the groups, i.e. differential item functioning (DIF; Penfield & Camilli, 2006), and in case of significance one may decide to leave out the item in the MG-model. Another aspect in the investigation of DIF items is to analyze the reasons for DIF on the level of skills.

For conducting a DIF-test in the CDM framework (Hou et al., 2014) the item parameters are estimated sequentially: In the estimation process of an item *j*, the item parameters of this item *j* are freed to vary between groups whereas the item parameters of all remaining $J - 1$ items are constrained to be invariant between groups (cf. the procedure of the likelihood ratio test for detecting DIF in IRT; Penfield & Camilli, 2006). Proceeding that way, we obtain for each group g and each item j in the DINA model a vector of item parameters $\boldsymbol{\delta}_{j|g} = [\delta_{j0|g}, \delta_{j1|g}]$, where for two groups g_1 and g_2 the equation $\boldsymbol{\delta}_{j|g_1} = \boldsymbol{\delta}_{j|g_2}$ can be violated. Then each item *j* may be tested for exhibiting DIF (de la Torre & Lee, 2013) using the null hypothesis

$$H_0\colon \boldsymbol{\delta}_{j|1} = \boldsymbol{\delta}_{j|2} = \ldots = \boldsymbol{\delta}_{j|G} \ .$$

This null hypothesis can equivalently be written as $H_0\colon C_j \cdot \boldsymbol{\delta}_j = 0$, with

$$C_j = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

By adapting of the Wald statistic of the G-DINA model (de la Torre, 2011; Hou et al., 2014), the Wald Statistic *W* for the DINA model is formed as

$$W = [\, C_j \cdot \boldsymbol{\delta}_j \,]' \{\, C_j \cdot \mathrm{Var}(\boldsymbol{\delta}_j) \cdot C_j{}' \,\}^{-1} [\, C_j \cdot \boldsymbol{\delta}_j \,]$$

$$\text{where } \boldsymbol{\delta}_j = [\boldsymbol{\delta}_{j|1}, \boldsymbol{\delta}_{j|2}, \ldots, \boldsymbol{\delta}_{j|G}]' \text{ and } \mathrm{Var}(\boldsymbol{\delta}_j) = \begin{bmatrix} \mathrm{Var}(\boldsymbol{\delta}_{j|1}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathrm{Var}(\boldsymbol{\delta}_{j|G}) \end{bmatrix}.$$

If H_0 holds, the test statistic W is assumed to be asymptotically χ^2 distributed with $2 \cdot (G - 1)$ degrees of freedom (de la Torre & Lee, 2013). For implementing the Wald test, $\boldsymbol{\delta}_j$ and $\mathrm{Var}(\boldsymbol{\delta}_j)$ are replaced by their sample counterparts $\widehat{\boldsymbol{\delta}}_j$ and $\widehat{\mathrm{Var}}(\boldsymbol{\delta}_j)$.

For illustrational purposes consider $G = 2$ groups with, for example, group $g_1 = 1$ representing boys and group $g_2 = 2$ the girls. Then the null hypothesis "item j exhibits no DIF between boys and girls" is

$$H_0\colon \boldsymbol{\delta}_{j|1} = \boldsymbol{\delta}_{j|2}\,,$$

which means that $\delta_{j0|1} = \delta_{j0|2}$ and $\delta_{j1|1} = \delta_{j1|2}$.

For computing the Wald statistic we define

$$\boldsymbol{\delta}_j = \left[\boldsymbol{\delta}_{j|1}, \boldsymbol{\delta}_{j|2}\right]' = \left[\delta_{j0|1}, \delta_{j1|1}, \delta_{j0|2}, \delta_{j1|2}\right]',$$

$$\mathrm{Var}\left(\boldsymbol{\delta}_j\right) = \begin{bmatrix}\mathrm{Var}(\boldsymbol{\delta}_{j|1}) & 0 \\ 0 & \mathrm{Var}(\boldsymbol{\delta}_{j|2})\end{bmatrix} \quad \text{and}$$

$$C_j = \begin{bmatrix}1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1\end{bmatrix}.$$

Note that we have presented a method for analyzing DIF separately for each item. Another approach, which is not considered here, is to simultaneously analyze DIF for all items of the test, i.e. to assume non-invariant item parameters across groups (Johnson et al., 2013).

In large sample sizes, the Wald statistic may become significant for items exhibiting small DIF effects considered irrelevant from a substantial point of view. Hence, we propose to use an effect size measure in addition, which is based on the difference of the item response functions between groups:

$$\mathrm{UA}_j = \sum_{l=1}^{L} w(\boldsymbol{\alpha}_l) \cdot \left|P\left(X_j = 1\middle|\boldsymbol{\alpha}_l, g_1\right) - P\left(X_j = 1\middle|\boldsymbol{\alpha}_l, g_2\right)\right|,$$

where

$$w(\boldsymbol{\alpha}_l) = \tfrac{1}{2}\left[P(\boldsymbol{\alpha}_l|G = g_1) + P(\boldsymbol{\alpha}_l|G = g_2)\right].$$

This DIF effect size measure UA_j is an adoption of the unsigned area (UA) originally introduced by Raju (1990) and novel to the framework of cognitive diagnosis modeling. In context of the three parameter IRT model, Jodoin and Gierl (2001) suggest as a rule of thumb values of .059 to distinguish negligible from moderate DIF and .088 to distinguish moderate from large DIF. We suggest adapting this rule for the UA measure in the framework of CDMs, too.

From a theoretical point of view, Roussos and Stout (1996) distinguish two dimensions which may cause DIF: Firstly, an auxiliary dimension, which is intended to be measured in the test (i.e. which is construct relevant) and secondly, a nuisance dimension, which is not intended to be measured (i.e. construct irrelevant). In our analyses, educational experts decide for each empirically detected DIF item, if the DIF is caused by construct relevant or construct irrelevant factors. Only in case the expert identifies construct irrelevant DIF, the respective items are removed from further analyses.

3 Data

The data reanalyzed in this article consists of 71464 Austrian grade 8 students' responses to a mathematics test, which was employed in the framework of educational standards testing in 2012 (Bildungsstandards-Mathematik 8; BIST-M8; Breit & Schreiner, 2012). The test population splits into 51 % boys and 49 % girls. One third of the students are attending the academic school (AHS), and the remaining 67 % are attending the general secondary school (HS, NMS).

The test comprises 72 items arranged in 6 test booklets according to a partially balanced incomplete block design (for BIST test designs: Kuhn & Kiefer, 2013). Each individual student responded to 48 items in one of the test booklets. The test booklets are mutually comparable concerning length, difficulty and content of the items.

