2015 ■ Vol. 11 ■ no. 3

DOI: 10.20982/tqmp.11.3.p189

Cognitive Diagnosis Models in R: A Didactic

Ann Cathrice George[a,✉] and Alexander Robitzsch[a]

[a]Federal Institute for Educational Research, Innovation and Development of the Austrian School System; Salzburg, Austria

Abstract ■ Cognitive diagnosis models (CDMs) are a class of discrete latent variable models which may be used to estimate the possession of attributes underlying a global ability or attitude. The present article explains the basics of CDMs to researchers not acquainted with this family of models. Two worked out examples show how to apply a CDM using the R package **CDM**.

Keywords ■ Cognitive Diagnosis Models, R, Tutorial, Competence Modeling

✉ a.george@bifie.at

Introduction

Consider the case where a researcher wants to assess mathematical abilities of students of a given age. For that purpose, the researcher consults experts in didactics and they determine the basic mathematical operations of students underlying the mathematical abilities to assess. Then, they create items, each requiring one or more of these operations to be solved. Based on the students' responses on the items, the researcher wants to determine, which of the previously specified mathematical operations were successfully applied by the students. This is the point where Cognitive Diagnosis Models (CDMs; DiBello, Roussos, & Stout, 2007; Henson, Templin, & Willse, 2009) come into play.

Before delving into the technical details, let us consider an example: The item $3+5*4=?$ requires the knowledge of (a) *"add small numbers"*, (b) *"multiply small numbers"*, and the knowledge that (c) *"multiplying precedes adding"*. So, if a student correctly responds 23, we may conclude that he disposes of all three capacities – and this is pretty close to what a CDM accomplishes: CDMs are a class of discrete latent variable models, allowing to trace back the respondent's answer to an item to his possession of basic features underlying the domain covered by the items.

Our toy example already illustrates the currently dominating area of application, i. e. educational assessment. Here, the algebraical operations, to which the items were traced back, are frequently called skills. A famous example in this context is the fraction subtraction analysis of Tatsuoka (1984): First, he broke down the domain of fraction subtraction into eight skills (e. g. (a) *"convert a whole number to a fraction"*, (b) *"separate a whole number from a fraction"*, or (c) *"simplify before subtraction"*). Then, he developed items, each covering selected skills and presented these items to the respondents. Finally, a CDM analysis yielded the percentage of students possessing the skills (e. g. 58 % of the students possess *"convert a whole number"*) and individual skill profiles for each student (e. g. Lisa is able to *"convert a whole number"* but not able to *"separate a whole number from a fraction"*). These profiles might then serve as an orientation for targeted coaching of the respondents (i. e. Lisa could be instructed how to separate a whole number from a fraction).

CDMs have sufficient generality to be applied to other areas of research as well. Templin and Henson (2006), for example, used a CDM in a clinical context where they analyzed pathological gambling. The authors traced back the domain of pathological gambling to behavior patterns like (a) *"needs to gamble with increasing amounts of money in order to achieve the desired excitement"*, (b) *"gambles as a way of escaping from problems or of relieving a dysphoric mood"*, or (c) *"has repeated unsuccessful efforts to control, cut back, or stop gambling"*. These patterns were analyzed by 41 items in the so-called Gambling Research Instrument (Feasel, Henson, & Jones, 2004), to which the examinees responded. With a CDM, they were able to estimate the percentage of examinees showing each behavioral pattern (e. g. 30 % of the examinees *"need to gamble with increasing amounts of money"*) and yields for each person a behavioral profile (e. g. Mark does not *"need to gamble with increasing amounts of money"* but he *"gambles as a way of escaping from problems"*). These profiles formed the ground for further clinical statements: For example, counting the number of behavioral patterns an individual exhibits in a respondent's profile yields a sum, based on which clinicians could diagnose "pathological gambling" using a previously determined cut-off score.

Another non-educational application of CDMs is the study of de la Torre, van der Ark, and Rossi (2015). They proposed to use CDMs for assessing mental disorders using items of a relevant clinical questionnaire. For each item of the questionnaire, the authors specified, which of the

disorders (a) *"anxiety"*, (b) *"somatoform"*, (c) *"thought disorder"*, or (d) *"major depression"* are involved to respond positively. By means of a CDM, they were able to assess, which disorder is probably present in each patient (i. e. the individual disorder profile) and how prevalent the disorders and their combinations are in the population. The individual disorder profile could again be used for diagnosing a clinical disorder in terms of a DSM diagnosis (American Psychiatric Association, 2013) by means of a cut-score applied to the number of disorders in an individual disorder profile (for a similar approach see also Jaeger, Tatsuoka, Berns, & Varadi, 2006).

Note that the attribution of items to more basic elements (e. g. skills in the maths example and behavioral patterns or disorders in the clinical examples) was present before the items had been developed, i. e. it was not derived from the item responses. This is a key issue when applying CDMs: Before conducting the model, the researcher has to determine on theoretical grounds (frequently involving experts of the field), which fundamental element is involved in a given item. This mapping of skills to items takes place in a design (or weight) matrix (frequently termed $\boldsymbol{Q}$). Details regarding this aspect will be presented at the begining of the next section.

The present article explains the basics of CDMs in the next two sections and illustrates their application using two examples in the last section using the package **CDM** (George, Robitzsch, Kiefer, Groß, & Ünlü, in press) of R (R Core Team, 2015). Compared to Ravand and Robitzsch (2015) we (a) work with published data (available in the **CDM** package), hence the reader is able to reproduce all examples presented here, and we extend the focus to (b) the multiple group case and (c) the usage of sampling weights in large-scale assessments. Based on the steps presented here, readers may more easily find their way to complex CDM applications.

Input and Output of a CDM analysis

Let us first consider CDMs as a black box, to which we feed some input and obtain some output, not bothering how such a transformation is performed (cf. next section for that issue).

Input: Response Data and Q-Matrix

A basic specification of a CDM requires two elements: The response data to the items of the test and a weight matrix designed by experts from the field. To explain the basic principles of both elements we will use the dichotomous case; generalizations to the polytomous case are straightforward.

The response data is typically stored in a $I \times J$ item response matrix $\boldsymbol{X}$, in which the element x_{ij} in the i-th row and the j-th column indicates whether examinee i gave a positive (i. e., correct) response to item j ($x_{ij} = 1$) or not ($x_{ij} = 0$). There have been few concrete recommendations in the CDM literature regarding the minimum sample size for conducting CDM analyses. Rupp and Templin (2008) suggest that for simple CDMs a "few hundred students" responding each item are sufficient for successfully estimating the model if the number of skills is small (i. e., four to six). A systematic study investigating the minimum sample size for various numbers of skills is so far missing. In each case, the sample size should be larger than the number of model parameters, see next section.

