Dissertation

Research School Education and Capabilities
TU Dortmund University, Germany August 2013

Investigating CDMs: Blending theory with practicality

Ann Cathrice George

Advisors:
Prof. Dr. Wilfried Bos
Insitute for School Development Research, TU Dortmund University

PD Dr. Jürgen Groß
Department of Statistics, TU Dortmund University

Research School Education and Capabilities
educap

technische universität dortmund

"Oh, I get by with a little help from my friends
Mm, I get high with a little help from my friends
Mm, gonna try with a little help from my friends."

The Beatles, With a little Help from my Friends, 1967

Contents

1 Theory and aspects of research

1.1 CDMs in the light of educational standards

In recent years educational research in Germany was characterized by an increasing demand of complex information on students' achievement. This may be caused by only moderate performances of German students in international comparative studies like the Trends in International Science Study (TIMSS; Mullis, Martin, Ruddock, O'Sullivan, Arora & Erberer, 2008), the Progress in International Reading Study (PIRLS; Mullis, Martin, Kennedy & Foy, 2007) and the Program for International Student Assessment (PISA; OECD, 2010). It may also be caused by the social and ethnic disparities detected in these studies (e.g., Mullis et al., 2007).

Consequently, the standing conference of the ministers of education and cultural affairs (Kultusministerkonferenz der Länder der Bundesrepublik Deutschland; KMK) passed in 2003 the educational standards, which yield binding and unified performance requirements for all German federal states for the first time. According to the KMK (2004) these requirements should be considered as a norm for students' performances (cf. also Klieme, Avenarius, Blum, Döbrich, Gruber, Prenzel, Reiss, Riquarts, Rost, Tenorth & Vollmer, 2003). In other countries similar devolopments took place and comparable rules for educational performance standards have been developed as well. For example the ministry for education, arts and culture (Bundesministerium für Unterricht, Schule, Kunst und Kultur; Bmukk) in Austria introduced educational standards in 2009 (BGBl. II Nr 1/2009) and Sweden, Finland and the USA exhibit comparable concepts. The efforts of the OECD (OECD, 2004) in defining standards by introducing the PISA study should be mentioned in this context as well.

For transferring the educational standards of the KMK (and other institutions) to an adequate testing and learning culture, statistical methods have to be found for empirically evaluating statements about the students' actual competence profiles and about their acquisition of competences. With these statistical methods the norms defined before should be tested and the need for individual support should be identified. Cur-

rently, large-scale assessments as TIMSS, PIRLS and PISA are often accompanied by standardized tests combined with statistical item response theory (IRT; de Ayala, 2009) methods, which yield valuable results for the evaluation of educational systems. However, the diagnostic content of these types of assessments is often criticized. For example the National Research Council (National Research Council, 2001, p.27) stated

> "On the whole, most current large scale tests provide very limited information that teachers and educational administrators can use to identify why students do not perform well or to modify the conditions of instructions in ways likely to improve student achievement."

or educational researchers as de la Torre & Karelitz (2009, p.450) claimed that

> "Scores derived from this (i.e. the IRT) framework are useful in scaling and ordering students along a proficiency continuum, but these proficiency scores contain limited diagnostic information necessary for the identification of students' specific strength and weaknesses."

In the present work so-called cognitive diagnosis models (CDMs; Rupp, Templin & Henson, 2010) are reviewed, applied and enhanced, which allow diagnostic conclusions and hence targeted pedagogical interventions. It seems worth noting that the aim of this work is *not* to criticize present approaches in large scale assessments. The currently applied methods seem adequate as long as the goal of these studies is defined in a description and comparison of educational systems (as it is). Rather, the focus of the present work is to investigate alternative models, which may provide further information for diagnosing students' performances beyond the system and class level, but also on the individual student level. The following paragraphs indicate why CDMs fulfill some of the prevailing needs and demands of educational research on diagnostics of performances or, in other words, assessment of competences.

Recently, the concept of "competences" is often mentioned in the context of students' performances. In the expertise for national educational standards (Klieme et al., 2003), which was an important pillar in the development of Germany's educational standards (KMK, 2004), competences are defined according to Weinert (2001, p.27f) as

> "available or learnable cognitive capacities and abilities of individuals for solving specific problems, as well as the related motivational, volitional and social willingness and ability to responsibly and successfully apply the problem solving strategies in various situations."

Hence, competencies are seen as synonymous to the potential of solving problems from

specific topic areas in specific situations (cf. Kanning, 2003, p.12; Prenzel, Drechsel, Carstensen & Ramm, 2004, p.18). From a psychological point of view, this concept of competences is very broadly defined, as it includes not only cognitive psychological elements (i.e. capacities and abilities) but also motivational elements (i.e. motivational, volitional and social willingness). It therefore allows for several tests and questionnaires about achievement, personality and behavioral assessment (Kubinger, 2006; Rost, 2004) for connecting the measured constructs.

Such a broad definition opens up liberties in the composition of normative demands such as the educational standards. On the contrary, the broader the definition, the more challenging is finding an adequate operationalization of the individual constructs and of their nomological integration (Embretson, 1983). The precision which is necessary in the definition of the constructs has to be emphasized in order to ensure that the applied measurement instruments generate reliable data and permit valid statements about students. In this regard four characteristics can be carved out, which define the concept of competences in the present work:

(1) Competences generally represent coarsely defined abilities, which may also be seen as competence levels or skills. In this sense, the mathematical skill "handling of numbers and measures" can be regarded as a part of the students' mathematical abilities (Prenzel, Drechsel, Carstensen & Ramm, 2004, p.50). The splitting of abilities into skills is normally based on educational and subject oriented didactics (Niss, 2003).

(2) The description of skills and their connections to the total ability is often explained in so called competence models (cf. Campbell, Kelly, Mullis, Martin & Sainsbury, 2001, for reading; Peschek & Heugl, 2007, for math).

(3) The definition of separate skills does of course not exclude the possibility that students have to possess a combination of these skills for successfully mastering a test problem. From a didactical point of view, in the problem solving process of complex tasks this is even desired (Blum, Neubrand, Ehmke, Senkbeil, Jordan, Ulfig & Carstensen, 2004). Thus it is important to define whether the relationship among the skills is compensatory or non-compensatory, i.e. if in the process of problem solving a lack in one required skill can be compensated by the possession of another skill or not.

(4) The skills are obviously not directly observable and therefore they have to be distinguished from the observable test responses (Prenzel et al., 2004, p.19). This aspect is transfered to the statistical level in modeling skills with latent variables.

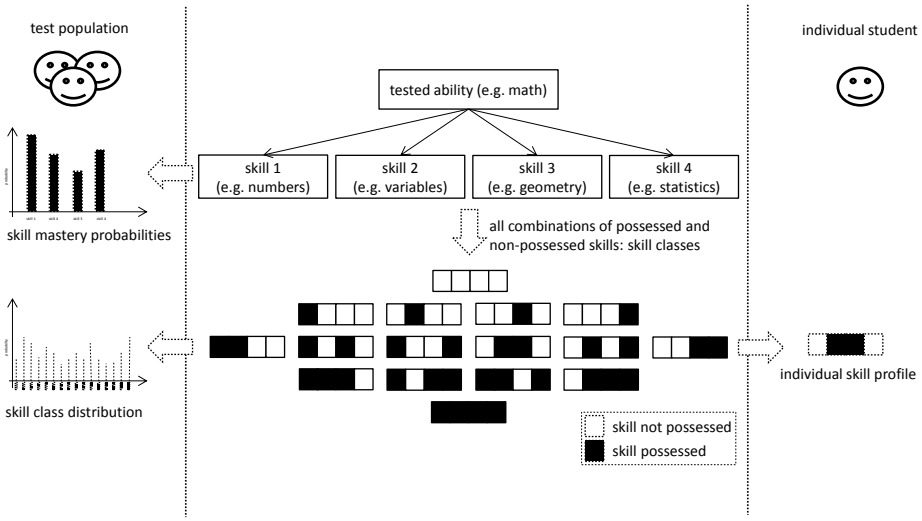

Figure 1.1.1: Illustration of population and individual oriented CDM results.

This theoretical confinement of the concept of competencies used in the education standards reveals a strong coherence to the CDM framework presented in this work: They (1) model latent skills with latent variables, which (2) can be combined in compensatory or non-compensatory ways, can therefore (3) represent complex competence structures and hence can (4) provide statements about individual skills.

Roughly spoken, the goal of CDMs is to classify students based on their observed response behavior in dichotomous latent skill classes, which predict the presence or absence of predefined skills underlying the tested ability. The main results obtained from CDMs are threefold: Firstly, the distribution of the skill classes in the test population allows for statements how many students possess certain combinations of skills. Secondly, the skill mastery probabilities in the population show how many students possess the individual skills. Thirdly, for each student an individual skill class is deduced, explaining the student's possession or non-possession of the individual skills. The concept of CDM results is illustrated in Figure 1.1.1.

Despite these possibilities, CDMs are not very well known in the empirical educational research so far. The reason may be grounded in various different statistically sophisticated modeling approaches or in the sparse number of successful CDM applications to empirical educational data (Templin & Henson, 2006). The present work is twofold: On the one hand it introduces some new statical aspects of CDMs and on the other hand it presents some new applications of CDMs to current educational data sets, for example the Austrian test of educational standards 2012 or the PIRLS 2006 study. Both statistical theory and practical applications are blended in using the statistical details

for the practical analyses and in illustrating the statical theory by real life examples. The research questions and their context are presented in detail in Section 1.3. For a better understanding of these topics at first the statistical theory of CDMs is reviewed and an embedding of CDMs in the context of related classification approaches is given in Section 1.2.

1.2 CDMs: Definition, estimation and related approaches

The origins of cognitive diagnosis models have not been conclusively established. Independent developments emerged from different directions: Firstly, from theory of classification, where basic ideas can be found in the mastery model by Macready & Dayton (1977) and in restricted latent class models by Haertel (1989). Secondly, from item response theory with initial approaches in the multicomponent model by Whitely (1980) and in the linear logistic test model by Fischer (1973), and thirdly from mathematical psychology, and here especially the field of knowledge space theory, see e.g. Doignon & Falmagne (1999). Based on the multitude of different approaches, CDMs have many names, as for example diagnostic classification models, cognitive psychometric models or structured item response theory models.

In all CDM approaches it is assumed that a set of basic skills (i.e. competencies) is underlying the tested ability. Furthermore, all CDM approaches determine the possession and non-possession of these skills (i.e. the skill classes) in the test population and for the individual students. Therefore, all approaches require the responses of examinees to (test) items and an expert assignment of the latent skills to these items. Even though CDMs may also be applied to psychological tests (Templin & Henson, 2006), in the present work we focus on CDMs for educational testing data.

The huge variety of CDMs differs basically in two aspects: Firstly, the combination in which students have to possess the skills for successfully mastering an item, i.e. the level of compensability. In some CDMs all assigned skills have to be possessed for mastering the items, in other CDMs just one of the assigned skills has to be possessed, and other CDMs require one of several specific combinations of the assigned skills. CDMs in which exactly one skill is required for mastering each item are called CDMs with between item dimensionality, whereas CDMs requiring more than one skill have a within item dimensionality. Secondly, CDM approaches differ in the way a stochastic component is

introduced into the model, i.e. students can slip or guess in items or skills or in both. In an achievement test each item may follow another CDM approach. These different approaches for the items are sometimes also called the items condensation rule or simply the item rule (DiBello, Roussos & Stout, 2007). A summary and discussion of prominent CDMs is for example given in DiBello et al. (2007), George (2010) or Rupp et al. (2010). Current research yields approaches which unify many different models in one framework as for example the Generalized-Deterministic Input Noisy-And-Gate (G-DINA; de la Torre, 2011) model, the General Diagnostic Model (GDM; von Davier, 2008) and the Log-linear Cognitive Diagnosis Model (LCDM; Henson, Templin & Willse, 2009).

In this section only models and frameworks are reviewed which are examined or applied in the present work: Firstly, the Deterministic Input Noisy And Gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model (cf. Section 1.2.2), which has achieved some popularity because of its simplicity in interpretation and its parsimony in establishing model parameters. Secondly, the Generalized-DINA (de la Torre, 2011, cf. Section 1.2.3), as it is the generalized framework build upon the DINA. Thirdly, Section 1.2.5 shows some important connections and equalities between the three generalized frameworks G-DINA, GDM and LCDM, and therefore justifies the subsequent priority in the use of the G-DINA model. Finally, CDMs are set in the context of related psychometric models (cf. Section 1.2.6).

The reviewed models are illustrated by data from the Austrian baseline testing 2009 of educational standards in math (Breit & Schreiner, 2010). In this study each test item (i.e. task in a test) is assigned to exactly one of the four content subcategories "numbers and measures", "variables and functional dependencies", "geometry" and "statistics" and to exactly one of the four operational subcategories "model building", "calculation", "interpretation" and "argumentation". In the present context the content and operational subcategories are used as basic skills underlying the tested mathematical ability of students in the eighth grade.

1.2.1 Basic components of CDMs, terminology and notation

Consider a test situation, in which I, $i = 1, \ldots, I$, students responded to J, $j = 1, \ldots, J$, items. A value of 1 indicates a correct response and a value of 0 an incorrect one. The binary empirical (manifest) response of student i, $i = 1, \ldots, I$, to item j, $j = 1, \ldots, J$, is denoted by X_{ij}. The responses of all I students to all J items are given in a $I \times J$ binary data matrix $\boldsymbol{X}$. The i-th row of $\boldsymbol{X}$ represents the answers of student i to all J

items, denoted by the J-dimensional response vector $\boldsymbol{X}_i$, which is called the response pattern of student i.

Educational experts define a set of K basic skills α_k, $k = 1, \ldots, K$, which the students have to possess for mastering all J items under consideration ($K \leq J$). The i-th student's dichotomous skill profile $\boldsymbol{\alpha}_i = [\alpha_{i1}, \ldots, \alpha_{iK}]$ denotes her possession and non-possession of the K predefined skills. Obviously the skill profiles are unknown. Furthermore the educational experts also define which skills are required to master which item through the $J \times K$ matrix $\boldsymbol{Q}$, the so-called Q-matrix (Tatsuoka, 1983): The (j,k)th element q_{jk} of $\boldsymbol{Q}$ is equal to 1 if skill k is relevant for the mastery of item j and equals 0 otherwise. Additionally the experts have to specify the items' condensation rules.

In the example of the Austrian baseline testing we consider one test booklet with $J = 36$ items which was administered to $I = 1308$ eight graders in Austria. Thus, the data matrix $\boldsymbol{X}$ has a size of 1308×36. Educational experts assigned each item to either exactly one content skill or to exactly one content and one operational skill. The first assignment with $K = 4$ content skills is summarized in a 36×4 matrix

$$\boldsymbol{Q}_{content} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \ldots & \ldots & \ldots & \ldots \end{bmatrix}$$

and the second assignment with $K = 8$ content and operational skills leads to

$$\boldsymbol{Q}_{content;operation} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix}.$$

CDMs assume that the manifest response X_{ij} of student i to item j arises as a result of her possessed skills $\boldsymbol{\alpha}_i$, the skills required for item j defined in the j-th row $\boldsymbol{q}_j$ of the Q-matrix and the j-th item's condensation rule (cf. Figure 1.2.2). Because the skills $\boldsymbol{\alpha}_i$ are unknown, a CDM algorithm deduces from the manifest responses, the Q-matrix and the condensation rules information on the K skills the student possesses.

From a statistical point of view this procedure has two steps: In the first step all students

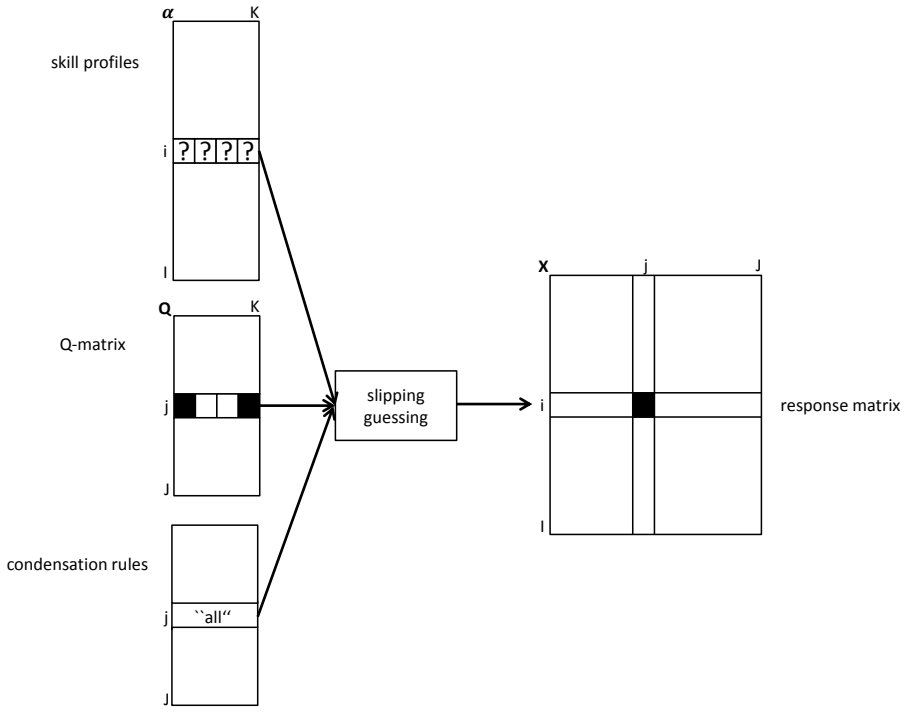

Figure 1.2.2: In CDMs the manifest response X_{ij} of student i to item j is assumed to arise as a result of the student's possessed skills $\boldsymbol{\alpha}_i$, the skills required for item j defined in the j-th row $\boldsymbol{q}_j$ of the Q-matrix and the j-th item's condensation rule.

are classified into skill classes $\boldsymbol{\alpha}_l$, $l = 1, \ldots, 2^K$, satisfying a global optimization criterion. Note, that all possible combinations of the assumed K skills yield the 2^K disjunctive skill classes $\boldsymbol{\alpha}_l$, $l = 1, \ldots, 2^K$. In our example, for $K = 4$ skills we obtain $2^K = 2^4 = 16$ skill classes, i.e. $\boldsymbol{\alpha}_1 = [0,0,0,0]$, $\boldsymbol{\alpha}_2 = [1,0,0,0]$, $\boldsymbol{\alpha}_3 = [0,1,0,0]$, $\boldsymbol{\alpha}_4 = [0,0,1,0]$, $\boldsymbol{\alpha}_5 = [0,0,0,1]$, $\boldsymbol{\alpha}_6 = [1,1,0,0]$, $\boldsymbol{\alpha}_7 = [1,0,1,0]$, $\boldsymbol{\alpha}_8 = [1,0,0,1]$, $\boldsymbol{\alpha}_9 = [0,1,1,0]$, $\boldsymbol{\alpha}_{10} = [0,1,0,1]$, $\boldsymbol{\alpha}_{11} = [0,0,1,1]$, $\boldsymbol{\alpha}_{12} = [1,1,1,0]$, $\boldsymbol{\alpha}_{13} = [1,1,0,1]$, $\boldsymbol{\alpha}_{14} = [1,0,1,1]$, $\boldsymbol{\alpha}_{15} = [0,1,1,1]$, $\boldsymbol{\alpha}_{16} = [1,1,1,1]$. Students who are classified in skill class $\boldsymbol{\alpha}_5 = [1,1,0,0]$ are predicted to have mastered the first and the second skill but not to possess the third and the fourth one. That is, they are predicted to be able to handle "numbers and measures" and "variables and functional dependencies" but not to master "geometry" and "statistics". From this first step the *distribution of the skill class probabilities*, i.e. the relative frequencies $P(\boldsymbol{\alpha}_l)$, $l = 1, \ldots, 2^K$, of students classified into the skill classes $\boldsymbol{\alpha}_l$, is obtained. If for example $P(\boldsymbol{\alpha}_5) = P([1,1,0,0]) = .13$, then 13 percent of the eight graders have mastered the first and the second skill. We also get the *skill mastery probabilities* $P(\alpha_k)$, $k = 1, \ldots, K$, giving for each skill α_k the relative frequency of students in possession of it. For example $P(\alpha_1) = 0.26$ means that 26 percent of all

students possess the first skill "numbers and measures". Obviously, $\sum_{l=1}^{2^K} P(\boldsymbol{\alpha}_l) = 1$ and $\sum_{k=1}^{K} P(\alpha_k) = 1$.

In a second step, the CDM algorithm deduces the skill classes which are optimal for each individual student i, $i = 1, \ldots, I$. The i-th student's vector of present and absent skills is also called the i-th student's (simplified) *skill profile* and is denoted by $\hat{\boldsymbol{\alpha}}_i = [\alpha_{i1}, \ldots, \alpha_{iK}]$. The skill profiles can be easily used as feedback for teachers or parents, providing a solid empirical base for further instruction and learning. In our example with $K = 4$ skills the skill profile of student $i = 137$ may be $\hat{\boldsymbol{\alpha}}_{137} = [1, 1, 0, 0]$ and thus the student should be supported in "geometry" and "statistics" because she does not master these skills yet.

Note that we have to distinguish between the skill classes $\boldsymbol{\alpha}_l$, $l = 1, \ldots, 2^K$, in the population and the individual skill profiles $\boldsymbol{\alpha}_i$, $i = 1, \ldots, I$, even though both are represented by K-length dichotomous vectors. Obviously, the 2^K possible skill classes cover all I skill profiles. For the ease of notation we use the same symbol. It will always become clear from the context whether $\boldsymbol{\alpha}_1$ refers to the first skill class or the first individual skill profile.

1.2.2 DINA

The Deterministic Input Noisy-And-Gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model is a very popular core CDM because of its simplicity and its parsimony in the use of model parameters. It was one of the first CDMs introduced as restricted latent class models by Haertel (1989, compare Section 1.2.6). The DINA model asserts that students have to possess all skills assigned to an item via the Q-matrix for successfully mastering it. To put it differently, the DINA model is completely non-compensatory, in that a lack in a single required skill can not be compensated.

The i-th student's probability to master the j-th item involves a deterministic one and a probabilistic component (cf. Figure 1.2.3). The former states whether the student is *expected* to master the j-th item given her possessed skills. If the student possesses all required (or even more) skills for item j, she is expected to master the item, whereas if she lacks at least one required skill, she is not expected to master the item. This deterministic component is expressed through the dichotomous latent response

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

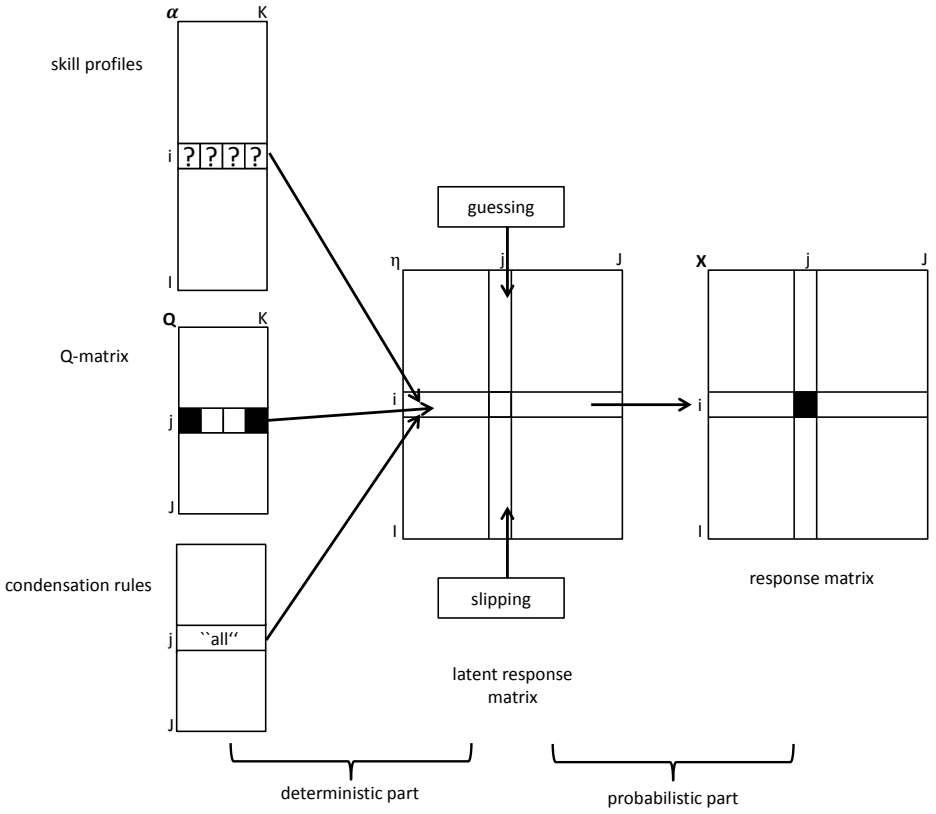

Figure 1.2.3: In the DINA model the manifest response X_{ij} of student i to item j is assumed to arise as a result of the student's possessed skills $\boldsymbol{\alpha}_i$, and *all* skills required for item j defined in the j-th row $\boldsymbol{q}_j$ of the Q-matrix. The stochastic component of the slipping and guessing errors is modeled on the level of items.

of student i with skill profile $\boldsymbol{\alpha}_i = [\alpha_{i1}, \ldots, \alpha_{iK}]$ to item j, where $[q_{j1}, \ldots, q_{jK}]$ denotes the q-th row of the Q-matrix. In case of $\eta_{ij} = 1$ the student is expected to master item j, in case of $\eta_{ij} = 0$ she is not. The latter, namely the probabilistic component, possible *deviates* from these expectations. On the one hand, if student i is expected to master the item (i.e. $\eta_{ij} = 1$), she may nevertheless slip and not solve the item. On the other hand, even if $\eta_{ij} = 0$ (i.e. she is not expected to master the item), she may succeed by luckily guessing the correct response. Thus

$$P(X_{ij} = 1|\boldsymbol{\alpha}_i) = (1 - s_j)^{\eta_{ij}} \cdot g_j^{(1-\eta_{ij})} = \begin{cases} 1 - s_1 & \text{for } \eta_{ij} = 1, \\ g_1 & \text{for } \eta_{i1} = 0. \end{cases}$$

Hence, for a given item j, $j = 1, \ldots, J$, all students have either the probability g_j to solve the item by lucky guess (conditional on not being expected to master the item, i.e.

$\eta_{ij} = 0$) or the probability $1 - s_j$ not to slip item j (conditional on being expected to master item j, i.e. $\eta_{ij} = 1$). As can be seen, the probabilities of guessing and slipping are modeled as item specific parameters.

Let us again consider the example in which the four content skills and the four operational skills in the testing of educational standards in math are analyzed. Let us further consider a student $i = 1$ with skill profile $\boldsymbol{\alpha}_1 = [1, 1, 1, 0, 0, 0, 1, 1]$ and recall the first row of the Q-matrix $\boldsymbol{Q}_{content;operation}$ with entries $\boldsymbol{q_1} = [0, 0, 1, 0, 0, 0, 1, 0]$. Because of

$$\eta_{11} = \prod_{k=1}^{K} \alpha_{1k}^{q_{1k}} = 1^0 \cdot 1^0 \cdot 1^1 \cdot 0^0 \cdot 0^0 \cdot 0^0 \cdot 1^1 \cdot 1^0 = 1$$

student 1 is *expected* to master item 1 in skill class $\boldsymbol{\alpha}_1$. Thus she is *likely* to master the item with probability $P(X_{11}|\boldsymbol{\alpha}_1) = 1 - s_1$, where s_1 is the slipping parameter of item 1. More generally spoken, the DINA model's two-probability constraint in item 1 of this example is

$$P(X_{i1} = 1|\boldsymbol{\alpha}_i) = \begin{cases} 1 - s_1 & \text{for all } \boldsymbol{\alpha}_i \text{ with } \alpha_{i3} = 1 \wedge \alpha_{i7} = 1 \text{ and thus } \eta_{i1} = 1, \\ g_1 & \text{for all } \boldsymbol{\alpha}_i \text{ with } \alpha_{i3} = 0 \vee \alpha_{i7} = 0 \text{ and thus } \eta_{i1} = 0. \end{cases}$$

1.2.3 G-DINA

Remember that the DINA model will always employ one of the two probabilities $1 - s_j$ or g_j for correctly solving item j, $j = 1, \ldots, J$:

$$P(X_{ij} = 1|\boldsymbol{\alpha}_i) = \begin{cases} 1 - s_j & \text{if all required skills are possessed,} \\ g_j & \text{if at least one required skill is not possessed.} \end{cases} \tag{1.2.1}$$

For relaxing this restrictive constraint, de la Torre (2011) introduced the Generalized-DINA (G-DINA) model, in which students exhibiting different sets of required skills have different probabilities of mastering item j. For that purpose, the G-DINA model employs the item response function

$$P\left(X_{ij} = 1 \middle| \boldsymbol{\alpha}_{j;i}^*\right) = \delta_{j;0} + \sum_{k=1}^{K_j^*} \delta_{j;k} \alpha_{j;ik}^* + \sum_{k=1}^{K_j^*-1} \sum_{k'=k+1}^{K_j^*} \delta_{j;kk'} \alpha_{j;ik}^* \alpha_{j;ik'}^* + \ldots + \delta_{j;12\ldots,K_j^*} \prod_{k=1}^{K_j^*} \alpha_{j;ik}^*. \tag{1.2.2}$$

Here $\boldsymbol{\alpha}_{j;i}^*$ is the shortened skill profile of student i, which includes only the skills relevant for the mastery of j-th item. Furthermore, $K_j^* = \sum_{k=1}^{K} q_{jk}$ represents the number of

skills necessary for the mastery of item j, i.e. K_j^* is the sum of ones in the j-th row of the Q-matrix. For notational convenience and without loss of generality, let the first K_j^* skills be the ones required for item j. The skill profiles $\boldsymbol{\alpha}_i$ decompose into different reduced skill profiles depending on the item j (i.e. on the skills required for item j), which necessitates the notation of an additional item index in each skill profile. For example, if only the first two skills are required for item j, the skill profile of student i for item j reduces from $\boldsymbol{\alpha}_{j;i} = [\alpha_{j;i1}, \ldots, \alpha_{j;iK}]$ to $\boldsymbol{\alpha}_{j;i}^* = [\alpha_{j;i1}, \alpha_{j;iK^*}] = [\alpha_{j;i1}, \alpha_{j;i2}]$. If the second and sixth skill are required for item m, the notation of the i-th student's skill profile reduces to $\boldsymbol{\alpha}_{m;i}^* = [\alpha_{m;i1}, \alpha_{m;i2}]$. If in the G-DINA framework only first-order effects $\delta_{j;k}$ are modeled (i.e. all other parameters are defined to be zero), the resulting models are called G-DINA 1way models or additive CDMs (A-CDM). G-DINA models with first-order effects and second-order interaction effects $\delta_{j;kk'}$ are called G-DINA 2way, and so on.

In the Austrian educational test in math with $K = 8$ skills exactly 2 skills are assigned to each item. The full model is represented by a G-DINA 2way model with second-order interaction effects between the 2 skills. For the first item the skill profile $\boldsymbol{\alpha}_{1;i} = [\alpha_{1;i1}, \ldots, \alpha_{1;i8}]$ of student i reduces to the second and seventh element, because these elements correspond to the required skills. The reduced skill profile is denoted by $\boldsymbol{\alpha}_{1;i}^* = [\alpha_{1;i1}, \alpha_{1,iK^*}] = [\alpha_{1;i1}, \alpha_{1,i2}]$. The G-DINA 2way model provides the following probabilities:

$$P(X_{i1} = 1|\boldsymbol{\alpha}_{1;i}^*) = \begin{cases} \delta_{1;0} & \text{for } \boldsymbol{\alpha}_{1;i}^* = [0,0], \\ \delta_{1;0} + \delta_{1;3} & \text{for } \boldsymbol{\alpha}_{1;i}^* = [1,0], \\ \delta_{1;0} + \delta_{1;7} & \text{for } \boldsymbol{\alpha}_{1;i}^* = [0,1], \\ \delta_{1;0} + \delta_{1;3} + \delta_{1;7} + \delta_{1;37} & \text{for } \boldsymbol{\alpha}_{1;i}^* = [1,1]. \end{cases}$$

As can be seen, in the G-DINA 2way model the response probability increases with every skill relevant for the item (i.e. $q_{jk} = 1$) and being possessed (i.e. $\alpha_{ik} = 1$).

The DINA model is a special case of the G-DINA framework and can be deduced in two ways: If exactly one skill k is required to master each item (e.g. if only the content skills are considered in the educational testing in math), the response function of the G-DINA model with $K_j^* = 1$ for all items j simplifies to

$$P\left(X_{ij} = 1 \middle| \boldsymbol{\alpha}_{j;i}^*\right) = \delta_{j;0} + \delta_{j;k}\alpha_{j;ik}^*.$$

In terms of the DINA parameters, $g_j = \delta_{j;0}$ and $1 - s_j = \delta_{j;0} + \delta_{j;k}$. If several skills are

required to master the items, the DINA model is deduced from the G-DINA model by setting all parameters except $\delta_{j;0}$ and $\delta_{j;12...,K_j^*}$ to zero. Then

$$P\left(X_{ij}=1\middle|\boldsymbol{\alpha}_{j;i}^*\right)=\delta_{j;0}+\delta_{j;12...,K_j^*}\prod_{k=1}^{K_j^*}\alpha_{j;ik}^*,$$

with $g_j=\delta_{j;0}$ and $1-s_j=\delta_{j;0}+\delta_{j;12...,K_j^*}$.

Instead of the identity link implicitly used in Equation (1.2.2), other versions of the G-DINA model use logit or log links for modeling the conditional response probability. For the logit link, the G-DINA response function is defined as

$$\text{logit}\left[P\left(X_{ij}=1\middle|\boldsymbol{\alpha}_{j;i}^*\right)\right]=\nu_{j;0}+\sum_{k=1}^{K_j^*}\nu_{j;k}\alpha_{j;ik}^*+\sum_{k=1}^{K_j^*-1}\sum_{k'=k+1}^{K_j^*}\nu_{j;kk'}\alpha_{j;ik}^*\alpha_{j;ik'}^*+\ldots+\nu_{j;12...,K_j^*}\prod_{k=1}^{K_j^*}\alpha_{j;ik}^*. \tag{1.2.3}$$

Many common CDM models can be deduced and new model variants can be defined by using the different link functions, by in- and excluding parameters, or by setting constraints on parameters, which makes the G-DINA a general CDM framework. For more details see de la Torre (2011). For a comparison between the different link functions and parameter restrictions see also Section 2.3.4 of the present work.

1.2.4 Parameter estimation in DINA and G-DINA models

Parameter estimation of G-DINA models (i.e. also of DINA models, because they are included in the G-DINA framework) involves four parts and can be implemented using an expectation maximization (EM) algorithm (de la Torre, 2009). The process of parameter estimation is carried out in the same way if different items follow different condensation rules. The goal is the estimation of the item parameters $\boldsymbol{\delta}=[\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_J]$, with $\boldsymbol{\delta}_j=[\delta_{j;0},\delta_{j;1},\ldots,\delta_{K_j^*;1},\delta_{j;12},\ldots,\delta_{j;12...K_j^*}]$ being the item parameters for item j, the estimation of the skill class probabilities $P(\boldsymbol{\alpha}_l)$, $l=1,\ldots,2^K$, in the population and based on that the deduction of the skill mastery probabilities $P(\alpha_k)$, $k=1,\ldots,K$, and the estimation of the individual skill profiles $\boldsymbol{\alpha}_i$, $i=1,\ldots,I$.

(1) It is assumed that the responses $\mathbf{X}_i$ of student i to the different items are independent conditional on $\boldsymbol{\alpha}_i$ (local independence). Furthermore, it is assumed that examinees are mutually independent as well, because they are expected to represent a random sample of the population. Let $\boldsymbol{\alpha}=[\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_I]$ be the matrix of

skill profiles. Then the conditional likelihood of the observed data $\boldsymbol{X}$ is

$$\begin{aligned} L(\boldsymbol{X} \,|\, \boldsymbol{\alpha}, \boldsymbol{\delta}) &= \prod_{i=1}^{I} L(\boldsymbol{X}_i \,|\, \boldsymbol{\alpha}_i, \boldsymbol{\delta}) \\ &= \prod_{i=1}^{I} \prod_{j=1}^{J} P(X_{ij} = 1 | \boldsymbol{\alpha}_{j;i}^*)^{X_{ij}} \left[1 - P(X_{ij} = 1 | \boldsymbol{\alpha}_{j;i}^*)\right]^{1-X_{ij}} \end{aligned}$$

where $L(\boldsymbol{X}_i | \, \boldsymbol{\alpha}_i, \boldsymbol{\delta})$ is the likelihood contribution of $\mathbf{X}_i$ conditional on $\boldsymbol{\alpha}_i$ and $\boldsymbol{\delta}$ and $P(X_{ij} = 1 | \boldsymbol{\alpha}_{j;i}^*)$ is the probability of student i for correctly solving item j defined through the G-DINA framework (cf. Equation 1.2.2). Investigations in the context of item response models have shown that a joint estimation of item and ability parameters does not lead to consistent estimators (Baker & Kim, 2004, p. 108). Thus, in CDMs as well as for the item response models, the item parameters $\boldsymbol{\delta}$ and the skill class probabilities $\boldsymbol{\alpha}_l$, $l = 1, \ldots, 2^K$, are not jointly estimated, but parameter estimation is conducted with marginal maximum likelihood (MML) methods.

Up to here the probability $P(X_{ij} = 1 | \boldsymbol{\alpha}_{j;i}^*)$ is interpreted as probability of student i to master item j given her skills $\boldsymbol{\alpha}_{j;i}^*$. This notion facilitates the interpretation and the understanding of the models. But, strictly speaking, the students' skill profiles are unknown and our goal is to estimate them. Thus, more correctly, we should denote $P(X_{ij} = 1 | \boldsymbol{\alpha}_{j;l}^*)$ as probability of student i to master item j if she is classified in skill class l, $l = 1, \ldots, 2^K$. This notation is used for the following three steps.

(2) In preparation for the MML procedure, the probabilities $P(\boldsymbol{\alpha}_l)$, $l = 1, \ldots, 2^K$, are defined to follow a uniform distribution, i.e. $P(\boldsymbol{\alpha}_l) = \frac{1}{2^K}$, $l = 1, \ldots, 2^K$ are taken as starting values for the estimation. Because the distribution of the skill classes is discrete, taking the weighted sum of the conditional likelihood across the 2^K possible skill classes is equivalent to integrating the conditional likelihood over the distribution of the parameters in the continuous case. The marginalized likelihood

$$L(\boldsymbol{X} \,|\, \boldsymbol{\delta}) = \prod_{i=1}^{I} L(\boldsymbol{X}_i \,|\, \boldsymbol{\delta}) = \prod_{i=1}^{I} \sum_{l=1}^{2^K} L(\boldsymbol{X}_i \,|\, \boldsymbol{\alpha}_l, \boldsymbol{\delta}) \, P(\boldsymbol{\alpha}_l).$$

depends only on the item parameters $\boldsymbol{\delta}$ and no longer on the skill classes. Maximizing the marginal likelihood $L(\boldsymbol{X} | \boldsymbol{\delta})$ over $\boldsymbol{\delta}$ leads to the *item parameter estimates* $\hat{\boldsymbol{\delta}} = [\hat{\boldsymbol{\delta}}_1, \ldots, \hat{\boldsymbol{\delta}}_J]$.

(3) For each student with response pattern $\boldsymbol{X}_i$ the probabilities $P(\boldsymbol{\alpha}_l \,|\, \boldsymbol{X}_i)$ of being classified into skill class $\boldsymbol{\alpha}_l$ are calculated by multiple applications of Bayes' theorem

$$P(\boldsymbol{\alpha}_l \,|\, \boldsymbol{X}_i) = \frac{P(\boldsymbol{X}_i \,|\, \boldsymbol{\alpha}_l)\, P(\boldsymbol{\alpha}_l)}{\sum_{l=1}^{2^K} P(\boldsymbol{X}_i \,|\, \boldsymbol{\alpha}_l)\, P(\boldsymbol{\alpha}_l)}, \quad l = 1, \ldots, 2^K.$$

By applying the formula of total probability, the so called *distribution of the skill class probabilities* in the population is calculated

$$P(\boldsymbol{\alpha}_l) = \sum_{i=1}^{I} P(\boldsymbol{\alpha}_l \,|\, \boldsymbol{X}_i)\, P(\boldsymbol{X}_i), \quad l = 1, \ldots, 2^K$$

and the *skill mastery probabilities* are defined as

$$P(\alpha_k) = \sum_{l:\alpha_{lk}=1} P(\boldsymbol{\alpha}_l), \quad k = 1, \ldots, K.$$

(4) Based on the probabilities $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)$, $l = 1, \ldots, 2^K$, $i = 1, \ldots, I$, the *individual student classifications* or *individual skill profiles* can be deduced with three methods: Firstly, according to maximum a priori (MAP) classification, the largest value of $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)$ for all $l = 1, \ldots, 2^K$ gives the skill class into which student i is classified:

$$\hat{\boldsymbol{\alpha}}_{i;MAP} = \arg\max_l \{P(\boldsymbol{\alpha}_l \,|\, \boldsymbol{X}_i)\}.$$

Secondly, an individual classification of student i based on maximum likelihood estimation (MLE) is obtained by maximizing

$$\hat{\boldsymbol{\alpha}}_{i;MLE} = \arg\max_l \{P(\boldsymbol{X}_i \,|\, \boldsymbol{\alpha}_l)\}.$$

Thirdly, for a classification based on expected a posterior (EAP) the marginal skill probabilities $P(\alpha_k \,|\, \boldsymbol{X}_i)$ of student i for mastering skill k are computed as the sum of all $P(\boldsymbol{\alpha}_l \,|\, \boldsymbol{X}_i)$ corresponding to mastery of skill k (i.e., having a 1 as the k-th element)

$$P(\alpha_k \,|\, \boldsymbol{X}_i) = \sum_{l:\alpha_{lk}=1} P(\boldsymbol{\alpha}_l \,|\, \boldsymbol{X}_i), \quad k = 1, \ldots, K.$$

Then, the i-th student's EAP skill profile is estimated by

$$\tilde{\boldsymbol{\alpha}}_{i;EAP} = [\, P(\alpha_1 \,|\, \boldsymbol{X}_i), \ldots, P(\alpha_K \,|\, \boldsymbol{X}_i)\,].$$

For deducing the simplified dichotomous skill profile $\hat{\boldsymbol{\alpha}}_{i;EAP}$ of student i, each marginal skill mastery probability $P(\alpha_k|\boldsymbol{X}_i)$, $k = 1, \ldots, K$, smaller than 0.5 is set to 0, whereas each one larger or equal to 0.5 is set to 1. For a comparison among MAP, MLE and EAP classification methods see Huebner & Wang (2011).