Following the competence model of Peschek and Heugl (2007), mathematical competence in the eighth grade can be divided into four operational sub-competencies "Representation" (α_1), "Calculation" (α_2), "Interpretation" (α_3), and "Argumentation" (α_4) and four content sub-competencies namely "Numbers and Measures" (α_5), "Variables and functional Dependencies" (α_6), "Geometry" (α_7), and "Statistics" (α_8). In the present study, the four operational and four content sub-competencies are used as the $K = 8$ basic skills underlying the tested mathematical competence in the eighth grade. According to educational experts, the mastery of each item in the standards test requires exactly one operational and one content skill. As a summary, Table 1 shows the number of items in each of the 6 test booklets requesting the 16 possible combinations of one content and one operational skill: For example the operational skill α_1 is required in combination with the content skill α_5 for the mastery of 3 items in the first test booklet.

Table 1:
Number of items requiring a specific combination of operational and content skills in each of the 6 test booklets and for the whole item pool

test-	α_1 and				α_2 and				α_3 and				α_4 and				
booklet	α_5	α_6	α_7	α_8	α_5	α_6	α_7	α_8	α_5	α_6	α_7	α_8	α_5	α_6	α_7	α_8	$\sum$
1	3	3	3	3	3	4	2	3	2	1	6	3	4	4	1	3	48
2	3	4	2	3	2	3	3	4	4	1	4	3	3	4	3	2	48
3	3	3	3	3	4	5	1	2	3	1	5	3	2	3	3	4	48
4	3	3	3	3	3	4	2	3	4	2	3	3	2	3	4	3	48
5	4	4	2	2	2	4	2	4	4	1	5	2	2	3	3	4	48
6	4	3	3	2	2	4	2	4	3	2	5	2	3	3	2	4	48
item pool	5	5	4	4	4	6	3	5	5	2	7	4	4	5	4	5	72

4 Methods

Remember that our goal is to estimate and discuss three models for the BIST-M8 data: One model analyzing the eight mathematical skills of the whole student sample and two models emphasizing differences in the skill possession between subgroups of students, i.e. boys compared to girls and students of the academic school type compared to students attending the general school type.

4.1 Q-matrix

The first three rows of the original Q-matrix which is underlying the assignment of the 8 skills $\alpha_1, \ldots, \alpha_8$ to the items is given in Table 2.

However, this Q-matrix would lead to non-identifiable skill class distribution, i.e. the marginal skill probabilities of the operational skills could be estimated independently of the content skills (see also Carstensen & Rost, 2007). As a simplified heuristic explanation for the non-identifiability of the skill classes in the original Q-matrix one may again consider the Likelihood in (1). For every skill class $\boldsymbol{\alpha}_l$ the second term may also be written as

$$P(\boldsymbol{\alpha}_l|G_i,\gamma) = P(\alpha_{oc}|G_i,\gamma) = p(\alpha_o|G_i) \cdot p(\alpha_c|G_i) \cdot \rho_{oc} = p_o \cdot p_c \cdot \rho_{oc},$$

where α_o represents the operational skill included in skill class $\boldsymbol{\alpha}_l$ and α_c represents the content skill and ρ_{oc} the correlation between both. Because we only estimated the joint probability $P(\alpha_{oc}|G_i,\gamma)$, with an appropriate constant b it may also hold

$$P(\boldsymbol{\alpha}_l|G_i,\gamma) = \frac{1}{b} p(\alpha_o|G_i) \cdot b\, p(\alpha_c|G_i) \cdot \rho_{oc} = \tilde{p}_o \cdot \tilde{p}_c \cdot \rho_{oc} = P(\boldsymbol{\alpha}_l|G_i,\tilde{\gamma})$$

Since the first term $P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l;\boldsymbol{\delta})$ of the likelihood in (1) depends only on the item parameters $\boldsymbol{\delta}$, the value of the Likelihood does not change even though the skill probabilities $p(\alpha_o|G_i)$ and $p(\alpha_c|G_i)$ were redefined. Therefore, not all parameters of the skill class distribution can be uniquely identified.

We therefore apply an alternative matrix $\mathbf{Q}$ (Table 3). In this 16-columns matrix each combination between an original operational skill α_o (i.e., $\alpha_1, \ldots, \alpha_4$) and an original content skill α_c (i.e., $\alpha_5, \ldots, \alpha_8$) is established as a combined skill $\alpha_o\alpha_c$. For example, items which load on the original combination α_1 and α_5 have a one in the first column of the redefined 16-columns Q-matrix; for further examples compare Tables 2 and 3.

Table 2:
Original Q-matrix

	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
Item 1	1	0	0	0	1	0	0	0
Item 2	1	0	0	0	0	0	1	0
Item 3	0	1	0	0	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 3:
Redefined Q-matrix $\mathbf{Q}$

	$\alpha_1\alpha_5$	$\alpha_1\alpha_6$	$\alpha_1\alpha_7$	$\alpha_1\alpha_8$	$\alpha_2\alpha_5$	$\alpha_2\alpha_6$	$\alpha_2\alpha_7$	$\alpha_2\alpha_8$	$\alpha_3\alpha_5$	$\alpha_3\alpha_6$	$\alpha_3\alpha_7$	$\alpha_3\alpha_8$	$\alpha_4\alpha_5$	$\alpha_4\alpha_6$	$\alpha_4\alpha_7$	$\alpha_4\alpha_8$
Item 1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Item 2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Item 3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Even though an application of the new Q-matrix solves the methodological problem of non-identifiability, this application also changes the skills employed in the models. While our goal is to emphasize relationships between the students' manifest response behavior and the eight original skills, the models applied in the following yield results in the level of the newly defined skills $\alpha_o\alpha_c$. Section 4.3 describes how we establish the link between the 8 original and the 16 new skills.

The presented strategy of estimating two skill facets (i.e. operational and content) through building all combinations between the two facets is also applied by Carstensen and Rost (2007) and recently by Harks, Klieme, Hartig, and Leiss (2014) for multidimensional item response models with continuous variables. Alternatively one may apply a hierarchical factor model involving the two facets (Rijmen, 2011), which is a special case of a multitrait-multimethod model (Eid, Lischetzke, & Nussbeck, 2006) .

4.2 Estimation

In the following, (MG-)DINA models are fitted to the data. This seems to be adequate as (a) the redefined Q-matrix $\mathbf{Q}$ holds a simple loading structure (between item dimensionality), and (b) the DINA model provides a simple partitioning of students into masters and non-masters for each skill. In both multiple group variants, i.e. the gender and the school track comparison, the full model would require the estimation of $2 \cdot (2^{16} - 1) = 131070$ skill class parameters and $2 \cdot 72 = 144$ item parameters. Because the number of model parameters in these full models exceeds the number of students ($I = 71464$), we only estimate $2 \cdot \left(1 + 16 + 16 \cdot \frac{(16-1)}{2}\right) = 274$ parameters of the skill class probability distribution by log-linear smoothed form of the skill space (see Section 2.4). For avoiding memory overflow, we also approximated the original 2^{16} skill classes by $L = 4000$ skill classes determined by a quasi Monte Carlo integration (see also Section 2.4). All statistical models and DIF tests are estimated with the R (R Core Team, 2014) package CDM (George, Robitzsch, Kiefer, Groß, & Ünlü, submitted; Robitzsch, Kiefer, George, & Ünlü, 2014). Here, we do not report standard errors for the skill class probabilities (cf. Johnson et al., 2013), and leave their calculation as an aspect of future research.