The construction of the weight matrix $\boldsymbol{Q}$ (usually called Q-matrix) involves qualitative preliminary work of experts: First, the experts subdivide the tested overall domain into a few skills according to a well-established qualitative relationship between the skills (e. g. a competence model). In CDM terminology these skills are termed α_k, $k = 1, \ldots, K$. Secondly, based on the relationship between the skills, the experts specify which skills α_k are required for giving a positive response in each test item. This specification is denoted in a binary $J \times K$ weight matrix $\boldsymbol{Q}$, in which q_{jk} expresses whether skill k is needed ($q_{jk} = 1$) or not ($q_{jk} = 0$) for enabling examinees to positively respond to item j. Thus, the Q-matrix reflects the essential theory of how skills contribute to responding to each item. A CDM infers the examinees' possession of the K skills from the examinees' response vectors.

Consider, for example, the Examination for the Certificate of Proficiency in English (ECPE) developed by the English Language Institute of the University of Michigan. The ECPE consists of four major sections, however we restrict our example to the grammar section. This section comprises $J = 28$ multiple-choice items (cf. first example of the last section), in which syntactically correct sentences are presented with one word omitted. Students have to select the missing word from a list of four (cf. Table 1). The data, including 2992 students, has already been analyzed using CDM methodology in Templin and Bradshaw (2014) and Templin and Hoffman (2013).

For the $J = 28$ items, educational experts (cf. Buck & Tatsuoka, 1998) identified $K = 3$ underlying skills presented in Table 2. The experts decided which item requires which skill to be solved and thus specified the Q-matrix. Table 3 shows a portion of this Q-matrix for the ECPE items (Henson & Templin, 2007). The second row of the Q-matrix in Table 3 ($\boldsymbol{q}_2 = [0,1,0]$) describes the example item of Table 1. The item requires students to apply *"cohesive rules"* (α_2), but not *"morphosyntactic rules"* (α_1) or *"lexial rules"* (α_3).

In line with this example, Li, Hunter, and Lei (2015) recently emphasized the relevance of the CDM methodology in the domain of language testing.

Table 1 ■ Example item of grammar section in the Examination for the Certificate of Proficiency in English (ECPE).

Item	Mary had to lean _____ the counter to open the window.	
Response options	(a)	above
	(b)	over
	(c)	after
	(d)	around

Table 2 ■ Sub-competencies underlying the overall ability of understanding English grammar.

Parameter	Skill	Description
α_1	morphosyntactic rules	word formation; combination of words into larger units such as phrases and sentences
α_2	cohesive rules	grammatical and lexical linking within a text or sentence
α_3	lexical rules	modification of argument structures of lexical text elements (i. e., verbs and declensions)

Output: Skill Distribution, Skill Class Distribution, and Individual Skill Profiles

CDMs aim at inferring the examinees' possession of the K skills. Dichotomous variants of CDMs assume respondents to either possess a skill k or not, therefore 2^K different patterns arise from building all possible skill combinations. These patterns are called skill classes $\boldsymbol{\alpha}_l = [\alpha_{l1},\ldots,\alpha_{lK}]$, $l = 1,\ldots,2^K$. Each element α_{lk} denotes whether or not members of skill class l dispose of skill k (i. e., $\alpha_{lk} = 1$ or $\alpha_{lk} = 0$, respectively). The ECPE data in our example comprises $K = 3$ skills and therefore allows allocating the students into $2^K = 2^3 = 8$ different skill classes: $\boldsymbol{\alpha}_1 = [0,0,0]$, $\boldsymbol{\alpha}_2 = [1,0,0]$, $\boldsymbol{\alpha}_3 = [0,1,0]$, $\boldsymbol{\alpha}_4 = [0,0,1]$, $\boldsymbol{\alpha}_5 = [1,1,0]$, $\boldsymbol{\alpha}_6 = [1,0,1]$, $\boldsymbol{\alpha}_7 = [0,1,1]$, $\boldsymbol{\alpha}_8 = [1,1,1]$.

Strictly speaking, in the CDM context the formulation "the examinees' possession of the skills" covers three different questions, which are addressed in the model's output:

(Q1) The *population skill possession* question:
"What is the proportion of examinees possessing a specific skill α_k?"
The skill distribution $P(\alpha_k)$, $k = 1,\ldots,K$, quantifies this question.

(Q2) The *population skill class distribution* question:
"What is the proportion of examinees possessing a specific combination $\boldsymbol{\alpha}_l = [\alpha_{l1},\ldots,\alpha_{lK}]$ of skills?"
The skill class distribution $P(\boldsymbol{\alpha}_l)$, $l = 1,\ldots,2^K$, answers this question.

(Q3) The *individual skill possession* question:
"Which skills does the i-th individual examinee possess?"
The i-th examinee's skill profile $\boldsymbol{\alpha}_i = [\alpha_{i1},\ldots,\alpha_{iK}]$, $i = 1,\ldots,I$, provides this information.

The last section presents two examples of application and interpretation of these three questions.

Selecting and Estimating a Specific CDM

This section refers to two basic topics lying between input and output: (1) the selection of a specific CDM. This choice is essential as it manages the rules according to which the CDM transforms the input to achieve the output. This choice also determines the statistical model equation and the model's likelihood.[1] (2) the statistical methods and algorithms for estimating the model parameters.

Selecting a Specific CDM

So far, explanations referred to the basic principles of CDM analyses. However, several variations of CDMs exist defining a CDM family of models. Let us consider an illustrative example: Some CDMs assume that students have to possess all required skills (as defined in row j of the Q-matrix; cf. Table 3) in order to positively respond to item j. In contrast, other CDMs assume that the possession of at least one required skill is sufficient for positively responding the item.

The model selection should be driven by experts because it involves fundamental questions of how the possession of the skills determines the response behavior. The example above defined the compensability of the skills: Can students compensate a lack in one required skill through the possession of another skill? Some further distinctions concern interaction of skills possessed together or the possibility of multiple strategies. These and other characteristics are defined and explained in detail in e. g. DiBello et al.

[1] Actually, the choice of which CDM to apply should have been made at the time of developing the items and defining the Q-matrix.

Table 3 ■ Portion of the Q-matrix for grammar section of ECPE test.

	Skill		
Item	α_1	α_2	α_3
1	1	1	0
2	0	1	0
3	1	0	1
4	0	0	1
⋮			
$J = 28$	0	0	1

(2007) or Rupp and Templin (2008).

Table 4 gives an overview of important CDMs. The upper part contains specific CDMs and the lower part lists general CDM frameworks. These general frameworks embrace the specific CDMs through setting parameter restrictions.

The **CDM** package in R covers all model formulations given in Table 4. Table 5 provides an overview of the essential function calls.