Standard errors of the estimated item parameters $\hat{\boldsymbol{\delta}}$ are computed from the Fisher-Information matrix

$$I(\boldsymbol{\delta}) = -\mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{X})}{\partial^2 \boldsymbol{\delta}}\right],$$

where $l(\boldsymbol{X}) = \log\prod_{i=1}^{I} L(\boldsymbol{X}_i) = \sum_{i=1}^{I} \log L(\boldsymbol{X}_i)$ is the marginal log-likelihood of the data. Instead of computing the expectation, the information matrix is approximated by evaluating it at $\hat{\boldsymbol{\beta}}$ using the observed $\boldsymbol{X}$, thus resulting in $I(\hat{\boldsymbol{\delta}})$. Finally, the inverse $I^{-1}(\hat{\boldsymbol{\delta}})$ provides an approximation of $\mathrm{Cov}(\hat{\boldsymbol{\delta}})$, and the square roots of its diagonal elements represent the standard errors $\mathrm{SE}(\hat{\boldsymbol{\delta}})$.

1.2.5 Some connections between G-DINA, LCDM and GDM

In this subsection some similarities and differences between the three general CDM frameworks G-DINA (de la Torre, 2011), LCDM (Henson, Templin & Willse, 2009) and GDM (von Davier, 2008) are clarified. It will be shown, that in the case of dichotomous data and skills and under the usage of the logit link all frameworks are equivalent concerning their representation of compensatory models.

The *Log-linear Cognitive Diagnosis Models* (LCDM; Henson et al., 2009) for dichotomous data and dichotomous skills is, as inherent in it's name, based on log-linear models. Many common CDMs can be subsumed under this framework, and it also allows for defining new CDMs by setting model constraints "in between" the constraints of the common CDMs. In the LCDM framework the i-th student's probability of correctly solving item j conditional on her skill profile $\boldsymbol{\alpha}_i$ is given by

$$P(X_{ij} = 1|\boldsymbol{\alpha}_i) = \frac{exp\left(\lambda_{j,0} + \boldsymbol{\lambda}_j' \cdot h(\boldsymbol{\alpha}_i, \mathbf{q}_j)\right)}{1 + exp\left(\lambda_{j,0} + \boldsymbol{\lambda}_j' \cdot h(\boldsymbol{\alpha}_i, \mathbf{q}_j)\right)}.$$

In this notation $\lambda_{j,0}$ is an intercept parameter for item j and $\boldsymbol{\lambda}_j$ is a $(2^K - 1)$ dimensional vector including the parameters for K main effects and all (up to Kway) interaction effects between the K skills, thus

$$\boldsymbol{\lambda_j} = [\underbrace{\lambda_{j,1}, \ldots, \lambda_{j,K}}_{\text{main effects}}, \underbrace{\lambda_{j,12}, \ldots, \lambda_{j,1K}, \ldots, \lambda_{j,(K-1)K}}_{\text{2way interaction effects}}, \ldots, \underbrace{\lambda_{j,1\ldots K}}_{\text{Kway interaction effect}}].$$

Furthermore, $h(\boldsymbol{\alpha}_i, \mathbf{q}_j)$ is a vector of size $2^K - 1$ with its components being linear combinations of $\boldsymbol{\alpha}_i$ and $\mathbf{q}_j$, where $\mathbf{q}_j$ again denotes the j-th row of the Q-matrix. By defining $\boldsymbol{\lambda_j}$ and h the condensation rule for item j can be specified, where h gives the level of compensability.

For example, one may define $h(\boldsymbol{\alpha}_i, \boldsymbol{q}_j)$ by

$$h(\boldsymbol{\alpha}_i, \boldsymbol{q}_j) = [\underbrace{\alpha_{i1} \cdot q_{j1}, \ldots, \alpha_{iK} \cdot q_{jK}}_{\text{for main effects}}, \underbrace{\alpha_{i1}q_{j1} \cdot \alpha_{i2}q_{j2}, \ldots, \alpha_{i(K-1)}q_{j(K-1)} \cdot \alpha_{iK}q_{jK}}_{\text{for 2way interaction}}, \ldots, \underbrace{\prod_{k=1}^{K} \alpha_{ik}q_{jk}}_{\text{for Kway interaction}}].$$

Then it holds

$$\lambda_{j,0} + \boldsymbol{\lambda}_j' \cdot h(\boldsymbol{\alpha}_i, \mathbf{q}_j) = \lambda_{j,0} + \sum_{k=1}^{K} \lambda_{j,k} \cdot \alpha_{ik}q_{jk} + \sum_{k=1}^{K} \sum_{k'=k+1}^{K} \lambda_{j,kk'} \cdot \alpha_{ik}q_{jk} \cdot \alpha_{ik'}q_{jk'} + \ldots$$

with $\lambda_{j,0}$ being the intercept term for item j, $\lambda_{j,k}$ being the main effects for item j with respect to skill k, and $\lambda_{j,kk'}$ being the two way interaction effects for item j with respect to skills k and k'. In defining $h(\boldsymbol{\alpha}_i, \boldsymbol{q}_j)$ this way, the similarity to the G-DINA framework, which only includes the skills relevant for the mastery of item j (i.e. $q_{jk} = 1$), can already be seen.

The following example illustrates how DINA parameters can be deduced from the LCDM framework: Assume a DINA model with $K = 2$ skills and an item j for which both skills are required. All main effects $\lambda_{j,k}$, $k = 1, 2$, are defined to be zero (because a DINA model assumes that the items are only mastered in case of possession of all relevant skills) and the interaction effect $\lambda_{j,12}$ with respect to skills 1 and 2 has to be estimated. The response probability for a student possessing no skill is

$$\begin{aligned} P(X_{ij} = 1 | \boldsymbol{\alpha}_i = [0,0]) &= \frac{exp\left(\lambda_{j,0} + [0,0,\lambda_{j,12}]'[\alpha_{i1}q_{j1}, \alpha_{i1}q_{j1}, \alpha_{i1}q_{j1} \cdot \alpha_{i2}q_{j2}]\right)}{1 + exp\left(\lambda_{j,0} + [0,0,\lambda_{j,12}]'[\alpha_{i1}q_{j1}, \alpha_{i1}q_{j1}, \alpha_{i1}q_{j1} \cdot \alpha_{i2}q_{j2}]\right)} \\ &= \frac{exp\left(\lambda_{j,0} + [0,0,\lambda_{j,12}]'[0,0,0]\right)}{1 + exp\left(\lambda_{j,0} + [0,0,\lambda_{j,12}]'[0,0,0]\right)} \\ &= \frac{exp\left(\lambda_{j,0}\right)}{1 + exp\left(\lambda_{j,0}\right)} \\ &=: g_j. \end{aligned}$$

Analogously, if both skills are possessed

$$\begin{aligned} P(X_{ij}=1|\boldsymbol{\alpha}_i=[1,1]) &= \frac{exp\left(\lambda_{j,0}+[0,0,\lambda_{j,12}]'[\alpha_{i1}q_{j1},\alpha_{i1}q_{j1},\alpha_{i1}q_{j1}\cdot\alpha_{i2}q_{j2}]\right)}{1+exp\left(\lambda_{j,0}+[0,0,\lambda_{j,12}]'[\alpha_{i1}q_{j1},\alpha_{i1}q_{j1},\alpha_{i1}q_{j1}\cdot\alpha_{i2}q_{j2}]\right)} \\ &= \frac{exp\left(\lambda_{j,0}+[0,0,\lambda_{j,12}]'[1,1,1]\right)}{1+exp\left(\lambda_{j,0}+[0,0,\lambda_{j,12}]'[1,1,1]\right)} \\ &= \frac{exp\left(\lambda_{j,0}+\lambda_{j,12}\right)}{1+exp\left(\lambda_{j,0}+\lambda_{j,12}\right)} \\ &=: 1-s_j. \end{aligned}$$

Thus students with skill profiles $[0,0]$, $[1,0]$ or $[0,1]$ have a probability of g_j to solve item j and only students in possession of both skills master the item with probability $1-s_j$. This reflects the two-probability constraint of the DINA model. In summary, the DINA model can be deduced from the LCDM by defining $g_j := \text{logit}(\lambda_{j,0})$ and $s_j := 1-\text{logit}(\lambda_{j,0}+\lambda_{j,1\ldots k^*})$, where k^* denotes the number of required skills for item j.

The main ideas of the LCDM framework originate from the *General Diagnostic Model* (GDM; von Davier, 2008). The GDM framework includes a general log-linear class of models for polytomous data, which also allows polytomous skills (i.e. skills with more than the two proficiency levels of mastery and non-mastery) as well as continuous skills. An instance of this class, the GDM for partial credit data, contains many well-known models, such as univariate and multivariate extensions of the Rasch model (Rasch, 1960), the two parameter logistic item response theory model (Birnbaum, 1968), the generalized partial credit model (Muraki, 1992), as well as a variety of skill profile approaches like latent class models and the compensatory version of the RUM model (Hartz, 2002).

Let the response data be polytomous with $x_{ij} \in \{0,1,\ldots,m_j\}$, let the skill levels be polytomous and user-specified with $\alpha_{ik} \in \{s_{k1},\ldots,s_{ko},\ldots,s_{kO_k}\}$ and let the Q-matrix be a $J \times K$ matrix with real-valued entries q_{jk}. For each non-zero response category $x \in \{1,\ldots,m_j\}$, the class of general diagnostic models is given by

$$P(X_{ij}=x|\boldsymbol{\alpha}_i) = \frac{exp\left(\beta_{xjg}+\boldsymbol{\gamma}'_{xjg}h(\boldsymbol{q}_j,\boldsymbol{\alpha}_i)\right)}{1+\sum_{y=1}^{m_j} exp\left(\beta_{yjg}+\boldsymbol{\gamma}'_{yjg}h(\boldsymbol{q}_j,\boldsymbol{\alpha}_i)\right)},$$

where β_{xjg} are real-valued difficulty parameters and $\boldsymbol{\gamma}_{xjg} = [\gamma_{xjg1},\ldots,\gamma_{xjgK}]$ is a K-dimensional slope parameter. The index g is a population indicator that allows formulating the GDM as multiple group model.

If we reduce the GDM to $g = 1$ group, dichotomous data and dichotomous skills, then

$$P(X_{ij} = 1|\boldsymbol{\alpha}_i) = \frac{exp\left(\beta_j + \boldsymbol{\gamma}_j' h(\boldsymbol{q}_j, \boldsymbol{\alpha}_i)\right)}{1 + exp\left(\beta_j + \boldsymbol{\gamma}_j' h(\boldsymbol{q}_j, \boldsymbol{\alpha}_i)\right)}.$$

This formulation is extremely similar to the LCDM. For the reduced case above, the difference between the GDM and the LCDM is the following: In the LCDM $\boldsymbol{\lambda}_j$ is a $(2^K - 1)$ dimensional vector including the parameters for K main effects and all interaction effects between the K skills. On the contrary, in the GDM $\boldsymbol{\gamma}_j$ is a K dimensional vector which only includes parameters for the main effects. In the case of the GDM, the DINA model can not be deduced because it is not possible (as there is no suitable parameter) but required to estimate the interaction between all required skills, while setting all other effects to zero.

Table 1.2.1 summarizes the comparison between the three frameworks G-DINA, LCDM and GDM. As can be seen, the GDM framework is the most flexible one concerning different data and skill formats. On the other hand, the GDM framework does not include some specific model types, as for example the non-compensatory DINA model or the additive CDM (de la Torre, 2011), which is defined with an identity link. A strength of the G-DINA framework is the application of different link functions, allowing for a definition of many core CDMs. However, the G-DINA framework does not enable the user to specify a function $h(\boldsymbol{q_j}, \boldsymbol{\alpha_i})$ for defining the influence of (non-)possessed and (non-)required skills to the items' response probabilities. The LCDM framework for dichotomous data and skills can be regarded as a mixture of both frameworks: It allows for a user-defined specification of $h(\boldsymbol{q_j}, \boldsymbol{\alpha_i})$ and includes non-compensatory models.

For the purpose of working with dichotomous data and skills (as done in the present work), both the G-DINA and the LCDM would be appropriate. However, because the most commonly used DINA model is structurally related to the G-DINA, the latter will be considered further. Additionally, the free definition of $h(\boldsymbol{q_j}, \boldsymbol{\alpha_i})$ in the LCDM seems to be most useful for polytomous data. Here, h may be defined in such a way that it defines a sufficient level for skill k on item j. Then a higher skill level will not increase the probability for mastering item j, whereas a lower skill level results in a lower probability for mastering item j.

		G-DINA	LCDM	GDM
data	dichotomous	✓	✓	✓
	polytomous			✓
skills	dichotomous	✓	✓	✓
	polytomous			✓
	continuous			✓
Q-matrix	dichotomous	✓	✓	✓
	real-valued			✓
model	compensatory	✓	✓	✓
	non-compensatory	✓	✓	
link	identity	✓		
	log	✓		
	logit	✓	✓	✓
$h(\boldsymbol{q_j}, \boldsymbol{\alpha_i})$	$q_{jk}\alpha_{ik}$	✓	✓	✓
	user-defined		✓	✓

Table 1.2.1: Comparison between G-DINA, LCDM and GDM frameworks.

1.2.6 Related approaches

In this subsection CDMs are linked to some related approaches: First, it is shown how CDMs are deduced from latent class models (LCM). Second, the connections to mathematical psychology, more precisely to knowledge space theory (KST) are shown, and third, the striking difference between item response theory (IRT) and factor analysis (FA) as opposed to CDMs is shown. All three approaches may be seen as the basis for the the development of CDMs: In LCM and KST students are also classified into groups with respect to their response behavior, but the basics of both approaches do not consider skills underlying the items. IRT and FA models are much more prominent for identifying students abilities and thus they are more often applied than CDM models. Finally, in this subsection the link between CDMs and the Rule Space Method (RSM) is explained, as RSM can be seen as a related approach, which gains some attention in recent literature.

Latent Class Analysis The goal of a Latent Class Analysis (LCA; Lazarsfeld, 1950) is to identify unobservable latent classes of students which have similar properties in their response behavior. More precisely, each student i, $i = 1, \ldots, I$, is classified into one latent class l, $l = 1, \ldots, L$, according to her response pattern $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{iJ}]$. The method is conducted in three steps:

(1) The conditional probabilities $P(\boldsymbol{x}_i | \, l)$ of observing the i-th student's response pattern in class l, $l = 1, \ldots, L$, are determined:

$$P(\boldsymbol{x}_i \,|\, l) = \prod_{j=1}^{J} {p_{jl}}^{x_{ij}} (1 - p_{jl})^{(1-x_{ij})}, \quad l = 1, \ldots, L. \tag{1.2.4}$$

Here p_{jl} are the unknown probabilities of students in class l who correctly respond item j, $j = 1, \ldots, J$.

(2) The marginal probability of observing a response pattern $\boldsymbol{x}_i$ is calculated as

$$P(\boldsymbol{x}_i) = \underbrace{\sum_{l=1}^{L} \pi_l}_{\text{stuctural model}} \underbrace{\prod_{j=1}^{J} {p_{jl}}^{x_{ij}} (1 - p_{jl})^{(1-x_{ij})}}_{\text{measurement model}}, \tag{1.2.5}$$

with π_l being the unknown relative frequency of class l, $l = 1, \ldots, L$, with

$$\sum_{l=1}^{L} \pi_l = 1.$$

(3) Via Bayes' Theorem

$$P(l \,|\, \boldsymbol{x}_i) = \frac{\pi_l \cdot P(\boldsymbol{x}_i \,|\, l)}{P(\boldsymbol{x}_i)}$$

where $P(\boldsymbol{x}_i | l)$ is the probability of observing response pattern $\boldsymbol{x}_i$ in class l. Each student is classified into the class for which $P(l | \boldsymbol{x}_i)$, $l = 1, \ldots, L$, is maximal.

The parameters π_l and p_{jl}, $l = 1, \ldots, L$, $j = 1, \ldots, J$, are determined with the help of an EM-algorithm (e.g. Formann, 1978; Goodman, 1979). Because both parameters are unknown, they are set to arbitrary starting values at the beginning of the algorithm, which fulfill $\pi_l, p_{jl} \in (0, 1)$, $l = 1, \ldots, L$, $j = 1, \ldots, J$, and $\sum_{l=1}^{L} \pi_l = 1$. In each step of the algorithm the parameter values are adapted until a stopping criterion is satisfied. Note that the number of classes L is not an estimable model parameter but has to be chosen in advance (McLachlan & Peel, 2000).

p_{jl}	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	π_l
Class 1	0.83	0.77	0.90	0.56	0.24	0.43	0.43
Class 2	0.33	0.28	0.45	0.75	0.81	0.69	0.28
Class 3	0.90	0.86	0.59	0.59	0.77	0.40	0.17
Class 4	0.32	0.22	0.09	0.19	0.31	0.29	0.12

Table 1.2.2: Probabilities p_{jl} for students in class l, $l = 1, \ldots, 4$, to correctly answer item j, $j = 1, \ldots, 6$, and class sizes π_l, $l = 1, \ldots, 4$.

Table 1.2.2 gives a possible solution for the π_l and p_{lj} in a LCM with 4 classes and 6 Items. The estimated values of p_{jl} comprise information about the response behavior of students in each class. For example, students in class 1 have high probabilities for mastering items 1 and 2 and low probabilities for solving items 5 and 6. On the opposite, students in class 2 have low probabilities for solving items 1 and 2 but high probabilities for answering items 4 and 5.

In CDMs each student i, $i = 1, \ldots, I$, is classified into a latent class $\boldsymbol{\alpha}_l$, $l = 1, \ldots, 2^K$, based on her response pattern $\mathbf{x}_i$. Here, the latent classes indicate presence or absence of K skills underlying the items. Analogously to LCA, in the CDM framework the probability of observing a response pattern $\boldsymbol{x}_i$ is defined as

$$P(\boldsymbol{x}_i) = \underbrace{\sum_{l=1}^{2^K} P(\boldsymbol{\alpha}_l)}_{\text{stuctural model}} \underbrace{\prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_l)^{x_{ij}} (1 - P_j(\boldsymbol{\alpha}_l))^{(1-x_{ij})}}_{\text{measurement model}} \tag{1.2.6}$$

with $P_j(\boldsymbol{\alpha}_l)$ being the conditional probability of correctly answering item j in skill class $\boldsymbol{\alpha}_l$, $P(\boldsymbol{\alpha}_l)$ being the probability of skill class $\boldsymbol{\alpha}_l$ and $\sum_{l=1}^{2^K} P(\boldsymbol{\alpha}_l) = 1$.

In comparing LCMs (cf. Equation 1.2.6) and CDMs (cf. Equation 1.2.5) the following statements hold:

(1) *CDMs are confirmatory LCA models with 2^K classes.* In CDMs the examinees are classified in 2^K classes according to their (non-)possession of K skills. Because the number and the structure of these skills (i.e. the Q-matrix) is defined prior to to parameter estimation, the model has a confirmatory character.

(2) *Measurement model: CDMs are restricted LCA models.* The restriction from LCA to CDMs evolves because, contrary to p_{lj}, $P_j(\boldsymbol{\alpha}_l)$ is not estimated independently for each class l and each item j. CDMs demand that students in skill classes including

all (or more than the) required skills for item j exhibit an equal probability $p_j(\boldsymbol{\alpha}_l)$ for mastering the item. More sophisticated CDMs, as for example the G-DINA, also demand that the probability of mastering item j increases in equal steps each time a specific required skill is possessed. For example, if item j requires skills 1 and 2, then the response probabilty equally increases from [1001] to [1101] and from [1000] to [1100].

Consider an example with $K = 4$ skills and $J = 3$ items in which a DINA model is developed by restricting the p_{jl} parameters: According to the Q-matrix

$$\boldsymbol{Q} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} .$$

students require skills 1 and 2 for mastering item 1. That is, in terms of a DINA condensation rule, item 1 is mastered with the same probability $1 - s_1$ in all skill classes l in which skills 1 and 2 are possessed, and is mastered with probability g_j in all skill classes in which at least one of the skills 1 or 2 is missing. Thus

$$p_{1l} = \begin{cases} 1 - s_1 & \text{for all } \boldsymbol{\alpha}_l \text{ with } \alpha_{l1} = 1 \text{ and } \alpha_{l2} = 1 \\ g_j & \text{for all } \boldsymbol{\alpha}_l \text{ with } \alpha_{l1} = 0 \text{ or } \alpha_{l2} = 0 \ . \end{cases}$$

Restrictions for the items 2 and 3 are defined analogously. Table 1.2.3 yields a possible parameter constellation for this example. Note that this example is for illustrational purposes only, as in practice a CDM with only 3 items but 4 skills would not be estimable.

(3) *Structure model: LCA structural parameters can be used for modeling skill hierarchies in CDMs.* By structuring the LCA parameters π_l, that is the CDM probabilities $P(\boldsymbol{\alpha}_l)$, ambiguous skill classes can be avoided (see Chapter 3 of the present work or Groß & George, 2012). Further, attribute hierarchies can be defined (e.g. Groß & George, 2013; Leighton & Gierl, 2007; Tatsuoka, Varadi & Jaeger, 2013).

The development from latent class models to CDMs has begun with the mastery model by Macready & Dayton (1977) and was extended to more items, more classes and more response occasions by Macready & Dayton (1980) and Dayton & Macready (1983). A first core extension came from Haertel (1989), which leads to a model that later has been called the DINA model (Junker & Sijtsma, 2001). Note that based on the LCA

p_{jl}	skill class $\boldsymbol{\alpha_l}$	Item 1	Item 2	Item 3
Class 1	[0,0,0,0]	0.23	0.16	0.04
Class 2	[1,0,0,0]	0.23	0.16	0.04
Class 3	[0,1,0,0]	0.23	0.16	0.04
Class 4	[0,0,1,0]	0.23	0.16	0.04
Class 5	[0,0,0,1]	0.23	0.81	0.04
Class 6	[1,1,0,0]	0.75	0.16	0.04
Class 7	[1,0,1,0]	0.23	0.16	0.04
Class 8	[1,0,0,1]	0.23	0.81	0.04
Class 9	[0,1,1,0]	0.23	0.16	0.04
Class 10	[0,1,0,1]	0.23	0.81	0.04
Class 11	[0,0,1,1]	0.23	0.81	0.04
Class 12	[1,1,1,0]	0.75	0.16	0.04
Class 13	[1,1,0,1]	0.75	0.81	0.04
Class 14	[1,0,1,1]	0.23	0.81	0.04
Class 15	[0,1,1,1]	0.23	0.81	0.78
Class 16	[1,1,1,1]	0.75	0.81	0.78

Table 1.2.3: LCA probabilities p_{lj} restricted according to a DINA condensation rule in which item 1 is mastered in possession of skills 1 and 2, item 2 is mastered in possession of skill 4 and item 3 is mastered in possession of skills 2, 3, and 4.

framework with parameters p_{jl} and π_l, $j = 1, \ldots, J$, $l = 1, \ldots, L$ it can be explained that for example the model parameters of a DINA model are composed of skill class probability parameters $p(\boldsymbol{\alpha}_l)$, $l = 1, \ldots, 2^K$, and the item parameters g_j and s_j, $j = 1, \ldots, J$.

Knowledge Space Theory Knowledge Space Theory (KST) is a set and order theoretical approach for describing how respondents acquire and retain knowledge in a knowledge domain, with the domain being characterized by a set of items on which the students are tested (cf. Albert & Lukas, 1999; Doignon & Falmagne, 1999; Falmagne & Doignon, 2010). On the one hand, KST allows for representing the knowledge state

of an individual learner, that is her actual status of knowledge. On the other hand, a main goal of KST is to provide methods to significantly reduce the number of possible knowledge states and learning histories by introducing a hierarchy among the items. The a-priori information for the derivation of these hierarchies is based on qualitative methods including psychological theories and principles developed by domain experts. Two aspects observe attention: Firstly, the basic ideas of CDMs and KST, that is classifying students in learning states and including hierarchies between these states are the same. However, the main difference between the two approaches is that the classification in basic KST approaches is done on the item and not on the skill level. Only extended KST approaches, as for example the research of Düntsch & Gediga (1995) considers skills. A second difference is that original KST is a qualitative, discrete mathematical approach. That is, in the basic approaches of KST no probabilities for mastering the items are assumed, instead an item is deterministically mastered or not. Only extensions of KST lead to probabilistic models, such as the basic local independence model (Doignon & Falmagne, 1999, Chapter 7) and the newer approach of learning spaces (Falmagne & Doignon, 2010). For a more detailed analysis of the connections between the models for describing the knowledge states (i.e. skill profiles) in KST and CDMs, see for example George (2010), George & Ünlü (2011) and Schrepp (2005).

IRT, M-IRT, and CFA As stated earlier, CDMs are discrete latent variable models, which provide direct statistically driven classifications of the respondents into disjunctive, prior to estimation defined skill classes. In contrast, item response theory models (IRT; e.g. Van der Linden & Hambleton, 1997) and multidimensional item response theory (M-IRT; e.g. de Ayala, 2009) models or confirmatory factor analysis (M-CFA; e.g. McDonald, 1999) models contain continuous latent variables. When data is scaled with (M-)IRT or CFA methods the classification of students is only possible through post-hoc procedures such as standard settings (e.g. Cizek, Bunch & Konns, 2004). This classifications are based on consensual cut-scores on the continuous scales. For an extensive comparison of the CDM DINA model and the one-dimensional IRT Rasch model (Rasch, 1960) see Chapter 4 of the present work. For a discussion about the difference between CDM and M-IRT models see Section 5.3.4.

Rule Space Method The Rule Space Method is a method to classify students in clusters according to their response patterns and their possessed skills. The Rule Space Method was introduced by Tatsuoka (e.g. Tatsuoka, 2009, 1995, 1983) in two steps:

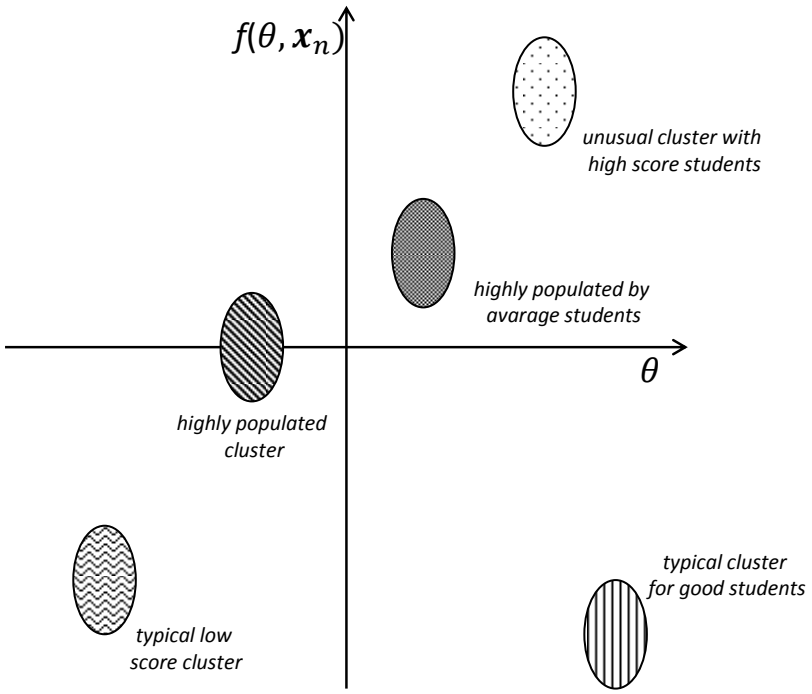

Figure 1.2.4: Interpretation of clusters with standardized caution indices in the rule space (Figure from Tatsuoka (2009), page 192).

(1) A traditional unidimensional IRT model is fitted to the response data. That is, IRT functions $P_j(\theta)$ for each item j, $j = 1, \ldots, J$, and an average IRT function $T(\theta)$ are computed, with $T(\theta) = \frac{1}{J} \sum_{j=1}^{J} P_j(\theta)$. In this step no information about the skills required to master the items is used.

(2) The students are clustered into groups according to their achievement and according to the unusualness of their response patterns. Therefore Tatsuoka (1983) introduced the so called two dimensional rule space $\{(\boldsymbol{\theta}, f(\boldsymbol{\theta}, \mathbf{X})\}$, with the first dimension build up by the person parameters $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_I]$ of the beforehand computed IRT model and the second dimension defined through the so called caution indices $f(\theta_i, \boldsymbol{x}_i)$, $i = 1, \ldots, I$. The caution indices measure the unusualness of the i-th student's response pattern $\mathbf{x}_i$:

$$f(\theta_i, \boldsymbol{x}_i) = -\sum_{j=1}^{J} (P_j(\theta_i) - T(\theta_i)) x_{ij} + \sum_{j=1}^{J} P_j(\theta_i)(P_j(\theta_i) - T(\theta_i)).$$

Figure 1.2.4 gives some interpretations for different locations of student clusters in the rule space.

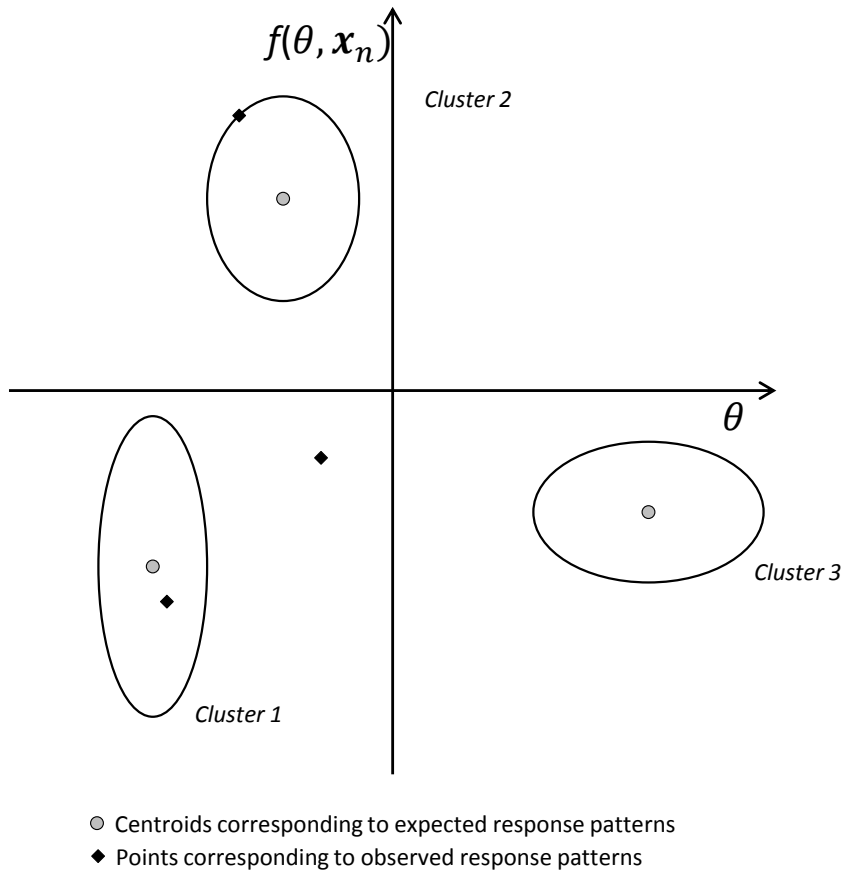

Figure 1.2.5: Illustration of a rule space with unstandardized caution indices (Figure from Rupp et al. (2010), page 104).

Step 2 is conducted as follows: The rule space includes a two dimensional coordinate point $(\theta_i, f(\theta_i, \boldsymbol{x}_i))$ for each observed response pattern $\boldsymbol{x}_i$, $i = 1, \ldots, I$. Additionally it contains two dimensional coordinate points for each expected response pattern, with the expected response patterns being the deterministic responses (i.e. responses without errors) of students within certain skill classes. The variances corresponding to the person parameters and caution indices are displayed as centroids around the coordinate points of the expected response patterns. Centroids with standardized variances are called standardized caution indices. For classification purposes, from each observed response pattern the Mahalanobis distance to all expected response patterns is computed and, finally, the observed response pattern (i.e. the respective student) is classified into the cluster of the expected response pattern with shortest distance. For an illustration see Figure 1.2.5.

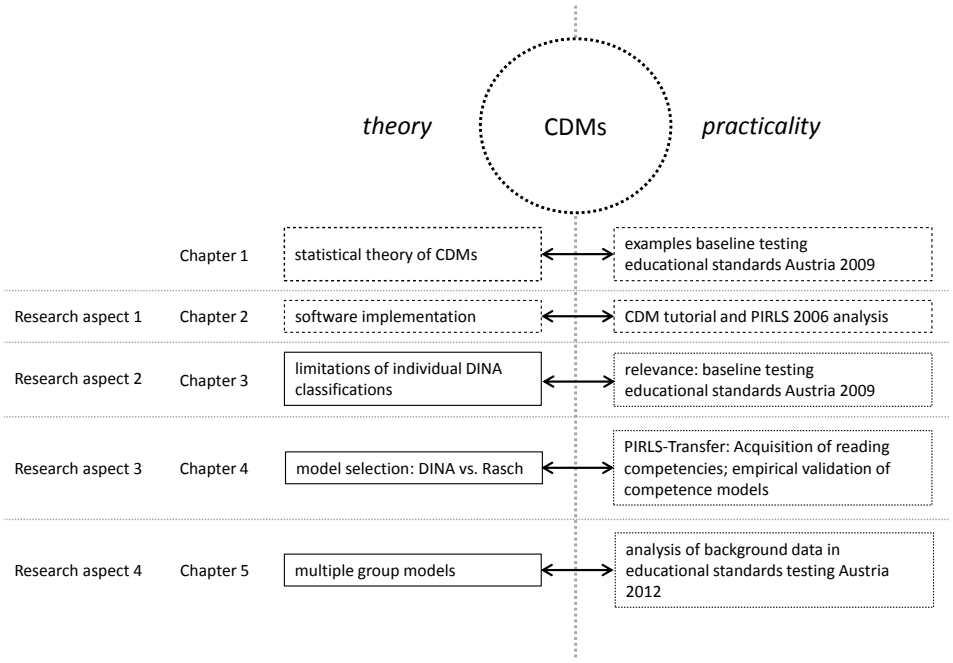

Figure 1.3.6: Structure of present work: Blending theory with practicality.

The first dimension of the rule space constitutes the main difference between the RSM and the CDM framework, as in CDMs no unidimensional ability scores are involved. However, the basic idea for classification in clusters of expected response patterns (i.e. latent response patterns) in the second dimension is used in both approaches. Further research of Tatsuoka shows how to reduce the number of expected response patterns by taking skill hierarchies into account. For more information see for example Gierl, Leighton & Hunka (2000) or Tatsuoka (2009) and also Groß & George (2013) as an application for CDMs.

1.3 Aspects of research

While recent CDM research mostly splits up into theoretical (i.e. statistical) based (e.g. de la Torre, 2011; von Davier & Yamamoto, 2004) and application-oriented parts (e.g. DeCarlo, 2011; Park & Lee, 2011) the present work blends both parts, compare Figure 1.3.6. This work is divided into four aspects of research: Software implementation, description and solution of statistical limitations in individual DINA model classifications, model selection and analysis of background data.

1.3.1 Aspect 1: Software implementation

Working with CDMs requires an adequate software program for estimating the model parameters. Currently a handful of different programs by various authors supports CDM parameter estimation, for example the G-DINA procedure in Ox (Doornik, 2002) by de la Torre, the LCDM framework in SAS (SAS Institute Inc., 2007) and a function by Templin, Henson, Douglas and Hoffman in Mplus (Muthén & Muthén, 2010), or the mdltm stand alone program by Von Davier. All these programs differ in mainly two aspects: In their programing framework (and thus in their availability and price) and in the model frameworks implemented. In recent years the programming framework R (R Core Team, 2013) has become more and more important in social sciences (Alexandrowicz, 2012; Kubinger, Rasch & Yanagida, 2011), as it is freely available and very flexible. R supports many ready to use methods, but beyond this the user has the possibility to code up any method that is needed. Nevertheless, an implementation of CDM algorithms in R was missing so far and thus became the first research aspect of the present work. The R package CDM (George, Kiefer, Robitzsch, Groß & Ünlü, 2013) is introduced in Chapter 2 in a kind of tutorial. It is illustrated and discussed using PIRLS 2006 data. Furthermore, a review of existing software for estimating CDMs is given.

1.3.2 Aspect 2: Limitations of individual DINA classifications

One of the most important results obtained from CDMs is the set of individual skill profiles, because they can easily be used as empirical base for feedback and further instructions. We expect that the CDM algorithm classifies the students in their true (but unknown) skill profiles. Consider again the baseline test of educational standards in math with the four underlying skills measures, functions, geometry and statistics: It is of course expected that the estimated skill profile of a student i with true skill profile $\boldsymbol{\alpha}_i = [0, 0, 0, 1]$ is $\hat{\boldsymbol{\alpha}}_i = [0, 0, 0, 1]$, i.e. she possesses only the skill statistics and is actually predicted to possess only statistics.

A basic prerequisite for achieving an accurate classification is that each student is assigned to a unique skill profile based on her manifest response $\boldsymbol{X}_i$, the Q-matrix and the items' condensations rules. For example, if student m solves the 36 items of the baseline test with

$$\boldsymbol{X}_m = [1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0]$$

she is desired to be uniquely classified into a skill profile

$$\boldsymbol{\alpha}_m = [1, 0, 1, 1].$$

On the contrary, an ambiguous classification of student m in classes

$$\boldsymbol{\alpha}_{m_1} = [1, 0, 1, 1] \text{ or } \boldsymbol{\alpha}_{m_2} = [1, 0, 0, 1]$$

is undesired, as such an ambiguous classification may result in incorrect feedback or improper recommendations for supporting the student.

In Chapter 3 of the present work it is shown that DINA models do not necessarily lead to unique student classifications. For the case of given data and Q-matrix a statistical solution is introduced. Implications for the interpretation of the model are described and illustrated by data of the Austrian test in educational standards in math. Furthermore it is discussed that the problem of ambiguous skill classifications can be avoided in the test construction phase by using an appropriate Q-matrix.

1.3.3 Aspect 3: Model selection

Which of various statistical models should be fitted to the data can not only be evaluated based on the absolute model fit but also based on several other measures as for example the relative model fit, item and person fit or classification criteria. Another aspect in the selection of a statistical model for the description of the manifest response data are model inherent presuppositions: For example a Rasch model (Rasch, 1960) includes the assumption that the modeled competencies are hierarchically ordered. On the contrary, an unrestricted CDM DINA model assumes non-ordered parallel skills. This difference between various statistical model approaches can be exploited if the order between the skills or the number of skills underlying the data is *not* known: First different statistical models can be build which mirror different theoretical assumptions about the connections between and the number of the skills. In a second step, by empirically comparing the different statistical models, the different theoretical competence concepts are validated. In Chapter 4 various theories about the connection between reading skills are evaluated based on the PIRLS-Transfer data. In recent literature there are many studies analyzing reading competences with CDMs (cf. e.g. Jang, 2009; Li, 2011; Svetina, Gorin & Tatsuoka, 2011; Wang & Gierl, 2011), however in all of them the deployed competence model is already predefined.

1.3.4 Aspect 4: Analyses of background data

In recent years CDMs have not only been applied to smaller studies (cf. e.g. DeCarlo, 2011; Tatsuoka, 1984) but also to large scale assessment data (e.g. Chiu & Minhee, 2009, for PIRLS data; Park & Lee, 2011, for TIMSS data). In these studies some beneficial information was found about how specific skill mastery effects student performances. A fundamental goal of large scale assessments is to perform international comparisons between countries (i.e. different educational systems) or national comparisons between federal states. Some first CDM studies also employ these kind of comparisons, see for example Birenbaum, Tatsuoka & Yamada (2004), Dogan & Tatsuoka (2008) or Lee, Park & Taylan (2011). Another fundamental goal of large scale studies are comparisons between different groups of students, as they enable predicting students' abilities (cf. the PIRLS framework, Bos et al., 2007, p.22). Grouping variables on the student level are for example sex, social background, the socio economic status (SES; including the migration status) and, on the structural level, for example the federal state or the school form. In the context of debates about equal opportunities, predicting students' abilities based on grouping variables may help to develop funding programs and to improve teaching and learning conditions.

Chapter 5 of the present work introduces and illustrates some possibilities of multiple group DINA models applied to the Austrian educational test of math 2012. The study is twofold: Firstly, the results obtained for the group comparisons with standard 2PL IRT models (Bruneforth & Lassnigg, 2013) are reproduced with DINA models. Secondly, the reported differences are broken down to the skill level and refined information is obtained, which provides an empirical basis for establishing the following questions: Are there skills with respect to which migrants perform better than non-migrants even if their general ability is lower? Or: Are there at least skills in which the differences in the mastery between migrants and non-migrants are much smaller as the mean difference? Other group comparisons discuss differences in skill mastery of boys and girls, students with strong and weak social background, or comparisons between students from different federal states in Austria.

2 Analyzing CDMs with the R Package CDM - A didactic

2.1 Introduction

2.1.1 Objectives of the R package CDM

The objective of the software package CDM (George, Kiefer, Robitzsch, Groß & Ünlü, 2013) is to provide an extensive, easy manageable and open source tool for CDM analyses. In achieving this aim we benefit from the advantages of R (R Core Team, 2013): Firstly, R provides a free programming framework and secondly it is object-orientated (i.e. a CDM model is treated as an object with which all steps of the analysis may be performed). An additional argument for choosing the R framework is its increasing popularity for research in the social sciences (Alexandrowicz, 2012; Kubinger, Rasch & Yanagida, 2011), combined with a lack of implemented CDM algorithms. The R package CDM is available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=CDM.

Until now the R package CDM supports estimation and subsequent analysis of DINA, DINO and G-DINA models. It seems worthwhile noting that the class of G-DINA models involves other prominent CDM models as for example the NIDA or the R-RUM model (cf. Sections 1.2.3 and 2.3.4). The R package CDM also supports analyses in which we define a different model for each item. An expansion to the class of GDMs is in preparation, for first prospects see Section 2.8. A short review and a comparison of other software packages for the estimation of CDMs is given in Section 2.1.3.

The composition of the R package CDM is two-sided: On the one hand it allows for a simple and straightforward introduction to software based CDM analyses. These simple analyses only require the user to specify the data, the Q-matrix and the model. On the other hand, the R package provides advanced methods and techniques for analyzing

CDMs, as for example the analysis of correlations between skills (cf. Section 2.4.3), possibilities to avoid ambiguous skill classes in DINA models (cf. Section 2.3.2) or the examination of item fit indices (see Section 2.4.4). Furthermore, the package can deal with datasets from large scale analysis employing a block design, it can perform multiple group analysis and provides tools for simulation studies.

The present chapter may serve as tutorial for the usage of the R package CDM and, at the same time, as a tutorial for CDM analyses. As a thorough review of CDMs and their objectives is given in Chapter 1, it is left out at this point and is assumed to be known. However, Section 2.1.2 briefly reviews the goals of and the different types of model parameters in a CDM analysis.

In the following, R codes or R objects are printed in `typewriter font`. For illustrating how to access features of DINA, DINO and G-DINA models (i.e. of objects of the class `din` or `gdina`), the notation `model` or `qmatrix` is used by referring to a general exemplary CDM model or Q-matrix. The tutorial is illustrated by DINA, DINO and G-DINA models for the PIRLS 2006 data. For a description of the data see Section 2.1.4. The tutorial proceeds according to the steps of a CDM analysis, with Sections 2.3, 2.6, 2.7 and 2.8 yielding deepening aspects.