4.3 Summary and overview of measures

The global mathematics ability (see Figure 1, Level 1) has been split into the eight original skills reflecting the underlying competence concept (Figure 1, Level 2). A further partition into the sixteen new skills was necessary because of methodological reasons (Figure 1, Level 3). As a consequence, the results obtained by the (MG-)CDMs applied in the following are located on the level of the sixteen new skills (Level 3). On this level, differences in skill mastery between two groups g_1 and g_2 are reported using the appropriate differences $\Delta P_{\alpha_o\alpha_c} = P(\alpha_o\alpha_c|g_1) - P(\alpha_o\alpha_c|g_2) = P_1 - P_2$ in the skill probabilities between the two groups. These differences are transformed into the widely accepted Cohen's d effect size (Cohen, 1988) by

$$d = \frac{\Delta P}{s^*} ,$$

where $s^* = \sqrt{[P_1 \cdot (1 - P_1) + P_2 \cdot (1 - P_2)]/2}$ denotes the pooled standard deviation. Comparing Cohen's d to ΔP, one can show for medium $P_1 \approx P_2 \approx .5$ that $d = 2 \cdot \Delta P$ and for extreme p, say $P_1 \approx P_2 \approx .1$ (or = .9), that $d = 3.33 \cdot \Delta P$.

Despite the detailedness of the results obtained on Level 3, our main interest is still to find relationships between the students' manifest response behavior, the eight original skills, and the two domains (i.e. operation and content) they belong to (Level 2). We establish the link between Level 3 and Level 2 with the help of two alternative methods:

Firstly, in retransforming the sixteen marginal skill probabilities $P(\alpha_o\alpha_c|g)$ to the eight original skills, we back-reference from Level 3 to Level 2. In detail, we define the skill mastery probability of an original skill α_k for group g as the mean of the four combined skill mastery probabilities $P(\alpha_o\alpha_c|g)$ including α_k. For example for skill α_1 Statistics in group g it holds

$$P(\alpha_1|g) = \tfrac{1}{4}[\, P(\alpha_1\alpha_5|g) + P(\alpha_1\alpha_6|g) + P(\alpha_1\alpha_7|g) + P(\alpha_1\alpha_8|g)\,],$$

which means that the four content skills α_5 to α_8 are equally weighted. Note that every other linear combination of $P(\alpha_1\alpha_5|g)$ to $P(\alpha_1\alpha_8|g)$ would also be possible. The summary of skill mastery probabilities we have chosen is compensatory: It allows students to compensate a lack in one combined skill mastery probability (say $P(\alpha_1\alpha_5|g) = .12$) through a large probability in another combined skill (say $P(\alpha_1\alpha_8|g) = .78$). One could also apply a completely compensatory rule, i.e. an original skill α_k is mastered if at least one of the combined skills $\alpha_o\alpha_c$ including α_k is mastered, or a completely non-compensatory rule, i.e. an original skill α_k is mastered if all new skills $\alpha_o\alpha_c$ including α_k are mastered. Due to their strictness, completely compensatory or non-compensatory rules generally require a strong theoretical fundament.

Secondly, we analyze the impact of the operational and content domain on the sixteen skill mastery probabilities $P(\alpha_o\alpha_c)$ and their group-differences $\Delta P_{\alpha_o\alpha_c}$. In conducting analyses of variance, we can describe, if the variability in the skill mastery probabilities (Level 3) is mostly attributed to the operational domain, to the content domain or the

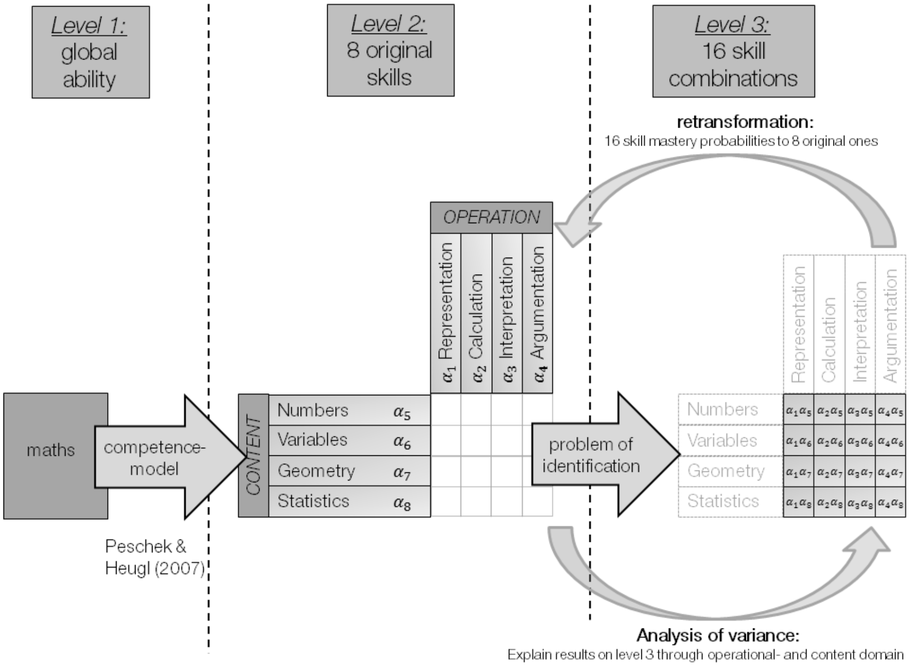
Level 1:
global ability
Level 2:
8 original skills
Level 3:
16 skill combinations
retransformation:
16 skill mastery probabilities to 8 original ones
OPERATION
α_1 Representation
α_2 Calculation
α_3 Interpretation
α_4 Argumentation
maths
competence-model
CONTENT
Numbers α_5
Variables α_6
Geometry α_7
Statistics α_8
problem of identification
Representation
Calculation
Interpretation
Argumentation
Numbers
Variables
Geometry
Statistics
$\alpha_1\alpha_5$ $\alpha_2\alpha_5$ $\alpha_3\alpha_5$ $\alpha_4\alpha_5$
$\alpha_1\alpha_6$ $\alpha_2\alpha_6$ $\alpha_3\alpha_6$ $\alpha_4\alpha_6$
$\alpha_1\alpha_7$ $\alpha_2\alpha_7$ $\alpha_3\alpha_7$ $\alpha_4\alpha_7$
$\alpha_1\alpha_8$ $\alpha_2\alpha_8$ $\alpha_3\alpha_8$ $\alpha_4\alpha_8$
Peschek & Heugl (2007)
Analysis of variance:
Explain results on level 3 through operational- and content domain

Figure 1:
Relations between levels of analysis

interaction of both domains (Level 2). In the analyses of variance, the operational and content domains are treated as factors and the skill mastery probabilities (or their differences) as dependent variable.