Model Equation

The decisions qualitatively justified in the previous subsection have to be implemented quantitatively in the form of model equations and likelihoods for enabling parameter estimation and thus responding to the Questions Q1-Q3 seen earlier. We illustrate that point using the example of the DINA model. The DINA model has two main properties: P1 the DINA model is non-compensatory, i. e., examinees cannot compensate for a lack in one skill with a surplus in another skill. P2 the examinees' probability of responding an item positively increases only if examinees possess all skills required for the respective item, i. e., the probability increases only if an interaction term becomes non-zero. The model equation employs these two properties as follows:

(P1) If we knew the exact skill profile of student i, his expected response to item j would be expressed through

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \in \{0,1\}.$$

That is, examinee i is only expected to respond to item j positively ($\eta_{ij} = 1$) if he possesses all skills required for the item, i. e. all skills assigned to item j in $\boldsymbol{Q}$. Analogously, the examinee is not expected to respond to the item positively ($\eta_{ij} = 0$) if he lacks at least one required skill.

For example, the examinees may fail to produce a positive response, although they are expected to positively respond to the item because they possess the required skills (a case which is termed "slip"). Or the other way round, they may succeed by luckily guessing the correct response, although they are not expected to positively respond an item. These probabilistic error components are modeled as item specific parameters g_j (for guessing item j) and s_j (for slipping in item j).

(P2) Combining the expected response and the probabilistic error component, the response probabilities in the DINA model are expressed through

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

That is, the DINA model involves only two probabilities g_j and $1 - s_j$ for responding item j. The probability of positively responding to item j increases from g_j to $1 - s_j$ only if examinees possess all required skills and thus η_{ij} is non-zero.

We would advise against interpreting the item parameters as guessing or slipping rates. Rather, we suggest the following two lines of interpretation:

Firstly, based on the models' assumptions, examinees who possess all relevant skills for a correct response to an item, are expected to master the item. Analogously, examinees who do not possess all relevant skills, are not expected to master the item. Following these expectations, the manifest item responses allow for a sharp separation of examinees possessing the relevant skills or not. The item parameters g_j and s_j reflect the portion of examinees showing responses other than the expected. Large values of g_j indicate the portion of correct responses, although the examinees were not expected to master the item, and large values of s_j indicate the portion of incorrect responses, although the examinees possess all of the required skills. Hence, the larger the sum of item parameters $g_j + s_j$ (i. e. the smaller $1 - g_j - s_j$), the less the skills were suitable to predict the actual item responses. Items exhibiting large values of $g_j + s_j$ were therefore not sufficiently capable of correctly separating examinees possessing the relevant skills from those who do not. This interpretation can be seen similar to the

Table 4 ■ Selected CDMs and their major references. The upper part of the table involves specific CDMs, the lower part consists of general CDM frameworks.

Model	Model name	Reference
DINA	Deterministic Input Noisy "And" Gate	Haertel (1989)
DINO	Deterministic Input Noisy "Or" Gate	Templin and Henson (2006)
RRUM	Reduced Reparameterized Unified Model	Hartz (2002)
HO-DINA	Higher Order DINA	de la Torre and Douglas (2004)
MS-DINA	Multiple strategies DINA	de la Torre and Douglas (2008)
MC-DINA	Multiple-choice DINA	de la Torre (2009a)
GDM	General Diagnostic Model	von Davier (2005)
G-DINA	Generalized DINA	de la Torre (2011)
LCDM	Log-Linear CDM	Henson, Templin, and Willse (2009)

Table 5 ■ Functions and arguments in the functions for specifying and estimating CDM models with the R package **CDM**.

Model	Function in **CDM**	Arguments
DINA	`din()`	`rule="DINA"` (default)
DINO	`din()`	`rule="DINO"`
RRUM	`gdina()`	`rule="RRUM"`
HO-DINA	`gdina()`	`HOGDINA=1`
MS-DINA	`slca()`	with appropriate design matrix
MC-DINA	`mcdina()`	
GDM	`gdm()`	
G-DINA	`gdina()`	
LCDM	`gdina()`	`linkfct="logit"`

item discrimination parameter in the item response theory framework (cf. Birnbaum, 1968), where values close to or greater than one indicate a good separation of examinees with low abilities from examinees with high abilities. Thus we might interpret $\omega_{1j} = 1 - g_j - s_j$ in a similar fashion as the item discrimination parameter (see also Lee, de la Torre, & Park, 2012).

Secondly, g_j describes the probability of correctly responding in case of not possessing all relevant skills, while $1 - s_j$ describes the probability of correctly responding in case of having all skills. Thus, the term $\left[g_j + (1 - s_j)\right]/2$ might be seen as the average probability of correctly responding the item. In case g_j decreases or s_j increases (i. e. $1 - s_j$ decreases) the average probability of a correct response decreases. In this respect, one may compare $\omega_{2j} = \left[g_j + (1 - s_j)\right]/2$ to the idea of item easiness and thus ω_{2j} corresponds to the item p-values (i. e. the percentage of examinees who responded the item correct). For an example of both interpretations see first example of the last section.

Likelihood Function

In practice, the individual skill profiles $\boldsymbol{\alpha}_i$ are not known, therefore the item parameters $g_1, \ldots, g_J, s_1, \ldots, s_J$ and the skill profiles have to be estimated. One option is joint maximum likelihood estimation. However, as in traditional item response models, joint maximization of the item parameters (structural parameter) and the skill profiles (incidental parameter) may yield inconsistent estimations of the item parameters (Neyman and Scott, 1948, for CDMs: de la Torre, 2009b). Instead we maximize the marginal likelihood

$$\begin{aligned}\log L(\boldsymbol{g}, \boldsymbol{s}, \boldsymbol{\gamma}) &= \sum_{i=1}^{I} \log L(\boldsymbol{X}_i; \boldsymbol{g}, \boldsymbol{s}, \boldsymbol{\gamma}) \\ &= \sum_{i=1}^{I} \log \left[\sum_{l=1}^{L} P\left(\boldsymbol{X}_i | \boldsymbol{\alpha}_l; \boldsymbol{g}, \boldsymbol{s}, \boldsymbol{\gamma}\right) \times P(\boldsymbol{\alpha}_l | \boldsymbol{\gamma}) \right]\end{aligned}$$

with respect to the guessing parameters $\boldsymbol{g} = [g_1, \ldots, g_J]$, the slipping parameters $\boldsymbol{s} = [s_1, \ldots, s_J]$ and the $L = 2^K$ skill class probabilities $P(\boldsymbol{\alpha}_l)$, where $\boldsymbol{\gamma} = [P(\boldsymbol{\alpha}_1), \ldots, P(\boldsymbol{\alpha}_{2^K})]$ describes a vector of skill class probabilities. Note that $P(\boldsymbol{\alpha}_{2^K}) = 1 - \sum_{k=1}^{2^K - 1} P(\boldsymbol{\alpha}_k)$ and thus the model has only $2^K - 1$ instead of 2^K skill class probabilities which need to be estimated. In cases where models have almost as many parameters as observations, which, consequently, would lead to weakly or non identifiable skill classes, Xu and von Davier (2008) proposed to change from the estimation of all skill class probabilities $P(\boldsymbol{\alpha}_l)$, $l = 1, \ldots, 2^K$, to a log-linear reduced form (called reduced skill space).