2.1.2 Goals and parameters in CDM analyses

The aim of CDMs is to identify dichotomous skill profiles; that is, to perform multiple classifications of students based on their observed response patterns with respect to features (i.e. skills) that are assumed to derive the probability of correct responses. Which skills are required to master the items is predefined by educational experts in the so called Q-matrix.

CDMs employ two types of parameters, the item and the person parameters. Item parameters describe characteristics of the items with regard to the students' response probabilities. In some items the possession of almost all necessary skills directly leads to a high probability of success. However, in other items there may be large chances of slipping or the response probabilities remain small until students possess a specific combination of skills. Person parameters describe characteristics of the students with regard to their possession and non-possession of the skills. We distinguish between population and individual oriented person parameters: The set of population oriented classification parameters includes the distribution of the skill classes in the population and the population's skill mastery probabilities. These parameters are mainly used for the in-

terpretation of large scale educational assessments, in which the model must not hold for each individual student. The set of individual oriented person parameters includes the individual students' skill profiles. These parameters are established in studies which have their focus on individual diagnosis and feedback.

Because in the estimation process the individual classification parameters are obtained based on the item parameters and the population oriented classification parameters, only the two last sets have an influence on the models' identification. They are also called model parameters (cf. also the definition of CDMs as restricted latent class models, Section 1.2.6).

2.1.3 Review of existing software for CDM parameter estimation

The software for the estimation of CDMs reviewed in this section basically differs in the embedded programming framework and in the CDM frameworks they are able to estimate. The following list describes different programs in terms of the supporting operational systems, the input of code and the file format of the output. Additionally, a comparison of the programs in terms of their possibilities with regard to e.g. the adaption of the output, the estimation of the parameters, the provided fit statistics and the usage of different sampling designs is given in Table 2.1.1. Explicitly not listed are software packages for the estimation of log-linear models or latent class models, even if the estimation of these models in a restricted form leads to CDM parameters as well (cf. Section 1.2.6). An example of this software is the free program LEM (Vermunt, 1997) for latent equation modeling.

LCDM with SAS and Mplus LCDM estimation can be conducted with a set of stand-alone macros (Templin, Henson, Douglas & Hoffman, 2009) for the commercial package SAS (SAS Institute Inc., 2007). After specification of the data and Q-matrix (either as external files or as SAS data sets) the adapted SAS script generates Mplus (Muthén & Muthén, 2010) code and calls Mplus, which runs the estimation of the LCDM parameters by using marginal maximum likelihood (MML) methods. Finally, the parameter information is returned to SAS in form of SAS data files. The output includes the estimated item parameters with their standard errors, information about the classification reliability, some fit indices like item and person fit statistics and information criteria for evaluating the model fit.

GDM with mdltm Von Davier (2005) implemented the GDM framework in the standalone software for multidimensional discrete latent traits models (mdltm). For research purposes, the mdltm software is available free of charge from the Educational Testing Service and works on Windows, Unix and Mac OS systems. The software comes with a beta version of a graphical user interface, which allows editing the control file and entering the data, the Q-matrix, an optional IRT person parameter file, and it of course allows starting the estimation. The estimation of person and item parameters is conducted via MML methods, individual classification is accomplished by either EAP or MAP estimation. The ASCII output file contains item and person parameter estimates, item and person fit indices and classification information. Goodness of fit can be assessed via χ^2 and RMSEA measures, and the program yields information criteria for overall model-data fit and model selection.

G-DINA with Ox De la Torre implemented the G-DINA framework using a console version of Ox (Doornik, 2002). Ox and the Ox editor can be downloaded free of charge for academic research purposes, the program code has to be requested from its authors. After a modification of the code concerning the data set, the Q-matrix, the number of students, items and skills (up to $K = 15$) and the convergence criterion, the Ox procedure estimates the parameters of the G-DINA model with identity link function by conducting MML methods. The output is provided as Exel file.

NC-RUM with Arpeggio System Software The NC-RUM (or fusion model) is implemented by Bolt, Chen, DiBello, Hartz, Henson, Roussos, Stout and Templin in the commercial Arpeggio System software (DiBello & Stout, 2008). The software is called from a DOS command window and requires the user to specify a response data file, a Q-matrix file, an IRT person parameter file, and a run parameter file as input. It then estimates the model parameters of the NC-RUM model including a continuous latent ability component, the skill class probabilities and skill classification consistency indices using a Markov Chain Monte Carlo (MCMC) procedure. Because the MCMC procedure is not feasible for individual student classification in large datasets, this part of the estimation is accomplished by another component of the Arpeggio system, the so called Fast Classifier. Using the calibrated NC-RUM parameters and a likelihood approach, the Fast Classifier yields individual EAP and MAP classification of the students.

RSM with C++ The rule space method is implemented by Kikumi and Curtis Tatsuoka in a C++ (Sun Microsystems, 2001) procedure that only runs on Linux systems

(Tatsuoka & Yan, 2001). The C++ procedure requires as input a Q-matrix and IRT difficulty parameters for each item. It first uses a Boolean descriptive function to generate the expected patterns and then, second, parameter estimation and the analysis of the model are accomplished by assuming a latent class model on the partially ordered network of the generated expected response patterns. The model output of the RSM program provides the coordinates for each observed response pattern in the rule space and its four closest expected response patterns with their coordinates. Additionally, measurement and classification errors are computed.

CDMs with R The R package CDM (George et al., 2013) is an open source software package which can be downloaded at the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/web/packages/CDM/index.html. The package does not only allow the estimation of one CDM framework or model but rather of the two general main frameworks G-DINA and GDM. Thus it also allows the estimation of all common CDMs by specifying parameters of the general frameworks. Almost all methods for analyzing CDMs which are included in the other software packages (i.e. global fit measures, item fit, classification accuracy,...) are contained in the R package as well. Additionally the R package CDM provides a simulation tool for DINA and G-DINA models.

2.1.4 The PIRLS 2006 data

The Progress in International Reading Literacy Study (PIRLS; Mullis, Martin, Kennedy & Foy, 2007) is a large scale assessment study for analyzing and providing information about the reading achievement of fourth graders. The data includes 126 items in 10 booklets and students from 35 countries around the world. Following the PIRLS test design (i.e. a partial incomplete balanced block design), each student worked 2 test booklets, i.e. on between 22 to 26 items.

All test booklets include multiple choice and open format items. For the purposes of this example, the students' responses are recoded as follows: Only completely correct responses received a 1, all other response categories a 0. Missing responses were coded as 0 and not administered items were coded as `NA`.

For the PIRLS study a competence model was conceptualized (Campbell, Kelly, Mullis, Martin & Sainsbury, 2001) according to which a student's possession of a general reading ability is divided into the mastery of the four reading processes "focus on and retrieve explicitly stated information", "make straightforward inferences", "interpret and integrate

	LCDM	GDM	G-DINA	NC-RUM	CDM
software					
name of framework	SAS & Mplus	mdltm	Ox	Arpeggio	R
availability	commercial	on inquiry	on inquiry	commercial	free download
model types	LCDM	GDM	G-DINA	NC-RUM	G-DINA & GDM
estimation	MML	MML	MML	MCMC	MML
user interface		✓			
input					
Q-matrix	✓	✓	✓	✓	✓
response data	✓	✓	✓	✓	✓
IRT parameters		optional		✓	
required adaption of code	✓		✓	✓	
adaption of function parameters		✓			✓
output					
item parameters with SE	✓	✓	✓	✓	✓
skill class distribution	✓	✓	✓	✓	✓
skill mastery probabilities	✓	✓	✓	✓	✓
individual classification	✓	✓	✓	✓	✓
plots					✓
fit statistics					
model fit	✓	✓	✓		✓
item fit	✓	✓			✓
person fit	✓	✓			
LR tests					✓
classification accuracy	✓			✓	✓
classification reliability	✓			✓	✓
sampling design					
missing data	✓	✓	✓	✓	✓
sample weights		✓			✓
block design		✓			✓
multiple group design		✓			✓
simulation studies					
support of					✓

Table 2.1.1: Comparison between different softeware frameworks for estimation of CDMs.

ideas and information; make complex inferences" and "examine and evaluate content, language, and textual elements". According to educational experts, each item in PIRLS is based on exactly one of these four processes. In the following the much-discussed question whether to treat the four reading processes as parallel or as hierarchically ordered will also be discussed. A deeper discussion of this topic is given in Chapter 4.

2.2 Data, Q-matrix and sample size

Before starting a CDM analysis three things have to be prepared: The data has to be arranged in the right form, the Q-matrix has to be built and the issue of sample size has to be considered. Additionally, the number of model parameters is deduced, as it is required for determining information criteria like AIC or BIC.

2.2.1 Data

The data contains the manifest dichotomous responses of I students to J items. Missing values (responses) are allowed, they have to be coded as `NA`. In large scale studies, which employ a partially balanced incomplete block design (Bose & Nair, 1939), the items that are not administered to parts of the students have to be coded as `NA` as well. The $I \times J$ data matrix `data` has to be of class `matrix` or `data.frame`.

Example The PIRLS 2006 data `pirls` for Germany includes $I = 7899$ students and $J = 126$ items partitioned to 10 test booklets. Each student worked on between 22 to 29 items. The students' responses were coded as 0 or 1 and missing responses were coded as `NA`. Not administered items were coded as `NA` as well.

2.2.2 Q-matrix

The $J \times K$ binary Q-matrix contains for each item the skills which have to be possessed by the students in order to solve it. Of course each item has to be assigned to at least one skill, that is each row in the Q-matrix has to comprise at least one 1. Different educational theories concerning the skills underlying the tested ability imply different Q-matrices, especially the number of skills, and thus the number of columns in the Q-matrix may vary. If the number and manner of the underlying skills is predefined, different experts may nevertheless assign different skills to the items. This phenomenon is called

inter-rater disagreement. For a deeper discussion of this topic see for example Rupp & Templin (2008b). Models with all sorts of Q-matrices can be evaluated concerning their fit to the response data and the best fitting model - or Q-matrix - may be chosen. In the R package CDM the Q-matrix object `q.matrix` has to be of class `matrix` or `data.frame`.

Example The four reading processes in PIRLS may be considered as parallel (i.e. to have the same level of difficulty), because it is possible to construct simple and difficult items for each process. On the contrary, these processes may be considered as ordered in a linear hierarchical form, as it is plausible to assume that the process "focus on and retrieve explicitly stated information and ideas" is easier than the process "make straightforward inferences", which itself is easier than "interpret and integrate ideas and information" and than "examine and evaluate content, language, and textual elements". The 126×4 Q-matrix `Q_RC` for the reading concept assuming no order between the reading processes has the form

$$\texttt{Q\_RC} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \dots & \dots & \dots & \dots \end{bmatrix},$$

whereas the 126×4 Q-matrix `Q_H` for the linear hierarchical order of the skills has the form

$$\texttt{Q\_H} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ \dots & \dots & \dots & \dots \end{bmatrix}.$$

2.2.3 Sample size

There have been few concrete recommendations in the CDM literature (as well as in the LCA literature) regarding the minimum sample size for conducting CDM analyses. Rupp & Templin (2008b) suggest that for simple models such as the DINA a "few hundred" students responding each item are sufficient for convergence if the number of skills is small (four to six). A systematic study investigating minimum sample size for various numbers of skills is so far missing. A related and also not yet investigated issue is that

of parameter identifiability in the sense of achieving a unique set of item parameters for a given data set. Von Davier (2005) states that models diagnosing more than eight skills are likely to have problems with identifiability. Related issues about identifiability of students' skill profiles are also discussed in Chapter 3 of the present work.

2.2.4 Number of model parameters

DINA or DINO models (without constraints on the parameters) employ $(2^K - 1) + 2 \cdot J$ model parameters, i.e. the number of skill classes minus one (as they sum up to 1) and 2 parameters (guessing and slipping) per item. The number of parameters in G-DINA models is significantly larger and depends on the number of skills assigned to the items. For example, a G-DINA 1way model has $(2^K - 1) + J + \sum_{j=1}^{J} \sum_{k=1}^{K} q_{jk}$ parameters. In the R package CDM the number of parameters is accessible via `model$Npar` after the estimation of a model.

Example Independent of the Q-matrix `Q_RC` or `Q_H`, DINA or DINO models for the PIRLS data with 4 skills employ $(2^4 - 1) + 2 \cdot 126 = 267$ parameters. For the Q-matrix `Q_RC`, in which only between item dimensionality is considered, the G-DINA 1way model is equivalent to the DINA model and thus employs 267 parameters as well. A G-DINA 1way model based on `Q_H` needs $(2^K - 1) + J + \sum_{j=1}^{J} \sum_{k=1}^{K} q_{jk} = 432$ parameters. A G-DINA 2way based on Q_H has 165 additional parameters, that is 597 parameters, because it needs one additional parameter for each two-way interaction in each item considering more than one skill.

2.3 Further settings prior to model estimation

Besides the selection of a specific model (cf. Section 2.4) some additional, more elaborate settings can be defined prior to the model estimation. They may influence the accuracy of the parameter estimates, the identifiability of the model or the computing time.

2.3.1 Convergence criteria

The convergence criteria define when and how the estimation process terminates. More strictly chosen criteria may increase the computation time but also the accuracy of the estimated parameters.

The R package CDM provides three types of convergence criteria: `maxit`, `conv.crit` and `dev.crit`. Following the first criterion the estimation process terminates if a maximal number `maxit` of iterations is reached. Concerning the second criterion the process ends if the maximal parameter change between successive iterations is below `conv.crit`. Maximal change means the maximum of the changes between model parameters of the same type (e.g. in DINA models the maximal parameter change may emerge either in the guessing, or the slipping or skill class parameters). The third criterion causes the termination of the process if the relative difference between the deviances of the models fitted is below `dev.crit`. Here "deviance" is defined as $-2 \log L$, with L being the likelihood of the model.

The whole estimation process terminates if the maximal number of iterations is reached, or if the `conv.crit` and the `dev.crit` criterion are both true. There may be data sets and Q-matrices in which `conv.crit` would lead to a termination of the estimation process after the first or second iteration, because the differences between the parameters in consecutive steps of iteration are very small. To reach convergence in these cases it is more appropriate to consider the `dev.crit`.

In applications where the exact value of the parameter estimates (including the positions after the digital point) is relevant, as for example in simulation studies for the detection of the true parameter values, it may be reasonable to choose a more stringent setting of the convergence criteria as the one given in the default setting. A graphical visualization of the progress in the log-likelihood or in the parameter values can be obtained by plotting `model$param.history`.

Example For a DINA model on the PIRLS data Figure 2.3.1 shows the convergence history of the likelihood and of an item's guessing and slipping parameter. The convergence history of the likelihood is included in the object

`model$param.hist$likelihood.hist` ,

the convergence history of the slipping parameters in

`model$param.hist$slip.history`

and the convergence history of the guessing parameters in

`model$param.hist$guess.history`.

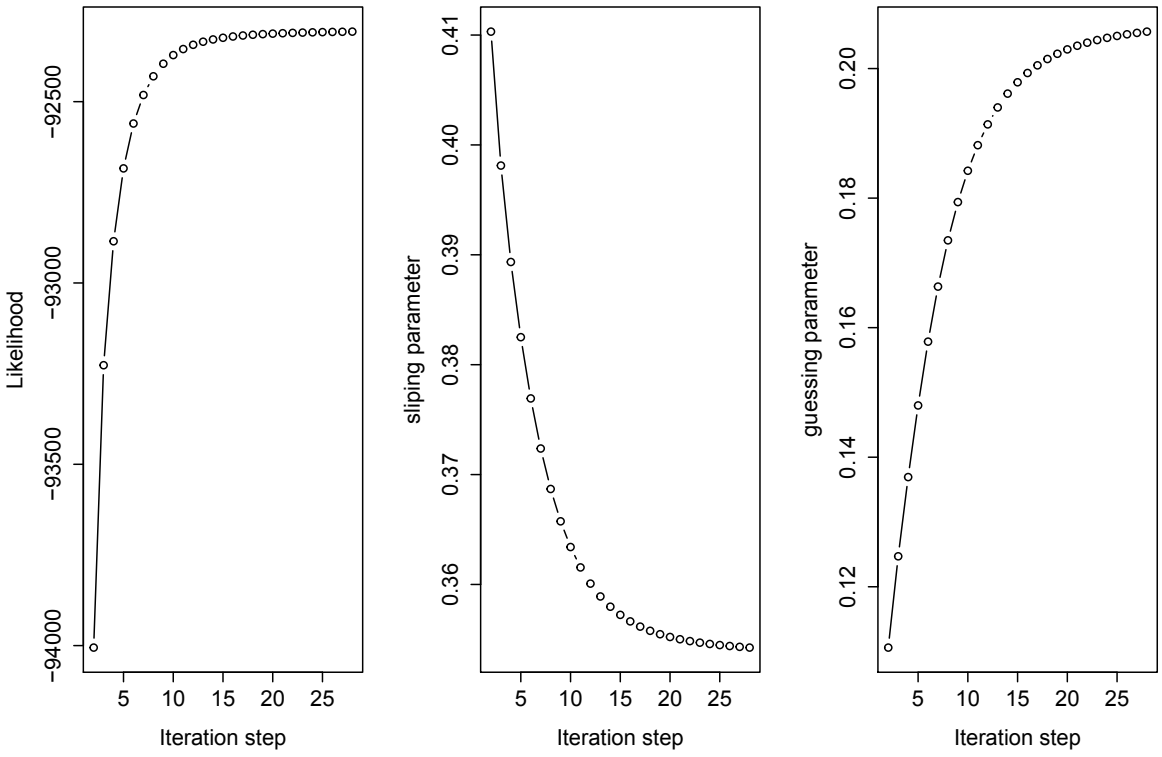

Figure 2.3.1: Convergence history of the likelihood (left hand side), a slipping parameter (middle part) and a guessing parameter (right hand side) for a DINA model on the `pirls` data.

2.3.2 Reducing the skill space

A reduction of a model's skill space which is characterized by the distribution of the skill classes can have three goals: First, to reach unambiguous classifications of individual students, second to test hypothesis about the students' acquisition of skills and third, to reduce the number of parameters.

In DINA models students may not be unambiguously classified into one of the 2^K skill classes. Rather they are classified into an equivalence class of skill classes consisting of skill classes leading to the maximal value of the likelihood. The larger these equivalence classes get, the less specific becomes the students' classification. For further details see Groß & George (2013) and Chapter 3 of the present work.

The R package CDM offers the opportunity to check how many of the 2^K skill classes in a DINA model are distinguishable by applying the function `din.equivalent.class (qmatrix)`. In this function the Gini coefficient (Gini, 1921) is used as a measure of

the concentration of skill classes, that is, it measures the number of uniquely identified skill classes and the size of the equivalence classes. For example, if all skill classes are distinguishable the Gini coefficient becomes 0. On the contrary, if there is only one equivalence class including all skill classes, the Gini coefficient becomes $\frac{2^K-1}{2^K}$. Additionally, the `din.equivalent.class`-function returns the equivalence classes.

We can avoid ambiguous classifications of students in a DINA by conducting two steps (cf. Chapter 3): Firstly, one representative skill class in each equivalence class has to be chosen. Secondly, the likelihood probabilities of all other skill classes than the representative ones are set to zero. In this way, we do not allow classification in another class than the representative one, and thus achieve a unique classification for each student.[1] Setting likelihood probabilities of skill classes to zero is possible by `zeroprob.skillclasses`, which is a vector of integers between 1 and 2^K identifying the zero-skill-classes. The skill classes are ordered according to binary principles[2]. Another method is to determine all skill classes for which the likelihood should *not* be set to zero in the matrix `skillclasses`.

Setting likelihood probabilities of skill classes to zero can not only be used for obtaining a non-ambiguous classification, but also for testing hypothesis about the skill class distribution (cf. Chapter 4). There may be applications in which it is for example reasonable to assume that only linear hierarchical skill classes occur (e.g. if the acquisition of the tested ability follows a developmental theory). The comparison of the model with specific selected skill classes (e.g. the linear hierarchical ones) and the full model (i.e. the model employing all 2^K skill classes) may then yield an idea about the true learning theory underlying the tested ability. The possibilities concerning model comparison are discussed in detail in Section 2.5.

In G-DINA models a reduction of the skill space is mainly established to control for the number of model parameters. Choosing `reduced.skillspace` for G-DINA models reduces the model's skill space based on a method by Xu & von Davier (2008) in which the skill space is modeled through use of tetrachoric correlations. A tetrachoric correlation is the correlation between two underlying normally distributed variables (with zero mean and unit variance) that have both been dichotomized by cut-point parameters specific to each variable. Extrapolating from the bivariate distribution of any pair of given skills to the joint distribution of all skill patterns, the tetrachoric model presumes underlying continuous multivariate normal variables with a zero mean vector and a tetra-

[1] Note that the representative classes are of course not unique for the model interpretation as they still represent all skill classes included in the equivalence classes.

[2] The ordering can be found for example in the first column of the object `model$attribute.patt`.

choric correlation matrix. Then, for dichotomous attributes, the cut-point parameters represent the marginal probability of an individual mastering a skill in the population. The method reduces the number of person parameters from $2^K - 1$ to $K + \frac{K(K-1)}{2}$, although it does not considerably affect the accuracy of the parameter estimates compared to the estimates obtained in a full model (von Davier, 2008). Furthermore, the CDM package allows users to define specific tetrachoric correlations between skills by specifying the `Z.skillspace` matrix. This part of the method can be employed similar to the `zeroprob.skillclasses` method for DINA models concerning the aspects of identifiability (i.e. non-ambiguous skill classes) and hypothesis testing.

Example Applying the function `din.equivalent.class(Q_RC)` for the Q-matrix `Q_RC` representing the parallel reading processes yields the result

```
16 Skill classes | 16  distinguishable skill classes |
                   Gini coefficient = 0.
```

That is, in a DINA model based on the Q-matrix `Q_RC` all 16 skill classes are distinguishable, which is also expressed by a Gini coefficient of zero. For the linear hierarchical Q-matrix `Q_H` the command `din.equivalent.class(Q_H)` leads to

```
16 Skill classes | 5  distinguishable skill classes |
                   Gini coefficient = 0.425,
```

which means that only 5 of $2^4 = 16$ skill classes are distinguishable. The equivalence classes are given by `din.equivalent.class(Q_H)$skillclasses[,c(1,3)]`

```
    skillclass distinguish.class
1  Skills_0000                 1
2  Skills_1000                 2
3  Skills_0100                 1
4  Skills_1100                 3
5  Skills_0010                 1
6  Skills_1010                 2
7  Skills_0110                 1
8  Skills_1110                 4
9  Skills_0001                 1
10 Skills_1001                 2
```

```
11 Skills_0101                1
12 Skills_1101                3
13 Skills_0011                1
14 Skills_1011                2
15 Skills_0111                1
16 Skills_1111                5.
```

Because of that the skill space for the DINA model based on `Q_H` may be restricted to 5 skill classes (e.g. the 5 linear hierarchical ones) representing the 5 equivalence classes. This is in accordance to the reading literacy concept, as it seems to be reasonable to assume that students acquire the 4 reading skills in a linear hierarchical order. Hence, we define a model which classifies the students only into the 5 linear hierarchical skill profiles by specifying the `skillclasses` matrix:

$$\texttt{skillclasses} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

2.3.3 Constraining item parameters

In some cases it may be desired to constrain the item parameters, for example they have been calibrated during a pre-test and and should be utilized in the post-test. In DINA and DINO models the items' guessing and slipping parameters may be constrained by making use of the commands `contrained.guess` and `constrained.slip`. In G-DINA models item parameter constraints may be inserted by `delta.designmatrix`. Defining equality constraints of parameters in DINA and DINO models is possible with the commands `guess.equal` and `slip.equal`, and in G-DINA models by defining the matrices `Mj`. For a detailed definition of `Mj` see de la Torre (2011).

2.3.4 Establishing the link function in G-DINA models

Establishing the link function in G-DINA models has two effects: Firstly, G-DINA models with different link functions include other prominent CDM models. Secondly, the link function regulates the impact of skill mastery to the response probabilities.

As defined by de la Torre (2011) the R package CDM also allows defining G-DINA models with three different link functions, namely the identity, the logit and the log link. G-DINA models with all link functions include DINA and DINO models. Models with identity link involve the Additive CDM (A-CDM), which is equivalent to the G-DINA 1way model. The logit link formulation includes the linear logistic model (LLM; Maris, 1999) and the log link formulation contains the NIDA model (Junker & Sijtsma, 2001), a generalized form of the NIDA model (G-NIDA) which is equivalent to a model discussed in Maris (1999), and the R-RUM (Hartz, 2002). The parameter constraints needed to achieve the LLM, NIDA, G-NIDA and R-RUM models are given in detail in de la Torre (2011).

Although the A-CDM, LLM and the G-NIDA have the same number of parameters, they assume different underlying processes, and therefore will not provide an identical model-data fit: In the A-CDM (as in all models implying the identity link) skill mastery has an direct additive impact on the response probabilities, in the LLM it has a direct additive impact on the logit of the response probabilities (i.e. an indirect impact on the response probabilities) and in the G-NIDA skill mastery has a direct multiplicative impact on the response probabilities. The direct impact makes the interpretations of the A-CDM and G-NIDA model, particularly the former, more straightforward. Another point is that of the three links only for the logit link the item mastery probabilities are automatically constrained to be between 0 and 1. On the contrary, probability estimates resulting from the identity and the log link need appropriate constraints (e.g., $0 \leq P(\alpha_{lj}) \leq 1$). In the R package `CDM` the link functions of the G-DINA model may be addressed by the command `linkfct`, with the identity link function being the default one.

2.4 Estimation and interpretation

2.4.1 Conducting the model estimation

The main part of the parameter estimation process relies on marginalized maximum likelihood (MML) methods, in which in a first step the item parameters and then in a second step the population orientated classification parameters are estimated. Technically, this part of the estimation is conducted with an expectation maximization (EM) algorithm, which is implemented according to de la Torre (2009). Hereafter, the individual classification is accomplished by maximum likelihood estimation (MLE), maximum a posteriori (MAP) or expected a posteriori (EAP) estimation. In the MLE case students

are classified in the skill class exhibiting the maximum likelihood value, whereas they are classified in the skill class with the maximum a posteriori or maximum expected a posterior value in the MAP or EAP case, respectively. Under the precondition of a uniform prior distribution of the skill classes (which is the default setting) MLE and MAP methods yield the same results. For a comparison of the classification methods see Huebner & Wang (2011).

As already mentioned in Chapter 1, the selected items rules (e.g., DINA, DINO or G-DINA) do not have to be the same for all items. In these cases the `rule` argument is specified as a vector of character strings specifying the model rule that is used for each item, e.g. `rule = c("DINA", "DINA", "DINO", ...)`.

Example For an illustration of the before discussed models just run

```
DINRC<-din(pirls,Q_RC)
```

for a DINA model based on `Q_RC`,

```
DINH<-din(pirls,Q_H,skillclasses=skillclasses)
```

for a DINA model based on `Q_H` which classifies the students only into linear hierarchical skill classes,

```
oneskill<-din(pirls,matrix(rep(1,ncol(pirls)),ncol=1))
```

for a DINA model which only differentiates between masters and non-masters of the reading ability,

```
GDIN1H <- gdina(pirls,Q_H,rule="GDINA1",reduced.skillspace=FALSE)
```

for a G-DINA 1way model based on `Q_H`,

```
GDIN1Hred<-gdina(pirls,Q_H,rule="GDINA1")
```

for a G-DINA 1way model based on `Q_H` with reduced skill space and

```
GDIN2Hred<-gdina(pirls,Q_H,rule="GDINA2")
```

for a G-DINA 2way model based on `Q_H` with reduced skill space.

2.4.2 Item parameters

The models' estimated item parameters can be accessed by `model$coef`. For DINA and DINO models this object contains for each item the used model rule (i.e. DINA or DINO) and the estimated guessing and slipping parameters together with their standard errors and an item fit measure (cf. Section 2.4.4). For G-DINA 1way models the object includes for each item the condensation rule (i.e. DINA, DINO, GDINA1, ACDM,...), the used link function, an intercept parameter and main effect parameters for the skills assigned to the item. In G-DINA 2way models the set of G-DINA 1way model parameter estimates is supplemented by two-way interaction effects between assigned skills. All parameters come along with estimated standard errors. The list of parameters is completed by an item fit measure.

In DINA and DINO models the additional constraint $g_j < 1 - s_j$ should be satisfied for each item, with g_j being the j-th items guessing and s_j the j-th items slipping parameter. This constraint ensures that a student who possesses all required skills for item j has a higher chance of mastering the item without slipping than a student who lacks in at least one of the required skills and masters the item by a lucky guess. With the item discrimination index $IDI_j = 1 - s_j - g_j$ (Lee, de la Torre & Park, 2012) it can be checked if the items fulfill the additional constraint, as negative IDI values signalize a violation of it[3]. The IDI may also be seen as diagnostic index, reporting for each item how it discriminates between students possessing all skills (i.e. having a response probability of $1-s_j$) and students lacking in at least one skill (i.e. guessing with probability g_j). Thus, IDIs close to 1 signalize a good discrimination or "diagnosticity" of the item, whereas IDI values close to 0 detect items with a low discrimination. In the R package CDM the items' IDIs may by accessed by `model$IDI` or the model's guessing and slipping parameters and the values of the IDI may be plotted by the command `plot(model,display.nr = 1)`. Items exhibiting negative or low IDIs may be excluded from further analysis or the guessing and slipping parameters of these items may be constrained before the estimation of the model. It should be noted that the IDI values are not used as item fit measures, as the response data has no direct influence on that index. A possibility to evaluate the item fit is given in Section 2.4.4.

Example Figure 2.4.2 shows the item parameters and the IDIs for the `DINRC` model. This plot is obtained by the command `plot(DINRC, display.nr = 1)`. For each of the 126 items the guessing parameter is illustrated as grey bar, the slipping parameter is

[3] In these cases the din-function will also end with a warning.

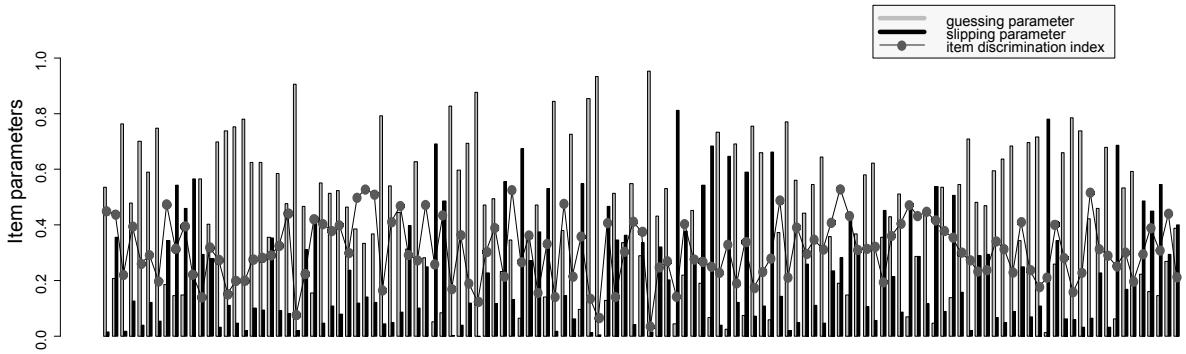

Figure 2.4.2: Item parameters and IDIs for `DINRC` model on `pirls` data.

drawn as red bar and the IDIs are depicted as solid black line. There are no items with negative IDIs, but some items have strikingly low IDIs. The low IDI values are caused by large guessing parameters: Item 64 has a guessing parameter of .95 and additional 6 items have guessing parameters above .8. If we have an influence on the test construction phase, these items should be checked concerning their task formulation and may be changed or excluded from further analysis. The values of the guessing and slipping parameters with their standard errors may be addressed by `DINRC$coef`. This yields for the first 6 items

```
           type guess se.guess  slip se.slip rmsea
R011C01C_R DINA 0.536    0.014 0.015   0.038 0.033
R011C02C_R DINA 0.206    0.008 0.354   0.072 0.282
R011C03C_R DINA 0.762    0.016 0.017   0.012 0.155
R011C04M_R DINA 0.479    0.013 0.126   0.039 0.052
R011C05M_R DINA 0.700    0.015 0.037   0.018 0.056
R011C06C_R DINA 0.590    0.014 0.119   0.223 0.097
...        ...  ...      ...   ...     ...   ...
```

For the G-DINA 1way model `GDIN1Hred` based on `Q_H` the item parameters are accessible via `GDIN1Hred$coef` and we obtain for the first two items

```
    link       item nr tp   rule   est    se partype.attr
identity R011C01C_R  1  0 GDINA1 0.453 0.042
identity R011C01C_R  1  1 GDINA1 0.521 0.054 focus on explicitly stated ideas
identity R011C02C_R  2  0 GDINA1 0.038 0.041
identity R011C02C_R  2  1 GDINA1 0.479 0.037 focus on explicitly stated ideas
identity R011C02C_R  2  2 GDINA1 0.201 0.058 make straightforward inferences
...      ...         .  . ...    ...   ...   ...
```

Because of a lack of space the captions of the output are shortened: "itemno" became "nr" and "partype" became "tp". For the first item an intercept and a main effect for the possession of the first skill is estimated. For the second item, an intercept, and main effects for the first two skills are estimated, because both skills are assigned to the item (cf. `Q_H`).

2.4.3 Person parameters

Population oriented perspective

The population oriented skill class distributions in DINA, DINO, and G-DINA models may be accessed via the command `model$attribute.patt`. The population oriented skill mastery probabilities are included in the object `model$skill.patt`.

For DINA and DINO the population oriented skill class distribution may be plotted by the command `plot(model,display.nr = 3)`. The `top.n.skill.classes` exhibiting the largest frequencies are labeled in this plot. The population oriented skill mastery probabilities in DINA and DINO models can be plotted in the form of gray bars with the command `plot(model, display.nr = 2)`.

Another aspect of the model's population oriented interpretation are the tetrachoric correlations between skills. For DINA, DINO and G-DINA models, the correlation matrix may be invoked by the command `skill.cor(model)$cor.skills`. Skills with correlations exceeding .9 exhibit a large amount of similarity and it may be reasonable to merge them.

Example Figure 2.4.3 shows the population oriented skill mastery probabilities and the skill class distribution of the `DINRC` model (top) and the `DINH` model (bottom). The `DINRC$skill.patt` object contains the population oriented skill mastery probabilities of the `DINRC` model

```
                                            skill.prob
Skill_focus on explicitly stated information  0.7010765
Skill_make straightforward inferences         0.6690154
Skill_interpret information                   0.5692033
Skill_evaluate content                        0.6110003
```

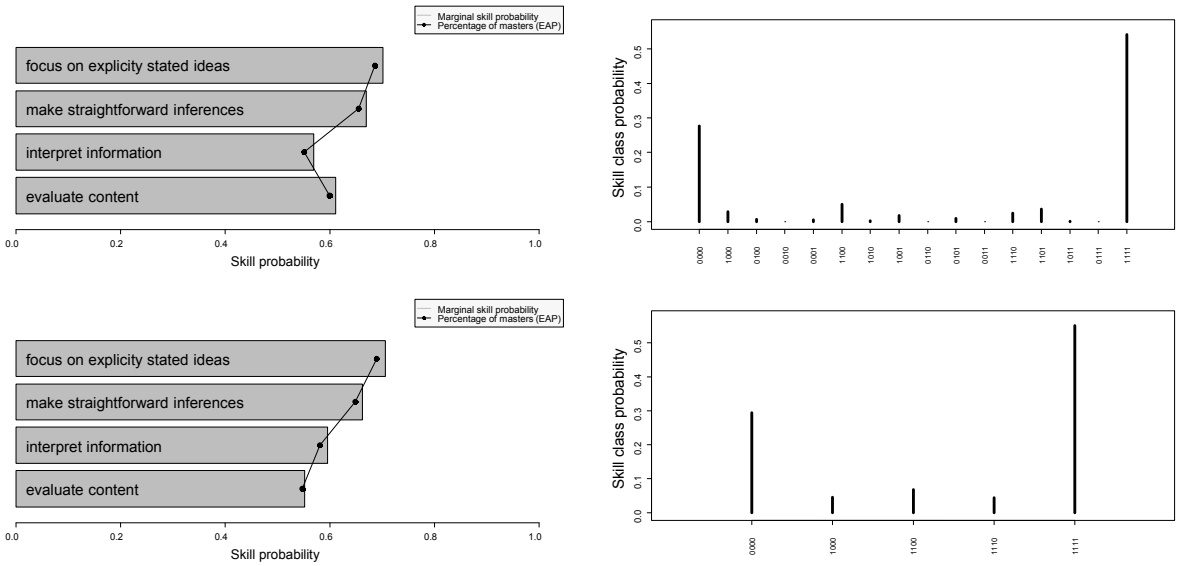

Figure 2.4.3: Population oriented skill mastery probabilities and skill class distribution for `DINRC` (top) and `DINH` model (bottom) on `pirls` data.

and, analogously, in `DINH$skill.patt` for the `DINH` model. In the `DINRC` model 70% of the students possess the skill "focus on explicitly stated information", 66% possess "making straightforward inferences", 56% of the students are able to "interpret information" and 61% are able to "evaluate the content". The mastery probabilities of the individual skills in the `DINH` model confirm the hierarchy assumption: Skill 1 is possessed by 70% of the students, skill by 66%, skill 3 by 59% and, at last, skill 4 by 55% of the students.

Via `DINRC$attribute.patt` we can access the population oriented skill class distribution of the `DINRC` model. The skill class distribution of the `DINH` model is given by `DINH$attribute.patt` object

```
     class.prob
0000 0.29428116
1000 0.04398327
1100 0.06665478
1110 0.04370169
1111 0.55137912
```

which only includes the linear hierarchical skill classes to which the skill space was constrained before (cf. Section 2.4.1). In the skill class distribution of the `DINH` model no hierarchical order between the skill classes can be identified, for example the class $[1, 1, 0, 0]$ is possessed by more students than the class $[1, 0, 0, 0]$.

Another point is that in both models the skill classes with the highest frequencies are the zeroclass $[0, 0, 0, 0]$ and the class $[1, 1, 1, 1]$. This indicates a strong coherence between the skills and makes it necessary to analyze the correlations between the skills. The `DINH` model is a unidimensional model in which one skill builds upon another and thus the skills correlate to a large extent. On the contrary, the `DINRC` model is assumed to be a four dimensional model, in which the reading skills do not provide a systematic order and because of that, they should not be highly correlated. Nevertheless, the tetrachoric correlation matrix of the skills shows extremely high correlations:

$$\texttt{skill.cor(DINRC)\$cor.skills} = \begin{bmatrix} 1 & & & \\ .98 & 1 & & \\ .99 & .98 & 1 & \\ .95 & .95 & .95 & 1 \end{bmatrix}.$$

Based on that, it is questionable if the PIRLS items are constructed to discriminate between the four reading processes and thus, if a CDM analysis of the PIRLS data is reasonable.

Individual oriented perspective

For DINA, DINO and G-DINA models the individual MLE and MAP classifications are contained in the object `model$pattern`. This object also includes a posterior skill probability for each student and each skill, that is the students' probabilities to master the skills conditional on their response pattern.

The K posterior probabilities of an individual student are also called posterior skill profiles. They offer a third possibility to classify individual students into skill classes: Students exhibiting a posterior probability smaller than .5 in a skill are not assumed to possess this skill, whereas they are assumed to possess the skill if the respective posterior skill probability is larger than .5. This procedure yields a skill profile for each student, which is also called the student's expected a posteriori (EAP) or simplified skill profile. Based on the EAP skill profiles, the frequency of students in the test population possessing an individual skill may be calculated. This frequency is depicted in the plot `plot(model, display.nr = 2)` as a solid black line. The plot allows for a comparison of an individual oriented classification method and the population oriented classification.

For DINA and DINO models it is possible to plot an individual student's posterior skill profile with the command `plot(model, pattern = "110100010", display.nr = 5)`

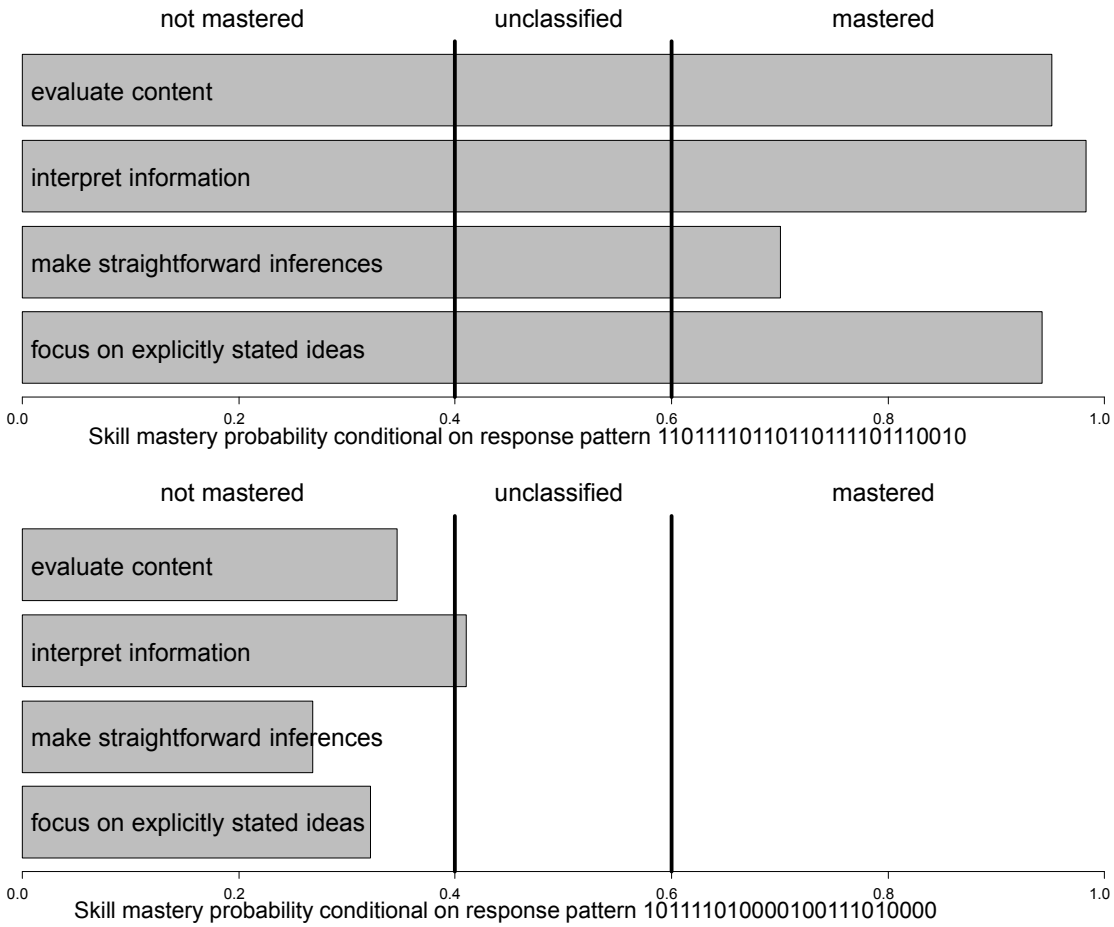

Figure 2.4.4: Posterior skill profiles of two individual students: The first student worked on 27 items in test booklets 9 and 10 and solved 18 of these items (top). The second student worked on 24 items in test booklets 4 and 8 and solved 11 of these items (bottom).

and a beforehand specified response pattern of the student (here: "`110100010`").