5 Results

5.1 Item parameters

In Figure 1 the item parameters of the DINA model analyzing the eight mathematical skills of the whole test population are presented. The plots on the diagonal of Figure 2 show the distributions of the item p-values (relative frequencies of solving), the guessing parameters and the slipping parameters. In the lower panels pairwise scatterplots with smoothed regression lines between the three variables are presented. The associated correlation coefficients are in the upper panels.

The item p-values ranged from 0.11 to 0.94 with a mean of 0.48 and a standard deviation of 0.24. The guessing (min=0.01, max=0.90, M=0.30, SD=0.23) and slipping parameter

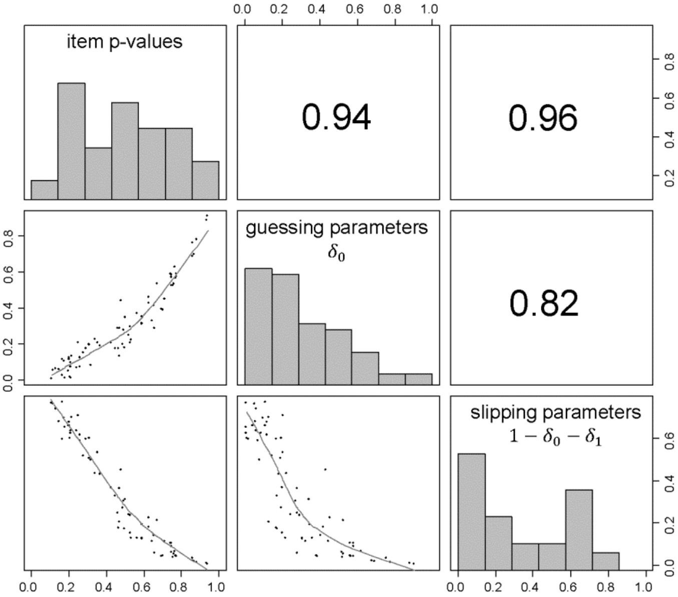

Figure 2:
Item p-values, guessing and slipping parameters for the DINA model on the whole sample: Histograms in the diagonal elements, pairwise scatterplots with smoothed regression lines in the lower panels and appropriate correlations in the upper panels

(min=0.09, max=0.67, M=0.37, SD=0.15) distributions are both skewed to the left and the parameters were highly correlated with $\rho = .82$. The correlations between the item p-values and the guessing parameters ($\rho = .94$) as well as between the item p-values and the slipping parameters ($\rho = .96$) were also very large. These high correlations were expectable given the derived relationships in de la Torre and Karelitz (2009) between the item p-values, the guessing and slipping parameters when a uni-dimensional IRT model with a continuous latent trait holds.

The three parameter distributions and correlations between the parameters may be seen as approximately representative for the two following multiple group models, even if the item p-values in these models of course differ between groups.

5.2 Skill parameters

A CDM based on the four operational skills α_1 to α_4 the four content skills α_5 to α_8 for the whole sample of 71464 eight-graders yields the following results (cf. Figure 3): In mean, the operational and the content skills are both mastered with a probability of $.492$. With regard to the marginal skill mastery probabilities, the operational skill Calculation

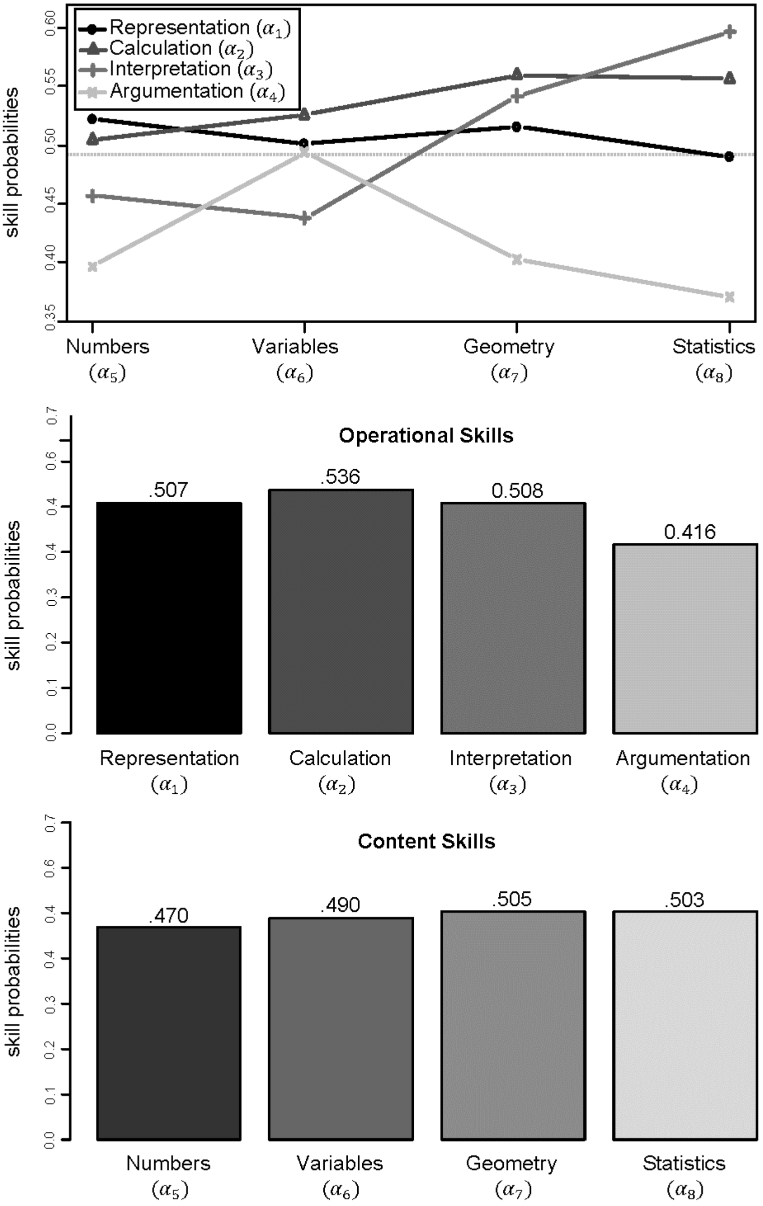

Figure 3:

Representation of skill probabilities for whole sample: Skill probabilities for content skills conditioned on operational skills (top), marginal skill mastery probabilities for operational (middle) and content skills (bottom). The dotted line in the graphic on the top illustrates the mean probability of skill possession

($P_{\alpha_2} = .536$) and the content skill Geometry ($P_{\alpha_7} = .505$) are mastered most often, whereas the operational skill Argumentation ($P_{\alpha_4} = .416$) and the content skill Variables ($P_{\alpha_6} = .490$) are the most difficult ones. In general, the content skills are mastered homogeneously, whereas the mastery of the operational skills is more unbalanced.