Estimation Algorithm

Because of desirable convergence behavior and simple estimation steps within iterations, the optimization is implemented via an expectation-maximization-algorithm (EM-algorithm; Dempster, Laird, & Rubin, 1977). Prior to the first iteration of the EM-algorithm, initial item parameters $[\boldsymbol{g}, \boldsymbol{s}]$ and skill class distribution parameters $P(\boldsymbol{\alpha}_l)$ have to be chosen. Then, the EM-algorithm alternates between the E-step and the M-step until convergence is achieved. The algorithm converges if a stopping criterion is fulfilled, e. g. if the maximal change between the parameter values or the relative change in the deviance is below a predefined value.

In the E-Step two types of expected counts are derived from the posterior

$$P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i) = \frac{P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l|\boldsymbol{\gamma})}{\sum_{m=1}^{L} P(\boldsymbol{X}_i|\boldsymbol{\alpha}_m)P(\boldsymbol{\alpha}_m|\boldsymbol{\gamma})}, \quad l = 1,\ldots,L.$$

The first count is the expected number of examinees which are classified into skill classes $\boldsymbol{\alpha}_l$ for item j, $l = 1,\ldots,L$, $j = 1,\ldots,J$. The second count gives the number of examinees classified in skill class $\boldsymbol{\alpha}_l$ while responding to item j positively.

In the M-Step the item parameters and the skill class distribution parameters are updated consecutively. Primarily, the first derivative of the log-likelihood with respect to the item parameters is set to zero. The derivative involves only the two counts obtained in the first step and thus allows for updating the item parameters. Secondly, the expected number of examinees in skill class $\boldsymbol{\alpha}_l$ is determined as a basis for updating the skill class distribution Q2 and the skill possession probabilities Q1.

Once the algorithm has converged, the individual skill profiles may be derived from the estimated model via maximum likelihood (MLE), maximum a posterior (MAP) or expected a posterior (EAP) classification.

Interested readers will find statistical details of the models and the estimation algorithm in e. g. de la Torre (2009b), George and Robitzsch (2014), or Xu and von Davier (2008).

Examples

CDMs may be estimated with the **CDM** package (George, Robitzsch, et al., in press) in R (R Core Team, 2015), a free open source software. Both the program and the package are available on the server cran.r-project.org. In this section, we present two applications of CDMs together with the respective R code. The first example analyzes the students' abilities in understanding grammatical rules of English. The analysis is based on the ECPE data, which were introduced in the first subsection of the previous section. The second example examines students' abilities in mathematics. This example is more complex in that it builds on a multiple matrix booklet design of the large scale Trends in International Science Study (TIMSS; Martin & Mullis, 2013).

A Simple Example - Analysis of ECPE Data

This example demonstrates the basic functionality of the **CDM** package, exemplified with the DINA model.

Response Data and Q-Matrix

For the ECPE, educational experts divided the ability to understand english grammatical rules into three skills, *"morphosyntactic rules"* (α_1), *"cohesive rules"* (α_2), and *"lexical rules"* (α_3). Furthermore, the experts decided that students require all skills assigned to an item in the Q-matrix to solve the respective item. Hence, the most appropriate model is the DINA. The R commands for the following analysis is given in Listing 1.

First (Listing 1, Line 1), we have to load the package **CDM** which automatically provides the example data set `data.ecpe`. It contains both the response data `data.ecpe$data` and the corresponding Q-matrix `data.ecpe$q.matrix` (cf. the description in the Section "Input and Output of a CDM analysis").

We may now immediately apply the model (Listing 1, Line 5), which is invoked with the function `din()`. This function requires at least two arguments, the name of the data set `data.ecpe$data` and the Q-matrix `data.ecpe$q.matrix`. Note that the data includes student IDs in the first column, which need not to be included in the response matrix for the `din()` function. The estimated model is now stored in the object `ecpe`.

Item and Skill Characteristics

Lines 7 to 21 in Listing 1 demonstrate how to extract item and skill related information from the `ecpe` results object. We obtain the item parameters (i. e. guessing g_j and slipping s_j), the item p-values, the statistics ω_{j1} and ω_{j2} (as introduced earlier), and the skill characteristics (according to Q1, Q2 and Q3 seen above). For that purpose, the `IRT.se` function extracts the model parameters along with their standard errors from the result object `ecpe` into the data frame `param`. Then, we use the `split` function to group the values by parameter type, resulting in a list named `p`. Finally, we extract the item parameters (Lines 13 and 14) and the skill parameters (Lines 19 and 20), and compute the item statistics (Line 12: p-values, Line 15: ω_{j1}, and Line 16 ω_{j2}). Furthermore, line 21 shows how to extract individual skill profiles according to MLE classification (cf. Q3 earlier).

The obtained information (Tables 6 to 9) may be interpreted as follows: Table 6 shows a summary of the item related values. The item p-values describe the percentage of

Listing 1 ■ R commands for analyzing the ECPE data.

```
# load package
library(CDM)

# estimate DINA model
ecpe <- din(data.ecpe$data[,-1], data.ecpe$q.matrix)

# parameters and standard errors of DINA model
param <- IRT.se(ecpe, extended=TRUE)
p <- split(param, param$partype)

# items characteristics (cf. Table 6)
pvalues <- colMeans(data.ecpe$data[,-1], na.rm=TRUE) # item p-values
p$guess                                              # guessing parameters
p$slip                                               # slipping parameters
omega1 <- 1 - p$guess$est - p$slip$est               # item discrimination
omega2 <- (p$guess$est + (1 - p$slip$est))/2         # item easiness

# skill characteristics (cf. Table 7 and 8)
p$margprobs                                  # skill distribution Q1
p$probs                                      # skill class distribution Q2
IRT.factor.scores(ecpe, type="MLE")[1:5,]    # individual skill profile Q3

# plot model parameters
par(mfrow=c(2,2))
plot(ecpe, pattern=data.ecpe$data[1,-1])

# correlation between skills
skill.cor(ecpe)$cor.skills

# various fit criteria
fit.ecpe <- IRT.modelfit(ecpe)

# new Q-matrices
newq13 <- newq23 <- data.ecpe$q.matrix
newq13[,4] <- 1*(newq13[,1]==1 | newq13[,3]==1)
newq23[,4] <- 1*(newq23[,2]==1 | newq23[,3]==1)
newq13 <- newq13[,c(2,4)]
newq23 <- newq23[,c(1,4)]

# define, estimate and derive model fit of competing models
ecpe13 <- din(data.ecpe$data[,-1], newq13)
fit.ecpe13 <- IRT.modelfit(ecpe13)
ecpe23 <- din(data.ecpe$data[,-1], newq23)
fit.ecpe23 <- IRT.modelfit(ecpe23)

# compare competing models
IRT.compareModels(fit.ecpe, fit.ecpe13, fit.ecpe23)
```

Table 6 ■ Summary of item characteristics for DINA model on ECPE data.