Example In Figure 2.4.4 posterior skill profiles of two different students are shown. The first student worked on 27 items in test booklets 9 and 10 and solved 18 of these items (top). Since all four posterior probabilities are larger than 0.5, she is classified as a master of all 4 reading processes and her EAP skill profile is $[1, 1, 1, 1]$. The second student worked on 24 items in test booklets 4 and 8 and solved 11 of these items (bottom). She did not master skills 1,2 and 4 and reached the uncertainty region for the classification of skill 3. As her posterior skill mastery probability of skill 3 is below .5, her EAP skill profile is $[0, 0, 0, 0]$. As mentioned in Section 2.4.3, the `DINRC` model tends towards classifying students into the skill profiles $[0, 0, 0, 0]$ and $[1, 1, 1, 1]$. Figure 2.4.3 shows

model	$(0, 0.05]$	$(0.05, 0.1]$	$(0.1, 1]$	mean
oneskill	125	0	0	**0.0004**
DINRC	44	55	26	0.0725
DINH	108	14	3	**0.0284**
GDIN1H	74	45	6	0.0499
GDIN1Hred	72	47	6	0.0511
GDIN2Hred	97	25	3	0.0358

Table 2.4.2: Number of items j with RMSEA_j in $(0, 0.05]$, $(0.05, 0.1]$, and $(0.1, 1]$ and mean RMSEA value for models on `pirls` data.

that the population oriented and the individual oriented EAP classification do not differ to a large extent.

2.4.4 Item fit

The R package CDM also provides an item fit statistic, the so called root mean square error of approximation (item-fit RMSEA; Kunina-Habenicht, Rupp & Wilhelm, 2009). An item fit measure indicates how good an item j suits the chosen model. Roughly spoken, the item-fit RMSEA for an item j compares the model-predicted item response probabilities $P(X_j = 1|\boldsymbol{\alpha_l})$ with the observed proportions of correct responses $N(X_j = 1|\boldsymbol{\alpha_l})$ for students in each skill class $\boldsymbol{\alpha_l}$:

$$\text{RMSEA}_j = \sum_{l=1}^{2^K} \pi(\boldsymbol{\alpha_l}) \left[P(X_j = 1|\boldsymbol{\alpha_l}) - \frac{N(X_j = 1|\boldsymbol{\alpha_l})}{N(X_j|\boldsymbol{\alpha_l})} \right]^2$$

Here $\pi(\boldsymbol{\alpha_l})$ is the frequency of students classified in skill class $\boldsymbol{\alpha_l}$ and $N(X_j|\boldsymbol{\alpha_l})$ is the observed number of responses (i.e. correct and incorrect ones) of students in skill class $\boldsymbol{\alpha_l}$ to item j. As a general guideline items with item fit indices below .05 show good fit, items with RMSEA values below .10 show moderate fit, whereas items with $\text{RMSEA}_j > .10$ indicate a poor fit (Kunina-Habenicht et al., 2009, p. 68). The item fit indices are included in the object `model$itemfit.rmsea`.

Example Table 2.4.2 shows the number of items j with RMSEA_j between 0 and 0.05 (i.e. items with good fit), between 0.05 and 0.1 (i.e. items with moderate fit) and between 0.1 and 1 (i.e. items with poor fit) and the mean RMSEA value for all models applied to the `pirls` data. The `oneskill` DINA provides the best item fit, that is this model predicts the students' response probabilities in the different items in the most accurate

model	#dim	#p	loglike	AIC	BIC
oneskill	1	251	-92474.15	185450.30	187200.90
DINRC	4	265	-92306.87	185143.75	186991.91
DINH	1	254	-92347.75	185203.51	**186975.03**
GDIN1H	1	427	-91750.48	184354.97	187333.07
GDIN1Hred	1	423	-91747.23	184340.46	187290.67
GDIN2Hred	1	670	-91304.80	**183949.61**	188622.52

Table 2.5.3: Number of dimensions (#dim), number of parameters (#p), loglikelihood (loglike), AIC and BIC for different models applied to the `pirls` data.

way. We have to note that the `oneskill` DINA only differentiates between students who *are* able to read and those who are *not*, and of course, students who are able to read are predicted to solve the items. A more differentiated analysis is given by the `DINH` model, which also provides a good item fit.

2.5 Model selection

A model may be evaluated concerning two aspects: Following the population oriented perspective, model fit is measured in terms of likelihood based criteria, whereas in the individual oriented perspective it may be more reasonable to assess the model's classification accuracy or classification consistency.

2.5.1 Likelihood based criteria

Different DINA, DINO and G-DINA models may be compared in terms of the information criteria AIC or BIC and, if the models are nested, by likelihood ratio tests. Nested CDM models are of different nature: A model that only involves a subset of skill classes (i.e. a model with restricted skill space) may be nested in the original full model which employs all skill classes, or a model that includes only a subset of skills may be nested in a model with a larger set of skills. The model's number of parameters may be accessed via `model$Npar`, the value of the loglikelihood is obtained by the command `model$loglike` and the information criteria AIC and BIC are included in the objects `model$AIC` and `model$BIC`. Likelihood ratio tests can be accomplished by `anova(model1,model2)`.

Example In Table 2.5.3 the number of dimensions, the number of model parameters, the loglikelihood and the AIC and BIC information criteria are listed for all models

which were estimated in Section 2.4 for the PIRLS data. As can be seen the G-DINA 2way model based on the Q-matrix `Q_H` provides the best model fit in terms of the AIC. Because of the large number of parameters included in this model, it does not have the lowest BIC value. In terms of the BIC, the DINA model `DINH` based on the Q-matrix `Q_H` provides the best model fit, but the `DINRC` model cannot be seen as considerably different (i.e. the difference in the BIC values is smaller than 20). Based on the PIRLS data, it seems to be hard to decide whether the reading skills follow a linear hierarchical order or not. By means of the large correlations between the skills, the low IDI values and the fact that the `oneskill` DINA provides the best item-fit, we assert again that the PIRLS items are not built to distinguish the four reading processes.

Likelihood ratio tests show that the `DINH` and the `DINRC` model fit the data significantly better than the `oneskill` DINA model: Both, `anova(DINHred, oneskill)`

```
    Model   loglike Deviance Npars      AIC      BIC    Chisq df  p
2 Model 2 -92474.15 184948.3   251 185450.3 187200.9 334.5562 14  0
1 Model 1 -92306.87 184613.7   265 185143.7 186992.0       NA NA NA
```

and `anova(DINRC, oneskill)` lead to a p-value of about zero. This means that the data includes more information than dividing students in masters and non-masters of reading, even if in both models, the `DINH` and the `DINRC`, only a low percentage of students is not classified into the extreme classes $[0, 0, 0, 0]$ or $[1, 1, 1, 1]$. On the contrary, a likelihood ratio test for the comparison of the `GDIN1H` and the `GDIN1Hred` model did not lead to a significant result

```
    Model   loglike Deviance Npars      AIC      BIC    Chisq df  p
1 Model 1 -91747.23 183494.5   423 184340.5 187290.7 -6.50149  4  1
2 Model 2 -91750.48 183501.0   427 184355.0 187333.1       NA NA NA
```

which underlines that the skill space reduction does not pose a severe restriction.

2.5.2 Classification criteria

For evaluating a model from the individual oriented perspective it might be useful to analyze the model's classification accuracy and classification consistency. Classification accuracy is a measure of how well individual students are correctly classified into their true competence levels, whereas classification consistency is a measure for the consistence of the classifications in two parallel test forms with the same items and parameters. In the R package CDM, the classification accuracy and consistency for DINA and DINO

	MLE				MAP			
model	ac	ac sim	con	con sim	ac	ac sim	con	con sim
oneskill	**.89**	**.96**	**.82**	**.92**	**.88**	**.96**	**.82**	**.93**
DINRC	.36	.47	.24	.29	.74	.82	.81	.82
DINH	.49	.72	.37	.56	.76	.85	.80	.83

Table 2.5.4: Classification accuracy assessed via analytical method (ac) and via simulation (ac sim) and classification consistency assessed via analytical method (con) and via simulation (con sim) for different DINA models based on MLE and MAP classification methods.

models are assessed via simulation methods (cf. DiBello, Roussos & Stout, 2007) and analytically by the method of Cui, Gierl & Huang (2012). Concerning the former, the simulation is conducted with known guessing, slipping and skill class parameters (i.e. the parameters of the beforehand estimated model). For G-DINA models, classification accuracy and consistency can only be assessed analytically. Accuracy and consistency are estimated using MLE and MAP classification methods and may by accessed by the command `cdm.est.class.accuracy(model)`. We have to note, that both, the accuracy and consistency measures, rely on the assumption that the data is actually generated by the particular examined model.

Example Table 2.5.4 contains the classification accuracy and consistency assessed via analytical methods (ac and con) and via simulation (ac sim and con sim) for different DINA models based on MLE and MAP classification methods. The `oneskill` DINA model provides the best classification accuracy and consistency. For this model the measures may be accessed via `cdm.est.class.accuracy(oneskill, n.sims=10000)`

```
      P_a P_a_sim   P_c P_c_sim
MLE 0.888   0.955 0.818   0.919
MAP 0.884   0.958 0.822   0.928
```

However, as mentioned before, these measures rely on the assumption that the data is generated by the examined model. In our case this means, that the data is actually generated by the `oneskill` DINA model, which only differs between masters and non-masters of reading and thus facilitates the classification.

-

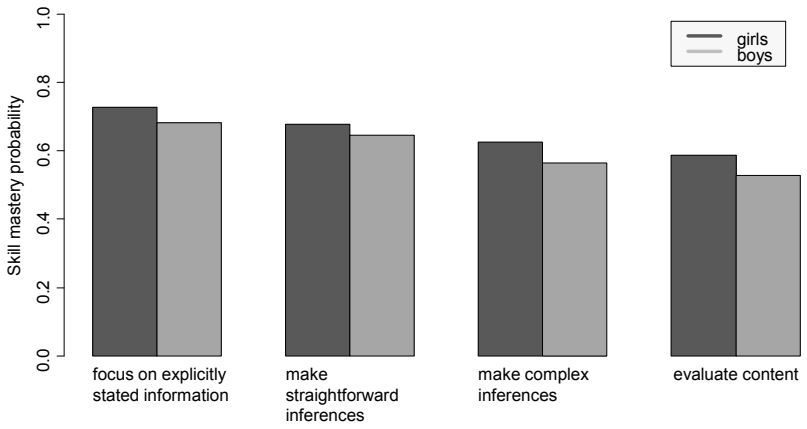

Figure 2.6.5: Population oriented mastery probabilities of the 4 reading skills for boys and girls in the PIRLS data.

2.6 Specific models

Two specific methods which are often used two analyze data in large scale assessments are multiple group analysis and the inclusion of sample weights.

2.6.1 Multiple group analysis

In educational tests it may be desirable to compare different groups of students concerning their abilities. For example, it could be of interest if boys possess the skills in another form than girls, or if migrants have particular difficulties in specific skills. For detailed analysis of this topic, i.e. the differences in possession of mathematical skills, see Chapter 5 of the present work. Conducting a multiple group analysis in the R package CDM is possible by using the `group` argument in the `gdina` function. The statistical theory of multiple group analysis in CDMs is also introduced in Chapter 5 of the present work.

Example Figure 2.6.5 shows the population oriented mastery probabilities of the 4 reading skills for boys and girls in the `DINRC` model. Within the group-vector students are assigned to the group of boys or girls. Girls are coded by 1 and boys by 2 in the data `background`, which gives information about the students taking part in PIRLS.

```
group <- background[,"ITSEX"]
zero <- c(3,4,5,7,8,9,10,11,13,14,15)
mod2 <- gdina(pirls, Q_RC, rule="DINA",
```

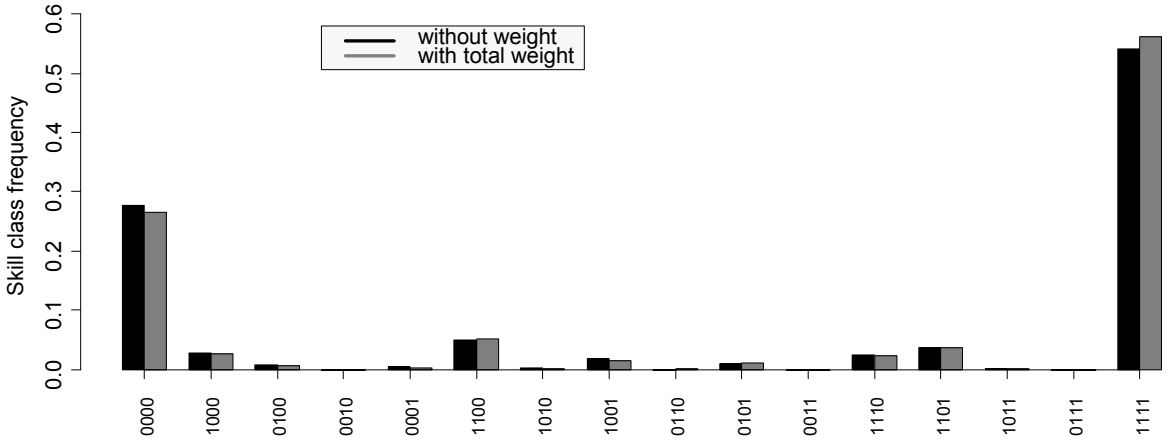

Figure 2.6.6: Population oriented skill class frequencies of `DINRC` model without weights and `DINRC` model with PIRLS sample weights.

```
zeroprob.skillclasses = zero, group = group)
```

As can be seen, girls perform slightly better in each reading skill but the differences do not seem to be significant.

2.6.2 Sample weights

Many large scale assessments include student specific sample weights `tot.wgt` to balance the sampling design (Levy & Lemeshow, 1999). In the R package CDM it is possible to include these weights in the analysis, for example `din(data, weights=tot.wgt)`.

Example Figure 2.6.6 shows the difference in the population oriented skill class distribution preserved from the `DINRC` model without weights and the `DINRC` model with PIRLS sample weights, respectively. The differences are not that large. The largest differences can be seen in the mostly occupied classes $[0,0,0,0]$ and $[1,1,1,1]$.

2.7 Simulation studies

For analyzing theoretical aspects of DINA, DINO or G-DINA models it is often helpful to work with datasets for which we know the true data generating model and the true model parameters. Simulated data may be created based on known item parameters (e.g. the slipping and guessing parameters in DINA models) or based on the known

(simplified) skill profiles of students. The first method may be extended by specifying mean values of skill mastery and correlations between the individual skills. Of course both methods can be combined. Simulated DINA data is obtained by using the function `sim.din` and simulated G-DINA data is generated by use of the function `sim.gdina`.

Example To simulate data based on item parameters (i.e. response data for 125 items), mean values of skill mastery and skill correlations from the `DINRC` model run

```
sim.guess <- DINRC$guess[,1]
sim.slip <- DINRC$slip[,1]
sim.mean <- DINRC$skill.patt[,1]
sim.cor <- skill.cor(DINRC)$cor.skills
sim.rc <- sim.din(1000, Q_RC, guess=sim.guess, slip=sim.slip,
                mean=sim.mean, Sigma=sim.cor, rule="DINA")
DINRCSIM <- din(simrcdata$dat, Q_RC, dev.crit= 10^(-8),
                conv.crit = 10^(-5))
```

In this example a data set with 1000 responses was created. If a hundred of these data sets are simulated and fit them by the DINA `RCSIM` model with Q-matrix `Q_H`, the true parameters (i.e. the parameters with which we started the simulation) and the estimated parameters (i.e. the parameters obtained from the fitted models) can be compared which results in Figure 2.7.7. This figure shows the distribution of the maximal differences between the true parameters and the estimated parameters for the guessing and slipping parameters, the skill mastery and the skill class probabilities. For example, the first value used for the distribution of the maximal differences in the guessing parameters is the largest of the $J = 125$ differences between the items' true and estimated guessing parameters in the first of the 100 simulated data sets. As Figure 2.7.7 shows, the maximal differences between true and estimated guessing parameters are slightly larger than the maximal differences in the slipping parameters. However, even the differences in the guessing parameters, which have a mean value below 0.08, cannot be assessed as serious. The situation is slightly different when it comes to the evaluation of the maximal differences between the true and estimated skill mastery and the skill class probabilities, which exhibit mean values between .14 and .16. A detection of those parameters in the simulated data sets seems to be difficult because of the large correlations between the skills.

It is also possible to create DINA data based on the individual skill profiles of the `DINRC` model: Firstly, the individual skill profiles (e.g. based on MAP classification) of the

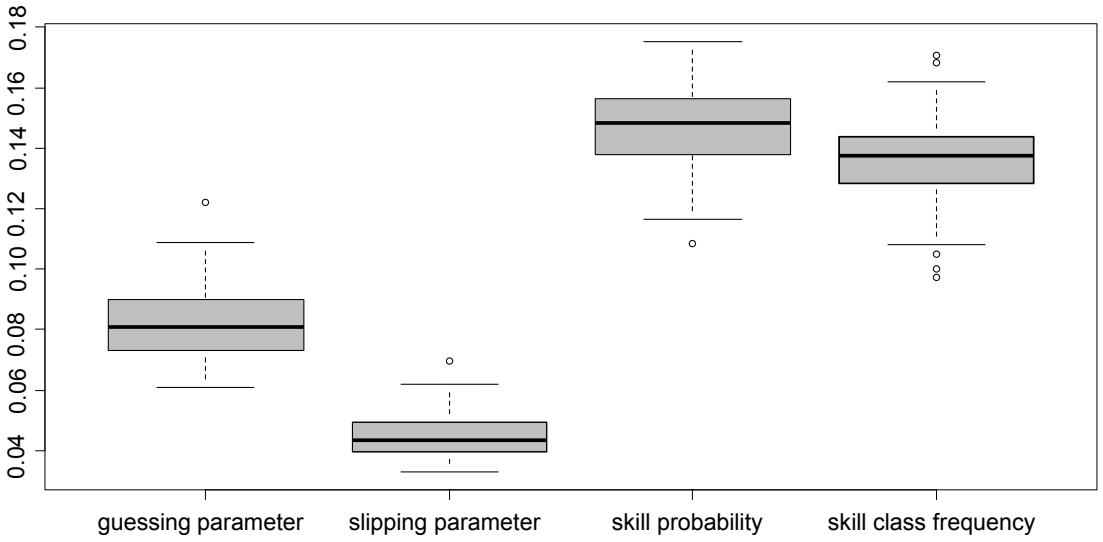

Figure 2.7.7: Distribution of maximal differences between true and estimated parameters for 100 data sets simulated according to the `DINRC` model.

`DINRC` model are imported to a 7899×4 matrix `est.skills.map` (the dimension of the that matrix is a result of the 7899 students and the 4 reading skills in the `pirls` data). Secondly, with the help of

```
alpha <- est.skills.map
simrcclass <- sim.din(q.matrix=Q_RC, alpha=alpha)
```

7899 response patterns are simulated.

The construction of 1000 simulated response patterns from a model with analogous features as the `GDIN2Hred` model may be performed by the following code:

```
# preparing necessary skills for items
rp <- sim.gdina.prepare(Q_H)
necc.attr <- rp$necc.attr

# preparing item parameters
delta<-GDIN2Hred$delta
Aj <- GDIN2Hred$Aj
Mj <- GDIN2Hred$Mj

# preparing skill mastery probabilities and skill correlations
```

```
thresh.alpha <- GDIN2Hred$skill.patt[,1]
cov.alpha <- skill.cor(GDIN2Hred)$cor.skills

sim.gdin2 <- sim.gdina(n=1000, q.matrix=Q_H, delta=delta,
             link = "identity", thresh.alpha=thresh.alpha,
             cov.alpha=cov.alpha, Mj=Mj, Aj=Aj,
             necc.attr=necc.attr).
```

2.8 Future prospects: The GDM model

An extension of the R package CDM to the class of GDM models (von Davier, 2008) is currently work in progress. The class of GDMs includes nearly all common CDM models (cf. Chapter 1 of the present work), but can also be applied to polytomous data. Furthermore, with GDMs not only dichotomous skills can be established, but also polytomous and continuous ones. Hence, the class of GDMs also includes a partial credit model for polytomous response data as well as uni- and multidimensional IRT models. Furthermore, in this class Q-matrices with polytomous entries can be handled.

As in DINA, DINO and G-DINA models, the estimation of GDMs is based on marginal maximum likelihood methods and is implemented by an EM-algorithm based on Xu & von Davier (2008). In GDMs as in DINA, DINO and G-DINA models, model parameters are estimated and individual skill parameters are determined (with MLE, MAP and EAP classification methods). Basic components for the analysis of GDMs are available: item fit indices, model fit criteria, likelihood ratio tests, reductions of the skill space, multiple group designs and sample weights.

The estimation of uni- and multidimensional IRT models opens the possibility to compare IRT and CDM models in terms of their model fit. Of course model fit is not the only substantial part in the selection of a statistical model. We should always thoroughly analyze the goals of a study and the quality of the data.

Example In the R package CDM estimation of GDMs is implemented in the function `gdm`. We can define the specific model via `irtmodel`. The default `irtmodel = 2PL` corresponds to a 2PL model in which the item slopes on all dimensions are equal for all item categories. If item-category slopes should be estimated, one may use `irtmodel = 2PLcat`. If no item slopes should be estimated `irtmodel = 1PL` can be selected.

2.9 Discussion

This chapter describes various steps in the analysis of response data with CDMs. All steps are illustrated by CDM models applied to the PIRLS 2006 data, and software code has been provided to reconstruct the steps with the R package CDM. As all substantial parts in a CDM analysis are supported and all common CDM models are included in this package, it can be seen as an alternative to existing programs as for example M-plus (Muthén & Muthén, 2010), Latent Gold (Vermunt & Magidson, 2005), lem (Vermunt, 1997), the mdltm package by von Davier or the G-DINA routines by de la Torre (cf. Section 2.1.3 of the present work for a review and comparison of these software packages). Estimation of CDMs with these software packages leads to similar results as the estimation with the R package CDM. While working with the CDM package it has been shown that it supports practical applications of CDMs as well as theoretical analysis of CDM characteristics. In future work the CDM package should be extended by some functions to increase the user-friendliness, as for example direct routines for conduction of NIDA, NIDO and R-RUM models. Furthermore, it is planned to extend the plot function for G-DINA models and GDMs and to implement measures of person fit (cf. Lui, Douglas & Henson, 2009).

The CDMs analyses applied to the response data of the PIRLS 2006 study showed that the PIRLS items do not seem to be constructed to distinguish the four reading processes but rather consider a general unidimensional reading ability. This is underlined by the official PIRLS analyses (Martin, Mullis & Kennedy, 2007) using unidimensional 2PL and partial credit models. A detailed comparison of student classifications via one dimensional 1PL (i.e. Rasch) models and four dimensional DINA models can be found in Chapter 4 of the present work. An approach to construct test items designed to distinguish on the one hand as much skill classes as possible and on the other hand the four reading processes is given in Chapter 3 of the present work.

3 Limitations of individual classifications in DINA models

3.1 Problem

A main goal of CDM analyses is to accurately and uniquely estimate the students' individual skill profiles, which then are used as empirical base for feedback and further instruction. This is obviously also true for the DINA model, which is applied in many practical CDM applications (cf. e.g. DeCarlo, 2011; Lee et al., 2012, 2011; Templin & Henson, 2006). Assume student i solves the 36 items of the baseline test of educational standards in math (cf. Section 1.2) with

$$\boldsymbol{X}_i = [1,0,0,1,1,0,0,0,1,1,0,1,0,0,1,0,1,0,0,1,1,1,1,1,1,1,0,0,0,0,1,0,0,1,1,1,0].$$

It is desired that the DINA model uniquely classifies student i into a skill profile

$$\hat{\boldsymbol{\alpha}}_i = [1,0,1,1]$$

predicting possession and non-possession of the four underlying skills "measures", "functions", "geometry" and "statistics". In this example the student is predicted to be no master of the skill α_2 "functions". Therefore she should be supported in this content domain. On the contrary, it is not desired, but might nevertheless happen, that the applied DINA model yields an ambiguous classification of student i in classes

$$\hat{\boldsymbol{\alpha}}_{i_1} = [1,0,1,1] \quad \text{or} \quad \hat{\boldsymbol{\alpha}}_{i_2} = [1,0,0,1].$$

In the first case the student should be only supported in the domain "functions", while in the second case she should be supported in "functions" and "geometry". Even more

relevant are situations in which student i is ambiguously classified into

$$\hat{\boldsymbol{\alpha}}_{i_1} = [1, 1, 0, 0] \quad \text{or} \quad \hat{\boldsymbol{\alpha}}_{i_2} = [0, 0, 1, 1]$$

given her manifest response $\boldsymbol{X}_i$. Here it remains completely unspecific whether to support her in "geometry" and "statistics" or "measures" and "functions".

In the following Section 3.2 the reasons and implications of ambiguous skill classifications in DINA models are shown. It will come out that many of the mentioned problems are connected to the construction of the Q-matrix. Typically we have no impact on this construction as it is defined by educational experts, rather we have to take it as given. In Section 3.3.1 a statistical solution for given Q-matrices and given data is introduced, while Section 3.3.2 discusses how to construct tests (i.e. Q-matrices) which avoid ambiguous skill classifications.

3.2 Theory

3.2.1 Individual skill classes

Remember that the iterative CDM estimation process consists of two steps: In the first step the item parameters and the population oriented values (i.e. the skill class distribution and the skill mastery probabilities) are determined. Based on that, in the second step, the individual skill profiles are deduced (for details see Section 1.2.2). After each iteration (including steps one and two) all parameters are adapted. The iteration terminates if a stopping criterion is fulfilled. For details see Section 2.3.

The individual classification in the second step of the estimation step may be conducted via MLE, MAP or EAP methods. The cases of MLE and MAP classification are described in this subsection while EAP classification is discussed in Subsection 3.2.3.

The MLE case

In the case of MLE classification a student i, $i = 1, \ldots, I$, is allocated in that class $\hat{\boldsymbol{\alpha}}_l$, $l = 1, \ldots, L$, for which

$$\hat{\boldsymbol{\alpha}}_{l;MLE} = \max_{\boldsymbol{\alpha}_l : l=1,\ldots,2^K} P(\boldsymbol{X_i}|\boldsymbol{\alpha}_l). \tag{3.2.1}$$

Here,

$$P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l) = \prod_{j=1}^{J} P(X_{ij} = 1|\boldsymbol{\alpha}_l)^{X_{ij}} \left(1 - P(X_{ij} = 1|\boldsymbol{\alpha}_l)\right)^{(1-X_{ij})}$$

is the probability of observing the i-th student's manifest response pattern $\boldsymbol{X}_i$ if she would be classified into class $\boldsymbol{\alpha}_l$. The student's probability

$$P(X_{ij} = 1|\boldsymbol{\alpha}_l) = g_j^{(1-\eta_{lj})}(1 - s_j)^{\eta_{lj}}$$

for correctly mastering item j, $j = 1, \ldots, J$, in class $\boldsymbol{\alpha}_l = [\alpha_{l1}, \ldots, \alpha_{lK}]$ depends only on her latent non-stochastic response

$$\eta_{lj} = \prod_{k=1}^{K} \alpha_{lk}^{q_{jk}} \in \{0, 1\} \tag{3.2.2}$$

to item j (i.e. the item parameters g_j and s_j have already been estimated in the first step and thus may be considered as constant here).

If now two (or more) skill classes $\boldsymbol{\alpha}_{l_1}$ and $\boldsymbol{\alpha}_{l_2}$, $l_1, l_2 = 1, \ldots, 2^K$, provide an equal latent response $\eta_{l_1 j} = \eta_{l_2 j}$ for item j, then

$$P(X_{ij} = 1|\boldsymbol{\alpha}_{l_1}) = P(X_{ij} = 1|\boldsymbol{\alpha}_{l_2}).$$

Note that equal latent responses are no exceptional cases: The dichotomous latent response $\eta_{l_1 j}$ may be regarded as a combination of zeros and ones provided by the entries of the skill class vector $\boldsymbol{\alpha}_{l_1}$ and of the j-th Q-matrix row $\boldsymbol{q}_j$, e.g.

$$\eta_{l_1 j} = \alpha_{l_1 1}^{q_{j1}} \cdot \alpha_{l_1 2}^{q_{j2}} \cdot \ldots = 1^1 \cdot 1^0 \ldots.$$

While the entries of the Q-matrix (i.e. the exponents) are given, the bases vary by selecting different skill classes. Obviously there may be several combinations of different bases and given exponents leading to the same response of either 1 or 0.

Furthermore, if two skill classes $\boldsymbol{\alpha}_{l_1}$ and $\boldsymbol{\alpha}_{l_2}$ lead to equal latent responses $\eta_{l_1 j} = \eta_{l_2 j}$ for all test items j, $j = 1, \ldots, J$, i.e. $\boldsymbol{\eta}_{l_1} = \boldsymbol{\eta}_{l_2}$, then even

$$P(\boldsymbol{X}_i|\boldsymbol{\alpha}_{l_1}) = P(\boldsymbol{X}_i|\boldsymbol{\alpha}_{l_2}) \tag{3.2.3}$$

holds, because

$$\begin{aligned} P(\boldsymbol{X}_i|\boldsymbol{\alpha}_{l_1}) &= \prod_{j=1}^{J} P(X_{ij}=1|\boldsymbol{\alpha}_{l_1})^{X_{ij}} \left(1-P(X_{ij}=1|\boldsymbol{\alpha}_{l_1})\right)^{(1-X_{ij})} \\ &= \prod_{j=1}^{J} P(X_{ij}=1|\boldsymbol{\alpha}_{l_2})^{X_{ij}} \left(1-P(X_{ij}=1|\boldsymbol{\alpha}_{l_2})\right)^{(1-X_{ij})} \\ &= P(\boldsymbol{X}_i|\boldsymbol{\alpha}_{l_2}). \end{aligned}$$

Equal latent responses for all items are naturally more seldom than equal latent responses for individual items. However they are no artificial cases as well, which will be shown in the next Section 3.2.2. Note that Equation (3.2.3) holds for all response patterns $\boldsymbol{X}_i$, $i = 1, \ldots, I$.

But if then for the specific skill class $\boldsymbol{\alpha}_l$ it holds

$$\hat{\boldsymbol{\alpha}}_{l_1;MLE} = \max_{\boldsymbol{\alpha}_l : l=1,\ldots,2^K} P(\boldsymbol{X_i}|\boldsymbol{\alpha}_l)$$

in Equation (3.2.1), then also

$$\hat{\boldsymbol{\alpha}}_{l_2;MLE} = \max_{\boldsymbol{\alpha}_l : l=1,\ldots,2^K} P(\boldsymbol{X_i}|\boldsymbol{\alpha}_l)$$

is true because of Equation (3.2.3). That is, there is no unique maximum in Equation (3.2.1), rather two (or even more) skill classes provide the same maximal value. This implies that student i can not be uniquely classified.

The MAP case

In the case of MAP classification, student i is assigned to class $\hat{\boldsymbol{\alpha}}_l$, $l = 1, \ldots, 2^K$, satisfying

$$\hat{\boldsymbol{\alpha}}_{l;MAP} = \max_{\boldsymbol{\alpha}_l : l=1,\ldots,2^K} P(\boldsymbol{\alpha}_l|\boldsymbol{X_i}), \tag{3.2.4}$$

with

$$P(\boldsymbol{\alpha}_l|\boldsymbol{X_i}) = \frac{P(\boldsymbol{X_i}|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l)}{\sum_{l=1}^{L} P(\boldsymbol{X_i}|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l)}.$$

By default the estimation process starts with $P(\boldsymbol{\alpha}_l) = \frac{1}{2^K}$ for all l, $l = 1, \ldots, 2^K$. Then, for two (ore more) skill classes $\boldsymbol{\alpha}_{l_1}$ and $\boldsymbol{\alpha}_{l_2}$ with equal latent responses it holds in the

first step of the iteration

$$P(\boldsymbol{\alpha}_{l_1}|\boldsymbol{X_i}) = \frac{P(\boldsymbol{X_i}|\boldsymbol{\alpha}_{l_1})P(\boldsymbol{\alpha}_{l_1})}{\sum_{l=1}^{L} P(\boldsymbol{X_i}|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l)} \overset{MLE}{=} \frac{P(\boldsymbol{X_i}|\boldsymbol{\alpha}_{l_2})P(\boldsymbol{\alpha}_{l_2})}{\sum_{l=1}^{L} P(\boldsymbol{X_i}|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l)} = P(\boldsymbol{\alpha}_{l'}|\boldsymbol{X_i}).$$

For the subsequent step of the iteration the skill class probabilities are adapted:

$$P(\boldsymbol{\alpha}_{l_1}) = \sum_{i=1}^{I} P(\boldsymbol{\alpha}_{l_1}|\boldsymbol{X_i})P(\boldsymbol{X_i}) = \sum_{i=1}^{I} P(\boldsymbol{\alpha}_{l_2}|\boldsymbol{X_i})P(\boldsymbol{X_i}) = P(\boldsymbol{\alpha}_{l_2}).$$

Thus, skill classes with equal latent responses always exhibit equal skill probabilities $P(\boldsymbol{\alpha}_{l_1})$ even if they may change in each step of the iteration. For this reason

$$P(\boldsymbol{\alpha}_{l_1}|\boldsymbol{X_i}) = P(\boldsymbol{\alpha}_{l_2}|\boldsymbol{X_i}), \quad i = 1, \ldots, I$$

is true for all steps and if $\boldsymbol{\alpha}_{l_1}$ fulfills $\max_{\boldsymbol{\alpha}_l : l=1,\ldots,2^K} P(\boldsymbol{\alpha}_l|\boldsymbol{X_i})$ then $\boldsymbol{\alpha}_{l_2}$ does as well. Hence the MAP classification in Equation (3.2.4) does not provide a unique maximum either.

3.2.2 Examples

A Contrived Example

Suppose a test consists of $J = 6$ items, and $K = 3$ skills are required to master these items. Furthermore, let the assignment of skills to the items be given by the Q-matrix

$$\boldsymbol{Q} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Altogether, 3 items are build upon skill α_1, 3 items are assigned to skill α_2 and all items request skill α_3. Note again that we have no impact on the construction of the Q-matrix.

Let us further assume that 10 students responded to the test items as given in the matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

The DINA model classifies each student i, $i = 1, \ldots, 10$, in one of the $2^K = 2^3 = 8$ possible skill classes $\boldsymbol{\alpha}_l$, $l = 1, \ldots, 8$. Note that, compared to the rules of thumb about sample size and convergence in CDMs, which demand a "few hundred" students (Rupp & Templin, 2008b), the sample size of $I = 10$ in this example is extremely small. Even if we may not relay on the convergence of the algorithm, the artifacts described above can be illustrated here as well and are the same as in larger data sets.

Table 3.2.1(a) gives the latent responses $\boldsymbol{\eta}_l$ of an arbitrary student in skill class $\boldsymbol{\alpha}_l$, $l = 1, \ldots, 8$, i.e. the response patterns $\boldsymbol{X}_i$ do not affect the latent responses, see Equation (3.2.2). As can be seen, the skill classes $[1, 0, 0]$, $[0, 1, 0]$, $[1, 1, 0]$ yield the same latent response as the skill class $[0, 0, 0]$, being $\boldsymbol{\eta} = [0, 0, 0, 0, 0, 0]$. That is, students in these four skill classes are (independently of their manifest response $\boldsymbol{X}_i$) not expected to master any of the 6 items.

Tables 3.2.1(b) and (c) give the probabilities $P(\boldsymbol{X_2}|\boldsymbol{\alpha}_l)$ and $P(\boldsymbol{\alpha}_l|\boldsymbol{X_2})$, $l = 1, \ldots, 8$, for student 2 with response pattern $\boldsymbol{X}_2 = [0, 0, 0, 0, 1, 0]$. Both probabilities are listed after the first and the last iteration step of the DINA estimation algorithm (implemented in the R package CDM, see Chapter 2). In the MLE classification case the estimated skill profile $\hat{\boldsymbol{\alpha}}_{2;MLE}$ of student 2 is obtained through the class $\boldsymbol{\alpha}_l$ which has the largest value $P(\boldsymbol{X_2}|\boldsymbol{\alpha}_l)$ amongst all l, $l = 1, \ldots, 8$, in the last step of the iteration. Analogously, in the MAP case the class $\boldsymbol{\alpha}_l$ with largest value $P(\boldsymbol{\alpha}_l|\boldsymbol{X_2})$ is chosen to define the students skill profile $\hat{\boldsymbol{\alpha}}_{2;MAP}$. As usually, the estimation process starts with $P(\boldsymbol{\alpha}_l) = \frac{1}{2^K} = \frac{1}{8}$ for all l, $l = 1, \ldots, 8$.

As can be seen in Table 3.2.1 skill classes $\boldsymbol{\alpha}_l$ with equal latent responses lead to equal values of $P(\boldsymbol{X}_2|\boldsymbol{\alpha}_l)$ and $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_2)$ in both the first and the last step of the iteration.

$\boldsymbol{\alpha}_l$	(a) $\boldsymbol{\eta}_l$	(b) $P(\boldsymbol{X}_2\|\boldsymbol{\alpha}_l)$		(c) $P(\boldsymbol{\alpha}_l\|\boldsymbol{X}_2)$	
		first step	last step	first step	last step
$[0,0,0]$	[0,0,0,0,0,0]	0.066	**0.222**	0.118	0.181
$[1,0,0]$	[0,0,0,0,0,0]	0.066	**0.222**	0.118	0.181
$[0,1,0]$	[0,0,0,0,0,0]	0.066	**0.222**	0.118	0.181
$[0,0,1]$	[0,0,0,0,1,0]	0.262	0.221	0.470	**0.274**
$[1,1,0]$	[0,0,0,0,0,0]	0.066	**0.222**	0.118	0.181
$[1,0,1]$	[1,0,0,1,1,0]	0.016	< 0.001	0.029	< 0.001
$[0,1,1]$	[0,1,1,0,1,0]	0.016	< 0.001	0.029	< 0.001
$[1,1,1]$	[1,1,1,1,1,1]	0.000	0.000	0.001	0.000

Table 3.2.1: Skill classes $\boldsymbol{\alpha}_l$, latent responses $\boldsymbol{\eta}_l$, probabilities $P(\boldsymbol{X}_2|\boldsymbol{\alpha}_l)$ and $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_2)$ (in the first and last step of the estimation process) for student 2 with response pattern $\boldsymbol{X}_2 = [0,0,0,0,1,0]$.

In the case of MLE classification the maximal value 0.222 of $P(\boldsymbol{X}_2|\boldsymbol{\alpha}_l)$, $l = 1,\ldots,2^K$, arises four times, i.e. for all skill classes providing a latent response of $\boldsymbol{\eta}_l = [0,0,0,0,0,0]$. Consequently, student 2 is arbitrarily classified into one of the corresponding skill classes $[0,0,0]$, $[1,0,0]$, $[0,1,0]$ or $[1,1,0]$, which differ in the student's possession of skills 1 and 2. The largest difference with regard to the possession of skills is located between skill classes $[0,0,0]$ and $[0,1,1]$: The first class confirms the student's possession of no skills, while the second assigns possession of skills α_1 and α_2. On the contrary, on the basis of the Q-matrix $\boldsymbol{Q}$, we would rather expect a student who solved item $j = 5$ (i.e. student 2) to possess skill α_3, as the mastery of item 5 requires only skill α_3.

In the case of MAP classification the maximal value 0.274 of $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_2)$, $l = 1,\ldots,2^K$, is unique and the second student's estimated skill profile is $\hat{\boldsymbol{\alpha}}_{2;MAP} = [0,0,1]$. The essence from this example is *not* that MAP delivers unique classifications in contrast to MLE. Indeed, in our example, student 2 is not ambiguously classified, but other students are (i.e. all students for whom $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_2)$, $l = 1,2,3,5$, is maximal).

The baseline test of educational standards in math

Consider again the Austrian baseline testing 2009 of educational standards in math (Breit & Schreiner, 2010) presented in Chapter 1. Each of the $J = 36$ test items is assigned to exactly one of the four content subcategories α_1 : "numbers and measures", α_2 : "variables and functional dependencies", α_3 : "geometry" and α_4 : "statistics", and on exactly one of the four operational subcategories α_5 : "model building", α_6 : "calculation", α_7 : "interpretation" and α_8 : "argumentation". By conducting a DINA analysis, each

content				operation			
α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
10	10	12	4	8	13	10	5

Note: α_1 : numbers, α_2 : variables, α_3 : geometry, α_4 : statistics, α_5 : model building, α_6 : calculation, α_7 : interpretation, α_8 : argumentation.

Table 3.2.2: Number of items assigned to the 8 skills in the Austrian baseline test of educational standards in math 2009.

of the $I = 1308$ eight grades is classified into one of the $2^K = 2^8 = 256$ possible skill classes. Table 3.2.2 gives the number of items which are assigned to the 8 skills. The 256 skill classes $\boldsymbol{\alpha}_l$ lead to 196 different latent responses $\boldsymbol{\eta}_l$, amongst others 33 skill classes provide the zero latent response $\boldsymbol{\eta} = [0, 0, 0, \ldots, 0, 0]$. The latter are listed in Table 3.2.3. If, for an arbitrary student i, one of these skill classes in Table 3.2.3 yields the maximal value of $P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)$ or $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)$, then all other listed classes lead to this maximal value as well. That is, student i is classified into one of the 33 classes by chance, though they differ strongly in their prediction: The classification in class $[0, 0, 0, 0, 0, 0, 0, 0]$ means that the student is predicted to possess no skills, in classes $[1, 0, 0, 0, 0, 0, 0, 0]$ to $[1, 1, 1, 1, 0, 0, 0, 0]$ the student is predicted to possess combinations of content but no operational skills, and finally in classes $[0, 0, 0, 0, 1, 0, 0, 0]$ to $[0, 0, 0, 0, 1, 1, 1, 1]$ she is likely to possess combinations of operational but no content skills. With regard to feedback it would be careless to confirm a student a skill profile $\hat{\boldsymbol{\alpha}}_i = [0, 1, 0, 1, 0, 0, 0, 0]$ and advise her to practice all operational skills and the content domains numbers and geometry, although her true skill profile is $\boldsymbol{\alpha}_i = [0, 0, 0, 0, 1, 1, 1, 1]$, meaning that she should rather practice all content skills instead of the operational ones.