In more detail, the mastery of the operational skills shows most variability in the content domain of Statistics (cf. Figure 3 top), ranging from $P_{\alpha_4 \alpha_8} = .370$ to $P_{\alpha_3 \alpha_8} = .596$. On the contrary, the mastery of the operational skills is, with a range from $P_{\alpha_3 \alpha_6} = .437$ to $P_{\alpha_2 \alpha_6} = .525$, most homogenous in the content domain of Variables (α_6). Concerning the operational skills, it can be seen that Representation is mastered most homogeneously ($P_{\alpha_1 \alpha_8} = .490$ to $P_{\alpha_1 \alpha_5} = .522$) with regard to the content domains, whereas the mastery of Interpretation (α_3) yields a large range from $P_{\alpha_3 \alpha_6} = .437$ to $P_{\alpha_3 \alpha_8} = .596$, again because of being easier in Statistics (α_8).

To analyze the impact of the operational and content domain on the sixteen skill mastery probabilities $P(\alpha_o \alpha_c)$ we conducted an analysis of variance: Accordingly, most variability in the skill probabilities can be attributed to the operational skills ($\eta^2 = .542$) and the interaction of both factors ($\eta^2 = .406$). The amount of the explained variance of the content skills is negligible ($\eta^2 = .052$). In this line, it can be stated that the mastery of the content skills is equally difficult for the students, which may be explained by the curriculum giving teachers guidelines about contents of mathematical education in the eighth grade. On the contrary, for the operational skills a rough hierarchy of difficulty may be derived: Calculation seems to be easier than Representation, followed by Interpretation and Argumentation.

5.3 Gender comparisons

Prior to conducting a multiple group model for analyzing the differences in the achievement of mathematical skills between boys and girls, we tested the 72 items for exhibiting gender DIF at the level of skills. In the Wald test 40 out of 72 items turned out to be significant at the 5 % significance level. As we already had the reasonable suspicion that this result is due to the large sample size, we also calculated the UA effect size measure for all significant items (cf. Table 4). It was found that 4 items exhibit moderate gender DIF (.059 < UA < .088) and 2 items show large DIF (UA > .088). These 6 items are unsystematically spread over the different content and operational skills and an educational expert considered these items relevant for the construct (cf. Roussos & Stout, 1996). Thus, the multiple group model for assessing gender differences is conducted with all items.

With a sample size of 35133 female students (49.1 %) and 36331 male students (50.9 %) the multiple group DINA model yields the following results: In mean the difference in the possession of mathematical skills between boys and girls is $.029$ favoring boys. With regard to the marginal skill mastery probabilities (cf. Figure 4), the gender differences in the operational skill Calculation ($\Delta P_{\alpha_2} = .016$) and the content skill Variables ($\Delta P_{\alpha_6} = -.010$) are the smallest, whereas the differences in the operational skill Representation ($\Delta P_{\alpha_1} = .043$) and the content skill Statistics ($\Delta P_{\alpha_8} = .055$) are the largest.

Table 4:
Number of items exhibiting no significant gender DIF and UA effect measure for significant items

	Non-significant	Significant		
		negligible UA<.059	*moderate* .059 < UA < .088	*large* UA >.088
Gender	32	34	4	2
School form	25	30	5	10

Going into detail, two specific aspects are noticeable: The first thing to be mentioned is the content skill Variables (Figure 4, top left): If Variables (α_6) is combined with the operational skills Calculation ($\Delta P_{\alpha_2\alpha_6} = -.044$), Argumentation ($\Delta P_{\alpha_4\alpha_6} = -.037$) or Representation ($\Delta P_{\alpha_1\alpha_6} = -.007$) girls achieve better results than boys. On the contrary, if Variables is combined with Interpretation ($\Delta P_{\alpha_3\alpha_6} = .049$) boys performed better. Second, as counterpart to the content skill Variables, in the content skill Statistics (α_8) boys outperformed girls, independently of the operational skill. All of these differences are ranging from $\Delta P_{\alpha_3\alpha_8} = .043$ to $\Delta P_{\alpha_4\alpha_8} = .079$ and thus are larger than the mean difference between boys and girls (0.029).

Apart from that, the differences exhibit a relatively wide range from $\Delta P_{\alpha_2\alpha_6} = -.044$ to $\Delta P_{\alpha_1\alpha_5} = .089$ and seem to show no systematic effect for neither content nor operational skills. This finding is also confirmed by the analysis of variance, according to which most variability in the skill probabilities can be attributed to the interaction between operational and content skills ($\eta^2 = .773$). In contrast, the amount of the explained variance of only the content skills ($\eta^2 = .082$) or of only the operational skills ($\eta^2 = .143$) is very small.

In summary: the size of the gender differences seems to be neither explainable only by operational nor only by content skills, but only by the interaction between both skill domains. Noticeable is the girls' strength in Variables: In three out of four combinations of Variables with one content skill girls have slight advantages. We also found that girls possess Calculation to approximately the same extent than boys (cf. also Budde, 2009).

5.4 School type comparisons

Again, prior to conducting the multiple group model for analyzing differences between students in academic and general school types, we tested the $J = 72$ items for exhibiting school form DIF at the level of skills. The Wald test found 45 to be significant (cf. Table 4), out of which 30 items show negligible school form DIF (UA < .059), 5 items exhibit moderate (.059 < UA < .088) and 10 items large DIF (UA > .088). An educational expert of mathematics again analyzed the latter 15 items concerning the involved content and