Type	Min	Max	Mean	SD
item p-values	.43	.90	.72	.13
guessing parameters g_j	.19	.82	.55	.17
slipping parameters s_j	.04	.37	.15	.09
item discrimination $\omega_{1j} = 1 - g_j - s_j$	.12	.50	.29	.10
item easiness $\omega_{2j} = (g_j + (1 - s_j))/2$	.44	.88	.70	.13

Table 7 ■ Skill distribution $P(\alpha_k)$ and respective standard errors SE for DINA model on ECPE data, cf. Q1 in the second section.

	α_1	α_2	α_3
$P(\alpha_k)$	.49	.61	.63
SE	.02	.02	.01

Note. α_1: morphosyntactic rules; α_2: cohesive rules; α_3: lexical rules

students solving each of items: 43% of the students solved the most difficult item, whereas 90% of the students solved the easiest item. On average, 72% of the students solved the items, hence the test seems not too difficult. As discussed in the third section, there exists a correspondence between the item p-values and the item easiness parameter ω_{2j}. The item guessing parameters range from .19 to .82 (SD = .17) and have a maximal standard error of .02. Note again that the object `p$guess` includes the guessing parameters as well as their standard errors. The item slipping parameters range from .04 to .37 (SD = .09) and have a maximal standard error of .01. The item discriminations ω_{1j} range from .12 to .50 with a mean value of .29. These values indicate that the items do not separate very good between examinees possessing the relevant skills and those who do not.

Table 7 shows the skill probabilities $P(\alpha_k)$ in the sense of Q1 of the second section). We see that only 49% of the students possess *"morphosyntactic rules"* (α_1), whereas 61% of the students obtain *"cohesive rules"* (α_2), and 63% of the students acquire *"lexical rules"* (α_3).

Table 8 shows the skill class distribution $P(\boldsymbol{\alpha}_l)$ (in the sense of Q2 in the second section) and allows an analysis of skill combinations. We learn from the skill class distribution that most students possess either all skills ($P([1,1,1]) = .45$) or none of the skills ($P([0,0,0]) = .31$). Furthermore, students possessing skill α_1 often also possess skill α_3 ($P([1,0,1]) + P([1,1,1]) = .02 + .45 = .47$) in contrast to students possessing skill α_1 but not skill α_3 ($P([1,0,0]) + P([1,1,0]) = .01 + .01 = .02$). In the same line, students possessing skill α_2 often also possess skill α_3 ($P([0,1,1]) + P([1,1,1]) = .01 + .45 = .46$) in contrast to students possessing skill α_2 but not skill α_3 ($P([0,1,0]) + P([1,1,0]) = .04 + .01 = .05$). The results indicate that possessing *"lexical rules"* (α_3) could be a prerequisite of acquiring either *"morphosyntactic rules"* (α_1) or *"cohesive rules"* (α_2).

Table 9 shows the individual skill profiles of the first five respondents in the data set (cf. Listing 1, Line 21). According to MLE classification, respondent $i = 2$ possesses skills 1 and 3 (i. e. *"morphosyntactic rules"* (α_1) and *"lexical rules"* (α_3), cf. Table 2), while the remaining four respondents possess all three skills.

Graphical Repesentation

The **CDM** package allows plotting some of the parameter values obtained above (Listing 1, Lines 24 and 25). The plot in Figure 1 includes the item parameters (Figure 1, top left hand side), the skill distribution (Figure 1, top right hand side), the skill class distribution (Figure 1, bottom left hand side) and, if specified in the plot command `pattern`, an individual EAP skill profile (Figure 1, bottom right hand side).

Correlations between Skills

We may further inspect the correlations between the skills (Listing 1, Line 28), which are given in Table 10. From the correlation matrix we can see that the correlations between α_1 and α_3 ($\text{Cor}(\alpha_1, \alpha_3) = .915$) and α_2 and α_3 ($\text{Cor}(\alpha_2, \alpha_3) = .914$) are large. These correlations are in line with the results we have inferred from the skill class distribution, i. e. students who possess skill α_1 tend to possess skill α_3 as well.

Model Fit

As a criterion of absolute model fit (Listing 1, Line 31) the item pairwise χ^2 measures of Chen, de la Torre, and Zhang (2013) may be considered. Without going into statistical details (cf. Groß, Robitzsch, & George, 2015), we can reject the adequacy of the model if the p-value of the maximal

Table 8 ■ Skill classes resulting for the ECPE test incorporating $K = 3$ skills with according skill class probabilities $P(\boldsymbol{\alpha}_l)$ and standard errors SE of skill class probabilities, cf. Q2 in the second section.

	$\boldsymbol{\alpha}_1$ [0,0,0]	$\boldsymbol{\alpha}_2$ [1,0,0]	$\boldsymbol{\alpha}_3$ [0,1,0]	$\boldsymbol{\alpha}_4$ [0,0,1]	$\boldsymbol{\alpha}_5$ [1,1,0]	$\boldsymbol{\alpha}_6$ [1,0,1]	$\boldsymbol{\alpha}_7$ [0,1,1]	$\boldsymbol{\alpha}_8$ [1,1,1]
$P(\boldsymbol{\alpha}_l)$	.31	.01	.04	.05	.01	.02	.01	.45
SE	.02	.01	.01	.01	.01	.01	.01	.02

Table 9 ■ Individual skill profiles $P(\boldsymbol{\alpha}_i)$ for the first five respondents $i = 1,\ldots,5$ in the ECPE data set, cf. Q3 in the second section.

Respondent i	α_{i1}	α_{i2}	α_{i3}
1	1	1	1
2	1	0	1
3	1	1	1
4	1	1	1
5	1	1	1

Note. α_1: morphosyntactic rules; α_2: cohesive rules; α_3: lexical rules

Figure 1 ■ Plot for DINA model on ECPE data including item parameters (top left hand side), skill mastery probabilities (top right hand side, cf. Q1 in the second section), skill class distribution (bottom left hand side, cf. Q2) and individual EAP skill profile for student 1 (bottom right hand side, cf. Q3).