$[0, 0, 0, 0, 0, 0, 0, 0]$	$[1, 1, 0, 1, 0, 0, 0, 0]$	$[0, 0, 0, 0, 1, 1, 1, 0]$
$[1, 0, 0, 0, 0, 0, 0, 0]$	$[0, 0, 1, 1, 0, 0, 0, 0]$	$[0, 0, 0, 0, 0, 0, 0, 1]$
$[0, 1, 0, 0, 0, 0, 0, 0]$	$[1, 0, 1, 1, 0, 0, 0, 0]$	$[0, 0, 0, 0, 1, 0, 0, 1]$
$[1, 1, 0, 0, 0, 0, 0, 0]$	$[0, 1, 1, 1, 0, 0, 0, 0]$	$[0, 0, 0, 0, 0, 1, 0, 1]$
$[0, 0, 1, 0, 0, 0, 0, 0]$	$[1, 1, 1, 1, 0, 0, 0, 0]$	$[0, 0, 0, 0, 1, 1, 0, 1]$
$[1, 0, 1, 0, 0, 0, 0, 0]$	$[0, 0, 0, 0, 1, 0, 0, 0]$	$[0, 0, 0, 0, 0, 0, 1, 1]$
$[0, 1, 1, 0, 0, 0, 0, 0]$	$[0, 0, 0, 0, 0, 1, 0, 0]$	$[0, 0, 0, 0, 1, 0, 1, 1]$
$[1, 1, 1, 0, 0, 0, 0, 0]$	$[0, 0, 0, 0, 1, 1, 0, 0]$	$[0, 0, 0, 0, 0, 1, 1, 1]$
$[0, 0, 0, 1, 0, 0, 0, 0]$	$[0, 0, 0, 0, 0, 0, 1, 0]$	$[0, 0, 0, 0, 1, 1, 1, 1]$
$[1, 0, 0, 1, 0, 0, 0, 0]$	$[0, 0, 0, 0, 1, 0, 1, 0]$	$[0, 0, 1, 0, 1, 0, 1, 0]$
$[0, 1, 0, 1, 0, 0, 0, 0]$	$[0, 0, 0, 0, 0, 1, 1, 0]$	$[1, 0, 0, 0, 0, 0, 0, 1]$

Table 3.2.3: Skill classes leading to zero latent response in math baseline test.

3.2.3 Individual skill mastery probabilities

For each individual student i the probability of mastering skill α_k, $k = 1, \ldots, K$, is calculated as the sum of her probabilities to master all skill classes $\boldsymbol{\alpha}_{l_1}, \ldots, \boldsymbol{\alpha}_{l_S}$ which contain skill α_k:

$$\begin{aligned} P(\alpha_k|\boldsymbol{X}_i) &= \sum_{l:\boldsymbol{\alpha}_{lk}=1} P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i) \\ &= P(\boldsymbol{\alpha}_{l_1}|\boldsymbol{X}_i) + \ldots + P(\boldsymbol{\alpha}_{l_s}|\boldsymbol{X}_i) + \ldots + P(\boldsymbol{\alpha}_{l_S}|\boldsymbol{X}_i) \end{aligned} \tag{3.2.5}$$

If there exists another skill class $\boldsymbol{\alpha}_{l_t}$ leading to the same latent response $\boldsymbol{\eta}_{l_t}$ than $\boldsymbol{\alpha}_{l_s}$, i.e. $\boldsymbol{\eta}_{l_t} = \boldsymbol{\eta}_{l_s}$, the two skill classes $\boldsymbol{\alpha}_{l_s}$ and $\boldsymbol{\alpha}_{l_t}$ are not distinguishable as shown in Section 3.2.1. Thus the skill mastery probability $P(\alpha_k|\boldsymbol{X}_i)$ might incidentally include the probability $P(\boldsymbol{\alpha}_{l_s}|\boldsymbol{X}_i)$ or $P(\boldsymbol{\alpha}_{l_t}|\boldsymbol{X}_i)$. Because furthermore $P(\boldsymbol{\alpha}_{l_s}|\boldsymbol{X_i}) = P(\boldsymbol{\alpha}_{l_t}|\boldsymbol{X_i})$ for all i, $i = 1, \ldots, I$, and in each step of the iteration, the value of $P(\alpha_k|\boldsymbol{X}_i)$ does not change in dependence of $\boldsymbol{\alpha}_{l_s}$ or $\boldsymbol{\alpha}_{l_t}$:

$$\begin{aligned} P(\alpha_k|\boldsymbol{X}_i) &= \sum_{l:\boldsymbol{\alpha}_{lk}=1} P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i) \\ &= P(\boldsymbol{\alpha}_{l_1}|\boldsymbol{X}_i) + \ldots + P(\boldsymbol{\alpha}_{l_s}|\boldsymbol{X}_i) + \ldots + P(\boldsymbol{\alpha}_{l_S}|\boldsymbol{X}_i) \\ &= P(\boldsymbol{\alpha}_{l_1}|\boldsymbol{X}_i) + \ldots + P(\boldsymbol{\alpha}_{l_t}|\boldsymbol{X}_i) + \ldots + P(\boldsymbol{\alpha}_{l_S}|\boldsymbol{X}_i). \end{aligned} \tag{3.2.6}$$

Now it might happen that skill class $\boldsymbol{\alpha}_{l_s}$ contains skill α_k, i.e. $\boldsymbol{\alpha}_{l_s k} = 1$, but skill class $\boldsymbol{\alpha}_{l_t}$ does not, i.e. $\boldsymbol{\alpha}_{l_t k} = 0$. Then the skill mastery probability in Equation (3.2.6) includes probabilities of skill classes *not* including skill α_k. If in this case we would accumulate only the probabilities of skill classes actually including skill α_k in (3.2.6), we end up with a lower skill mastery probability for α_k than in (3.2.5).

This can be explained as follows: Based on the given Q-matrix some skill classes (as $\boldsymbol{\alpha}_{l_s}$ and $\boldsymbol{\alpha}_{l_t}$) are not distinguishable. It is ambiguous if students possess the skills included in $\boldsymbol{\alpha}_{l_s}$ or in $\boldsymbol{\alpha}_{l_t}$, and thus in our example if the students possess skill α_k or not. However, calculating the individual skill mastery probabilities as in Equation (3.2.5) requires a differentiation between these skill classes, as only the probabilities of the skill classes actually including skill α_k should be added up.

The EAP case

For classifying student i, $i = 1, \ldots, I$, based on EAP

$$\tilde{\boldsymbol{\alpha}}_{i;EAP} = [P(\alpha_1|\boldsymbol{X}_i), \ldots, P(\alpha_K|\boldsymbol{X}_i)]$$

is dichotomized at the threshold 0.5. But, given (3.2.6) the estimated skill class probabilities $P(\alpha_k|\boldsymbol{X}_i)$ may be a lot larger than they actually are and thus the chance of obtaining 1 instead of 0 increases. Thus student i is rather classified into a skill class containing too many skills.

A Contrived Example

Consider again the example from Section 3.2.2 with Q-matrix $\boldsymbol{Q}$ and student 2 with manifest response pattern $\boldsymbol{X}_2 = [0, 0, 0, 0, 1, 0]$. According to (3.2.5) and Table 3.2.1 (c), her probability for possessing the first skill is

$$\begin{aligned} P(\alpha_1|\boldsymbol{X}_2) &= \sum_{l:\boldsymbol{\alpha}_{l1}=1} P(\boldsymbol{\alpha}_l|\boldsymbol{X}_2) \\ &= P([1,0,0]|\boldsymbol{X}_2) + P([1,1,0]|\boldsymbol{X}_2) + P([1,0,1]|\boldsymbol{X}_2) + P([1,1,1]|\boldsymbol{X}_2) \\ &= 0.181 + 0.181 + 0.103 \cdot 10^{-5} + 0.000 \\ &= 0.362. \end{aligned}$$

That is, although student 2 did not solve any item requesting skill α_1, she has a probability of .36 to possess that skill. This is much higher than we would expect. Because the skill classes $[0,0,0]$, $[1,0,0]$, $[0,1,0]$ and $[1,1,0]$ are not distinguishable judged by $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_2)$, i.e.

$$P([0,0,0]|\boldsymbol{X}_2) = P([1,0,0]|\boldsymbol{X}_2) = P([0,1,0]|\boldsymbol{X}_2) = P([1,1,0]|\boldsymbol{X}_2),$$

it is possible to rewrite $P(\alpha_1|\boldsymbol{X}_2)$ as

$$P(\alpha_1|\boldsymbol{X}_2) = P([0,0,0]|\boldsymbol{X}_2) + P([0,0,0]|\boldsymbol{X}_2) + P([1,0,1]|\boldsymbol{X}_2) + P([1,1,1]|\boldsymbol{X}_2).$$

Consequently, the students probability of possessing skill α_1 includes $P([0,0,0]|\boldsymbol{X}_2)$ twice, although the skill class $[0,0,0]$ does not include skill α_1.

3.2.4 State of research

Already Haertel (1989) notes that the DINA model may produce ambiguous skill classes. However, Heartel mainly discusses the problem of model identification, rather than dealing with the consequences of ambiguous skill classes for deducing individual student classifications and individual skill mastery probabilities. Heartel proposes to pool the ambiguous (i.e. unidentified) skill classes $\boldsymbol{\alpha}_{l_1}$ and $\boldsymbol{\alpha}_{l_2}$ into one class and estimate their joint probability $P(\boldsymbol{\alpha}_{l_1} + \boldsymbol{\alpha}_{l_2})$. This is in contrast to the approach presented in Section 3.3.1, in which $P(\boldsymbol{\alpha}_{l_1})$ will be estimated by defining $P(\boldsymbol{\alpha}_{l_2}) = 0$.

DeCarlo (2011) describes problems in the calculation of the individual skill mastery probabilities. More precisely, he assessed that students, who responded no item correctly, nevertheless yield large individual skill mastery probabilities with the DINA model. A possible reason for this problem is described in Section 3.2.3 and illustrated by example 3.2.3. A solution can be found in Section 3.3.1.

Obviously, improper or even wrong individual classifications in DINA models can also be a result of an ill specified Q-matrix (cf. e.g. DeCarlo, 2012; de la Torre, 2008; Rupp & Templin, 2008a). In contrast to that research, the present chapter deals with known Q-matrices (i.e. assuming that the entries are given and definitively correct). In real life approaches both problems appear and mix up.

In the ongoing chapter the skill class distribution $P(\boldsymbol{\alpha}_l)$, $l = 1, \ldots, L$, is kind of modeled by defining specific skill class probabilities as zero and several studies can be mentioned in connection with the modeling of the skill class distribution. The goal of our adaption is to reach unique individual classification and proper individual skill class probabilities. In contrast, most of the other studies deal with the population oriented skill class distribution and aim mainly at two goals: Modeling the population oriented skill class distribution in an accurate way by simultaneously reducing the number of model parameters (Hartz, 2002; Templin, Henson, Templin & Roussos, 2008; Xu & von Davier, 2008) and mirroring predefined hierarchies between the skills in the population oriented skill class distribution (Groß & George, 2013; Leighton & Gierl, 2007).

3.3 Solutions

The following section shows how to handle or to avoid ambiguous skill classes in two cases: Firstly, in the case of a given test, i.e. given Q-matrix and data (cf. Section 3.3.1) and secondly, in the case of test construction, i.e. if the test and therefore the Q-matrix

can be newly developed and adapted to one's needs (cf. Section 3.3.2). Mathematically both cases differ in that in the first case the Q-matrix (i.e. its rows $\boldsymbol{q}_j$) and therefore the latent responses $\boldsymbol{\eta}_l$, $l = 1, \ldots, 2^K$, are given, whereas in the case of test construction the $\boldsymbol{q}_j$, $j = 1, \ldots, J$, can be designed and thus the structure of $\boldsymbol{\eta}_l$ can be influenced.

3.3.1 The case of given data and Q-matrix

Let $\boldsymbol{\alpha}_l$, $l = 1, \ldots, 2^K$, be the $L = 2^K$ skill classes leading to the non-stochastic latent responses $\boldsymbol{\eta}_l$, $l = 1, \ldots, 2^K$, through application of Equation 3.2.2 for all items j, $j = 1, \ldots, J$. Furthermore, let $M \leq L$ be the number of different latent responses. Then, each of these distinguishable latent responses $\boldsymbol{\eta}_m$, $m = 1, \ldots, M$, is deduced through a set of skill classes $\boldsymbol{\alpha}_{m;1}, \ldots, \boldsymbol{\alpha}_{m;l_m}$, where m indicates the set and $l_m \geq 1$ the number of skill classes included in the m-th set. Obviously it holds $l_1 + \ldots + l_m + \ldots + l_M = L = 2^K$. In the following the set $\{\boldsymbol{\alpha}_{m;1}, \ldots, \boldsymbol{\alpha}_{m;l_m}\}$ is called the m-th equivalence class of skill classes. For handling ambiguous skill classes in the case of MAP classification the following procedure may be chosen:

(1) From all skill classes $\alpha_{m;1}, \ldots, \alpha_{m;l_m}$ in the m-th equivalence class which are leading to the same latent response $\boldsymbol{\eta}_m$ one representative skill class is chosen. In the following this class is denoted as $\alpha_{m;1}$. Obviously, in equivalence classes containing only one skill class, i.e. $l_m = 1$, this single skill class is chosen as the representative skill class.

(2) The starting values for the probabilities $P(\boldsymbol{\alpha}_{m;1}), \ldots, P(\boldsymbol{\alpha}_{m;l_m})$, $m = 1, \ldots, M$, are newly arranged: Whereas usually the starting values are set to

$$P(\boldsymbol{\alpha}_{m;1}) = \ldots = P(\boldsymbol{\alpha}_{m;l_m}) = \frac{1}{2^K}, \quad m = 1, \ldots, M,$$

now the probabilities for each representative class are fixed as

$$P(\boldsymbol{\alpha}_{m;1}) = \frac{1}{M}, \quad m = 1, \ldots, M$$

and all other probabilities are defined as

$$P(\boldsymbol{\alpha}_{m;2}) = \ldots = P(\boldsymbol{\alpha}_{m;l_m}) = 0, \quad m = 1, \ldots, M.$$

This solution for the starting values satisfies

$$\sum_{l=1}^{l_1} P(\boldsymbol{\alpha}_{1;l}) + \ldots + \sum_{l=1}^{l_m} P(\boldsymbol{\alpha}_{m;l}) + \ldots + \sum_{l=1}^{l_M} P(\boldsymbol{\alpha}_{M;l}) = \sum_{l=1}^{2^K} P(\boldsymbol{\alpha}_l) = 1.$$

Setting the probabilities of the non-representative classes to zero is like switching them off and forcing them not to occur. Note that if the probabilities of these classes have been defined to be zero in the first iteration of the algorithm they remain zero throughout the whole process.

By selecting only one of the non-distinguishable skill classes and setting all others to zero we avoid skill mastery probabilities which are much larger than expected. For an illustration see Example 3.2.3. In this example a large skill mastery probability for α_1 was obtained although no item requiring α_1 has been mastered correctly. On the contrary, after defining the priors as described above the skill mastery probability for α_1 decreases to almost zero (cf. Example 3.3.1), which is much more what we would expect because the student mastered no item requiring skill α_1.

(3) The representative skill class of each equivalence class may be chosen as the skill class within the equivalence class including the minimal number of skills (i.e. having the minimal number of ones). Mathematically, this selection seems reasonable as the skill class with minimal skills within an equivalence class is always unique. For a proof see below. From the perspective of learning this solution seems convenient as well: it is better to learn more than necessary than to learn less than necessary. However, from a didactic perspective, the solution of choosing the class with minimal skills is little sensible as students may become unmotivated by such feedback. The gap between knowledge transfer and students' motivation is discussed in detail in Seedhouse (2005).

(4) Resulting, according to MAP classification students are only classified into the M representative skill classes, inducing a unique classification. Note that the presented procedure does not influence the probabilities $P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)$ and thus the MLE classification results directly (i.e. their calculation is not influenced by the probabilities $P(\boldsymbol{\alpha}_l)$), rather the guessing and slipping parameters change through setting some $P(\boldsymbol{\alpha}_l) = 0$. The latter influences the MLE results but does not mend their ambiguity. Even in the mathematical unique MAP case, the interpretation of the classification of students into representative skill classes requires some sensibility: If a student is classified into such a representative class, her skill profile may be

allocated in each other class of the respective equivalence class, which should be noted in any case. However, in deducing the individual skill mastery probabilities, the advantages of the presented method and choosing the skill classes with minimal skills can be seen: The sum in Equation (3.2.6) only includes probabilities of skill classes which include a minimal number of skills. These probabilities can not be exchanged with probabilities of other skill classes including fewer skills and therefore perhaps not including the skill of interest.

Proof: Unique skill classes with minimal skills in each equivalence class

Let the score of a skill class be the sum of its elements. Then, if a specific set consists of all possible skill classes producing the same latent response, there is a unique skill class of minimal score within this set. This can be deduced from the two facts stated below. Let the intersection

$$\boldsymbol{\alpha}_l \wedge \boldsymbol{\alpha}_{l'}$$

of two skill classes $\boldsymbol{\alpha}_l$ and $\boldsymbol{\alpha}_{l'}$ be a binary operation carried out elementwise, such that $0 \wedge 0 = 0$, $0 \wedge 1 = 1 \wedge 0 = 0$ and $1 \wedge 1 = 1$.

Fact 1 If two different skill classes have identical score, then their intersection has strictly smaller score, since the intersection involves at least one operation $0 \wedge 1$ (or $1 \wedge 0$), otherwise the skill classes could not be different.
Fact 2 If two skill classes produce the same latent response, their intersection also produces this latent response.

To see Fact 2, it is enough to consider the latent response to a single item, since the argument is the same for all items. Let $q = [q_1, \ldots, q_K]$ denote a specific row in the Q-matrix (corresponding to a specific item) and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$ denote a skill class. Then the corresponding latent response to this specific item is $\prod_{k=1}^{K} \alpha_k^{q_k}$ and it is either 0 or 1.

Now, if the latent response of two skill classes l and l' to a specific item is 0, then for each skill class the above product must contain at least one factor 0^1, where this factor can occur at possibly different positions. Necessarily, the product for the intersection also involves at least one factor 0^1, thus also producing 0 as a latent response. If the latent response of two skill classes l and l' to a specific item is 1, then the above products for both skill classes cannot contain 0^1 as a factor. Thus, whenever the intersection has 0 as an element, the corresponding element in q must be 0, so that the product for the intersection does not contain a factor 0^1 and therefore produce 1 as latent response.

m	latent response $\boldsymbol{\eta}_m$	representative class $\boldsymbol{\alpha}_{m;1}$	included skill classes $\boldsymbol{\alpha}_{m;l}$, $l = 1, \ldots, l_m$
1	$\boldsymbol{\eta}_1 = [0,0,0,0,0,0]$	$\boldsymbol{\alpha}_{1;1} = [0,0,0]$	$\boldsymbol{\alpha}_{1;1} = [0,0,0]$, $\boldsymbol{\alpha}_{1;2} = [1,0,0]$, $\boldsymbol{\alpha}_{1;3} = [0,1,0]$, $\boldsymbol{\alpha}_{1;4} = [1,1,0]$
2	$\boldsymbol{\eta}_2 = [0,0,0,0,1,0]$	$\boldsymbol{\alpha}_{2;1} = [0,0,1]$	$\boldsymbol{\alpha}_{2;1} = [0,0,1]$
3	$\boldsymbol{\eta}_3 = [1,0,0,1,1,0]$	$\boldsymbol{\alpha}_{3;1} = [1,0,1]$	$\boldsymbol{\alpha}_{3;1} = [1,0,1]$
4	$\boldsymbol{\eta}_4 = [0,1,1,0,1,0]$	$\boldsymbol{\alpha}_{4;1} = [0,1,1]$	$\boldsymbol{\alpha}_{4;1} = [0,1,1]$
5	$\boldsymbol{\eta}_5 = [1,1,1,1,1,1]$	$\boldsymbol{\alpha}_{5;1} = [1,1,1]$	$\boldsymbol{\alpha}_{5;1} = [1,1,1]$

Table 3.3.4: Equivalence classes of skill classes with their included skill classes, their representative class and their respective latent response.

By combining Facts 1 and 2, it is now clear that a complete set of skill classes with identical latent response must also contain the intersection between these classes, and this unique intersection must have minimal score. The possible benefit from using classes with minimal score is demonstrated by the following example.

A Contrived Example

Consider again the contrived example from Sections 3.2.2 and 3.2.3. From Table 3.2.1 we learned that we have $M = 5$ different latent responses $\boldsymbol{\eta}_l$ out of $2^3 = 8$ possible latent classes. The five corresponding equivalence classes of skill classes are given in Table 3.3.4. According to the procedure presented in Section 3.3.1 the starting values of all non-representative classes are set to zero and the starting values of the representative classes are defined as

$$P([0,0,0]) = P([0,0,1]) = P([1,0,1]) = P([0,1,1]) = P([1,1,1]) = \frac{1}{M} = \frac{1}{5}.$$

The results of the estimation process for student 2 with manifest response pattern $\boldsymbol{X}_2 = [0,0,0,0,1,0]$ are given in Table 3.3.5: In comparison to Table 3.2.1 the MLE values $P(\boldsymbol{X}_2|\boldsymbol{\alpha}_l)$ only change because of different estimated guessing and slipping parameters. Skill classes providing equal latent responses still yield equal values of $P(\boldsymbol{X}_2|\boldsymbol{\alpha}_l)$. In the MAP case each probability $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_2)$ belonging to a non-representative class is zero, thus it is "not possible" to classify students into non-representative classes. In our example student 2 is classified into skill class $[0,0,1]$, which is the unique class in both the MLE and the MAP case.

According to the procedure presented before the individual skill mastery probability of

$\boldsymbol{\alpha}_l$	(a) $\boldsymbol{\eta}_l$	(b) $P(\boldsymbol{X}_2\|\boldsymbol{\alpha}_l)$		(c) $P(\boldsymbol{\alpha}_l\|\boldsymbol{X}_2)$	
		first step	last step	first step	last step
$[0,0,0]$	[0,0,0,0,0,0]	0.066	0.221	0.182	0.431
$[1,0,0]$	[0,0,0,0,0,0]	0.066	0.221	0.000	0.000
$[0,1,0]$	[0,0,0,0,0,0]	0.066	0.221	0.000	0.000
$[0,0,1]$	[0,0,0,0,1,0]	0.262	**0.222**	0.727	**0.569**
$[1,1,0]$	[0,0,0,0,0,0]	0.066	0.221	0.000	0.000
$[1,0,1]$	[1,0,0,1,1,0]	0.016	< 0.001	0.045	< 0.001
$[0,1,1]$	[0,1,1,0,1,0]	0.016	< 0.001	0.045	< 0.001
$[1,1,1]$	[1,1,1,1,1,1]	0.001	0.000	0.001	0.000

Table 3.3.5: Skill classes $\boldsymbol{\alpha}_l$, latent responses $\boldsymbol{\eta}_l$, probabilities $P(\boldsymbol{X}_2|\boldsymbol{\alpha}_l)$ and $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_2)$ (in the first and last step of the estimation process) for second student with response pattern $\boldsymbol{X}_2 = [0,0,0,0,1,0]$ and setting starting probabilities of non-representative classes to zero.

student 2 for skill α_1 is

$$\begin{aligned} P(\alpha_1|\boldsymbol{X}_2) &= \sum_{l:\boldsymbol{\alpha}_{l1}=1} P(\boldsymbol{\alpha}_l|\boldsymbol{X}_2) \\ &= P([1,0,0]|\boldsymbol{X}_2) + P([1,1,0]|\boldsymbol{X}_2) + P([1,0,1]|\boldsymbol{X}_2) + P([1,1,1]|\boldsymbol{X}_2) \\ &= 0.000 + 0.000 + 0.181 \cdot 10^{-6} + 0.000 \\ &= 0.181 \cdot 10^{-6}. \end{aligned}$$

This is much more what we would expect given that student 2 has not mastered any items requiring skill α_1.

3.3.2 The case of test construction

In the case of test construction the occurrence of ambiguous latent classes may already be avoided or at least their number limited before the conduction of the DINA analysis. In the test construction phase we have influence on (a) the number of items which request a skill and (b) the combinations of skills required to master the items. Together, (a) and (b) compose the rows $\boldsymbol{q}_j$, $j = 1, \ldots, J$, of the Q-matrix. Furthermore, because the latent responses are defined as $\eta_{lj} = \prod_{j=1}^{J} \alpha_{lj}^{q_{jk}}$ for all items j and the latent classes $\boldsymbol{\alpha}_l$ are given by default, only the design of the $\boldsymbol{q}_j$ influences the form of the latent classes $\boldsymbol{\eta}_l$. Thus $\boldsymbol{q}_j$, $j = 1, \ldots, J$ (i.e. $\boldsymbol{Q}$) may be structured in such a form that the skill classes $\boldsymbol{\alpha}_l$ lead to as much distinguishable latent classes as possible. For Q-matrices in which

(1) there exists at least one item for each skill which solely requests this skill and

(2) all other items may request the skills in an arbitrary combination

no equal latent responses and thus no ambiguous skill classes occur.

Apart from that specific Q-matrices we distinguish between Q-matrices which evoke many equivalence classes including few skill classes (preferably only one skill class) and Q-matrices which generate few equivalence classes including many skill classes. For classification purposes, the first sort of Q-matrices is preferred, as has been discussed extensively above. In the following we characterize the first sort of Q-matrices as Q-matrices evoking a low concentration of equivalence classes, and the second sort of Q-matrices as Q-matrices generating a high concentration of equivalence classes. The concentration may be measured by an adaption of the Gini coefficient (Gini, 1921):

$$G = \frac{2\sum_{m=1}^{M}(m)\cdot l_{(m)}}{M\sum_{m=1}^{M} l_m} - \frac{M+1}{M} \quad \text{with} \quad 0 \leq G < \frac{2^K-1}{2^K}.$$

Here, M is the number of equivalence classes, l_m the size of the m-th equivalence class (i.e. the number of skill classes included in the equivalence class), and (m) the m-th equivalence class ordered by size, with $(m) = (1)$ being the smallest class. If $G = 0$ each skill class leads to a different latent response and the Q-matrix evokes zero concentration. Contrary, if $G = \frac{2^K-1}{2^K}$ all skill classes would lead to the same latent response.

Thus, before developing suitable items, the concentration of the desired Q-matrix may be measured. Q-matrices with low concentration may be preferred because they avoid or diminish the number of ambiguous skill classes.

Examples

Table 3.3.6 includes three Q-matrices $\boldsymbol{Q}_1$, $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_3$, each constructed for $K = 4$ skills and $J = 7$ items. The first Q-matrix $\boldsymbol{Q}_1$ is constructed according to the above mentioned principle of having at least one item for each skill which solely requests that skill. This Q-matrix invokes no concentration: All $2^K = 16$ possible skill classes are distinguishable and thus $G = 0$. $\boldsymbol{Q}_2$ slightly violates this desire, as item 1 to 3 measure solely skill α_1 to α_3, but there exists no item measuring solely skill α_4: For $\boldsymbol{Q}_2$ 11 out of 16 skill classes are distinguishable and $G = 0.170$. In $\boldsymbol{Q}_3$ only skill α_1 is measured by an item for its own. Here only 8 out of 16 skill classes are distinguishable and G_3=0.375.

	$\boldsymbol{Q}_1$				$\boldsymbol{Q}_2$				$\boldsymbol{Q}_3$			
	1	0	0	0	1	0	0	0	1	0	0	0
	0	1	0	0	0	1	0	0	1	1	0	1
	0	0	1	0	0	0	1	0	1	0	1	0
	0	0	0	1	1	0	1	0	0	1	1	1
	1	0	1	0	0	1	1	1	1	1	1	1
	1	1	1	1	0	1	1	0	1	0	0	0
Gini-coefficient	G_1=0.000				G_2=0.170				G_3=0.375			
skill classes	16				16				16			
distinguishable classes	16				11				8			

Table 3.3.6: Three Q-matrices $\boldsymbol{Q}_1$, $\boldsymbol{Q}_2$ and $\boldsymbol{Q}_3$ leading to no, medium and high concentration of the equivalences classes of skill classes.

3.4 Discussion

In the light of the presented results about individual classifications in DINA models (partially arbitrary individual classification, unexpected individual skill mastery probabilities) it seems even more important to handle and interpret them sensibly. Lacking care about the present problems can produce inaccurate empirical bases for student feedback, which may end in disastrous misjudgments and students learning skills they already possess while believing to possess skills they are not able to master.

This chapter presents two approaches for the mentioned problems, one in the case of existing tests (i.e. Q-matrices) and one for tests to be designed. The approach for existing tests aims more at sensitizing for the problem than at finding a solution for the individual classification: non-distinguishable skill classes can not be made distinguishable. In the phase of test construction desired Q-matrices can be judged by an adapted form of the Gini coefficient, which measures the number and size of equivalence classes the Q-matrices evoke. Q-matrices with low coefficients are to be preferred. Here it might be helpful and is planned to develop a graphical tool (an adaption of the Lorenz-curve) to illustrate and thus evaluate the number and size of the equivalence classes.

4 Modeling reading abilities

4.1 Problem

The Progress in International Reading Literacy Study (PIRLS; Mullis, Martin, Kennedy & Foy, 2007) provides information about the reading achievement of fourth graders in 35 countries around the world. Comparing Germany's results with those of all other participating countries, they can be found in the upper middle part of the rank list. More alarming is that about 13.2% of the German students do not possess basic reading abilities, which means that they are not able to fulfill the demands of any secondary education (Bos, Lankes, Prenzel, Schwippert, Walther & Valtin, 2003, p. 118). On the opposite, 10.8% of the German students are classified as excellent readers. These differences in reading abilities of fourth graders should be taken seriously, as they may result in severe inequalities concerning economical, political, cultural and social conditions in the students' further lives (Bos et al., 2003). Thus, some effort seems indicated to raise the overall competence level and especially to maintain lower performing students. But obviously, before targeted methods for supporting students can be developed, the students' abilities have to be measured adequately.

In this section two model approaches which provide qualitative classifications of students abilities are presented. As it will turn out, the approaches are not only different in their statistical nature, but they also presuppose various underlying concepts of reading. Thus a quantitative comparison of the statistical models also includes an empirical validation of the different reading theories. The possibility of evaluating theoretical competence models (here: concepts of reading) empirically is rather new.

4.2 Theory

4.2.1 The preferred approach: Rasch model

At present, the preferred method to model dichotomous responses in educational tests is the Rasch model (RM; Rasch, 1960). Belonging to the family of Item Response Theory (IRT) models (e.g. Van der Linden & Hambleton, 1997), the RM delivers a unidimensional quantification of the items' difficulties and the respondents' abilities through real-valued parameters located on the same continuous latent scale.

Let $X_{ij} \in \{0, 1\}$ be the dichotomous response of student i, $i = 1, \ldots, I$, to item j, $j = 1, \ldots, J$. The probability of student i to correctly respond item j is given by

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)},$$

where θ_i is the ability parameter of student i and β_j the difficulty parameter of item j. Since item difficulty and student ability parameters are located on the same latent scale, only the difference between the two parameters is utilized for determining the response probability. This allows for ordering and comparing individual students and items with respect to their ability or difficulty, respectively: If, for example, $\beta_1 = 1.1$ and $\beta_2 = -0.5$, then item 1 is 1.6 Logits more difficult than item 2 for all respondents. If $\theta_1 = 0.3$ and $\theta_2 = 2.6$, student 2 is located 2.3 Logits above student 2 on the ability scale, irrespective of the chosen items. Moreover, mutual inferences between the student and item parameters can be drawn: For example, student 1 with ability $\theta_1 = 0.3$ will master the items with difficulty $\beta_j < 0.3$ with a probability exceeding .5, while the student's probability to master items with difficulty $\beta_j > 0.3$ is lower than .5. For an illustration see Figure 4.2.1.

Albeit the parameters of the RM are quantitative in nature, they may be transformed to obtain qualitative diagnostic information, as has been done for example in PIRLS (Martin, Mullis & Kennedy, 2007) or the National Assessment in Education Progress (NAEP) Study (Lee, Grigg & Dion, 2007). Three steps have to be taken (for an illustration see Figure 4.2.2):

(1) Discrete levels of ability are defined by discretizing the continuous parameter scale at cutpoints (benchmarks). The benchmarks are chosen to be the percentiles of the estimated student ability distribution (e.g. .25, .50, .75, .90), which are obtained by evaluating plausible values drawn from the RM. Note that individual person

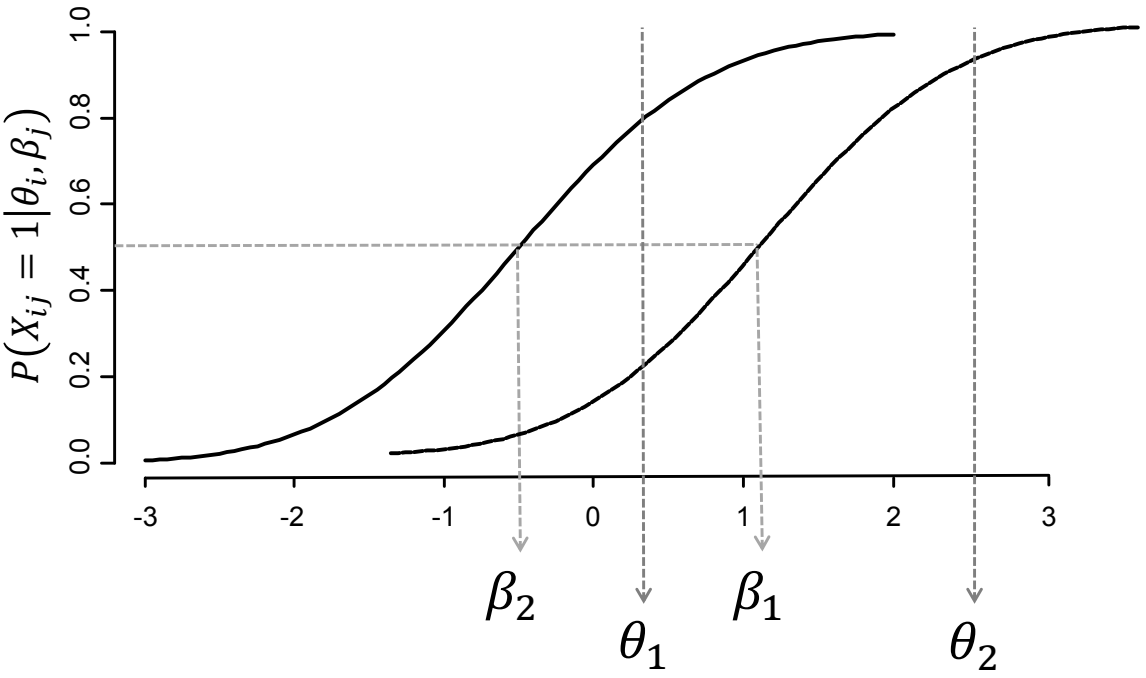

Figure 4.2.1: Illustration of item response curves for two items with difficulties and $\beta_1 = 1.1$ and $\beta_2 = -0.5$ in the Rasch model. The probability that student 1 with ability $\theta_1 = 0.3$ masters item 2 exceeds .5.

parameter estimates (like the weighted likelihood estimate, WLE; Warm, 1989) should not be chosen here: The distribution of the WLEs contains measurement error variance and thus may lead to a biased estimation of the percentiles (Wu, 2005). Defining 4 benchmarks (percentiles) yields 5 levels of ability.

(2) In the second step, we use the fact that both parameters reside on a common latent scale. Hence, discrete levels of difficulty are build by again using the percentiles of the estimated student ability parameter distribution. Similar to the procedure in PIRLS (Martin et al., 2007, Chapter 12) in the present study the 65 percent criterion is used to classify the items into these difficulty levels. That is, the RM item difficulties β_j are transformed into difficulties β_j^* such that a student with ability $\theta_i = \beta_j^*$ correctly solves this item with a probability of .65. Then all items with transformed difficulty parameters β_j^* below the .25 percentile of the estimated student ability distribution (first benchmark) are classified into difficulty level I, items with β_j^* between the .25 and .50 percentile of the estimated student ability distribution (first and second benchmark) are classified into difficulty level II, and so on. Then educational experts (try to) generalize the content of the items in each of the 5 difficulty levels to one type or description of a competence, which

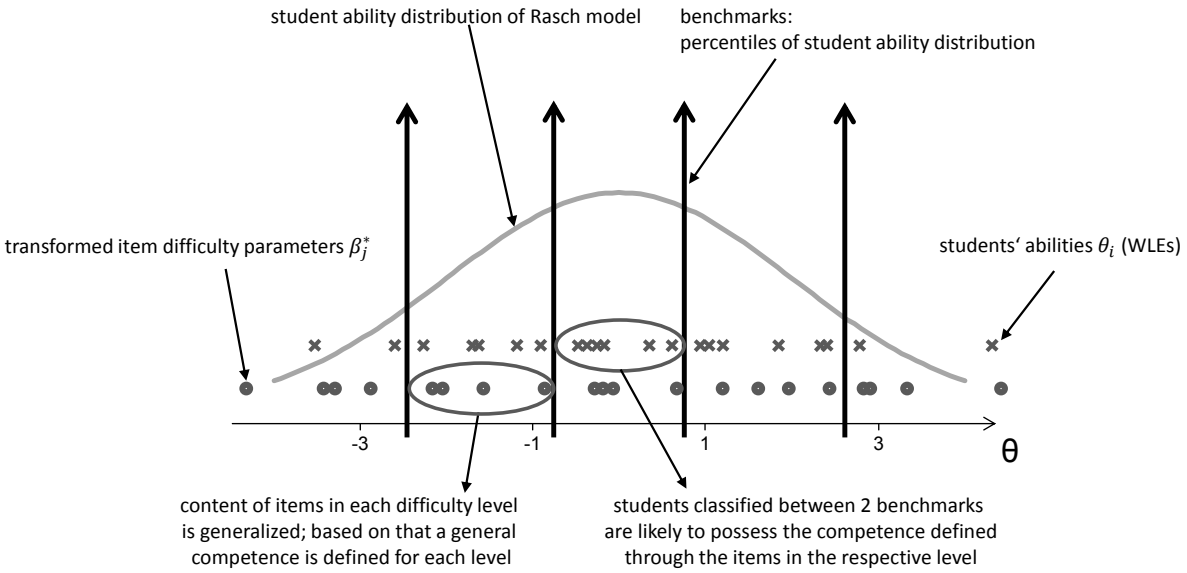

Figure 4.2.2: Illustration of qualitative competence levels obtained through quantitative parameters of Rasch model.

may represent the whole level, e.g. "recognize and repeat explicitly requested information" (Bos et al., 2007, p. 100) describes the first level in PIRLS.

(3) In the third step, switching back from the item difficulties to the students' abilities, the students are classified into the five ability levels according to their WLE ability estimates. For example, a student with estimated WLE student ability parameter below the .25 percentile of the estimated ability parameter distribution (first benchmark) is classified into ability level I, and a student with estimated ability between the .25 percentile and the .50 percentile (first and second benchmark) is classified into ability level II, and so on. Because of step (2) students in each ability level are assumed to possess the before defined general competence of the respective level, e.g. students in ability level I are likely to "recognize and repeat explicitly requested information".

Because for the whole transformation information about the items' difficulties and contents is used the obtained levels are called competence levels: Students classified in competence level I are assumed to be able to "recognize and repeat explicitly requested information" and items classified in competence level I require the students to "recognize and repeat explicitly requested information". It is worthwhile noting that the obtained competence levels are hierarchically ordered, because students in one competence level are likely to solve the items of lower competence levels as well. Sometimes such a hierarchy is called "linear" (de la Torre & Karelitz, 2009), because no bifurcation appears;

however, the term "linear" does not imply a linearity in competence gain across levels.

4.2.2 The not yet well-known approach: CDMs

As already discussed extensively, CDMs allow for the measurement of students' abilities not only on a general ability scale but rather on several basic underlying skills. According to their possession and non-possession of these skills, students are classified into dichotomous skill classes. The skill classes differ in three main aspects from the competence levels obtained through the RM:

(1) The basic skills underlying the general ability are defined through educational experts *before* the estimation of the CDM. On the contrary, in the RM the competence levels are defined based on the estimated model parameters (i.e. the estimated student ability distribution and the estimated item difficulty parameters).

(2) In CDMs educational experts define the skills which are required for the mastery of each item in the so-called CDM Q-matrix (Tatsuoka, 1984). The Q-matrix normally *does not* include any dependencies or orders between the skills. Each item may request an arbitrary combination of skills, see for example Q-matrix $\boldsymbol{Q}_1$. On the contrary, the Q-matrix *may* include hierarchies between the skills, see for example $\boldsymbol{Q}_2$. This Q-matrix assumes a linear hierarchy between the skills: Skill α_4 is the most difficult one as items requesting this skill presuppose the possession of skills α_1 to α_3. In the RM the competence levels are as model inherent hierarchically ordered, e.g. the mastery of an item with difficulty parameter in competence level III also requires students to master competence levels I and II.

$$\boldsymbol{Q}_1 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ \dots & \dots & \dots & \dots \end{bmatrix} \qquad \boldsymbol{Q}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

(3) In unrestricted CDMs students are classified into skill classes which allow for each combination of possessed and non-possessed skills, i.e. the skill classes assume no dependencies or order between the skills. For an example see Figure 4.2.3 on the left hand side: In an unrestricted CDM which assumes four underlying skills,

students are classified into all $2^4 = 16$ possible skill classes. On the contrary, it is possible and may be reasonable to restrict the applied CDM by classifying students in only these skill classes representing a special order between the skills. In Figure 4.2.3 on the right hand side the students are only classified in skill classes satisfying a linear hierarchical order between the skills.

To put it in a nutshell that means that CDMs do not include a model inherent hierarchy between the skills, but that it is possible to define such hierarchies or dependencies by restricting the model. On the contrary, because the competence levels in the RM are based on real-valued parameters they imply a model inherent hierarchy. The different definitions with respect the order of the skills or competence levels imply that the RM and the CDM model approaches emanate from different theoretical competence constructs: While the RM assumes that reading competencies are hierarchically ordered and hierarchically acquired, the CDM approach is less restrictive and does not assume any order. Thus the quantitative comparison of the two statistical model approaches includes a qualitative comparison of the different underlying reading concepts (cf. Section 4.4.2).

4.2.3 Developing the H-DINA

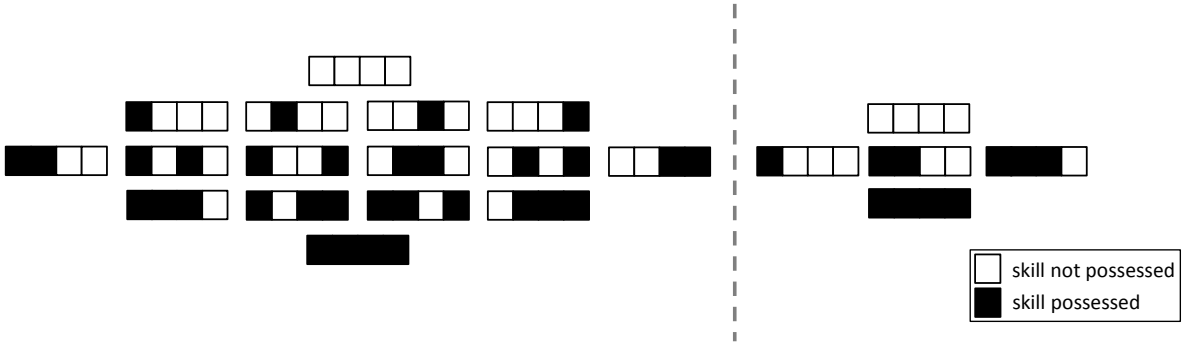

Figure 4.2.3: Illustration of skill classes including no hierarchy or dependencies between skills (left hand side) and skill classes including a linear hierarchy between the skills (right hand side).