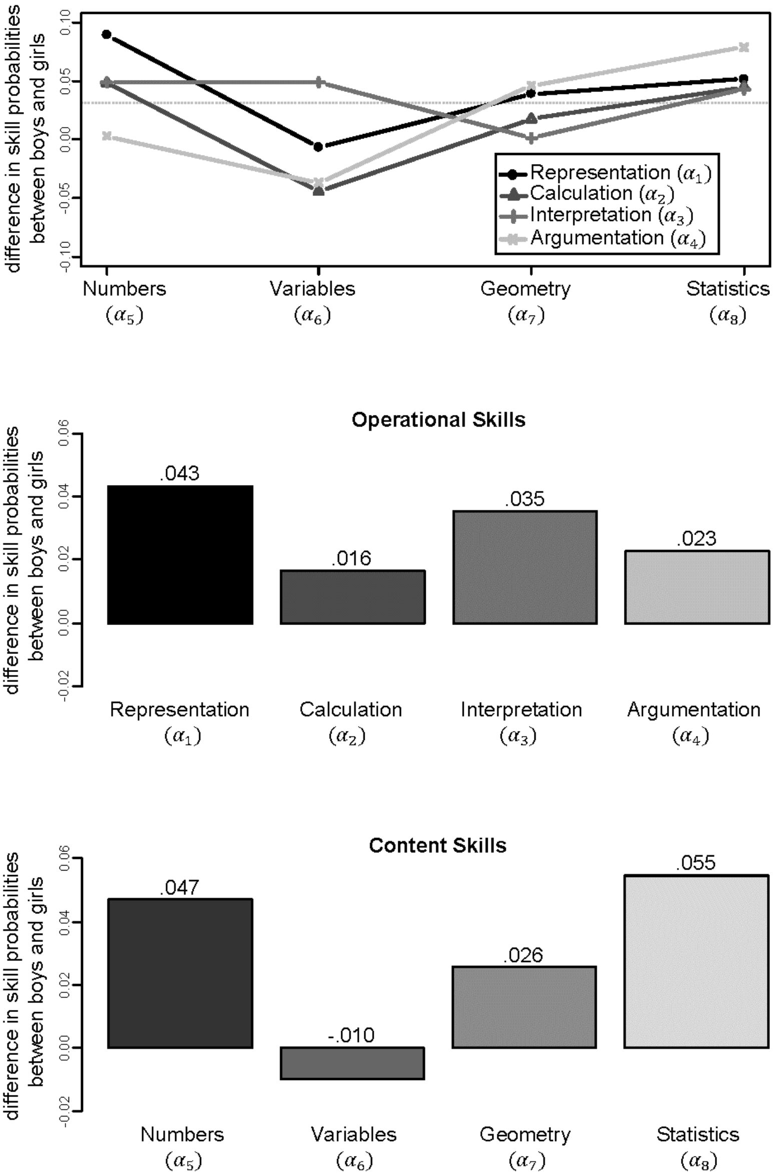

Figure 4:
Representation of differences in skill probabilities between boys and girls: Differences ΔP of skill probabilities for content skills conditioned on operational skills (top), differences of marginal skill mastery probabilities for operational (middle) and content skills (bottom). The dotted line in the graphic on the top illustrates the mean difference in skill possession

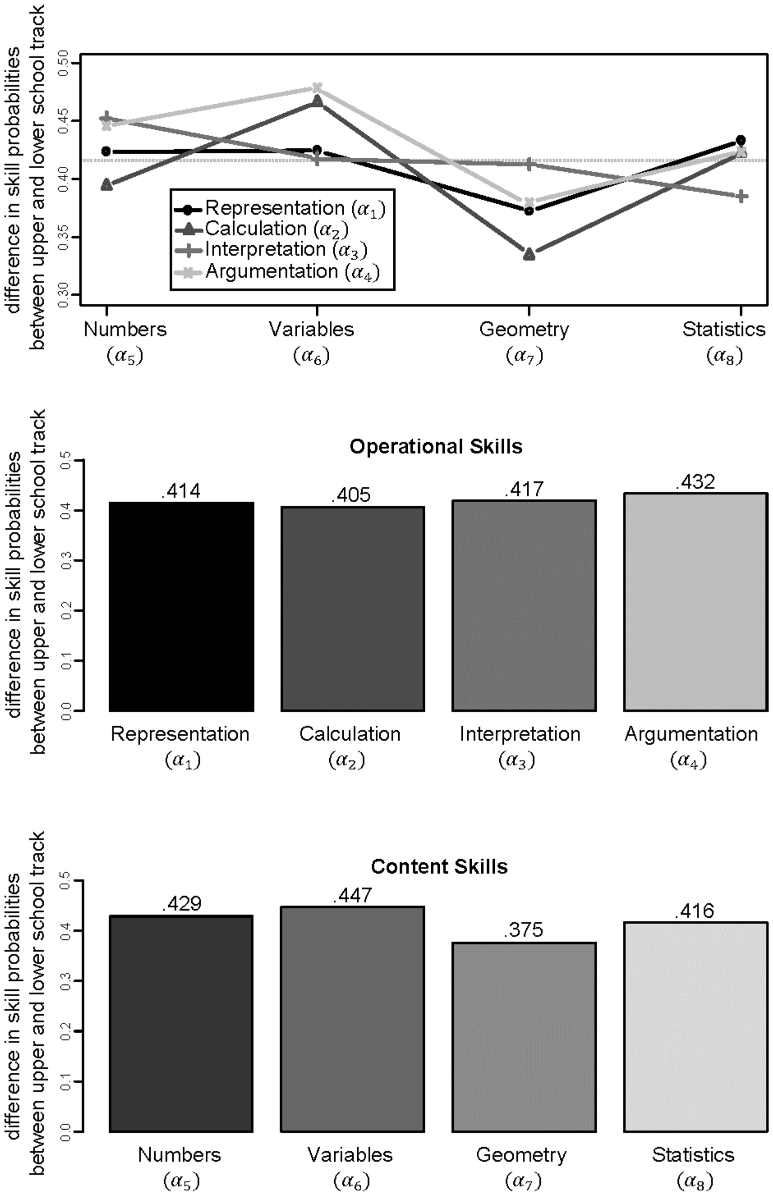

Figure 5:
Representation of differences in skill probabilities between academic and general school type: Differences ΔP of skill probabilities for content skills conditioned on operational skills (top), differences of marginal skill mastery probabilities for operational (middle) and content skills (bottom). The dotted line in the top graphic illustrates the mean difference in skill possession

operational skills and concerning possible construct irrelevant DIF. Because no conspicuous or common characteristics of the items were found, the multiple group model for school form differences is also conducted with all items.

On average, the difference in the possession of mathematical skills between the 23491 students in the academic school type (33 %) and the 47973 students (67 %) in the general school type is $.416$ favoring the academic type. This difference is considerably larger than the mean difference of $.029$ between boys and girls. The differences between students of the academic and the general school type in the possession of the operational skills are rather homogeneous (cf. Figure 5 top right), ranging from a difference of $\Delta P_{\alpha_2} = .405$ in Calculation to $\Delta P_{\alpha_4} = .432$ in Argumentation. On the contrary, in the differences of the content skill probabilities a larger variance can be seen (cf. Figure 5 bottom left): Whereas the difference in the mastery of Geometry is with a value of $\Delta P_{\alpha_7} = .375$ smaller than the mean difference between academic and the general school type, the difference in Variables $\Delta P_{\alpha_6} = .447$ is larger than the mean difference.