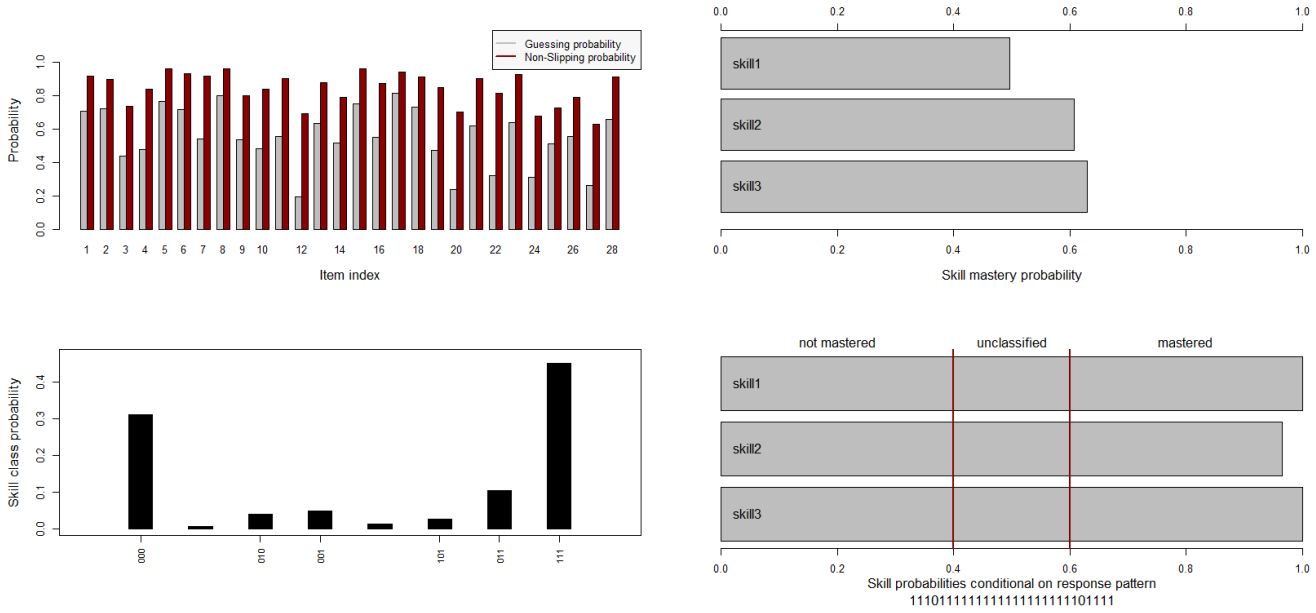

Table 10 ■ Correlations between the $k = 3$ skills for DINA model on ECPE data.

	α_1	α_2	α_3
α_1	1		
α_2	.888	1	
α_3	.915	.914	1

item pairwise χ^2 measure falls short of the desired significance level. Another criterion of absolute model fit is the standardized root mean square residual (SRMSR; Maydeu-Olivares, 2013). Maydeu-Olivares suggests that SRMSR values smaller than 0.05 indicate well-fitting models. Table 11 shows the results for the `ecpe` model, which are interpreted later on.

Because of the large correlations between the skills α_1 and α_3 or α_2 and α_3, respectively, one could argue for a model including only two skills instead of three. Thus, in the next step, we construct the Q-matrices of these two models and then compare the resulting three models with respect to their model fit. For the first model `ecpe13` including only two skills we subsume the two skills α_1 and α_3 to one new skill α_{13}. That is, we assign each item requiring either skill α_1 or skill α_3 the new skill α_{13} and end up in a two column Q-matrix `newq13` (Listing 1, Lines 34, 35 and 37). Analogously, the second model `ecpe23` including two skills is built by combining skills α_2 and α_3 in a new Q-matrix `newq23` (Listing 1, Lines 34, 36 and 38). After having estimated the two models (Listing 1, Lines 41 and 43), we calculate the model fit (Listing 1, Lines 42 and 44) and compare all three models (Listing 1, Line 47). Tables 11 and 12 are automatically generated by the latter command and yield the results of the model comparison.

Based on Table 11 the best model in terms of model fit is the `ecpe` model including all three skills, which has the highest loglike and the lowest AIC and BIC values of all three models. As shown in Table 12 a likelihood ratio test (for an introduction see Hélie, 2006) for the comparison of the `ecpe13` and `ecpe` model or the `ecpe13` and `ecpe` model, respectively, assigns the `ecpe` model with three skills a significantly better fit. However, the adequacy of all three models is rejected by the maximal item pairwise χ^2 measure. In contrast, all models' SRMSRs are smaller than 0.05 indicating at least satisfactory model fit.

Considering all results (skill class distribution, correlations, model comparisons) we found a strong coherence between skills *"cohesive rules"* (α_2) and *"lexical rules"* (α_3) and we have reasons to believe that students need to possess α_3 before they can acquire α_2.

A Complex Example - A Multi-Group Model on TIMSS Data

The second example illustrates the application of CDMs to complex data structures, like those typically appearing in large scale assessments.

Response Data

In this example we analyze students' abilities in mathematical sub-competences based on a part of the Austrian TIMSS 2011 data. Working with a data set coming from a large scale assessment study as TIMSS presents three challenges: (1) the data structure, (2) the weighting, and (3) the estimation of the parameters standard errors.

(1) The data structure specific to large scale assessments is called multiple matrix design and results from the fact that only a subset of all items involved in the study is presented to each of the students. The data we are working on includes responses of 1010 Austrian fourth grade students and a total of 47 items. The 47 items are divided up into three blocks, so called booklets. Two of the three booklets are presented to each student. Hence, the response data involves items which were not presented to the individual student (coded as missing `NA`), items, which the student had omitted or not reached (coded as false `0`), wrong item responses (also coded as false `0`) and correct item responses (coded as right `1`).

(2) The 1010 Austrian fourth graders constitute a sample drawn from all Austrian fourth graders. However, for cost reasons, students are drawn in a two-stage procedure: firstly, schools are drawn from the population of schools and secondly classes are drawn from all classes in the sampled schools. Because larger schools (i. e. their students) have a higher probability to be sampled, the individual students have unequal probabilities to be drawn. To compensate for these unequal selection probabilities and to compensate for non-response occurring because of students being not able to take part in the test (e. g. because of illness), the application of sample weights is a commonly used technique (cf. Technical Report on TIMSS; Martin & Mullis, 2013).

(3) For the complex data structures in large scale assessments it is hard or sometimes impossible to find a closed form (i. e. a formula) for determining the model parameters' standard errors. Because of that the standard errors are estimated using jackknife procedures (e. g. Friedman, Hastie, & Tibshirani, 2001). The jackknife procedure implies repeatedly drawing subsamples of schools, estimating the model for each subsample and deriving the parameters variance between the models (cf. Technical Report on TIMSS; Martin & Mullis, 2013).

The **CDM** package supports all three aspects (1), (2) and (3), which is shown in this example.

Table 11 ■ Model comparisons for DINA models on ECPE data. This table has been generated with the command `IRT.compareModels(fitecpe, fitecpe13, fitecpe23)`.

Model	Loglike	Npars	Nobs	AIC	BIC	$p\left(\max(\chi^2)\right)$	SRMSR
ecpe	-42843	63	2922	85813	86190	< 0.001	0.033
ecpe13	-42958	59	2922	86035	86387	< 0.001	0.034
ecpe23	-42864	59	2922	85847	86200	< 0.001	0.033

Table 12 ■ Likelihood ratio tests for DINA models on ECPE data.

Model1	Model2	χ^2	df	p
ecpe13	ecpe	229.65	4	< 0.001
ecpe23	ecpe	41.81	4	< 0.001

Q-Matrix

As we are interested in the students' possession of mathematical sub-competences we rely on the TIMSS competence model. In this model educational experts subdivided the overall ability of math in fourth grade into three content skills and three cognitive skills (Table 13).