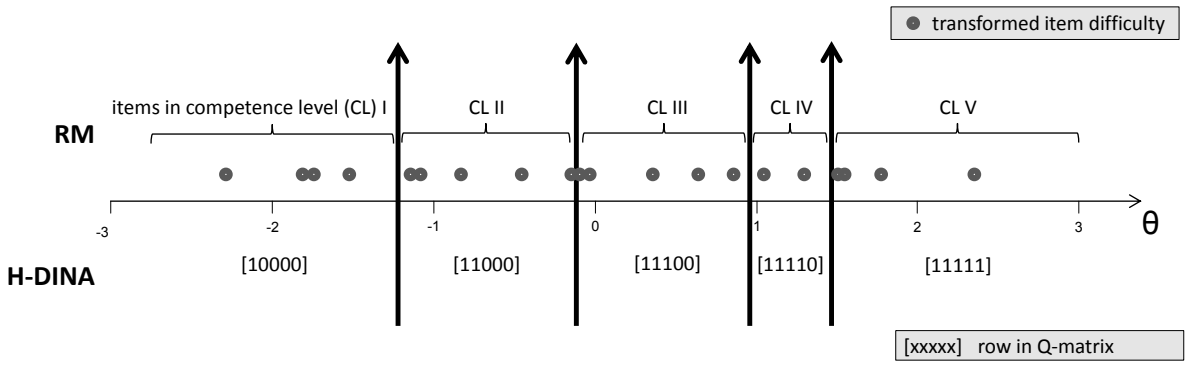

Figure 4.2.4: Q-matrix of H-DINA developed based on item difficulty parameters obtained through RM.

DINA (H-DINA) in the following. The H-DINA will be compared to the RM in terms of general model fit (cf. Section 4.4.2) and in terms of accordance between individual classifications (cf. Sections 4.2.3 and 4.4.2).

For the purpose of developing the DINA, we may assume that each competence level in the RM framework corresponds to a skill in the H-DINA model, e.g. skill 1 requires the students to possess the abilities of competence level I and so on. With this assumption, we achieve comparability by conducting two steps: First, the DINA model has to be constrained in such a way that the skills reflect the linear hierarchical structure of the competence levels in the RM (Groß & George, 2013; Leighton, Gierl & Hunka, 2004). Second, the set of skill classes into which the students are classified has to be restricted to those classes which also include the before defined linear hierarchy between the skills.

Concerning the first step, the H-DINA's Q-matrix needs not to be designed by experts, but evolves already from the model structure and the estimated item difficulty parameters of the RM (cf. Figure 4.2.4). For example, for solving an item with estimated RM item difficulty parameter in competence level I, students only have to master the first skill, which yields to a [10000] row in the Q-matrix of the H-DINA. For solving an item with estimated RM difficulty parameter in competence level III, students have to master skill 3, and, because of the hierarchy assumption, skills 1 and 2. The respective row of the Q-matrix has the entries [11100].

Concerning the second step, the complete set of step $2^5 = 32$ skill classes is reduced to the 6 skill classes satisfying the linear hierarchical structure, namely [00000], [10000], [11000], [11100], [11110], and [11111].

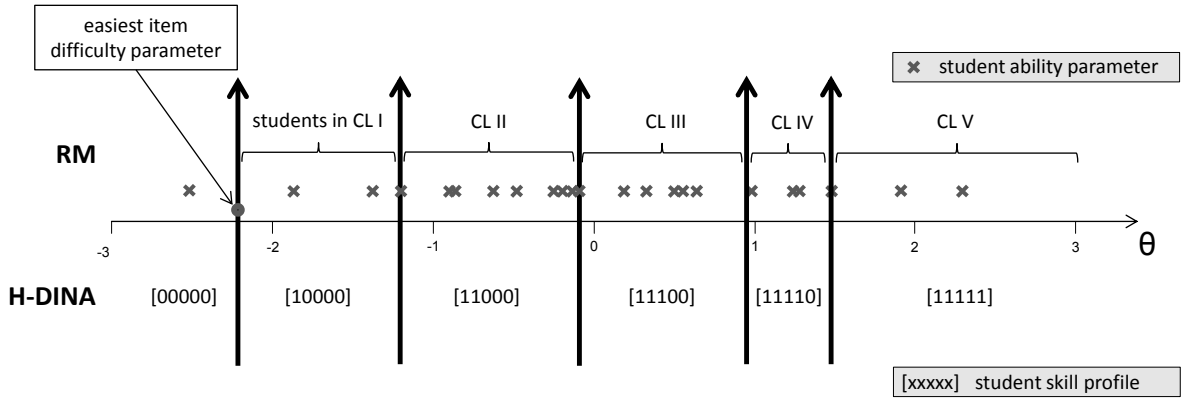

Figure 4.2.5: Transformation of individual competence levels obtained through RM into CDM skill profiles.

Comparison of individual classifications obtained from Rasch and H-DINA

Analogously to the approach of defining the H-DINA's Q-matrix, the individual students' competence levels obtained through the RM are transformed into CDM skill profiles (cf. Figure 4.2.5). For example, an individual student in the RM competence level III possesses the first three skills in the H-DINA model and gets the skill profile [11100] and so on. Furthermore, students having lower RM ability parameters than the easiest RM item parameter are handled analogously to students with skill profile [00000] which possess no skill in the H-DINA model.

4.2.4 State of research

Recently, several studies used CDMs to diagnose reading abilities: Jang (2009, 2008) and Li (2011) applied the Reduced Reparameterized Unified Model (Reduced RUM; Hartz, 2002) to L2 reading comprehension assessments. Kim (2011) also used the RUM to analyze English for academic purposes and Wang & Gierl (2011) and Svetina, Gorin & Tatsuoka (2011) used the Attribute Hierarchy Method (Leighton et al., 2004) or the Rule Space Method (Tatsuoka, 1983), respectively, for examining skills in critical reading. All authors agree in that CDMs can provide more fine-grained diagnostic information about the level of competency in reading than traditional aggregated-test scoring and the authors consequently used this information for feedback systems. Nevertheless, in all studies concerns were raised with regard to the uncertainty in the assignment of reading skills to the test items.

The major difference between the present study and the studies mentioned above is

that the latter work with presupposed competence models. These competence models are assumed to be true without any empirical verification. As opposed to this, the present study considers several different competence concepts reflected by the different conducted statistical models. By means of a quantitative comparison of the statistical model approaches, the connections between the competences (i.e. skills) and thus the competence concepts are empirically validated as well.

The topic of the second research question, i.e. the comparison of individual student classifications obtained through the RM and H-DINA, can be recovered in a simulation study by de la Torre & Karelitz (2009). The present study differs in the following aspects from the study by de la Torre & Karelitz (2009): Firstly the benchmarks are developed in different ways: de la Torre & Karelitz (2009) simulate IRT and CDM models in a way that enables them to use theoretically deduced benchmarks for building the IRT competence levels. In the present study the benchmarks are build upon the percentiles of the estimated student ability distribution. This new approach seems superior for practical applications because the distribution of the person parameters is available in empirical studies in contrast to the theoretical parameter structure applied by de la Torre & Karelitz (2009). The second difference between the two studies is the development of the Q-matrix for the CDM model: While de la Torre & Karelitz (2009) deduce the rows of the Q-matrix from the theoretical design of the simulation, in the present study the rows of the Q-matrix are derived from the estimated RM item difficulty parameters. This again puts the present study in a more practical relevant perspective. Finally, the third difference between the studies is that in the present one the item parameters of the H-DINA are not deduced from the RM item parameters. In the present study the students' individual classifications in both models should only be affected through the underlying competence model.

4.3 Data

As PIRLS is a study on the system level for educational monitoring, individual diagnosis and feedback of reading abilities is not a primary goal. Therefore, the German "PIRLS-Transfer" has been invoked in order to provide individual feedback and training opportunities as well. More precisely, three goals should be achieved: Firstly, to help individual learners in the second and third grade to improve their reading comprehension skills; secondly, to provide a solid empirical base for feedback systems, which inform teachers about the proficiency level in their classes; and thirdly, as a consequence,

to raise the overall reading competence level of German students. By now, the PIRLS-Transfer study contains two test booklets with literary stories (Nahberger, 2010), created according to the principles of PIRLS. Each test booklet consists of 21 items.

The PIRLS-Transfer data analyzed for this study include 153 second graders from the German district North Rhine-Westphalia responding to 21 multiple choice items of the test booklet named "Lockis adventures in the jungle" (Nahberger, 2007). The students' responses were coded dichotomously, i.e. 1 for a correct response and 0 for an incorrect one. Missing responses were allowed. Because pre-analysis showed that four items had a negative discrimination parameter (in the sense of the IRT 2PL model), the respective items were excluded from the analysis.

The sum score of the 17 remaining items had a reliability of .71, which can be considered to be sufficient. The test (i.e. the distribution of the sum scores) did not exhibit obvious floor or ceiling effects (M= 11.5, SD= 3.1). For legitimating the usage of unidimensional models, the degree of the test's multidimensionality is investigated with an exploratory factor analysis (EFA) based on tetrachoric correlations. The EFA results in five factors, but because 29.5 percent of the total variance is explained by the first factor and the ratio of the first and the second eigenvalue amounted to 2.0, the test could be considered as essentially uni-dimensional (Hattie, 1985). This finding is also confirmed by the following method: The model based reliability Omega total (Reise, Moore & Haviland, 2010) of the EFA model with five factors is .85.[1] Then a Schmid-Leiman transformation (Schmid & Leiman, 1957) is applied to the factor loadings of the EFA to obtain a bifactor model, which includes one general factor and specific factors. 59% of the variance explained by all factors could be attributed to the general factor, which also indicates a dominance of the general factor and provides another argument for using a unidimensional model (Reise et al., 2010). Because the conducted methods neither clearly prefer unidimensional nor multidimensional models, models of both types may be fitted to the data.

The item difficulty parameters for the RM are estimated with marginal maximum likelihood methods (MML) using the R package TAM (Kiefer, Robitzsch & Wu, 2013).[2] Ten plausible values are drawn in TAM for deducing the four benchmarks of the competence levels. The individual person parameters are estimated using WLEs. Classification ac-

[1] Note that because the data is dichotomous, the reliability measure has to be adjusted by the method of Green & Yang (2009).

[2] Note that absolute differences in the estimated item parameters using MML methods or distribution free conditional maximum likelihood (CML) methods (as e.g. implemented in the R package eRm; Mair & Hatzinger, 2007) are smaller than .05. Thus the deviations may be considered irrelevant for the results of this article. We decided to apply the estimation method which is used in PIRLS.

curacy and classification consistency are estimated in a simulation with known item and trait distribution parameters (i.e. the parameters estimated for the PIRLS-Transfer data).[3] The classification accuracy measure could also be assessed analytically by the method of Rudner (2001).

The DINA parameters and the distribution of the skill profiles are estimated with MML methods in the R package CDM (George, Kiefer, Robitzsch, Groß & Ünlü, 2013). For the prediction of the students' individual skill profiles Maximum Likelihood Estimations (MLEs) are used. The CDM package also allows a restriction of the skill profile space to linear hierarchical skill profiles. Classification accuracy and consistency of the MLEs is assessed by simulation (DiBello, Roussos & Stout, 2007) and analytically (Cui, Gierl & Huang, 2012). The simulation is conducted with known guessing, slipping and skill profile parameters (i.e. the parameters estimated for the PIRLS-Transfer data).

4.4 Results

4.4.1 Statistical models and underlying reading theories

In the following different statistical models are fitted to the PIRLS-Transfer data. The models differ in their dimension, the number of assumed skills (competences), the structure between these skills, the number of skill classes (competence levels) in which the students are classified and the structure between these skill classes (cf. Table 4.4.1). Thus each of the models presupposes a different concept of reading.

Rasch model For estimating the RM, the mean of the latent trait distribution is set to 0 and a standard deviation of .94 is obtained. The calculated benchmarks (cf. Figure 4.4.6) for the four competence levels are -0.59 (.25 percentile), 0.04 (.50 percentile) and 0.62 (.75 percentile). Because only 1 of 17 items is classified in each of the originally defined competence levels IV and V, these levels were merged for the analysis into one new competence level IV*, i.e. the .90 percentile is not been taken into consideration. By generalizing the content of the items included in the four levels educational experts defined the four competence levels based on the PIRLS standards: In competence level I students are likely to decode words and sentences, in competence level II students should

[3] As described before classification accuracy is a measure of how well individual students are correctly classified into their true competence levels, whereas classification consistency is a measure for the consistence of the classifications in two parallel test forms with the same items and parameters.

model	♯skills/ competences	skill/competence structure	♯dim	♯skill classes/ levels	skill class/level structure
RM	4	hierarchy	1	5	hierarchy
2PL	4	hierarchy	1	5	hierarchy
H-DINA	4	hierarchy	1	6	hierarchy
UN-DINA	4	hierarchy	1	16	no
1skill-DINA	1	no	1	2	no
4skill-DINA	4	no	4	16	no
3skill-DINA	3	no	3	9	no

Table 4.4.1: Number of assumed skills/competences (♯ skills/competences), structure between these skills/competences, dimension of model (♯dim), number of skill classes/competence levels in which students are classified (♯ skill classes/levels) and structure between those skill classes/levels for models fitted to PIRLS-Transfer data.

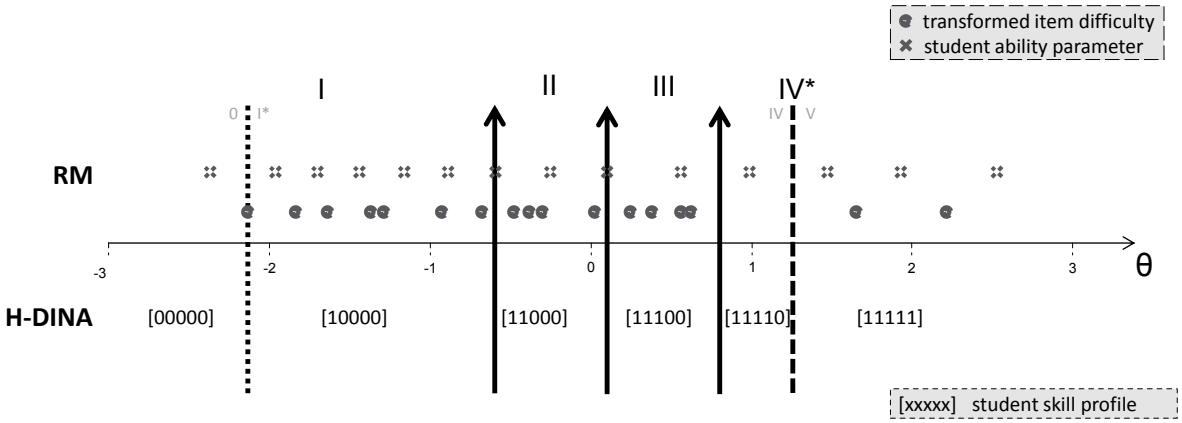

Figure 4.4.6: Competence levels with benchmarks, transformed item difficulty parameters and WLE student abilities for Rasch model on PIRLS-Transfer data.

know how to recognize and repeat explicitly given information, in competence level III students may possess the ability of finding relevant information and deducing simple conclusions and in competence level VI* students are predicted to find central actions and thoughts and to abstract, generalize, and justify preferences. As model inherent these competence levels are linear hierarchically ordered, i.e. students being classified into a higher competence level are assumed to possess the lower levels as well.

The third column of Table 4.4.2 yields the classifications of students into RM competence levels: A large percentage of students (i.e. 25%) is classified into level I, meaning that these students should be able to decode words and sentences, but that they are not very likely to tap into and acquire information from the text.

level		skill profile	RM		H-DINA		♯items
I	0	[1000]∗ [0000]	.25	.04	.20	.19	7
	I*	[1000]		.21		.01	
II		[1100]		.26		.37	4
III		[1110]		.23		.07	4
IV*		[1111]		.26		.36	2

Note: In the RM, classifications in level I are divided up into level 0 (students with estimated abilities lower than the easiest difficulty parameter) and level I* (students with abilities larger than the lowest difficulty parameter but lower than the .25 quantile). In the H-DINA model classifications in the skill classes [0000] and [1000] are merged to be directly comparable to the classification in RM level I.

Table 4.4.2: Relative classification frequencies of students in competence levels (RM) or skill classes (H-DINA) and number of items requesting the competences in each level (♯items) for the PIRLS-Transfer data.

2PL model The IRT 2PL model (Birnbaum, 1968) is fitted to the data as a matter of completeness. Like the RM the 2PL model assumes a linear hierarchical order between the competence levels. The only difference between the RM and the 2PL model is that the 2PL allows individual and thus different item discriminations for each item.

H-DINA The H-DINA CDM is build to take up the model inherent linear hierarchy assumption between the competence levels of the RM and 2PL. The fourth column of Table 4.4.2 yields the population oriented skill class distribution obtained from the H-DINA model: Low frequencies of students were classified into the skill profiles [1000] and [1110]. Students, who possess the first skill seem to possess the second as well, and the possession of the third skill seems to be adherent with the possession of the fourth skill. Relatively large frequencies of students were classified into the zero profile [0000], in which they do not possess any skill.

UN-DINA The so-called unrestricted-DINA (UN-DINA), checks the strength of the hierarchy restriction put on the H-DINA in reversing it. The UN-DINA keeps the assumption of linear ordered skill difficulties (i.e. for the mastery items requesting higher skills the lower skills have to be possessed as well) but it reverses the assumption of the H-DINA that students acquire the skills in a linear hierarchical way (i.e. that they are only classified into the hierarchical skill classes). That is, similar to the H-DINA model, the rows of the UN-DINA's Q-matrix are obtained through the RM item difficulties but,

in difference to the H-DINA model, the students are classified in all $2^4 = 16$ skill classes.

1skill-DINA The 1skill-DINA is a CDM which assumes just one single skill (i.e. the Q-matrix is a vector), which reflects whether a student is capable of reading or not.

4skill-DINA In the Rasch and H-DINA models discussed above a hierarchy between the competence levels is model inherent. From a linguistic point of view, the assumption of a hierarchical graduation of reading competencies has been doubted by Bremerich-Vos (1996). To shed light on this conflict, the 4skill-DINA is build upon a concept of reading, in which the competences are not assumed to be hierarchically ordered.

The before addressed reading concept builds on the cognitive psychology research of van Dijk & Kintsch (1983) and the psychometric approach of Kirsch & Mosenthal (1991) and is used for developing the items for PIRLS (Campbell, Kelly, Mullis, Martin & Sainsbury, 2001). Following this reading concept, the comprehension of texts is understood as a process of information processing, during which readers combine text immanent information with their previous and general knowledge. Finally, reading literacy is split up into four reading processes: α_1 "Focus on and retrieve explicitly stated information", α_2 "make straightforward inferences", α_3 "interpret and integrate ideas and information; make complex inferences" and "examine and evaluate content, language, and textual elements". For notational convenience the reading processes are called skills in the following. The reading skill α_1 "focus on and retrieve explicitly stated information" requires location of information explicitly given in the text, to understand that information and to link it to the question. The skill α_2 requires the reader to make straightforward inferences, that is to carry on thinking about information discussed in the text. Possessing α_3, the reader should be able to make complex inferences and substantiate them by statements given in the text. With skill α_4, readers should examine and critically evaluate contents, language, and textual elements. This is an ability on a meta-level, which requires critical thinking about the text itself. The skills α_1 to α_4 are assumed to underlie no order or structure, especially no hierarchy, as it is possible to construct items of different difficulty for each of the processes.

Every item in PIRLS is based on exactly one of these four reading skills. As PIRLS-Transfer is created according to the same principles as PIRLS, PIRLS-Transfer is also based on the same reading concept and thus each PIRLS-Transfer item is based on exactly one of the four reading skills as well. Which reading skill is required to master the items is summarized in the 18×4 expert Q-matrix, in which the rows reflect no order

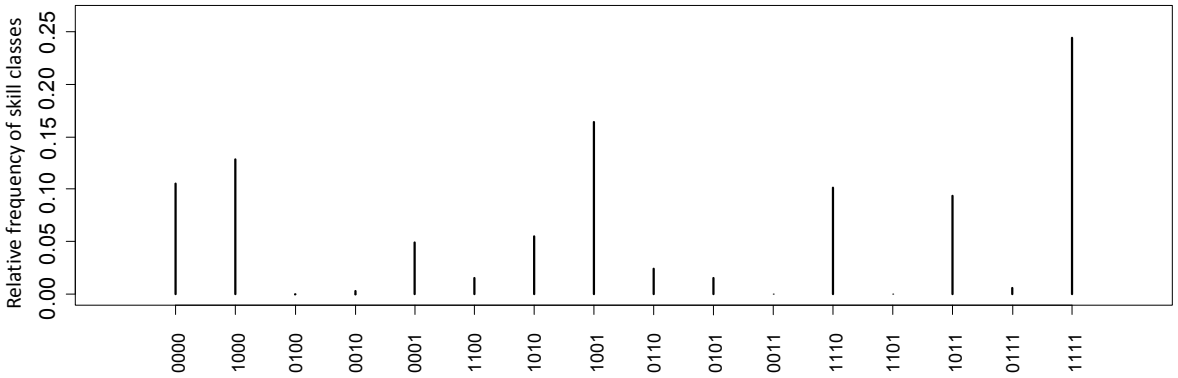

Figure 4.4.7: Skill class distribution of the 4skill-DINA: Some non-hierarchical skill classes like [1001] and [1011] prohibit large frequencies.

	α_1	α_2	α_3	α_4
α_1	.80			
α_2	.45	.44		
α_3	.61	.96	.53	
α_4	.47	.34	.35	.51

Table 4.4.3: Marginal probabilities of skills (diagonal elements) and tetrachoric correlations between skills in the 4skill-DINA for the PIRLS-Transfer data.

between the skills (contrary to the assumption in the H-DINA model). The respective multidimensional DINA, which represents the reading literacy concept (4skill-DINA), allows classification of students into all $2^4 = 16$ possible skill classes. In the 4skill-DINA skill α_1 is measured by 9 items, α_2 by 4 items, α_3 by 3 items and α_4 by 1 item. The average proportion correct values of the items (item p values) measuring skill α_1, α_2, α_3 and α_4 are .80, .50, .57 and .59, respectively, which means that items measuring skill α_1 are the easiest ones.

Figure 4.4.7 shows the population oriented skill class distribution of the 4skill-DINA, i.e. the estimated relative frequencies of the 16 possible skill classes. As can be seen, the non-hierarchical skill classes [1001] and [1011] have large frequencies as well. In both profiles students do not possess the skill α_2 "make straightforward inferences" but they are able to evaluate the text α_4. In total, 31% of the skill classes do not represent a hierarchical linear order. Altogether the probability of mastering skill α_1 is .80, skill α_2 is mastered with probability .44, α_3 with .53 and α_4 with .51 (cf. Table 4.4.3), which contradicts the assumption of hierarchical item difficulties in that items measuring skill α_2 are easier than items measuring skill α_4. This result is in accordance to the calculated item p values.

3skill-DINA By inspecting the association of the latent skills obtained in the 4skill-DINA, tetrachoric correlation coefficients between .34 and .61 are found (cf. Table 4.4.3). This indicates that the skills in the 4skill-DINA model are only moderately correlated. However, the high correlation of .96 between the skills α_2 and α_3 forms an exception, signaling that these two skills are hard to distinguish. For this reason, the 4skill-DINA was adapted in merging the skills α_2 and α_3, leading to a reduced three dimensional DINA model (3skill-DINA).[4] Like the 4skill-DINA, the 3skill-DINA assumes no order or structure between the three skills.

4.4.2 Model comparison

The different IRT models and CDMs (with their different underlying concepts of reading) are compared through likelihood ratio tests (if the models are nested) and through the information criteria AIC (Akaike, 1973) and BIC (Schwarz, 1978). These values describe a general population oriented model fit but do not specify how well the models perform in terms of individual classification. For the RM and the H-DINA the latter is analyzed by classification accuracy and consistency measures.

Hierarchical CDMs: H-DINA and UN-DINA The question behind the comparison of the H-DINA and the UN-DINA is the following: Given the assumption of linear hierarchically ordered skills (in both models), do students also acquire the reading skills in a linear hierarchal way (H-DINA) or not (UN-DINA)?

Figure 4.4.8 shows the population oriented skill class distributions of the H-DINA and the UN-DINA. There are only small differences between the skill class distributions or, more precisely, only 8% of the students in the UN-DINA are not classified in skill classes with a linear hierarchical order. Because the UN-DINA allows for a unrestricted classification of the students it has a significantly better fit than the H-DINA ($\chi^2(11) = 22.97$, p= .018). Nevertheless, because the hierarchy assumption in the H-DINA poses little restriction on the skill class distribution and, in addition, the H-DINA needs a lower number of parameters, one may prefer the H-DINA. Such a decision would be supported by the small difference between the AIC values of the two models (cf. Table 4.4.4). That is, given the assumption that the skills are linear hierarchical ordered, most students seem to acquire the skills in a linear hierarchical form as well.

[4] The high correlation is in line with the expert disagreement when assigning skills α_2 and α_3 to the items. Consider again, that the two skills cover the aspects "straightforward inferences" α_2 and "complex inferences" α_3, which are traceably difficult to distinguishing.

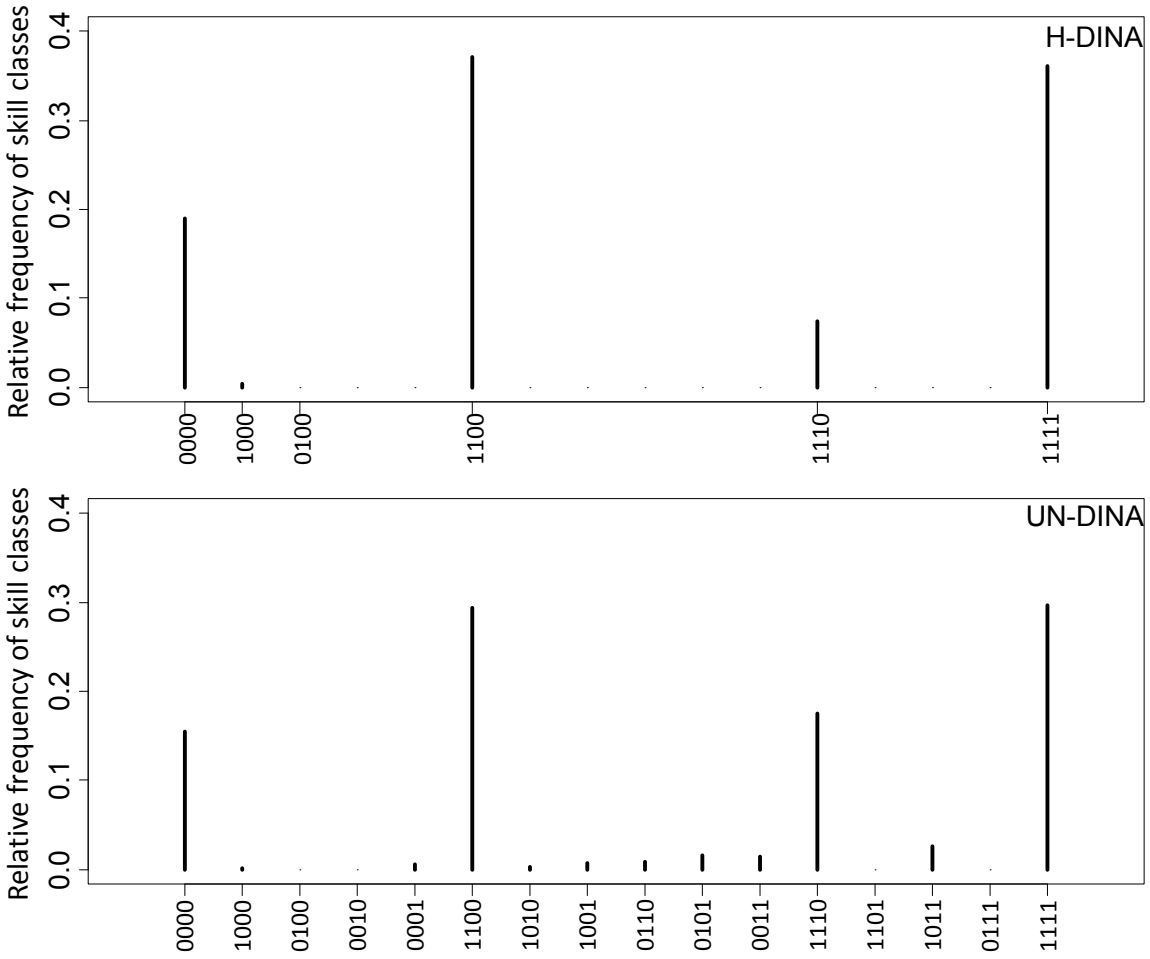

Figure 4.4.8: Population oriented skill class distributions of the H-DINA (top) and the UN-DINA (bottom) for PIRLS-Transfer data.

Hierarchical models: H-DINA and RM The H-DINA is build to reproduce the assumptions of the RM. The question behind the comparison of the two hierarchical models is if one may be preferred in terms of model fit (loglikelihood, AIC and BIC) or in terms of individual classifications (classification consistency and accuracy): With respect to the model fit, the H-DINA performs better in terms of the loglikelihood and the AIC, while the RM performs better in terms of the BIC. Note that the RM's low BIC value may be explained by the model's low number of parameters (compared to the other models' numbers of parameters). With respect to classification consistency and accuracy measures, the H-DINA model turned out to be more reliable: Whereas a simulation confirmed the H-DINA model a classification accuracy of .80 and a classification consistency of .67, the RM exhibited only a moderate accuracy of .58 and consistency of .49. Nevertheless, one has to consider that the accuracy and consistency measures rely on

Model	♯dim	♯par	loglike	AIC	BIC
RM	1	18	-1325.70	2687.39	2741.94
2PL	1	34	-1307.56	2683.12	2786.15
1skill-DINA	1	35	-1312.94	2695.88	2801.94
H-DINA	1	38	-1300.25	2676.51	2791.66
UN-DINA	4	49	-1288.78	2675.55	2824.04
4skill-DINA	4	49	-1296.39	2690.77	2839.27
3skill-DINA	3	41	-1298.11	2678.21	2802.46

Table 4.4.4: Number of dimensions (♯dim), number of parameters (♯par), value of log-likelihood (loglike), AIC and BIC for the models fitted to the PIRLS-Transfer data.

the assumption that the data is generated by the particular examined model.

Non-hierarchical CDMs: 1skill-DINA, 3skill-DINA and 4skill-DINA The comparison of the non-hierarchical CDMs is targeted at the dimension of the reading literary concept: Can students' reading abilities be described by only one general reading skill (1skill-DINA), the four reading processes (4skill-DINA) suggested by Campbell et al. (2001) or by three reading skills, which evolved from merging the second and third reading process (3skill-DINA)? All models are build under the assumption that the skills are not hierarchically ordered.

Likelihood ratio tests show, that the H-DINA model ($\chi^2(3) = 25.37$, $p < .001$) and the 4skill-DINA model ($\chi^2(14) = 33.10$, $p = .003$) fit the data significantly better than the 1skill-DINA model. That is, the data includes more information than only the differentiation of students being able to read or not. A likelihood ratio test for the comparison of the 4skill-DINA and the 3skill-DINA model does not lead to a significant result ($\chi^2(8) = 3.43$, $p = .904$). Therefore, one would not favor the 4skill-DINA, which underlines that three skills are sufficient to describe the students' abilities.

Hierarchical and non-hierarchical models In Table 4.4.4 all considered models are compared in terms of the model fit criteria AIC and BIC. The DINA model including four hierarchical reading skills without a hierarchy assumption in the students' acquisition (UN-DINA) performs best in terms of the AIC, but it is almost indistinguishable from the DINA model with four hierarchical skills and the assumption of a hierarchical acquisition (H-DINA) and from the DINA model representing the reduced reading literacy concept with three unordered skills (3skill-DINA). The BIC favors the RM because of its low number of parameters.

The remaining essential question is whether to prefer the H-DINA or the 3skill-DINA, with both models performing almost equally well. One may tend towards the 3skill-DINA because it is based on a fundamental reading theory and incorporates the information thereof. Because no clear empirical evidence for (H-DINA) or against (3skill-DINA) the hierarchy assumption was found, the present analysis is nondistinctive in whether the acquisition of reading competencies should be seen as a hierarchical process or not (in favor: Erikson (1950), Inhelder & Piaget (1958); against: Bremerich-Vos (1996)). It may be possible to find stronger empirical evidence for one of the two directions if, for example, each item is measured by the same number of skills, the items exhibit a high discrimination and mediator effects are controlled. If these and other aspects are considered in the test construction (cf. Henson & Douglas, 2005), possible problems in the estimation and classification process may be reduced (cf. de la Torre, 2009; de la Torre & Douglas, 2008; Rupp & Templin, 2008a).

Comparison of individual classifications obtained from RM and H-DINA

For the comparison of the students' classifications obtained through the RM and the H-DINA model the students' RM competence levels are transformed into skill profiles. Then the relative frequency of WLEs in each RM competence level is compared to the relative frequency of individual MLE classifications in the appropriate CDM skill class. For the comparison the skill classes [0000] and [1000] are merged to one new class [1000]* because the benchmark between [0000] and [1000] depends of the easiest test item.

The differences between the relative classification frequencies are relatively small in that they reach from .01 (levels II and IV*) to .07 (levels I and III). More detailed, for each of the 153 students it is analyzed into which level the student is classified in the RM and in the H-DINA model (cf. Table 4.4.5). For example, out of the 46 students being classified into competence level I, 32 students are classified into the appropriate skill class [1000]. Altogether, 90 of 153 students (59%) are classified within the same level in the RM and the H-DINA and 135 of 153 (88%) students are classified into the same or an adjacent level. The chance corrected kappa agreement measure is .44 ($z = 9.22$, $p < .01$), which signals a rough correspondence between the results of the two classifications methods. In general, the RM leads on average to lower competence levels (M=2.42) than the H-DINA model (M=2.56). However, a Wilcoxon matched pairs signed ranks test (Wilcoxon, 1945) did not reveal significant differences of the average levels of the RM and the H-DINA classification ($p = .06$).

On the whole, there seems to be no strong correspondence between Rasch and H-DINA

	Skill profile				
Level	[1000]	[1100]	[1110]	[1111]	marginal
I	**32**	*11*	2	1	46
II	*2*	**22**	*10*	9	43
III	0	*6*	**6**	*6*	18
IV*	0	6	*11*	**29**	46
marginal	34	45	29	45	153

Note: Out of the 46 students being classified into level I with the RM, 32 are classified into the skill profile [1000] in the H-DINA model, 11 into [1100], 3 into [1110] and 1 student into [1111]. The bold numbers signalize students classified within the same level in the RM and the H-DINA, italic numbers represent students classified in an adjacent level.

Table 4.4.5: Differences between individual classifications in RM levels (WLEs) and H-DINA skill profiles for all 153 students in the PIRLS-Transfer data.

classifications: Either in fitting both models to the PIRLS-Transfer data dependences between the population oriented skill class distributions were found (cf. Table 4.4.2), nor the comparison of the individual student classifications indicated a clear dependence (cf. Table 4.4.5). In the light of the literature, which suggests interpreting kappa values larger than .60 as agreement, the kappa value of .44 indicated no strong correspondence. Furthermore, the number of students classified in adjacent levels may not be overvalued, since for the PIRLS-Transfer data students are only classified into four levels. Moreover, increasing the sample size and the number of items in an additional simulation study yield no improvement in the classification agreement. Hence, the result appears to hold and may therefore not be attributed to sampling errors

4.5 Discussion

The first goal of this study was to build different statistical models, which all describe different theories about the acquisition of and structure between reading competences. By quantitatively comparing the statistical models the different underlying reading concepts are empirically validated. Altogether 7 models were analyzed: The H-DINA model with a linear hierarchical ordering of the skills and a DINA model which represents a three dimensional reading concept (3skill-DINA), yielded similar results in terms of the AIC. One may prefer the 3skill-DINA model because of its theoretical foundation. If reading is postulated to be a multidimensional concept, then reading competencies should

consequently be modeled with multidimensionaly (Goldstein, 1979).

Nevertheless, as noted by Holland (1990) the distinction between uni- and multidimensional item response models is difficult to evaluate in terms of likelihood based information criteria. If the uni-dimensional RM holds for a population of students it is nevertheless possible that individuals do not fulfill its assumptions. Students with this kind of person misfit may be characterized by deviating person response functions (Sijtsma & Meijer, 2001) which then may result in different person discriminations (Ferrando, 2004). The emerging variation in the person response functions and person discriminations may then be better described by a multidimensional model instead of the initial envisaged uni-dimensional one. In the multidimensional model not every student has to fulfill the assumptions of the uni-dimensional one, as in our case, in which not every student of the multidimensional UN-DINA has to fulfill the hierarchal acquisition of skills assumed in the uni-dimensional H-DINA model. Thus, there may be situations in which a multidimensional model (as the UN-DINA) is preferred to the uni-dimensional variant (the H-DINA) in terms of model fit because of relevant person misfit.

Another aspect of the model selection is that likelihood based approaches (like deviance tests and information criteria) may prefer high-dimensional item response models with low reliability of individual subscores (Haberman, 2008). This holds especially for CDMs. However, as emphasized in the "bandwidth-fidelity-dilemma" (DiBello, Roussos & Stout, 2007; Feldt, 1997), a decrease in reliability can sometimes be compensated by an increase in subscore validity. Therefore it has to be underlined that a unified perspective of reliability and validity for the assessment of statistical models and their use of test score definitions is needed (Kane, 1982). It may weaken the relevance of likelihood-based model selection. For further investigations about the sensitivity of various fit statistics for absolute or relative fit the discussion of Chen, de la Torre & Zhang (2013) should be considered.

The selection of the 3skill-DINA is fundamentally based on the model's Q-matrix. It has to be acknowledged that a different expert definitions of the Q-matrix might have led to different results. The structure of the Q-matrix is a crucial part of the model specification as it relies on valid expert judgments (Rupp & Templin, 2008a; Templin, Henson, Templin & Roussos, 2008). On the other hand, the RM assumes uni-dimensionality and parallel item response functions, which are known to be hard to achieve as well.

An important aspect in the discussion about a hierarchical or non-hierarchical acquisition of reading competencies seems to be the effect of mediators. From a linguistic point of view, acquisition of reading competences is often analyzed with theories stating a

hierarchy between items on the word-, the sentence, and the text-level (Bredel & Reich, 2008). Thus, further studies might analyze if the often postulated hierarchy between the reading processes (Campbell et al., 2001) only results because the easier items are located on the word level, whereas difficult items are mainly found on the text level.

The second goal of the present study was to compare individual student classifications resulting from the RM and a special DINA model, the H-DINA, which satisfies the assumptions of the RM competence levels concerning their dimensionality and their linear hierarchical ordering. Neither the population oriented skill class distributions (cf. Table 4.4.2) nor the individual classifications (cf. Table 4.4.5) showed a conspicuous accordance between the classifications. The lack of accordance may be traced back to the different theoretical fundaments of the two models (e.g. different forms of item response functions, cf. Chapter 5.3.4). This result is in accordance with a simulation study by de la Torre & Karelitz (2009), although a different definition of the benchmarks and a different concept for building the Q-matrix was developed. In a next step might it be meaningful to fit a 2PL model to the data because the item response functions of a 2PL model and the H-DINA have more similarities (e.g. both allow for different item discriminations) than those of the RM and the H-DINA. In the present study the RM was conducted, because most educational tests are scaled with this model.

5 Analyses of background data with CDMs

5.1 Problem

> "Because the school education effects the labor market participation, the vocational mobility and the quality of life, all countries insist on reducing differences caused by the educational system (OECD, 2001, p. 144)."

Obviously, before methods can be developed to reduce such differences, the differences firstly have to be uncovered. To reach this goal, in most large scale studies researchers attach student background questionnaires, in which student oriented context variables (also called background variables) like for example gender, age or migration status are captured. More precisely, the background questionnaires satisfy three tasks:

(1) *Descriptive task:* Background questionnaires provide information about the existence and extent of the student context variables, e.g. the percentage of females in the test population. In a subsequent step relational measures, i.e. factors, have to be found which describe the relationship between school achievement on the one hand and the student oriented context variables on the other hand (PIRLS framework model: Bos, Valtin, Voss, Hornberg & Lankes, 2007; PISA Konsortium, 2003).

(2) *Identification Task:* With the information obtained from background questionnaires it is possible to identify subgroups. The gained knowledge about conceptual dissimilarities between subgroups yields the possibility to initiate remedial actions, which themselves may stabilize the equality of educational opportunities. Additionally the reduction differences between subgroups may enhance the quality of schools or even of the educational system.

(3) *Explanation task:* For enhancing the quality of schools it is essential to find explanations for differences in student achievement between schools and between

classes with comparable determining factors. In the "fair comparison" the students' extracurricular situation (described through the student context variables) is considered as a factor which influences the students' achievement but which can not be influenced by the teacher or the school. Thus schools and classes are compared which reached the same level of ability under the same conditions (Ophoff, Koch, Hosenfeld & Helmke, 2006).

The common standard procedure for the descriptive and the identification task is to analyze the influence of the student context variables on a general ability, e.g. reading in PIRLS or math in TIMSS (cf. e.g. Bos, Valtin, Voss, Hornberg & Lankes, 2007; Mullis, Martin, Ruddock, O'Sullivan, Arora & Erberer, 2008; PISA Konsortium, 2003). For example, in many large scale studies (e.g. PIRLS, TIMSS, PISA) the students' migration status is identified as a context variable which has a strong influence on the students' achievement. However, we cannot expect that a context variable has an equal extent on the mastery of each basic skill underlying the general ability. For example, we may assume that migrants exhibit strong deficits in specific mathematical skills which are strongly related to the use of language (i.e. interpretation) while they may perform better in other skills (i.e. calculation). The present chapter yields methods and examples for empirically verifying assumptions about differences in the mastery of underlying skills for specific subgroups of students. These methods also allow specifying differences in skill mastery between specific subgroups of students, which do not have to be of equal extents in each skill. If once the descriptive task of background questionnaires is refined, then in the explanation task more concrete methods can be developed to reduce the existing differences and thus to ensure equal opportunities.

In current results from large scale studies amongst others the following student context variables turned out to be predictors of student achievement: gender, migration background, the parents' educational background, the number of books in the parents' household and the socio economic status (Bos et al., 2007; PISA consortium, 2001, p. 241). Thus in the following chapter comparisons of achievement between subgroups formed through the before mentioned variables are prioritized.

5.2 Data

The data reanalyzed in this chapter consists of the students' responses to a test of mathematics and to a background questionnaire. With a sample size of $I = 71464$ it is a complete survey of all Austrian eighth graders in 2012. Both, the test and

the questionnaire were originally employed in the framework of educational standards testing, i.e. a main goal was to check whether the students reach before defined standard norms of mathematical ability. The test consists of altogether $J = 72$ items arranged in 6 test booklets by a partially balanced incomplete block design (Bose & Nair, 1939). Each individual student responded to the items in one of the test booklets, with each test booklet including a number of 48 items. The test booklets are mutually comparable concerning length, difficulty and content of the items. In the following the test and the data are called BIST-M8 ("Bildungsstandards-Mathe 8"; mathematical educational standards in the eight grade).