Closer inspection shows a rough hierarchy in the possession of the content skills conditional on the operational skills (cf. Figure 5, top left): Whereas Variables seems to exhibit a large school type difference (independently of the combined operational skill), the differences in the skill mastery probabilities of the content skill Geometry (all four combinations of operational skills) are smaller than the mean difference between the school types. On the contrary, the size of the differences in each of the operational skills (cf. Figure 5 bottom right) varies unsystematically around the mean difference. The analysis of variance is in line with this result, as most of the variance in the skill probabilities is explained by the content skills ($\eta^2 = .545$). The remaining part of the variance is almost explained by the interaction between operational and content skills ($\eta^2 = .378$), whereas the part explained by the operational skills ($\eta^2 = .007$) is negligible.

5.5 Summary

In transforming the ΔP values to Cohen's d effect size (cf. Table 5), one determines small (d values between $-.02$ and $.11$) differences between genders, but quite large (d values between $.82$ and 1.01) differences between school types. The maximal gender difference in skill possession is observed in the content skill Statistics ($d = .11$), whereas the maximal difference for the comparison of school types is observed in the content skill Variables ($d = 1.01$). The differences between the skill probabilities show more variability in the comparison of school types than the differences in the gender comparison. A large part of this variability can by ascribed to the content skills.

6 Discussion

In the present article we proposed some methods for applying multiple group CDMs to large scale data and illustrated these by applying multiple group DINA models to the Austrian educational standards test in mathematics 2012 (71464 students, 72 items, 16 skills).

1) First, because of problems in identifying all $2 \cdot (2^{16} - 1) = 131070$ skill class parameters, we used a log-linear approach for modeling the skill space (Xu & von Davier, 2008b) and thus reduced the number of skill class parameters to $2 \cdot (1 + 16 + 16 \ \cdot (16 - 1)/2) = 274$. Nonetheless, even this representation of the appropriate high-dimensional posterior distributions did not prevent from memory overflow problems in the R software when applying the multiple group model to large scale assessment data. Therefore, we proposed to approximate the original $2^{16} = 65536$ skill classes by $L = 4000$ skill classes determined by a quasi Monte Carlo integration (Pan & Thompson, 2007).
2) Second, for applying multiple group CDMs, one has to assure the assumption of item parameter invariance between the groups, i.e. the items must not exhibit DIF on the level of skills. In large sample sizes, the appropriate Wald statistic for detecting DIF (de la Torre & Lee, 2013) may become significant even for very small DIF effects. Thus, we introduced a DIF effect size measure which is based on the unsigned area originally introduced by Raju (1990).

Because of methodological reasons (non-identifiability of skill classes) the Q-matrix with originally defined 8 mathematical skills in two domains (operation and content), had to be changed to a redefined Q-matrix incorporating all 16 combinations between the 4 operational and the four content skills (cf. Table 3). As a side effect, this step changes the

Table 5:
Skill mastery probabilities for whole population (P_{M8}), differences in skill probabilities for gender ($\Delta P_{\mathrm{gender}}$) and school track model ($\Delta P_{\mathrm{school}}$) with associated Cohen's d values (d_{gender} and d_{school}).

		P_{M8}	$\Delta\mathrm{P}_{\mathrm{gender}}$	d_{gender}	$\Delta\mathrm{P}_{\mathrm{school}}$	d_{school}
Operational Skills	Representation (α_1)	.507	.043	.09	.414	.92
	Calculation (α_2)	.536	.016	.03	.405	.91
	Interpretation (α_3)	.508	.035	.07	.417	.93
	Argumentation (α_4)	.416	.023	.05	.432	.96
Content Skills	Numbers (α_5)	.470	.047	.10	.429	.96
	Variables (α_6)	.490	-.010	-.02	.447	1.01
	Geometry (α_7)	.505	.026	.05	.375	.82
	Statistics (α_8)	.503	.055	.11	.416	.93

skills employed in conducted CDM models (cf. Figure 1). For establishing the link between the original 8 skills and the new 16 skills we

3) conducted analyses of variance to analyze the impact of the operational and content domain on the 16 skill mastery probabilities.
4) Furthermore, we retransformed the sixteen marginal skill probabilities to skill mastery probabilities of the eight original skills. For reporting the differences in skill mastery between the groups (on the level of the 8 and the 16 skills) we used differences in skill probabilities, which are afterwards transformed to the widely used Cohen's d effect size.

Apart from these methodological aspects, the chosen example of the Austrian educational standards test also showed an interesting substantial finding: Considering the MG-CDM for gender comparison we found neither the operational nor the content skills to sufficiently explain the differences between boys and girls, but only the interaction between both (the 16 skills). If we trace back the results in skill possession of the 16 skills to the original eight skills, we notice that in general boys exhibit higher skill mastery probabilities. However, two skills, Variables (α_6) and Calculation (α_2), formed an exception: Their effect size measures were close to zero (thus indicating no gender difference), in case of Variables (α_6) even of opposite sign. One could, therefore, argue that girls possess these two skills to the same extent as boys, or, as regards Variables, even to a larger extent. Hence, gender differences manifest themselves on the skill level rather than on the global mathematics domain. This differentiation may explain the diverging results of the PISA 2009 (target population: fifteen year old students in OECD countries) and the TIMSS 2003 (target population: eight graders in OECD countries) studies. While the former obtained an effect size of $d = .12$ (favoring boys overall mathematical capabilities), the latter failed to detect such differences at all ($d = .00$). But taking the items' content into account characteristic differences appear: The TIMSS items focus primarily on Calculation, while OECD items require the skills considered here to an equal extent (cf. Else-Quest, Hyde, & Linn, 2010). Based on the results of our study, it is little surprising that TIMSS found no gender differences in mathematics, because of testing primarily the skill we found of equal difficulty for boys and girls. However, the magnitude of the differences between groups and the interactions between skills needs further investigation, since no measures of standard error of these estimates were presented. Calculating standard errors for CDMs applied to data based on complex sample designs is an ongoing research topic. One pertinent approach is the application of jackknife methods for computing the standard errors (Johnson et al., 2013).

Moreover, the skill based approach of CDMs allows for a deeper understanding of the Austrian school type differences: an analysis of variance showed that the school type differences can be explained to an large extent by the differences in the possession of the content skills. Whereas all differences in the possession of the operational skills approximately have the size of the mean difference in skill possession between the two school types, in the Content skills strong advantages for the academic school type could be identified in Variables (α_6). Particularly with regard to both school types being bound to the same curriculum, the empirical findings about the school type differences open the expected and demanded possibility for a deeper analysis of learning cultures in the different Austrian school types (cf. also Eder & Mayr, 2001).

The present article has shown some approaches for handling typical problems which occur when applying CDMs to large scale data. However, we have to acknowledge that the results of the multiple group DINA models for the educational standards test may be limited in their explanatory power because they depend on the characteristics of the item pool. For example items requesting Interpretation (α_3) in Statistics (α_7) only require to "read" diagrams and thus are easier than interpreting dependencies among variables $(\alpha_3\alpha_6)$.