Mastering each of the TIMSS items requires the knowledge of exactly one content and one cognitive skill, i. e. a possible Q-matrix for this model would have exactly two ones in each row (cf. Table 14). For example mastering the first TIMSS item presumes students to possess the content skill *"numbers"* (α_N) and the cognitive skill *"applying"* (α_A).

George and Robitzsch (2014) showed that applying such Q-matrices including two facets (here content and cognition) renders the model non-identified. Thus, George and Robitzsch propose to apply an alternative matrix, which includes each skill combination between the facets. Table 15 includes the upper part of this Q-matrix.

Multiple-Group Model

Obviously, in this example we could perform the steps of analysis shown in the first example), resulting in the percentages of students possessing the 9 skills (cf. Table 15) and the combinations thereof. However now we focus on comparing the abilities of boys and girls. Listing 2 contains the R commands for the following analysis.

We start again with loading the required packages (Listing 2, Lines 1 and 2) and continue with loading the data (Listing 2, Line 5) into a data frame `timss.info`. This object also yields additional information as e. g. student IDs, school IDs, information on the student's gender, sample weights, jackknife zones and ability values. Hence, we extract the students' response data in the object `timss` (Listing 2, Line 11). This is done by the help of the item labels already defined in the Q-matrix `timssq` (Listing 2, Line 8; cf. Table 15). The response data is structured in a multiple matrix design, compare aspect (1) in the subsection "response data" of the second example.

Specifying a multiple group model for the comparison of boys and girls presumes the invariance of item parameters between the groups (i. e. boys and girls). Thus we first estimate a G-DINA model[2] `timss.all` with all students (Listing 2, Line 14). Note that the Q-matrix `timssq` again includes item IDs, which should be removed for the model's estimation. The desired multiple group model requires the models' item parameters `coef(timss.all)$est`, which have to be structured item-wise and stored in a list `param.all` (Listing 2, Line 17). Beyond that list of fixed item parameters `param.all`, the multiple group model `timss.mg` requires several other arguments (Listing 2, Line 20): The data `timss`, the Q-matrix `timssq`, the grouping vector `timss.info$female` and the sampling weights `timss.info$TOTWGT` (cf. aspect (2) in the subsection "response data" of the second example). The grouping vector and the sampling weights are given as additional information in the `timss.info` data frame. Note that the grouping vector of the multiple group models has to consist of the entries one (for the first group) and two (for the second group). The `timss.info$female` vector contains zeros (for the first group; i. e. boys) and ones (for the second group; i. e. girls), thus we have to calculate "+1" (Listing 2, Line 20).

Standard Errors for Model Parameters

For determining the parameters' standard errors we apply a jackknife procedure (Listing 2, Lines 23 to 27), compare aspect (3) in the subsection "response data" of the second example. A discussion of jackknife procedures (or, more generally, of replication methods) in the context of large scale assessments goes far beyond the scope of this article. For a general introduction, see George, Oberwimmer, and

[2] For Q-matrices involving only one entry per row DINA and G-DINA models are equivalent.

Listing 2 ■ R commands for analyzing TIMSS data.

```
library(CDM)
library(BIFIEsurvey)

# Data with additional information
timss.info <- data.timss11.G4.AUT.part$data

# Q-matrix involving combinations between content and cognitive skills
timssq <- data.timss11.G4.AUT.part$q.matrix1

# Response data
timss <- timss.info[,paste(timssq$item)]

# Model including all students for calibration of item parameters
timss.all <- gdina(timss, timssq[,-1])

# listwise form of item parameters for input in multiple group model
param.all <- split(coef(timss.all)$est, coef(timss.all)$itemno)

# Multigroup model with invariant item parameters between groups
timss.mg <- gdina(timss, timssq[,-1], weights=timss.info$TOTWGT, group=timss.info$
    female+1, delta.fixed=param.all)

# Generate replicate design for calculation of SEs (requires BIFIEsurvey)
repdes <- IRT.repDesign(data=timss.info, wgt="TOTWGT", jktype="JK_TIMSS", jkzone="
    JKCZONE", jkrep="JKCREP")

# Calculate SEs and define object param including modelparameters
jtimss.mg <- IRT.jackknife(timss.mg, repDesign=repdes)
param.mg <- jtimss.mg$jpartable

# Skill possession (Q1) of and between groups (group1=boys, group2=girls)
skill.dist <- param.mg[which(param.mg$partype=="margprobs"),]
skill.dist.boy <- skill.dist[grep("lev1_group1", skill.dist$parnames),]
skill.dist.girl <- skill.dist[grep("lev1_group2", skill.dist$parnames),]
skill.dist.diff <- skill.dist.boy$value - skill.dist.girl$value

# Define derived parameters (differences in skill possession between boys and girls
    ) for calculating their SEs
dp <- list(
"skilldiffDA" = ~ 0 + I(prob_skillCo_DA_lev1_group1 - prob_skillCo_DA_lev1_group2),
"skilldiffDK" = ~ 0 + I(prob_skillCo_DK_lev1_group1 - prob_skillCo_DK_lev1_group2),
...
"skilldiffNR" = ~ 0 + I(prob_skillCo_NR_lev1_group1 - prob_skillCo_NR_lev1_group2)
)

# Calculate SEs of drived parameters (differences between groups)
jtimssmg_dp <- IRT.derivedParameters(jtimssmg, derived.parameters=dp)
diffs <- summary(jtimssmg_dp)

# Test differences for beeing significantly different from zero
diffs$t <- diffs$value / diffs $jkse
diffs$sig <- "␣"
diffs$sig[1-pnorm(abs(diffs$t)) < 0.025] <- "*"
```

Table 13 ■ Sub-competencies underlying the TIMSS test for mathematical abilities. The upper part includes content specific skills, whereas the lower part involves cognitive skills.

Parameter	Skill	Description
α_D	data display	reading and interpreting displays of data; organizing data
α_G	geometric shapes and measures	identifying properties of lines, angles, and a variety of (two- and three-dimensional) geometric figures
α_N	numbers	understanding of place value, ways of representing numbers, relationships between numbers
α_A	applying	application of mathematical tools in a range of contexts
α_K	knowing	recall of mathematical language, basic facts and conventions of number, symbolic representation, and spatial relations
α_R	reasoning	capacity for logical, systematic thinking; intuitive and inductive reasoning based on patterns and regularities

Table 14 ■ Original Q-matrix for TIMSS data. This Q-matrix would render the model unidentified and is therefore not apt for analysis.