Following the competence model of Peschek & Heugl (2007) mathematical ability in the eighth grade can be divided into four operational sub-competencies namely "α_1: model building", "α_2: calculation", "α_3: interpretation" and "α_4: argumentation" and four content sub-competencies namely "α_5: numbers and measures", "α_6: variables and functional dependencies", "α_7: geometry" and "α_8: statistics" [1]. In the present study the four operational and four content subcategories are used as the $K = 8$ basic skills underlying the tested mathematical competence in the eighth grade. According to educational experts, for successfully mastering each of the items students require exactly one operational and one content skill. That is for mastering an item students require one of 16 possible combinations of one operational and one content skill, e.g. they have to possess α_1 in combination with α_7 for mastering item 1 in the first test booklet. Altogether, each skill is required for the mastery of 12 items in each test booklet. The specific combinations of skills required for the mastery of each item are defined in a Q-matrix. As a summary, Table 5.2.1 shows how many items in each test booklet request the 16 possible combinations of one content and one operational skill for their mastery: 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.

In the present study group specific differences in achievement are analyzed with respect to the following variables: gender, migration background, type of school, education of parents, number of books in the parents' household, HISEI index and the federal state in which the student is attending school. As already mentioned, these background variables are taken into account, because, with expectation of the federal state, they turned out to be predictors of school achievement in other larger scale studies like PIRLS, TIMSS and

[1] Note that the four content subcategories of the educational standards test are comparable to the content domains defined in the math test of the Trends in International Mathematics and Science Study (TIMSS; Mullis et al., 2008) for the eighth grade, which are: "numbers", "algebra", "geometry", and "data and chance".

test	α_1				α_2				α_3				α_4				
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

α_1: model building, α_2: calculation, α_3: interpretation, α_4: argumentation, α_5: numbers, α_6: functions, α_7: geometry, α_8: statistics

Table 5.2.1: Number of items requiring a specific combination of operational and content skills in each of the 6 test booklets. 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.

PISA (cf. e.g. Bos et al., 2007). The federal states are taken into account for national comparisons of student achievement, which is a usual procedure in common large scale studies as well (cf. e.g. Bos, Lankes, Prenzel, Schwippert, Valtin & Walther, 2004).

Here the considered variables from the student background questionnaire are presented and their categorizations for the subsequent analyses are introduced:

(1) *Gender:* Gender of the students, male or female.

(2) *Migration background:* Students are defined to have a migration background if both parents are born abroad and no migration background if at least one parent is born in Austria.

(3) *Type of school:* The type of school the students are attending, either AHS (Allgemeinbildende Höhere Schule) or BHS (Berufsbildende Höhere Schule). The AHS may be compared to the German grammar school (Gymnasium).

(4) *Education of Parents:* The parents' education is differentiated into three categories: compulsory school or vocational education, A level and university. The highest education of the parents is taken into account.

(5) *Number of books in parents' household:* The family is regarded as first educational world and most important socializing environment of children, where already from the point of birth, different basic competencies are deposed and formed (cf. e.g. Artelt, McElvany, Christmann, Richter, Groeben, Köster, Schneider, Stanat, Ostermeier, Schiefele, Valtin & Ring, 2007; Bos, Valtin, Voss, Hornberg & Lankes, 2007). The number of books serves as measure for the resources supporting read-

ing and learning in the parental household. Thereby "books" explicitly does not include magazines, newspapers or schoolbooks. The number of books is classified in five categories: 0 to 10 books, 11 to 25 books, 26 to 100 books, 100 to 200 books and more than 200 books.

(6) *HISEI:* The abbreviation HISEI signifies the Highest International Socio-Economic Index of Occupational Status and characterizes the maximal ISEI (International Socio-Economic Index) value of either the student's father or mother. The ISEI is a standardized measure for the socio economic status, which combines information about the profession, the income and the education. The ISEI is evaluated on the ISEI scale (Ganzeboom, De Graaf & Treiman, 1992): High ISEI values characterize a high socio economic status, for example the maximal ISEI value of 90 belongs to a legislator. On the contrary the minimal value of 16 belongs to unskilled laborers in agriculture or fisheries or cleaning personal. In the following analyses the HISEI values are divided into 4 categories: HISEI values below 30, values between 31 and 50, values between 51 and 70 and values above 70.

(7) *Federal State:* The federal state of Austria in which the students attend school: Burgenland (BL), Kärten (K), Oberösterreich (OÖ), Niederösterreich (NÖ), Salzburg (S), Steiermark (SM), Tirol (T), Voralberg (VA) and Wien (W).

Figure 5.2.1 yields a summary of the testpopulation and shows the relative frequencies of students in the different background categories.

5.3 Theory

5.3.1 Official scaling methods for BIST-M8

In the official BIST-M8 analysis (Bruneforth & Lassnigg, 2013) the data is scaled in four steps:

(1) After dichotomizing the student responses, the data is fitted with the Rasch model (Rasch, 1960). For a description of the Rasch model and its parameters see Section 4.2.1 of the present work.

(2) The individual student abilities θ_i, $i = 1, \ldots, I$, are estimated by conducting weighted likelihood estimation (WLE; Warm, 1989).

(3) For evaluating the students' abilities with respect to the before defined, normed and standardized educational performance requirements, the unidimensional Rasch

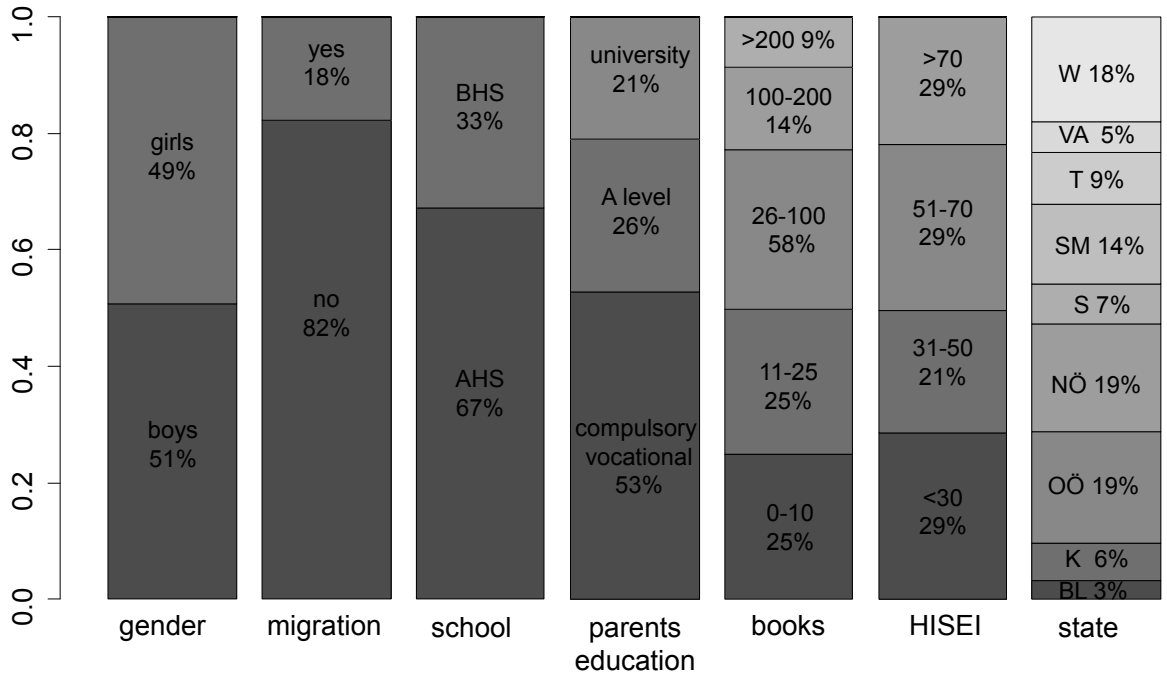

Figure 5.2.1: Percent distribution of students in background categories.

ability scale is discretized into four levels (for details see also Section 4.2.1). The cutpoints, i.e. the benchmarks, between the four levels and the interpretations of the levels are determined through a standard setting procedure (Cizek, Bunch & Konns, 2004). Depending on their individual WLE ability values the students are classified in the before defined four competence levels:

Students below level 1: Students which do not achieve the educational standards. Students are classified below level 1 if their WLE ability values are at most 439.

Students at level 1: Students which partly achieve the educational standards. These students possess basic knowledge in all parts of the math curriculum. They are able to manage reproductive tasks and routine work. Students are classified in level 1 if they have WLE ability values between 440 and 517.

Students at level 2: Students which achieve the educational standards. These students possess basic knowledge in all parts of the math curriculum and are able to use this knowledge in a flexible way. That is they are able to find appropriate strategies for solving the tasks and they are able to describe and justify their approaches. Students are classified in level 2 if they have WLE ability values between 518 and 690.

Students at level 3: Students which outperform the educational standards. These

students do not only possess basic knowledge in all parts of the math curriculum but also expanded knowledge which exceed the requirements of level 2. Particularly these students possess a distinct ability to abstract and to combine. Students are classified in level 3 if they have WLE ability values above 691.

(4) The student achievement in different subgroups is compared. Therefore percent distributions of specific groups of students in the four competence levels are determined. For example, in the group of boys, 16% do not achieve the educational standards, 26% partly achieved the standards, 52% achieved the standards and 6% outperformed the standards. This group specific student distributions are compared: For example if 16% of the boys do not achieve the educational standards and 6% outperform the educational standards, whereas 17% of the girls do not achieve the standards and 5% outperform them, then the comparison indicates that boys perform slightly better than girls.

5.3.2 Methods for reproducing official results with CDMs

The present section describes the statistical methods deployed to reproduce the results of the official group comparisons with CDMs:

(1) The data is fitted with different DINA and G-DINA models and the best fitting model in terms of the global fit indices AIC and BIC is chosen for the subsequent analysis.

(2) The individual CDM student classifications in dichotomous skill profiles are conducted with MLE methods. The MLE classifications may be seen as similar to the individual WLEs used in the official analysis.

(3) The unidimensional achievement scale θ and the benchmarks are recreated by forming three profile groups of students: The first group is characterized by students with low results who solved at most two arbitrary out of the eight analyzed skills. The second group is defined through students who achieved a moderate result, that is students who possess between three and six of the altogether eight skills. Finally, the third group includes students who have revealed a particularly good level, i.e. all students who possess all of the altogether eight skills or at least seven of the skills. Thus, the third group includes students who are classified in skill profiles such as [11111111], [01111111], [10111111] and so on. In forming the groups it was not important *which* skills the students possess but rather *how many* skills they posses. In this step the multidimensional construct of the students' CDM

skill profiles is broken down into a unidimensional construct, which is comparable to the Rasch ability scale. The partition of the skill profiles in three groups mirrors the idea of the benchmarks.

(4) The relative frequencies of students classified in the three profile groups in each specific student subgroup are calculated and compared. Illustrated by the example of comparing the achievement of girls and boys, the question of interest is "how large is the proportion of particularly good students in the group of girls compared to the proportion of particularly good students in the group of boys". That is, the frequencies

$$\begin{aligned} P(\text{ girl } \mid \text{ profile group 3 }) &= P(\text{ girl } |\boldsymbol{\alpha}_l = [11111111]) + P(\text{ girl } |\boldsymbol{\alpha}_l = [01111111]) + \\ &\quad P(\text{ girl } |\boldsymbol{\alpha}_l = [01111111]) + \ldots \\ &= \sum_{l:\sum_k \alpha_{lk} \geq 7} P(\text{ girl } |\boldsymbol{\alpha}_l) \end{aligned}$$

and

$$P(\text{ boy } \mid \text{ profile group 3 }) = \sum_{l:\sum_k \alpha_{lk} \geq 7} P(\text{ boy } |\boldsymbol{\alpha}_l)$$

are compared. With respect to profile group 2 the frequencies

$$\sum_{l:\sum_k \alpha_{lk} \in \{3,4,5,6\}} P(\text{ girl } |\boldsymbol{\alpha}_l) \quad \text{and} \quad \sum_{l:\sum_k \alpha_{lk} \in \{3,4,5,6\}} P(\text{ boy } |\boldsymbol{\alpha}_l)$$

are compared, and for profile group 1 the frequencies

$$\sum_{l:\sum_k \alpha_{lk} \in \{0,1,2\}} P(\text{ girl } |\boldsymbol{\alpha}_l) \quad \text{and} \quad \sum_{l:\sum_k \alpha_{lk} \in \{0,1,2\}} P(\text{ boy } |\boldsymbol{\alpha}_l)$$

are compared.

5.3.3 Refining official results on the skill level

In the following specific groups of students are not only compared on the level of a general ability but rather on the level of skills, i.e. on the level of the four operational and the four content domains. Therefore CDMs with separate skill class distributions and skill mastery probabilities for each group of students are estimated.

At first M groups $G_1, \ldots, G_M$ of students are defined. Each student i, $i = 1, \ldots, I$, belongs to exactly one of these groups, i.e. there exists exactly one m, $m = 1, \ldots, M$,

for each student i with $i \in G_m$. Furthermore, let g_m be the number of students belonging to group G_m and then obviously $\sum_{m=1}^{M} g_m = I$. For example, if the difference in achievement between girls and boys is examined, then $M = 2$ and G_1 corresponds to the group of girls and G_2 to the group of boys. If student $i = 1$ is a boy then it holds $i \notin G_1$ but $i \in G_2$ and accordingly g_1 is the number of tested girls and g_2 the number of tested boys. The procedure for the estimation of the group specific skill class distributions and skill mastery probabilities is similar to the general algorithm presented in Section 1.2.4. The differences to the general algorithm are described in the following:

(1) Based on the probabilities $P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)$, $i = 1, \ldots, I$, $l = 1, \ldots, 2^K$, the model likelihood is defined and the item parameters are estimated (cf. Section 1.2.4 step 1 and 2). According to de la Torre & Lee (2010) the item parameters are assumed to be invariant in the different groups.

(2) In difference to Section 1.2.4 for each group G_m, $m = 1, \ldots, M$, separate starting values $P(\boldsymbol{\alpha}_l|G_m) = \frac{1}{2^K}$, $l = 1, \ldots, 2^K$, $m = 1, \ldots, M$ are defined. Thus, the estimation algorithm starts with a uniform distribution over the probabilities $P(\boldsymbol{\alpha}_l|G_m)$ in each group.

(3) For each of the M groups the probabilities of student i in group G_m to be classified in skill class $\boldsymbol{\alpha}_l$ are calculated:

$$P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i, G_m) = \frac{P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l) \cdot P(\boldsymbol{\alpha}_l|G_m)}{\sum_{l=1}^{2^K} P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l) \cdot P(\boldsymbol{\alpha}_l|G_m)} \quad l = 1, \ldots, 2^K, \ m = 1, \ldots, M.$$

In this step all $2^K \cdot I \cdot M$ probabilities are calculated, i.e. it is irrelevant if student i actually belongs to group G_m.

(4) The group specific skill class distribution in group G_m is defined as

$$P(\boldsymbol{\alpha}_l|G_m) = \sum_{i:i \in G_m} \frac{P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i, G_m) \cdot P(X_i)}{g_m}, \quad l = 1, \ldots, 2^K, \ m = 1, \ldots, M,$$

where the weighted sum is only taken over the students belonging to group G_m. Based on that, the skill mastery probabilities in group G_m are given by

$$P(\alpha_k|G_m) = \sum_{l:\alpha_{lk}=1} P(\boldsymbol{\alpha}_l|G_m), \quad k = 1, \ldots, K, \ m = 1, \ldots, M.$$

Note that currently there exists only one common multiple group approach for CDMs, namely the one introduced in Xu & von Davier (2008). In contrast to the procedure

presented above, their approach assumes different item parameters per group. It is a legitimate subject for a debate, whether the difference in the item parameters also retains differences between the groups. We decided to assume invariant item parameters because of the procedure in the more common IRT framework: Here the comparison of different tests (e.g. pre and post test) or different groups is conducted with the help of so-called linking items which are assumed to have the same item parameters (i.e. difficulty and discrimination) in both tests or groups (cf. item linking and calibration, e.g. Kim & Cohen, 1998).

5.3.4 Discussion: CDM or M-IRT?

It should be noted here that in contrast to the obvious differences between the IRT Rasch model and the CDM DINA model (cf. Chapter 4), the differences between multidimensional item response models (M-IRT; cf. e.g. de Ayala, 2009) and CDMs are less striking. Instead of a CDM analyzing K skills one may also apply a K-dimensional M-IRT model in which each dimension describes one skill. Then CDMs and M-IRT models may be compared in the following aspects:

(1) *Q-matrix:* Comparably to the concept of the Q-matrix in CDMs it is also possible in M-IRT models to define on which dimension or dimensions an item loads (cf. e.g. Chalmers, 2012; Reckase, 2009)

(2) *Item response functions:* While in CDMs the item response functions are stepfunctions, in M-IRT models they have a logistic form. A major difference can be found between the logistic form of M-IRT item response functions and item response functions of non-compensatory CDMs, which exhibit only two levels (e.g. guessing and slipping in the DINA model).

(3) *Response probabilities:* In both models it is possible to calculate the students probabilities to master the different skills/dimensions based on their discrete/continuous K- dimensional vector of abilities.

(4) *Skill class distribution:* Comparable to the discrete skill class distribution in CDMs in M-IRT models a continuous ability distribution may be determined by evaluating plausible values (cf. Chapter 4 for the unidimensional case).

(5) *Individual classification:* The discrete individual classifications are directly obtained from a CDM. Similar classifications may also be obtained in an indirectly way from M-IRT models: After estimating the M-IRT model the K continuous

ability scales can be discretized at cutpoints (cf. Chapter 4 for the unidimensional case). The determination of these cutpoints may be challenging and needs expert driven methods like standard setting procedures.

Up to now specific studies investigating the differences between CDM and M-IRT models have been rare in number. Kunina-Habenicht, Rupp & Wilhelm (2009) compare skill mastery probabilities obtained from both model approaches based on an empirical data set. The authors of this article found no substantial differences between a variant of the GDM and a comparable M-IRT model. They emphasize the way of directly obtaining individual student classifications from CDMs as a feasible advantage. However, an accurate analysis of the differences between both models is still missing. Especially the differences between non-compensatory CDMs and M-IRT models may be of interest.

In fact it cannot be ruled out that the results of the group specific differences on the skill level obtained with CDMs may also be received through a comparable M-IRT model. Still (a) the procedure for multiple group models presented in this chapter is new and differs from the common approach, (b) CDM applications of multiple group models with background data have so far been sparse and (c) the BIST-M8 data has not yet been analyzed neither with CDMs nor with M-IRT models.

5.4 Results

5.4.1 BIST-M8 results

The results presented in a graphical or descriptive way in this section are taken from the BIST-M8 2012 (Schreiner & Breit, 2012) report about the educational achievement of Austrian eighth graders in 2012. Figure 5.4.2 shows the percent distribution of students in the four levels "educational standards not achieved", "educational standards partly achieved", "educational standards achieved" and "educational standards outperformed". Each bar represents students belonging to a different subgroup of students. The figure only includes the percent distributions of groups which are numerically documented in the BIST-M8 report. In the following all official results (i.e. the results given numerically and also those described in the text) obtained for the group comparisons of interest are documented:

(1) *Austria:* As a national result, in Austria 17% of the eighth graders do not achieve the educational standards in math, 26% partly achieve the educational standards,

52% achieve the standards and 5% of the students outperformed the standards.

(2) *Gender:* As can be seen in Figure 5.4.2 there is hardly a difference between the results of boys and girls. There are few more boys which outperformed the standards.

(3) *Migration background:* Students with migration background (cf. Figure 5.4.2) are clearly more often located below level 1 (35%) than students without migration background (13%).

(4) *School:* Students attending different types of schools (cf. Figure 5.4.2) also exhibit an apparent difference in achievement. Students attending BHS are much more often classified below level 1 (24%) than students attending AHS (1%).

(4) *Education of parents:* As stated in the BIST-M8 report (p. 44) parents of students classified below level 1 mostly exhibit a compulsory school graduation or a vocational training (28 %). On the contrary, parents of students which outperformed the educational standards mostly exhibit a university graduation (52%). Here an apparent relation between the education of the parents and the achievement of the students is captured.

(5) *HISEI:* As also stated in the BIST-M8 report (p. 45) students classified below level 1 exhibit clearly lower mean HISEI indices than students located in level 3.

(8) *Federal State:* As can be seen in Figure 5.4.2 the state of Wien exhibits the largest percentage of students below level 1 (25%) compared to the other federal states. On the contrary, the state of Oberösterreich exhibits the largest percentage of students which outperformed the educational standards (6%).

5.4.2 Reproduction of official results with CDMs

In a first step the response data is fitted with four different CDM models: A DINA model considering the four operational and the four content skills (full-DINA), a DINA model only analyzing the content skills (content-DINA), a DINA model taking only the operational skills into account (operational-DINA) and a G-DINA 1way model with content and operational skills (G-DINA 1way). In Table 5.4.2 the four models are compared in terms of the goodness of fit measures AIC and BIC and the mean item fit RSMEA (cf. Section 2.4.4). The G-DINA model shows the best fit in terms of the AIC and BIC, but it provides the highest (i.e. worst) mean RSMEA value of all models. The two DINA models which only include the operational or the content skills exhibit a

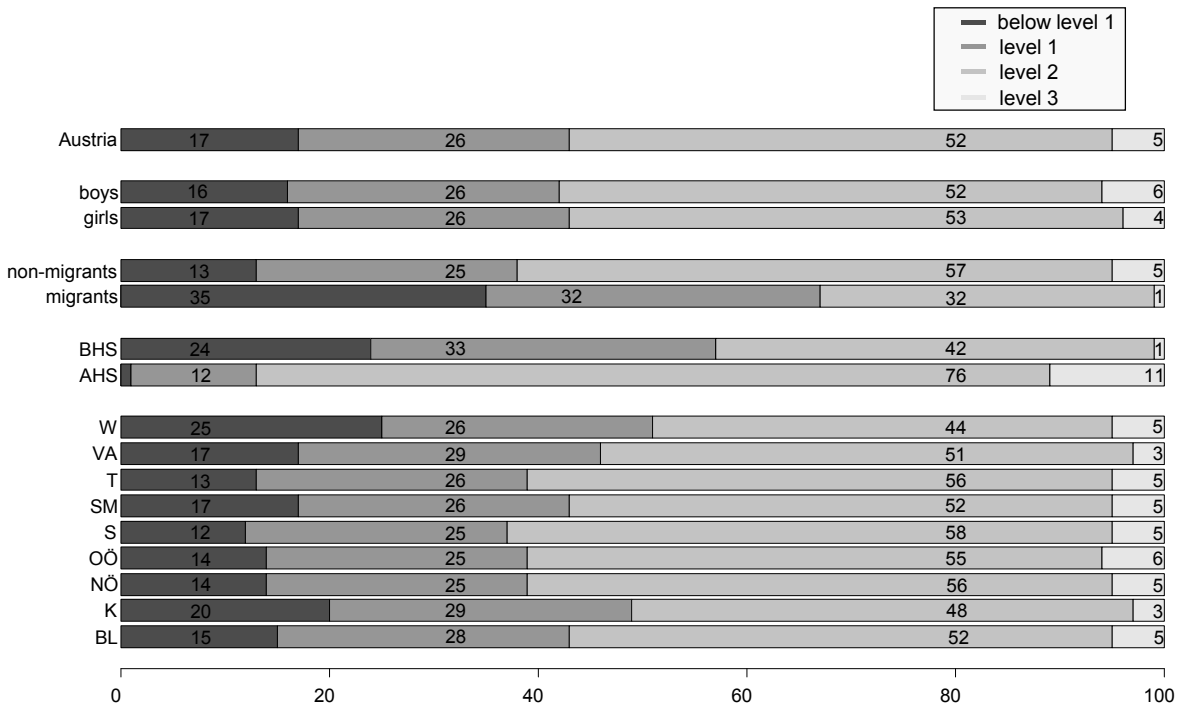

Figure 5.4.2: Percent distribution of students in competence levels obtained from official BIST-M8 2012 analysis.

worse model fit in terms of the AIC and BIC than the full DINA and the G-DINA, while they have the same mean RSMEA as the full DINA. As a compromise between model and item fit the BIST-M8 data is analyzed with the full-DINA model in the following.

Figure 5.4.3 shows the skill mastery probabilities of the eight mathematical skills obtained from the full-DINA model for the BIST-M8 data: Among the operational skills Austrian eight graders possess the skill "α_3: interpretation" best (with a probability of 72%) while they master the skill "α_4: argumentation" worst (54%). Among the content skills the students possess "α_8: statistics" best (69%) and "α_5: numbers" worst (60%). Altogether the students exhibit most problems in mastering "α_4: argumentation". This

model	competence	AIC	BIC	mean RSMEA
full-DINA	content, operation	3491861.01	3495522.61	0.07
content-DINA	content	3501439.87	3502899.01	0.07
operational-DINA	operation	3505077.87	3506537.00	0.07
GINA 1way	content, operation	3429253.48	3431575.25	0.10

Table 5.4.2: AIC, BIC and mean RSMEA for CDM models fitted to BIST-M response data.

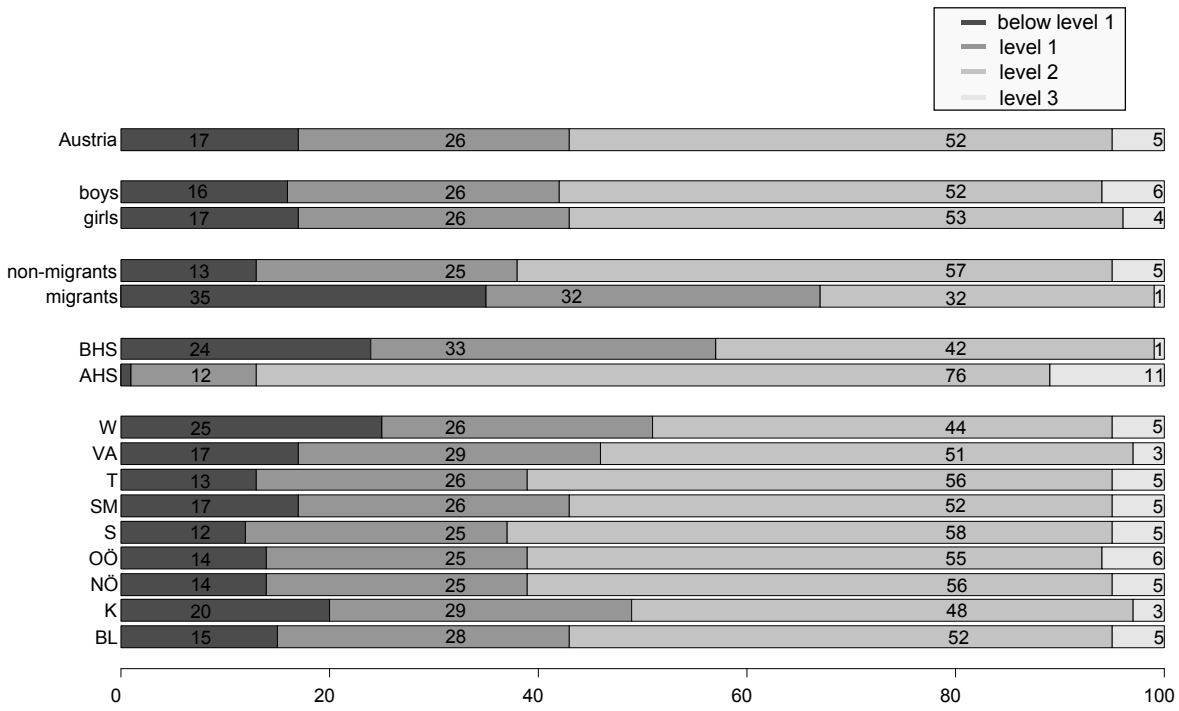

Figure 5.4.2: Percent distribution of students in competence levels obtained from official BIST-M8 2012 analysis.

worse model fit in terms of the AIC and BIC than the full DINA and the G-DINA, while they have the same mean RSMEA as the full DINA. As a compromise between model and item fit the BIST-M8 data is analyzed with the full-DINA model in the following.

Figure 5.4.3 shows the skill mastery probabilities of the eight mathematical skills obtained from the full-DINA model for the BIST-M8 data: Among the operational skills Austrian eight graders possess the skill "α_3: interpretation" best (with a probability of 72%) while they master the skill "α_4: argumentation" worst (54%). Among the content skills the students possess "α_8: statistics" best (69%) and "α_5: numbers" worst (60%). Altogether the students exhibit most problems in mastering "α_4: argumentation". This

model	competence	AIC	BIC	mean RSMEA
full-DINA	content, operation	3491861.01	3495522.61	0.07
content-DINA	content	3501439.87	3502899.01	0.07
operational-DINA	operation	3505077.87	3506537.00	0.07
GINA 1way	content, operation	3429253.48	3431575.25	0.10

Table 5.4.2: AIC, BIC and mean RSMEA for CDM models fitted to BIST-M response data.

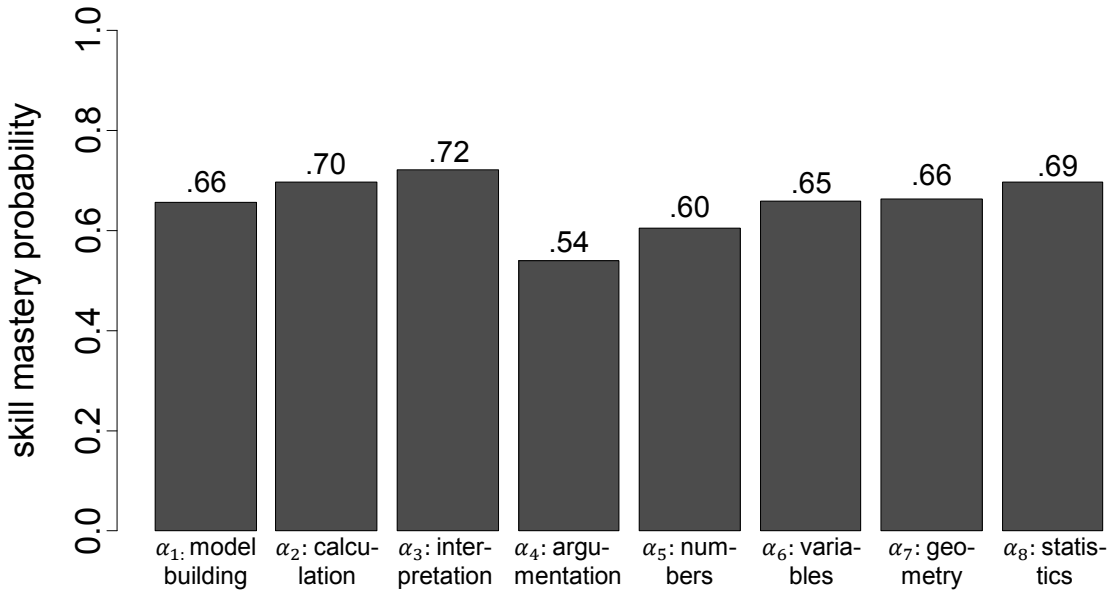

Figure 5.4.3: Skill mastery probabilities obtained from the full-DINA model for eighth graders in the test of educational standards in math.

may indicate that mathematical items in school textbooks and in school lessons often only require model building and calculation in specific well known and trained structures but they do not demand a justification of the used model framework. This is a well known phenomena in the research of mathematics educationalists (cf. e.g. Prediger, 2009; vom Hofe, 1995).

In the baseline study 2009 for mathematics in the eighth grade (Breit & Schreiner, 2010) students mastered the skill "statistics" worst of all content skills. Hence the mastery probability of this skill changed considerably. This may be caused by an increasing attention on mathematical tasks in the domain of statistics evoked by the results of the baseline study. On the contrary, it may also be true that teachers prepared their students for tasks in statistics only with regard to the test of educational standards ("teaching to the test").

Figure 5.4.4 shows the percent distribution of students belonging to different subgroups in levels of possessed skills obtained from the full-DINA model (cf. Section 5.3.2). The three levels are formed by students who have mastered less than three skills (profile group 1), students possessing between three and six skills (profile group 2) and students having more than six of the altogether eight skills (profile group 3). In the whole test population (cf. the topmost bar) and all subgroups of the test population a strikingly

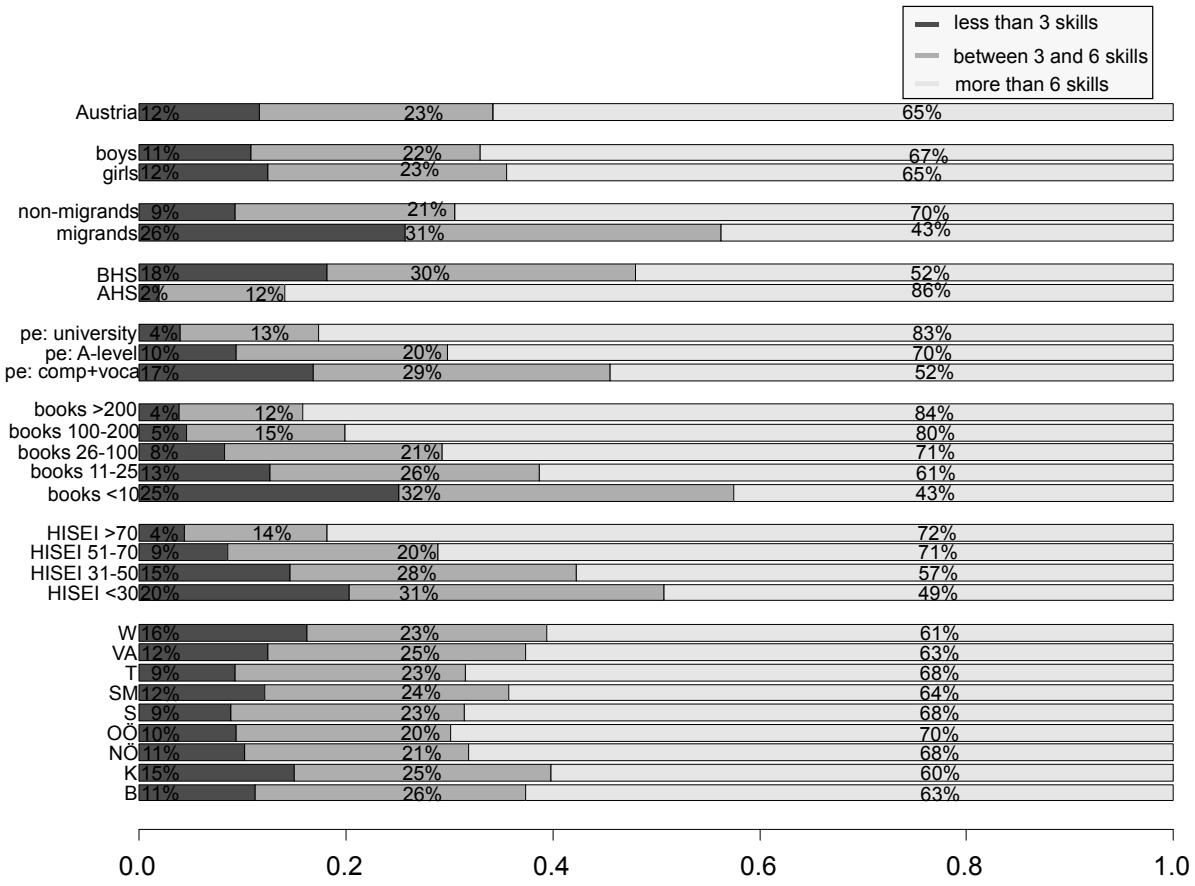

Figure 5.4.4: Percent distribution of students in levels of possessed skills obtained from the full-DINA model on the BIST-M8 data.

large percentage rate of students is classified in profile group 3. This group of students should not be compared to the group of students in level 3 of the official analysis who outperformed the educational standards (cf. Section 5.3.1). Nevertheless the results concerning group comparisons obtained from the official BIST-M8 analysis (cf. Figure 5.4.2 and Section 5.4.1) can be recovered in the percent distributions obtained from the full-DINA model (cf. Figure 5.4.4):

(1) *Gender:* As in the official analysis, in the results obtained from the full-DINA model no noticeable difference between the achievement of boys and girls can be found as well. Slightly more boys than girls possess all or at least seven of the eight skills.

(2) *Migration background:* In the full-DINA model students with migration background are far more frequently located in profile group 1 (25 %), i.e. in the group of students possessing at most 2 skills, than students without migration background (9%). This is in accordance with the official results.

(3) *School:* The results obtained from the full-DINA model also confirm the detected difference in achievement between students in the two types of school: Students attending BHS are much more often classified in profile group 1 (18%) than students attending AHS (2%).

(4) *Education of parents:* As already mentioned in the official results, in the full-DINA a connection between the education of the students' parents and the students' achievement is found as well: Students with parents exhibiting a university degree are less often classified in profile group 1 (4 %) than students with parents exhibiting a compulsory or vocational education (17 %).

(5) *Number of books in parents' household:* In addition to the official results, the full-DINA model provides evidence that students who have access to a huge variety of books in their parents' households are less often classified in profile group 1 than students who have access to only a limited number of books. 25% of the students whose parents own at least 10 books are classified in profile group 1 and, on the contrary, only 4 % of the students whose parents own more than 200 books are classified in profile group 1.

(6) *HISEI:* As captured in the official results, with the full-DINA it is detected as well, that students with higher HISEI are less often classified in profile group 1 than students with a low HISEI: Only 4% of the students with HISEI above 70, but 20% of students with HISEI below 30, are classified in profile group 1.

(7) *Federal States:* Finally the result of the official analysis on the level of the federal states is reproduced with the full-DINA as well: Wien is the state with the largest percentage of students in profile group 1 (16 %) and Oberösterreich has the largest percentage of students in profile group 3 (70 %).

To put it briefly, all results concerning group comparisons obtained in the official analysis with the help of a Rasch model are reproduced with the full-DINA model. These results are refined in the subsequent section.

5.4.3 Refined results

In order to measure the differences between groups of students not only on a unidimensional general ability scale, but on each of the eight mathematical skills multiple group DINA models are applied (cf. Section 5.3.3). Figure 5.4.5 shows the skill mastery probabilities for different subgroups of eighth graders. The results already obtained in

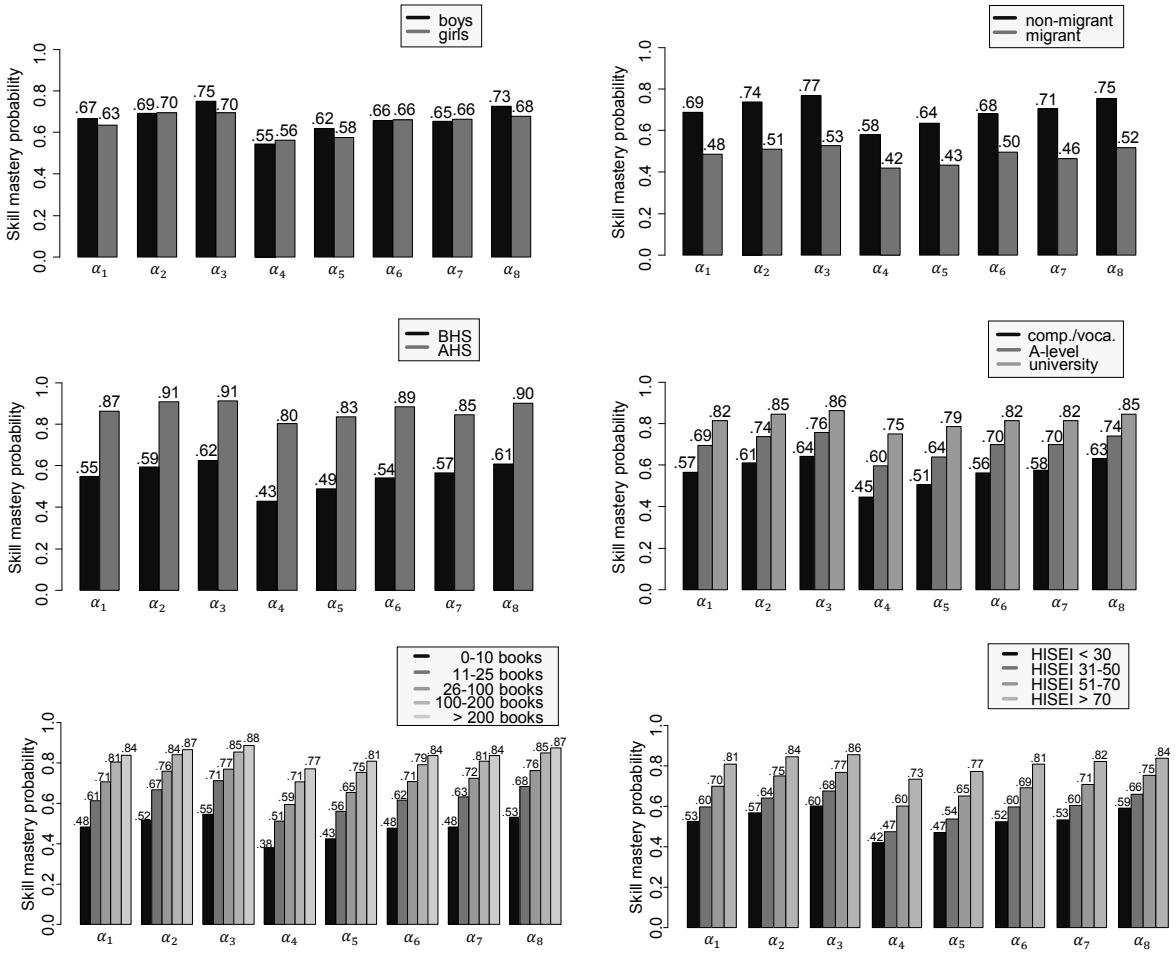

Figure 5.4.5: Comparisons between skill mastery probabilities for subgroups of eight graders in the BIST-M8 obtained through multiple group DINA models. Here "α_1: model building", "α_2: calculation", "α_3: interpretation", "α_4: argumentation", "α_5: numbers and measures", "α_6: variables and functional dependencies", "α_7: geometry" and "α_8: statistics".

Sections 5.4.1 and 5.4.2 can now be refined on the level of skills:

(1) *Gender:* There seems so be neither a difference between boys and girls with respect to a general overall ability nor with consideration of the eight skills. The largest difference in the skill mastery probabilities is located on skill "α_3: interpretation": The chance of boys to master α_3 is 6 percent points higher than the respective chance of girls.

(2) *Migration background:* The chance of non-migrants to possess the skills "α_3: interpretation", "α_7: geometry" and "α_8: statistics" is 24 percent points higher than the respective chance of migrants. The chance of non-migrants to possess the skill "α_4: argumentation" is "only" 16 percent points higher. This may be caused in language

problems, as tasks in geometry and statistics may be more sophisticated in their formulation than tasks in the domain of numbers. For the interpretation of results some advanced language knowledge is required as well. The fact that differences in achievement between non-migrants and migrants are caused in language problems is analyzed for example in Gürsoy, Benholz, Renk, Prediger & Büchter (2013) or Becker-Mrotzek, Schramm, Thürmann & Vollmer (2013), Chapter 3.