We are aware that in our modelling approach, no uncertainty in item allocation to skill dimensions (i.e. in the Q-matrix) is assumed. However, especially in retrofitting CDMs to existing data, different experts may propose different Q-matrices. The uncertainty in the Q-matrix entries can be accommodated by specifying appropriate prior distributions for these entries (DeCarlo, 2012) and may be evaluated by mixture cognitive diagnosis models (de la Torre & Douglas, 2008; Huo & de la Torre, 2014).

Furthermore, it should be empirically tested if the 16-dimensional skill representation (two facets of four dimensions of operational skills and four dimensions of content skills) proposed in this article may be reduced to a lower dimensional and hence more parsimonious CDM. A possible alternative is the higher order DINA model (de la Torre & Douglas, 2004), which includes eight (four operational plus four content skills) dichotomous skills loading on a higher order continuous ability. Combining both variants and proceeding like in two-tier models (Cai, 2010), one could sum up the two skill facets as two groups of operational and content skills. In such an approach, each of the content skills is assumed to be uncorrelated with each of the operational skills.

Finally, the proposed multiple group CDM could also be seen as a model-based classification approach. By pursuing this direction, the classification into mastered and non-mastered skills depends on the estimated guessing and slipping probabilities, which are interpreted as classification error probabilities. These classification errors could also be (implicitly) fixed in classification models based on cluster analysis (Chiu, Douglas, & Li, 2009) by specifying appropriate optimization functions. Consequently, the mastery of skills can be defined in a more formative and normative way and does not involve the CDM's inherent assumption of local independence; see also the "deterministic" classification approach of Chiu and Douglas (2013).

References

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York, NY: Springer.

Breit, S., & Schreiner, C. (2012). *Bundesergebnisbericht Standardüberprüfung Mathematik 2012, 8. Schulstufe [Report of the educational standards test in mathematics 2012, 8. grade]*. Salzburg, Austria: Bundesinstitut für Bildungsforschung Innovation und Entwicklung des österreichischen Schulwesens.

Budde, J. (2009). Mathematikunterricht und Geschlecht. Empirische Ergebnisse und pädagogische Ansätze. [Lessons in mathematics and gender. Empirical results and pedagogical approaches] Berlin: Bundesministerium für Bildung und Forschung.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*(4), 581–612.

Carstensen, C. H., & Rost, J. (2007). Multidimensional three-mode Rasch models. In M. von Davier & C. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 157–175). New York, NY: Springer.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419–437.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123–140.

Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification, 30*(2), 225–250.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*(4), 633–665.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cui, Y., Gierl, M. J., & Chuang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*(1), 19–38.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595–624

de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement, 46*(4), 450–469.

de la Torre, J., & Lee, Y.-S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement, 47*(1), 115–127.

de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*(4), 355–373.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-Matrix via a Bayesian extension of the DINA Model. *Applied Psychological Measurement, 36*(6), 447–468.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1–38.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, Volume 26, Psychometrics* (pp. 979-1030). Amsterdam, Netherlands: Elsevier.

Eder, F., & Mayr, J. (2001). Die Primadonna, das Aschenputtel und die Unschuld vom Lande: Vergleichende Befunde zu den Schulen der Zehn- bis Vierzehnjährigen. [Prima donna, Cinderella and innocent country maids: Comparing Schools of ten to fourteen year old students]. In F. Eder, G. Grogger & J. Mayr (Eds.), *Studien zur Bildungsforschung und*

Bildungspolitik: Sekundarstufe I [Studies of educational research and educational policy: Secondary schools]. Innsbruck, Austria: StudienVerlag.

Eid, M., Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation models for multitrait-multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of Multimethod Measurement in Psychology* (pp. 283–299). Washington, DC: American Psychological Association.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103–127.

Formann, A. K. (2007). (Almost) equivalence between conditional and mixture maximum likelihood estimates for some models of the Rasch type. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 177-189). New York: Springer.

Formann, A. K., & Kohlmann, T. (1998). Structural latent class models. *Sociological Methods and Research, 26*(4), 530-565.

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (submitted). *The R package CDM for Cognitive Diagnosis Models*.

Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions (RR-08-45)*. Princeton, NJ: Educational Testing Service.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301–321.

Harks, B., Klieme, E., Hartig, J., & Leiss, D. (2014). Separating cognitive and content domains in mathematical competence. *Educational Assessment, 19*(4), 243–266.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* University of Illinois, Urbana Champaign.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*(1), 98–125.

Huo, Y., & de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement, 38*(6), 464–485.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349.

Johnson, M., Lee, Y.-S., Sachdeva, R. J., Zhang, J., Waldman, M., & Park, J. Y. (2013, March). *Examination of gender differences using the multiple groups DINA model*. Paper presented at the 2013 Annual Meeting of the National Council on Measurement in Education, San Francisco CA.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.

Kuhn, J.-T., & Kiefer, T. (2013). Optimal test assembly in practice: The design of the Austrian educational standards assessment in mathematics. *Zeitschrift für Psychologie, 221*(3), 190–200.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35*(2), 64–70.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*(2), 99–120.

Pan, J., & Thompson, R. (2007). Quasi-Monte Carlo estimation in generalized linear mixed models. *Computational Statistics & Data Analysis, 51*(12), 5765–5775.

Penfield, R., & Camilli, G. (2006). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, Volume 26, Psychometrics* (pp. 125–167). Amsterdam, Netherlands: Elsevier.

Peschek, W., & Heugl, H. (2007). *Standards für die mathematischen Fähigkeiten österreichischer Schülerinnen und Schüler am Ende der 8. Schulstufe*: Alpen-Adria-University Klagenfurt, Austria.

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis of Guttman scaling. *Psychometrika, 35*(1), 73–78.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197–207.

Rijmen, F. (2011). Hierarchical factor item response theory models for PIRLS: Capturing clustering effects at multiple levels. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 4,* 59–74.

Robitzsch, A., Kiefer, T., George, A. C., & Ünlü, A. (2014). CDM: Cognitive Diagnosis Modeling. R Package version 3.4. Retrieved from http://CRAN.R-project.org/package=CDM

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355–371.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44*(4), 293–311.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307.

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement: Interdisciplinary Research and Perspectives, 7*(1), 67–74.

von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'diagnostic' classification models — A commentary. *Psychometrika, 79*(2), 340–346.

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*(3), 115–124.

Xu, X., & von Davier, M. (2008a). *Comparing multiple-group multinomial log-linear models for multidimensional skill distributions in the general diagnostic model (RR-08-35)*. Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008b). *Fitting the structured general diagnostic model to NAEP data (RR-08-27)*. Princeton, NJ: Educational Testing Service.