(3) *Type of school:* The largest difference in skill mastery between AHS and BHS students lays in skill "α_4: argumentation": The chance of AHS students to master α_4 is 38 percent points higher than the respective chance of BHS students. The smallest difference between AHS and BHS students is sought out in the skill "α_7: geometry". This seems reasonable as in prevocational types of school like BHS geometry may be of more importance than for example the usage of functions.

(4) *Education of parents:* The education of the parents mostly influences the students' chance to master the skill "α_4: argumentation": the chance to master α_4 is 31 percent points higher for students whose parents completed university than for students whose parents attended compulsory school. This result may be interpreted in connection with results in didactics of math (cf. Prediger, 2009) in which it is criticized that in school rather mathematical methods and tools are trained than the justification of the methods and models in practical applications. However, parents with a university degree may be apt to discuss these justifications and applications with their children.

(5) *Number of books in parents' household:* The difference between students who have access to many books and students who have access to only a small number of books is about the same in each of the eight skills: The chance of students who are provided with many books to master the skills is on average 35 percent points higher. This result suggests that students who are supported in voluntary reading exhibit generally less problems: they have less difficulties to understand the formulation of the items and they are more eloquent to express their responses.

(6) *HISEI:* The largest difference between students with high and low HISEI is again located in the skill mastery probability of "α_4: argumentation": The chance of students providing high HIISEI to master α_4 is 32 percent points higher than the respective chance of students with low HISEI.

(7) *Federal States:* The comparison between the Austrian federal states on the skill level (cf. Figure 5.4.6) reflects their comparison on the level of a general competence: Wien and Kärnten exhibit rather low results and Oberösterreich and

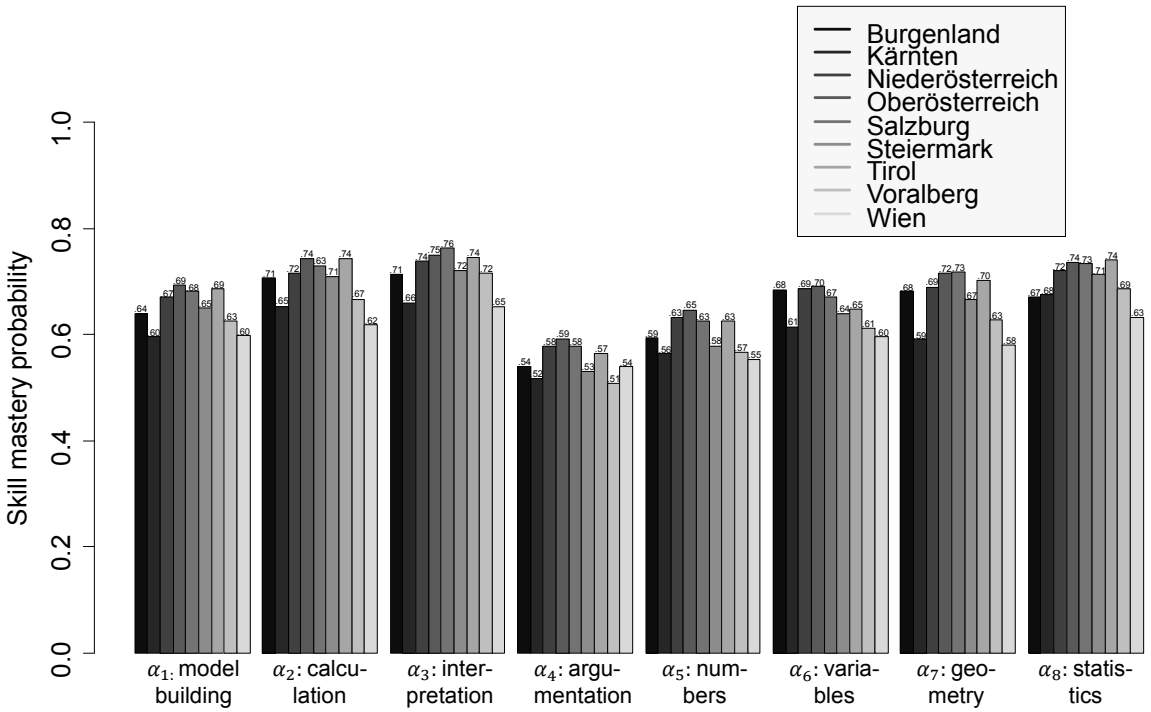

Figure 5.4.6: Comparisons between skill mastery probabilities for eight graders in different federal states in Austria based on BIST-M8 data.

Niederösterreich perform well. However, it is striking that Wien has a higher skill mastery probability in "α_4: argumentation" than expected and Burgenland shows lower results than expected in "α_8: statistics".

It has to be noted that the results of the different subgroups are not mutually independent because the groups mix up and correlate in different degrees. For example many students whose parents own many books and have a university degree attend AHS or many students with migration status also have a low HISEI.

Above some possible interpretations of the group comparisons are given. They may be rather considered as approaches and examples for demonstrating the possibilities of multiple group models with CDMs. Generally these models can be used as a substantiated empirical basis for further theoretical research about the reasons of differences between groups and for developing targeted methods to reduce these differences. For example, for a deeper analysis of the differences in achievement between migrants and non-migrants it might be helpful that linguists analyze the formulation of the items requiring the different skills for uncovering possible differences in the used language with respect to the number of technical terms or the length of the sentences.

5.5 Discussion

The present chapter is about comparisons of ability between different groups of Austrian eight graders. The knowledge about inequalities in achievement may facilitate the development of methods for reducing differences and therefore to offer each individual student optimal educational chances. These chances are fundamentally important to reduce further inequalities in economical, political, cultural and social conditions of the students in their future lives (cf. e.g. Bos et al., 2007).

The study in the present chapter is twofold: Firstly, by conducting a CDM model and conveniently merging skill classes, the results concerning group comparisons obtained from the official BIST-M8 analysis with a 2PL model are reproduced. Secondly, the abilities of specific subgroups are compared on the level of underlying skills by applying a newly developed multiple group approach for CDMs. In this second step particularly large differences between many subgroups are detected in the skill "argumentation". On the contrary, the differences between the content skills "numbers" and "functions" kept inconspicuous. As presented here with the BIST-M8 data, the results obtained from multiple group models for CDMs can be taken as substantiated empirical basis for further theoretical research about group differences.

Similar to the discussion about the application of unidimensional or multidimensional IRT models in large scale studies (e.g. Magnani, Monari, Cagnone & Ricci, 2006; Voss, Carstensen & Bos, 2005) there are arguments against and in favor of conducting a multidimensional model (i.e. a M-IRT model or CDM) for the BIST-M8 data. On the one hand the arguments against the application of multidimensional models point out the large correlations between the dimensions (i.e. skills) and as a result thereof that the main statement can already be captured on one dimension (e.g. AHS students perform better than BHS students). On the other hand, an argument in favor of applying a multidimensional model is the recovery of fine-grained nuances in the comparison of groups, which may enable even more targeted support (e.g. BHS students should be preferably supported in interpreting and arguing). Especially for large data sets like the present one the application of multidimensional models poses no statistical problems and is unproblematically accomplishable with current computers. The application of multidimensional models has already been inspired by Goldstein (1979) and is still excessively discussed (e.g. Gibbons, Immekus & Bock, 2007; Walker & Beretvas, 2003).

6 Summary and discussion

6.1 Summary and discussion

The present work deals with statistically and practically relevant issues of Cognitive Diagnosis Models (CDMs; e.g. DiBello, Roussos & Stout, 2007; Rupp, Templin & Henson, 2010) and blends both aspects. CDMs are a family of statistical models which allow diagnosing abilities of examinees in test situations. This work focuses on students' abilities in educational tests, but there are several other fields of application as for example psychology or biology (Carpenter, Just & Shell, 1990; Ivie & Templin, 2006; Levy & Mislevy, 2004; Templin & Henson, 2006). Roughly spoken, the analysis of students' abilities with CDMs is divided into three steps: Firstly, educational experts define basic abilities, the so-called skills, which are assumed to underly the tested ability. In a second step, the experts also define in a so-called Q-matrix which of these skills are relevant for the mastery of each item. Finally, based on the expert information and the manifest item responses, the students are classified into dichotomous skill classes, predicting their possession or non-possession of the underlying skills. These skill profiles allow a fine-grained diagnosis of the students' abilities and can be used as targeted empirical basis for further feedback or support.

Chapter 1 The content of the first chapter in the present work is twofold: As a first aspect, the practical relevance of CDMs for recent empirical educational research is demonstrated. It turns out that CDMs are in line with its demands: The National Research Council (National Research Council, 2001) as well as the OECD (OECD, 2004) and the KMK (KMK, 2004) claimed detailed diagnostic information, which teachers and educational administrators can use to identify why students do not perform as expected. Based on this knowledge, the organizations wish to modify the educational system and thus to reduce the resulting differences in economical, political, cultural and social conditions in the students' further lives. As discussed in Chapter 1, CDMs may yield this desired information. As a second aspect of Chapter 1, CDMs are embedded in their

statistical modeling framework. Many connections to other model approaches as latent class analysis (Lazarsfeld, 1950), item response models (de Ayala, 2009), knowledge space theory (Doignon & Falmagne, 1999) and the rule space approach (Tatsuoka, 1983) are pointed out.

Connections to further models as for example to so-called located latent trait models (Heinen, 1996), with the skill profiles corresponding to an ordinal ordered discretized ability distribution, or, more generally, to the generalized probabilistic Guttman model (Hanson, 2000; Proctor, 1970) and the latent distance model (Lazarsfeld, 1950) could be deduced. These connections may yield some further theoretical explanations: for example the difference between the item response functions of the Rasch model and the H-DINA model discussed in Chapter 4 may be clarified (i.e. in the Rasch model response behavior is modeled on two levels g_j and $1 - s_j$, $j = 1, \ldots, J$, whereas in the H-DINA model several qualitative levels are modeled). In connection with the comparison of student classifications obtained through the Rasch and the H-DINA model the work of Bartolucci (2007) could be considered as well. Bartolucci defines a discrete θ ability distribution in which a prespecified number of θ ability levels, the locations and probabilities of θ are estimated based on the empirical data. A completely new perspective in recent CDM research would be the classification of generalized CDM approaches (i.e. G-DINA, GDM or LCDM) into the framework of generalized linear models (McCullagh & Nelder, 1989) which could expedite the implementation of well-known and often used methods for CDMs as for example differential item functioning (Holland & Wainer, 1993) or new estimation methods and algorithms (cf. De Boeck & Wilson, 2004, for a classification of IRT into the framework of generalized linear models). Furthermore, the EM-algorithm for the estimation of the skill class distribution and the item parameters yields some new aspects: Is it possible to estimate the DINA model parameters through descriptive methods (cf. Chiu & Douglas, 2013; George & Ünlü, 2011)? Is the uniform prior of the skill class distribution a prior distribution in the sense of Bayesian methods (de la Torre, 2009) or is it a set of starting values in the sense of latent class analysis (Lazarsfeld, 1950)? If the second alternative is true, the EM algorithm has to be adapted in such a way that it starts with several different skill class distributions (i.e. stating values) and after the first steps of the iteration only continuous to consider the skill class distributions with largest likelihood values (cf. Linzer & Lewis, 2011, for the estimation of latnet class models).

Chapter 2 In the second chapter the R package CDM (George, Kiefer, Robitzsch, Groß & Ünlü, 2013) is introduced, which has been developed during to this work. The

R package CDM directly enables parameter estimation of DINA, DINO, G-DINA and GDM models. Through constraining parameters of the G-DINA approach, the package also allows parameter estimation of other prominent CDMs as NIDA, NIDO and RUM models. In Chapter 2 of the present work the handling of the R package CDM is described by running through the steps for analyzing student response data with CDMs. The tutorial like chapter includes descriptions of basic methods (e.g. goodness of fit, parameter interpretation, model comparisons via likelihood) as well as advanced methods of CDM analysis (e.g. reduction of skill space; establishment of link functions). Recent simulation studies showed that the estimation of the item parameters and the skill class distributions in the R package CDM is unbiased and that the RMSEA decreases with increasing sample size. The speed of the algorithm is at least similar to the calculation time of other free software packages as the Ox routine by de la Torre and the mdltm stand alone software by Von Davier. The development of the R package CDM has been continued up to now and will be further continued. New methods for analyzing response data with CDMs should be implemented as for example a person fit index (Lui, Douglas & Henson, 2009), methods for the empirical validation of the Q-matrix (de la Torre, 2008), additional graphical plot functions and methods for adaptive routines as for example presented in Chen, Xin, Wang & Chang (2012).

In Chapter 2 the handling of the R package CDM is demonstrated with student response data of PIRLS 2006 in Germany. As a question of practical relevance it is discussed which theoretical concept of reading (concerning dimensionality and connections between the reading skills) underlies the data. Based on the PIRLS 2006 data of Germany, no clear evidence for one of the discussed concepts was found, neither for the one assuming hierarchical ordered skills nor for the one emanating from the concept of four parallel reading processes. Based on a smaller data set, the topic of comparing several different theoretical reading concepts is discussed in more detail later on in Chapter 4 of the present work. In further analysis it might be beneficial to conduct these in-depth investigations on the larger PIRLS 2006 dataset. Furthermore, the new results could be compared in detail to the work of Voss, Carstensen & Bos (2005), who compared IRT models of different dimensions for the PIRLS 2001 data in Germany.

Chapter 3 In the third chapter of the present work it is shown that for awkward Q-matrices the individual student classification obtained from DINA models may be not meaningful or even wrong. For some Q-matrices the students' individual skill profiles are somehow randomly chosen: In these cases a whole set of skill classes yields equal probabilities after convergence of the EM algorithm, whereas typically the skill class

leading to the largest probability is chosen for the classification of the individual student. In Chapter 3 it is described how to handle these ambiguous skill classes in the case of given and unchangeable Q-matrices and how to avoid ambiguous skill classes in the case of a new Q-matrix construction. From that we can conclude that the Q-matrix is not only the most sensible part of DINA models if it is wrongly specified by educational experts (Henson & Douglas, 2005; Rupp & Templin, 2008a; Templin & Henson, 2009; Templin, Henson, Templin & Roussos, 2008) but also if it has an awkward form. Obviously it would be beneficial to expand the findings of this chapter to other and possibly more general CDMs.

Another aspect which arises from the discussion about ambiguous skill classes is the one about identifiability of CDMs with different Q-matrices. The basic and to be answered question is whether two CDM models which yield equal values of the likelihood after convergence necessarily must have equal skill class distributions and equal item parameters. Already Maris & Bechger (2009) showed the existence of equivalent NIDA models, i.e. NIDA models with different Q-matrices which are not distinguishable based on the items and the student responses. Similar but more general results were found by Bechger, Verhelst & Verstralen (2001) and Bechger, Verstralen & Verhelst (2002) for the linear logistic test model (Fischer, 1995; Scheiblechner, 1972), by Maris & Bechger (2004) for item response models with internal restrictions on item difficulty (Butter, De Boeck & Verhelst, 1998) and by Embretson & Yang (2013) for the multicomponent latent trait model (Whitely, 1980).

Chapter 4 Chapter 4 of the present work introduces first methods for empirically comparing different qualitative competence concepts. Different quantitative IRT and CDM models are build, which all involve different underlying competence models (here: concepts of reading). Whereas for example the Rasch model assumes a model inherent hierarchy between the defined reading competences and their acquisition, an unrestricted CDM DINA model may describe the assumption of non-ordered parallel reading skills. Based on the data of the PIRLS-Transfer study no clear preference for one model could be found, neither for one assuming hierarchically ordered competences nor for a model assuming no order. There exists at most a slight tendency towards a competence model assuming three parallel non-ordered reading skills, which are based on the four reading processes of Campbell, Kelly, Mullis, Martin & Sainsbury (2001). Again it should be underlined that the attempt to compare different competence concepts involved in different statistical model approaches is *not* aiming at criticizing the statistical methods used for the analysis of large scale studies. The main goal of large scale studies is not the

development of specific competence models, but rather the description, the comparison and the analyses of different educational systems. On the contrary, the present chapter provides a chance to empirically investigate different theoretical competence concepts and their dimensionality (for M-IRT models also compare Bartolucci et al., 2012; Hartig & Höhler, 2009). These considerations may evolve some affiliated aspects: Does reading work in the same way in all countries, in other words is reading based on the some competence construct in all countries? Do there exist mediator effects which have influence on the structure between the reading competences (e.g. if an item is constructed on the word, sentence or text level; cf. Bredel & Reich, 2008)? Are there connections between the linguistic complexity of an item task and the students' responses to the item?

Furthermore in Chapter 4 a DINA model is constructed which satisfies the model inherent assumptions of the Rasch model concerning dimensionality and hierarchy between the competences. Subsequently, the individual student classifications obtained from both models (after transformation) are compared but only rough similarities are found. From a statistical point of view the difference between the student classifications is explainable through the different forms of item response curves in the two models. It would be interesting to broaden the conducted comparison of Rasch and CDM DINA models to multidimensional IRT (M-IRT) models. Up to now there are only a few works which include a structured comparison of M-IRT and CDM models (e.g. Kunina-Habenicht et al., 2009). These comparisons could be extended to the measures of absolute and relative model fit, item and person fit and classification accuracy and reliability measures presented in the present work. A specific aspect of interest is the comparison of M-IRT and non-compensatory CDMs (as for example DINA models): Whereas M-IRT models always assume that students may compensate a lack in one skill with another possessed skill (Reckase, 2009) this is not possible in non-compensatory CDM models. However there may be situations of learning which assume non-compensatory skills, e.g. Tatsuoka's (1984) famous test of fraction subtraction test. In the end and irrespective of the researchers choice to conduct either IRT or CDM models, future studies may combine the advantages of both methods as already initiated by Tatsuoka's (2009) Rule-Space approach and recently continued by Bradshaw & Templin (2013).

Chapter 5 Chapter 5 of the present work is divided into 2 basic aspects, which are both illustrated with data from the Austrian educational standards testing in math 2012. On the one hand, it is shown that results of multiple group analysis obtained through a 2PL model can be reproduced by conducting a DINA model: It is recovered that there are only slight differences in the achievement of boys and girls and that considerably

less APS than AHS students reached the educational standards. Furthermore influences on students achievement correlated with the HISEI index, the number of books in the parents household and the parents education are rediscovered. On the other hand, a statistical method for multiple group analysis within the CDM framework is introduced. The added value multiple group analysis through CDMs instead of 2PL models is discussed: With CDM multiple group analysis groups of students are not only evaluated on one general ability (here: math) but rather on underlying basic skills which allows finer differentiations between the groups. Exploiting this advantage, it is for example found that, compared to the average difference between migrants and non-migrants, migrants have less difficulties in the domains of numbers and variables than in the fields of geometry and statistics. It may be supposed that the formulation of the item tasks in geometry and statistics requires more linguistic knowledge than the formulation of item tasks in the fields of numbers and variables. Furthermore it was shown that compared to the average difference between APS and AHS students, APS students have less difficulties in the domain of geometry. This may indicate a teaching style at APS schools which is more related to practice. Based on such differentiated differences in ability targeted feedback, fair comparisons (Ophoff et al., 2006) and systems for supporting students can be developed.

The results in Chapter 5 directly yield three research questions: The first question is whether the multiple group results obtained with a DINA model can also be preserved with a M-IRT model and if yes, which are the differences between both results. Because of the large sample size of the analyzed educational standards testing data this question could be investigated in a differentiated way. The second question is concerning the aspect of student classification in DINA models: In each group of the conducted DINA model high percentages of students are classified to possess all skills. This artifact of CDM models is known (Kunina-Habenicht et al., 2009) but it has not yet been explained. Similar reasons as the ones presented in Chapter 3 of the present work seem likely. However, it became obvious that the interpretations of the educational levels in the original standards testing analysis and the merged skill classes obtained through a CDM DINA model are not directly comparable. The third question opens up a topic which is interesting from a related practical point of view: The group of migrants may be divided into migrants with origins in the former Soviet region, in the Turkish and in the southern European region. Differences in achievement between these groups may be analyzed and class related disparities may be considered (cf. e.g. Baumert, Stanat & Watermann, 2006). Based on that targeted systems for supporting students from different backgrounds of migration may be developed (cf. e.g. Heckmann, 2008)

6.2 Main results

In the author's point of view the present work provides three main results:

(1) The usability and applicability of the CDM framework for educational researchers was increased with the development of the R package CDM. The package allows fitting CDMs without specialized stand alone software. Several simulation and retrofitting studies showed the accuracy of CDM results.

(2) The present work shows that through conducting and comparing different CDMs theoretical assumptions about the order and connections between underlying skills or competencies can be evaluated empirically.

(3) As mentioned in each of the chapters of this work, CDMs are fundamentally based on the construction and the form of the Q-matrix. There is an increased need of developing items which fit the descriptions of the Q-matrices or of essentially discussing and evaluating the entries of the Q-matrices. Some types of Q-matrices should be avoided completely. The proper specification of Q-matrices can only be achieved by an enhanced interdisciplinary cooperation of educationalists, linguists and psychometricians.

To put it in a nutshell: CDMs are a very sensible modeling approach. A correct usage of the model framework yields very differentiated, feasible and interesting results. On the contrary, if handled incorrectly, a CDM will unquestioningly process the most nonsensical input data and produce nonsensical output following the slogan "Garbage in, garbage out" (cf. e.g. Butler, Lidwell & Holden, 2010).

7 Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov, & F. Csake (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Albert, D., & Lukas, J. (1999). *Knowledge Spaces: Theories, empirical research, and applications*. Mahwah, NJ: Erlbaum.

Alexandrowicz, R. (2012). *R in 10 Schritten: Einführung in die statistische Programmierumgebung [R in 10 steps: Introduction to the statistical programming environment]*. Wien: UTB.

Artelt, C., McElvany, N., Christmann, U., Richter, T., Groeben, N., Köster, J., Schneider, W., Stanat, P., Ostermeier, C., Schiefele, U., Valtin, R., & Ring, K. (2007). *Förderung von Lesekompetenz – Expertise [Supporting reading abilities – An expersise]*. Bonn: Bundesministerium für Bildung und Forschung.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychmetrika*, *72*, 141–157.

Bartolucci, F., Montanari, G., & Pandolfi, S. (2012). Dimensionality of the latent structure and item selection via latent class multidimensional IRT models. *Psychmetrika*, *77*, 782–802.

Baumert, J., Stanat, P., & Watermann, R. (Eds.) (2006). *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit [Class related disparities in the educational system: diffentiated educational processes and equity]*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Bechger, T., Verhelst, N., & Verstralen, H. (2001). Identifiability of nonlinear logistic test models. *Psychmetrika*, *66*, 357–372.

Bechger, T. M., Verstralen, H., & Verhelst, N. (2002). Equivalent linear logistic test models. *Psychmetrika*, *67*, 123–136.

Becker-Mrotzek, M., Schramm, K., Thürmann, E., & Vollmer, H. J. (Eds.) (2013). *Sprache im Fach [Language in school lessons]*. Münster: Waxmann.

Birenbaum, M., Tatsuoka, C., & Yamada, T. (2004). Diagnostic assessment in TIMSS-R: Between-countries and within-country comparisons of eighth graders' mathematics performance. *Studies in educational evaluation*, *30*, 151–173.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M., Jordan, A., Ulfig, F., & Carstensen, C. (2004). Mathematische kompetenz [Mathematical competencies]. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekran, H.-G. Rolff, J. Rost, & U. Schiefele (Eds.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs [PISA 2003: Comepetences of German students – Results of the second international comparison]*. Münster: Waxmann.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R., & Walther, G. (Eds.) (2004). *IGLU: Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich [IGLU: National and international comparison of some federal states in Germany]*. Münster: Waxmann.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Walther, G., & Valtin, R. (Eds.) (2003). *Erste Ergebnisse aus IGLU: Schüllerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich [First results from IGLU: International comparison of students' achievements at the end of the fouth grade]*. Münster: Waxmann.

Bos, W., Valtin, R., Voss, A., Hornberg, S., & Lankes, E.-M. (Eds.) (2007). *IGLU 2006: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich [IGLU 2006: International comparison of students' reading achievement]*. Münster: Waxmann.

Bose, R. C., & Nair, K. R. (1939). Partially balanced incomplete block designs. *Sankhya: The Indian Journal of Statistics*, *4*, 337–372.

Bradshaw, L., & Templin, J. (2013). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, published online first.

Bredel, U., & Reich, H. H. (2008). Literale Basisqualifikation I und II [Literal basisqualifications I and II]. In K. Ehlich, U. Bredel, & H. H. Reich (Eds.), *Referenzrahmen zur altersspezifischen Sprachaneignung, Bildungsforschungsband 29/I [Reference frame for age specific language acquisition, Educational Research Volume 29/I]* (pp. 95–106). Bonn: Bundesministerium für Bildung und Forschung.

Breit, S., & Schreiner, C. (2010). *Bildungsstandards: Baseline 2009 (8. Schulstufe) [Educational standards test: Baseline 2009 (8th grade)]*. Technical Report Bundesinstitut für Innovation & Entwicklung des Österreichischen Schulwesens Salzburg, Austria.

Bremerich-Vos, A. (1996). Aspekte des Schriftspracherwerbs – Stufentheorie, das "Neue" und die Lehrer-Schüler-Interaktion [Aspects in written language accisition – Steptheory, "news" and teacher student interactions]. In A. Peyer, & P. Portmann (Eds.), *Norm, Moral und Didaktik – Die Linguistik und ihre Schmuddelkinder. Eine Aufforderung zur Diskussion [Norms, moral and didactics – linguistics and it's ragamuffins: A demand for discussion]* (pp. 267–290). Tübingen: Niemeyer.

Bruneforth, M., & Lassnigg, L. (Eds.) (2013). *Nationaler Bildungsbericht Österreich 2012 [Report of educational results in Austria 2012]*. Graz: Leykam.

Butler, J., Lidwell, W., & Holden, K. (2010). *Universal Principles of Design*. Rockport Publishers.

Butter, R., De Boeck, P., & Verhelst, N. (1998). An item response model with internal restrictions on item difficulty. *Psychmetrika*, *63*, 47–63.

Campbell, J., Kelly, D., Mullis, I., Martin, M., & Sainsbury, M. (2001). *Framework and specifications for PIRLS assessment 2001*. Chestnut Hill, MA: Boston College.

Carpenter, P., Just, M., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the raven progessive matrices test. *Psychological Review*, *97*, 404–431.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statsitical Software*, *48*, 1–29.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, published online first.

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychmetrika*, *77*, 201–222.

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

Chiu, C.-Y., & Minhee, S. (2009). Cluster analysis for cognitive diagnosis: An application to the PIRLS 2001 reading assessment. *IERI monograph series: Issues and methodologies in large-scale assessments*, *2*, 137–159.

Cizek, G. J., Bunch, M. B., & Konns, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, *23*, 31–50.

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

von Davier, M. (2005). *A general diagnostic model applied to language testing data (RR-05-19)*. Educational Testing Service Priceton, NJ.

Dayton, C. M., & Macready, G. (1983). Latent structure analysis of repeated classifications with dichotomous data. *British Journal of Mathematical and Statistical Psychology*, *36*, 189–201.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models*. New York: Springer.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343–362.

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

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

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

de la Torre, J., & Karelitz, T. (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*, 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*, 115–127.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26.

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

DiBello, L., Roussos, L., & 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, Psychometics* (pp. 979–1030). Amsterdam: Elsevier.

DiBello, L., & Stout, W. (2008). *Arpeggio Documentation and Analyst Manual*. Informative Assessment Research Enterprises.

Dogan, E., & Tatsuoka, K. K. (2008). An international comparison using a diagnostic testing model: Turkish students' profiles of mathematical skills on TIMSS-R. *Educational Studies in Mathematics*, *68*, 263–272.

Doignon, J. P., & Falmagne, J.-C. (1999). *Knowledge Spaces*. Berlin: Springer.

Doornik, J. A. (2002). *Object-oriented matrix programming using Ox*. Timberlake Consultants Press London.

Düntsch, I., & Gediga, G. (1995). Skills and knowledge structures. *British Journal of Mathematical and Statistical Psychology*, *48*, 9–27.

Embretson, S., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychmetrika*, *78*, 14–36.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.

Erikson, E. H. (1950). *Childhood and society*. New York: Norton.

Falmagne, J.-C., & Doignon, J. P. (2010). *Learning Spaces: Interdisciplinary Applied Mathematics*. Heidelberg: Springer.

Feldt, L. S. (1997). Can validity rise when reliability declines? *Applied Measurement in Education*, *10*, 377–387.

Ferrando, P. J. (2004). Person reliability in personality measurement: An item response theory analysis. *Applied Psychological Measurement*, *28*, 126–140.

Fischer, G. (1995). The linear logistic test model. In G. Fischer, & I. Molenaar (Eds.), *Raschmodels. Foundations, recent developments, and applications*. New York: Springer.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359 – 374.

Formann, A. K. (1978). A note on parameter estimation for Lazarsfeld's latent class analysis. *Psychmetrika*, *43*, 123–126.

Ganzeboom, H., De Graaf, P., & Treiman, D. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*, 1–56.

George, A. C. (2010). *Diagnostic knowledge structure and cognitive diagnosis models: A review and comparison*. Master's thesis TU Dortmund University.

George, A. C., Kiefer, T., Robitzsch, A., Groß J., & Ünlü, A. (2013). CDM: The R package for cognitive diagnosis models. Under revision for *Journal of Statistical Software*.

George, A. C., & Ünlü, A. (2011). Parameter estimation in skills-based knowledge space theory and cognitive diagnosis models: A comparison. In D. Conesa, A. Forte, A. López-Quílez, & F. Muñoz (Eds.), *Proceedings of the 26th International Workshop on Statistical Modelling* (pp. 258–262).

Gibbons, R. D., Immekus, J. C., & Bock, R. D. (2007). *The added value of multidimensional IRT models*. Technical Report Center for Health Statistics, University of Illinois at Chicago.

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2000). An NCME instructional module on exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice*, *19*, 34–44.

Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal*, *31*, 124–126.

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, *5*, 211–220.

Goodman, L. A. (1979). A note on the estimation of parameters in latent structure analysis. *Psychmetrika*, *44*, 123–128.

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*, 155–167.

Groß J., & George, A. (2012). On ambiguous skill classes in the DINA model. Submitted to *Methodology*.

Groß J., & George, A. (2013). On prerequisite relations between attributes in noncompensatory diagnostic classification. Under revision for *British Journal of Mathematical Psychology*.

Gürsoy, E., Benholz, C., Renk, N., Prediger, S., & Büchter, A. (2013). Erlös = Erlösung? – Sprachliche und konzeptuelle Hürden in Prüfungsaufgaben zur Mathematik [Proceeds = deliverance? – Linguistic and competence based obstacles in mathematical items]. *Deutsch als Zweitsprache [German as second language]*, *1/2013*, 14 –24.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–323.

Hanson, B. (2000). *IRT parameter estimation using the EM Algorithm*. Technical Report American College Testing Iowa City, Iowa. Http://www.b-a-h.com/papers/index.html.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competences. *Studies in Educational Evaluation*, *35*, 57–63.

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

Hattie, J. (1985). Methodology review: Assessing uni-dimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139–164.

Heckmann, F. (2008). *Education and migration: Strategies for integrating migrant children in European schools and societies*. Technical Report Network of Experts in Social Science of Education and training.

Heinen, T. (1996). *Latent class and discrete latent trait models*. Thousand Oaks, CA: Sage.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*, 262–277.

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

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577–601.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*, 407–419.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

Ivie, J., & Templin, J. (2006). Analysis of raven's progressive matrices (rpm) scale using skills assessment. Paper presented at the annual meeting of the National Council on Measurement in Education.

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117 – 131). Ames, IA: Iowa State University.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion model application to LanguEdge assessment. *Language Testing*, *26*, 31–73.

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

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, *6*, 125–160.

Kanning, U. P. (2003). *Diagnostik sozialer Kompetenzen. Kompendien Psychologische Diagnostik [Diagnostics of social competencies. Comepndiums psychological diagnostics]*. Wien: Hogrefe.

Kiefer, T., Robitzsch, A., & Wu, M. (2013). TAM: Test analysis modules. R package version 0.02-07. URL https://sites.google.com/site/irttam1.

Kim, S.-H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*, 131–143.

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing*, *28*, 509 – 541.

Kirsch, I., & Mosenthal, P. (1991). *Understanding documents. A monthly column appearing in the Journal of Reading*. Newark, DE: International Reading Association.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E., & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards: Eine Expertise [Comments on the development of the educational standards: An expertise]*. Berlin: Bundesministerium für Bildung und Forschung.

KMK (2004). *Bildungsstandards der Kultusministerkonferenz: Erläuterungen zur Konzeption und Entwicklung [Educational standards of the standing committee of the German ministers of culture: Comments on the conception and development]*. München: Wolters Kluwer.

Kubinger, K. D. (2006). *Klinische Diagnostik [Clinical diagnostics]*. Wien: Hogrefe.

Kubinger, K. D., Rasch, D., & Yanagida, T. (2011). *Statistik in der Psychologie [Statistics in psychology]*. Wien: Hogrefe.

Kunina-Habenicht, O., Rupp, 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*, 64–70.

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In S. A. Stouffler, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Studies in Social Psychology in World War II*. New York: John Wiley & Sons.

Lee, J., Grigg, W., & Dion, G. (2007). *The nation's report card: Mathematics 2007 (NCES 2007-494)*. Washington, DC: National Center for Education Statistics.

Lee, Y.-S., de la Torre, J., & Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: An empirical investigation. *Asia Pacific Educational Research*, *13*, 333–345.

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*, 144–177.

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and application*. Cambrige, UK: Cambrige University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205–237.

Levy, P., & Lemeshow, S. (1999). *Sampling of populations: Methods and Applications*. New York: Jon Wiley & Sons.

Levy, R., & Mislevy, R. (2004). Specifying and refineing a measurement model for computer-based interactive assessment. *International Journal of Testing*, *4*, 333–369.

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *9*, 17–46.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*, 1–29.

Lui, Y., Douglas, J., & Henson, R. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, *33*, 579–598.

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

Macready, G., & Dayton, C. M. (1980). The nature and use of state mastery models. *Applied Psychological Measurement*, *4*, 493–516.

Magnani, S., Monari, P., Cagnone, S., & Ricci, R. (2006). Multidimensional versus unidimensional models for ability testing. In H.-H. Bock, W. Gaul, & M. Vichi (Eds.), *Data Analysis, Classification and the Forward Search* (pp. 339–346). Springer.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*, 1–20.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212.

Maris, G., & Bechger, T. (2004). Equivalent mirid models. *Psychmetrika*, *69*, 627–639.

Maris, G., & Bechger, T. (2009). Equivalent diagnostic classification models. *Measurement: Interdisciplinary Research and Perspectives*, *7*, 41–46.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *PIRLS 2006 Technical Report*. Chestnut Hill, MA: Boston College.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. New York: Chapman & Hall.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Mullis, I., Martin, M. O., Ruddock, G., O'Sullivan, C., Arora, A., & Erberer, E. (2008). *TIMSS 2007 assessment frame*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's study of reading literacy achievement in primary schools*. Chestnut Hill, MA: Boston College.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Journal of Educational Measurement*, *16*, 159–176.

Muthén, L., & Muthén, B. (2010). *Mplus User's Guide*. Muthén and Muthén Los Angeles, CA (6th ed.).

Nahberger, G. (2007). *Lockis Abenteuergeschichten im Urwald [Locki's adventures in the jungle]*. Baltmannsweiler: Schneider-Verlag Hohengehren.

Nahberger, G. (2010). *Lesekompetenz Testen und Trainieren mit Lockis Abenteuergeschichten im Wilden Westen. Ein Buch für ErzieherInnen und GrundschullehrerInnen. [Testing and training reading competence with Locki's adventures in the Wild West. A book for educators and primary school teachers]*. Baltmannsweiler: Schneider-Verlag Hohengehren.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academic Press.

Niss, M. (2003). Mathematical competencies and the learning of mathematics: The Danish KOM project. In A. Gagatsis, & S. Papastavridis (Eds.), *Third Mediteranean conference on mathematical education. Athens – Hellas 3-5 January 2003*. Athens: The Hellenic Mathematical Society.

OECD (2001). *Learning for life. First results from PISA 2000*. Paris: OECD.

OECD (2004). *Leraning for tomorrow's world: First results from PISA 2003*. Paris: OECD.

OECD (2010). *PISA 2009 Results: What students know and can do*. Paris: OECD.

Ophoff, J. G., Koch, U., Hosenfeld, I., & Helmke, A. (2006). Ergebnisrückmeldungen und ihre Rezeption im Projekt VERA [Feedback and receptions of the project VERA]. In H. Kuper, & J. Schneewind (Eds.), *Rückmeldung und Rezeption von Forschungsergebnissen [Feedback and receptions of research projects]* (pp. 19–40). Münster: Waxmann.

Park, Y., & Lee, Y. (2011). Diagnostic cluster analysis of mathematical skills. *IERI monograph series: Issues and methodologies in large-scale assessments*, *4*, 75–107.

Peschek, W., & Heugl, H. (2007). *Standards für die mathematischen Fähigkeiten Österreichischer Schülerinnen und Schüler am Ende der 8. Schulstufe [Defining mathematical standards for the for Austrian students at the end of the 8th grade]*. Technical Report Alpen-Adria-University Klagenfurt.

PISA consortium (2001). *PISA 2000: International comparison of students' basic competencies*. Opladen: Leske und Budrich.

PISA Konsortium (2003). *PISA 2000: Ein diffenzierter Blick auf Bundesländer der Republick Deutschland [PISA 2000: A sophisticated view on the federal states of Germany]*. Opladen: Leske und Budrich.

Prediger, S. (2009). Inhaltliches Denken vor Kalkül – Ein didaktisches Prinzip zur Vorbeugung und Förderung bei Rechenschwierigkeiten [Prefering contentual thinking to calculation – A didactical principe for prevention and support of difficulties in caluculation]. In A. Fritz, & S. Schmidt (Eds.), *Fördernder Mathematikunterricht in der Sekundarstufe I [Lessons in mathematics to support students]* (pp. 213–234). Weinheim: Beltz Verlag.

Prenzel, M., Drechsel, B., Carstensen, C., & Ramm, G. (2004). PISA 2003 – eine Einführung [PISA 2003 – An introduction]. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekmn, H.-G. Rolff, J. Rost, & U. Schiefele (Eds.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs [PISA 2003: Comepetences of German students – Results of the second international comparison]*. Münster: Waxmann.

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

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0.

Rasch, G. (1960). *Probabilitstic models for some intelligence and attainment tests*. Ph.D. thesis University of Chicago Chicago, IL.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544–559.

Rost, J. (2004). *Lehrbuch Testtheorie und Testkonstruktion*. Bern: Huber.

Rudner, L. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, *7*.

Rupp, A., & Templin, J. (2008a). The effects of Q-matrix misspecification on the parameter estimates and the classification accurancy in the DINA model. *Educational Psychological Measurement*, *68*, 78–96.

Rupp, A., & Templin, J. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, *6*, 219–262.

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

SAS Institute Inc. (2007). *User's Guide for SAS Software Navigator*. SAS Institute Inc. Cary, NC.

Scheiblechner, H. H. (1972). Das Lernen und Lösen komplexer Denkaufgaben [Abount learning and solving of complex thinking tasks]. *Zeitschrift für Experimentelle und Angewandte Psychologie*, *19*, 476.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61.

Schreiner, C., & Breit, S. (2012). *Standardüberprüfung 2012 Mathematik, 8. Schulstufe: Bundesergebnisbericht [Testing of educational standards in math 2012 in the eighth grade: Report of results]*. Technical Report Bundesinstitut für Innovation und Entwicklung des Österreichischen Schulwesens.

Schrepp (2005). About the connection between knowledge structures and latent class models. *Methodlogy*, *1*, 93–103.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Seedhouse, P. (2005). The case oft the missing "no": The relationship between pedagogy and interaction. *Language Learning - A Journal of Research in Laguage Studies*, *51*, 347–385.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, *66*, 191–207.

Sun Microsystems, I. (2001). *C++ User's Guide*. Palo Alto, CA.

Svetina, D., Gorin, J., & Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach. *International Journal of Testing*, *11*, 1–23.

Tatsuoka, C., Varadi, F., & Jaeger, J. (2013). Latent partially ordered classification models and normal mixtures. *Journal of Educational and Behavioral Statistics*, *38*, 267–294.

Tatsuoka, K. (1984). *Analysis of errors in fraction addition and subtraction problems*. Final report for NIE-G-81-0002 University of Illinois, Urbana-Champaign.

Tatsuoka, K., & Yan, F. (2001). *Guide to use the rule space III program*. Educational Testing Service.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule-space method*. Florence, KY: Routledge.

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

Templin, J., Henson, R., Templin, S., & Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement*, *32*, 559–574.

Templin, J., & Henson, R. A. (2009). Practical issues in using diagnostic estimates: Measuring the reliability of diagnostic estimates. Paper presented at the annual meeting of the Natioanl Council on Measurement in Education in San Diego.

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

van Dijk, T., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Vermunt, J. (1997). *LEM 1.0: A general program for the analysis of categorical data.*. Tilburg University Tilburg.

Vermunt, J., & Magidson, J. (2005). *Latent GOLD 2.0 User's Guide*. Statistical Innovations Inc. Boston.

vom Hofe, R. (1995). Vorschläge zur Öffnung normativer Grundvorstellungskonzepte für deskriptive Arbeitsweisen in der Mathematikdidaktik [Proposals for opening normative concepts to descriptive methods in mathematics]. In H.-G. Steiner, & H.-J. Vollrath (Eds.), *Neue problem- und praxisbezogene Forschungsansätze [New problem and practice related research aspects]* (pp. 42–50). Köln: Spektrum Akademischer Verlag.

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

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement*, *28*, 389–406.

Voss, A., Carstensen, C., & Bos, W. (2005). Textgattungen und Verstehensaspekte: Analyse von Leseverständnis aus den Daten der IGLU-Studie [Types of textes and aspects of comprehension: Analysis of reading comprehension with data from IGLU]. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin, & G. Walther (Eds.), *IGLU: Vertiefende Analysen zu Leseverständnis, Rahmenbedingungen und Zusatzstudien [IGLU: Deepining analyses about reading comprehension, framework and additional studies]* (pp. 1–36). Münster: Waxmann.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement*, *40*, 255–275.

Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, *48*, 165–187.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – Eine umstrittene Selbstverständlichkeit [Comparing achievements in schools – A controvesial implicitness]. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen [Measuring achievement in schools]* (pp. 17–31). Weinheim: Beltz.

Whitely, S. E. (1980). Multicomponent latent-trait models for ability tests. *Psychometrika*, *45*, 479 – 494.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*, 80–83.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*, 114–128.

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

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

Statement of authorship and verification

I hereby declare that this dissertation has been composed by myself and describes and constitutes my own work unless otherwise acknowledged in the text. All references and verbatim extracts have been distinguished by quotation marks and all sources of information have been specifically acknowledged.

This dissertation is handled in to the Department 12 "Education and Sociology" of the TU Dortmund University for receiving a doctoral degree. This work has not been submitted for any other examinations.

_____
Ann Cathrice George, August 2013