Item	α_D	α_G	α_N	α_A	α_K	α_R
1	0	0	1	1	0	0
2	0	0	1	0	0	1
3	0	0	1	0	0	1
⋮						⋮
109	1	0	0	1	0	0

Note. α_D data display; α_G geometry; α_N numbers; α_A applying; α_K knowing; α_R reasoning

Itzlinger-Bruneforth (in press), technical details regarding their application in the TIMSS technical report (Martin & Mullis, 2013). Before we can apply the jackknife to our data, we have to prepare a so-called replicate design (`repdes`). For purpose of determining the replicate design, the **CDM** package provides the function `IRT.repDesign` (based on package **BIFIEsurvey**, BIFIE; 2015), which is called in Line 23 of Listing 2. The option `jktype="JK_TIMSS"` specifies that we use the standard method of TIMSS (others are available, e. g. the Bootstrap). In the TIMSS context the replicate design combines, roughly spoken, two kinds of information: Firstly, a design how to draw subsamples of schools in order to get "good" estimates while not determining all possible subsamples (this is structured through the so-called jackknife zones). Secondly, a rule on how to weight the schools in the subsamples (the replication weights). Usually, data sets of large scale studies deliver supplemental information (i. e. specific columns) required for the execution of the jackknife. In our case the `timss.info` data frame contains the columns JKCZONE and JKCREP, which are used for defining the jackknife zones and the replication weights. The jackknife method itself is invoked with the function `IRT.jackknife` (Listing 2, Lines 26), using the data set `timss.mg` and the replicate design `repdes`. Note that this function call may take a while to finish for it comprises separate model estimation runs for each jackknife sample. The entire results of the jackknife procedure are stored in the `jtimss.mg` object, from which we extract the parameter estimates along with their estimated standard errors (`param.mg`) in Line 27.

We extract the marginal probabilities from `param.mg` by choosing `param.mg$partype=="margprobs"`. By that, we obtain the percentages of students possessing the nine skills (cf. Table 15) with respect to Q1 (cf. the second section) and store them in the data frame `skill.dist` (Listing 2, Line 30). From the data frame `skill.dist` we can now tag the percentages of skill possession for boys (group1) and for girls (group2). That step is conducted with the `grep` command, which collects all information about `lev1` (possession of skill; in contrast `lev0` non-possession) in `group1` (Listing 2, Line 31) or `group2` (Listing 2, Line 32), respectively. Based on the skill possession of boys (`skill.dist.boys`) and girls (`skill.dist.girls`), we can compute the differences in skill possession `skill.dist.diff` (Listing 2, Line 33). Note that the result objects `skill.dist.boys` and `skill.dist.girls` also contain the according standard errors, whereas the standard errors of the derived differences have to be computed in a separate step (see next subsection).

Table 15 ■ Final Q-matrix for TIMSS data.

Item	α_{DA}	α_{DK}	α_{DR}	α_{GA}	α_{GK}	α_{GR}	α_{NA}	α_{NK}	α_{NR}
1	0	0	0	0	0	0	1	0	0
2	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	1
⋮					⋮				⋮
109	1	0	0	0	0	0	0	0	0

Note. α_{DA} *"data and applying"*, α_{DK} *"data and knowing"*, α_{DR} *"data and reasoning"*, α_{GA} *"geometry and applying"*, α_{GK} *"geometry and knowing"*, α_{GR} *"geometry and reasoning"*, α_{NA} *"numbers and applying"*, α_{NK} *"numbers and knowing"*, α_{NR} *"numbers and reasoning"*.

Table 16 holds the percentages of boys and girls possessing the nine skills (cf. Q1 in the second section) together with the associated standards errors. Boys succeed best in mastering *"geometry and applying"* (α_{GA} = .60) and worst in *"numbers and knowing"* (α_{NK} = .40). In contrast, girls show the highest abilities in *"data and reasoning"* (α_{DR} = .58) and the lowest abilities in *"geometry and reasoning"* (α_{GR} = .36) and *"numbers and reasoning"* (α_{NR} = .36).

t-Tests

In order to determine whether boys and girls differ significantly in their skill possession, we apply t-tests (Listing 2, Lines 48 to 50). In the `R` commands first the t-values are calculated and (Listing 2, Line 48) then the significance is evaluated at the 5 percent level and stored in a new row of the `diffs` object (Listing 2, Lines 49 and 50), where asterisks indicate significance.

Table 17 holds the derived differences in skill possession between boys and girls along with their standards errors. In the significance row asterisks indicate significant differences in skill possession.

Standard Errors for Derived Parameters

For determining the standard errors of the differences in skill possession between the groups we need two futher steps: First, we have to define a list `dp` of parameters derived from the model parameters (Listing 2, Lines 36 to 41). In our case we define as derived parameters the differences in skill possession between boys and girls (e. g. for *"data and applying"* the difference `skilldiffDA`) based on the model parameters, i. e. the skill possession in each group. For accessing the model parameters we make use of their parameter names, which can be found in the `skill.dist` data frame. Hence `prob_skillCo_DA_lev1_group1` denotes the skill mastery probability of boys in α_{DA} *"data and applying"* and analogously `prob_skillCo_DA_lev1_group2` stands for the skill mastery probability of girls in α_{DA}. Note that the syntax for specifying the derived parameters follows the `R` formula syntax, which is for example applied in the linear model `lm()` function. In Lines 37 to 40, the differences in three of the nine skills are defined, the notation of the remaining six differences appears by interchanging the two letters defining the skill (e. g. DA or DK). After the definition of the derived parameters their standard errors can be calculated using the `IRT.derivedParameters` method (Listing 2, Lines 44) and subsequently applying the summary command (Listing 2, Lines 45).

On a descriptive level, we find some revealing tendencies: With positive differences in skill possession indicating advantages for boys, we see boys superior in six out of the nine mathematical skills (α_{DK}, α_{GA}, α_{GK}, α_{GK}, α_{NA}, α_{NR}). Specifically, boys outperform girls in all skills including geometry (α_{GA}, α_{GK}, α_{GK}). The skill *"numbers and reasoning"* (α_{NR}) holds the largest difference between boys and girls: 47% of boys compared to only 36% of girls possess this skill.

However, only the largest difference in the skill possession (in *"numbers and reasoning"*, α_{NR}) is significantly different from zero. That means also that for all other skills the hypothesis of equal differences between boys and girls cannot be rejected from a statistical point of view.

Discussion

The two examples of the last section have shown that CDMs provide a handy means to trace back complex processes to basic features: A correct usage of the CDM framework leads to very differentiated results about students' skill possession. However we have to keep in mind that in any CDM the quality of the output strongly relies on the quality of the input, i. e. the experts' qualitative specification of the Q-matrix.

There might be cases in which fit measures indicate good fit and skills are highly correlated. Then, the CDM approach, which is a multidimensional one, is not necessarily the optimal one. One could rather think of applying a unidimensional model, like the Rasch model (Rasch, 1960), for

Table 16 ■ Population based skill distribution (in the sense of Q1 in the second section) for possession of TIMSS skills for boys $P(\alpha_k|\text{boys})$ and girls $P(\alpha_k|\text{girls})$ with according standard errors SE.

	α_{DA}	α_{DK}	α_{DR}	α_{GA}	α_{GK}	α_{GR}	α_{NA}	α_{NK}	α_{NR}
$P(\alpha_k\|\text{boys})$	.44	.58	.52	.60	.44	.43	.52	.40	.47
SE	.04	.04	.04	.03	.03	.04	.04	.04	.04
$P(\alpha_k\|\text{girls})$	.44	.56	.58	.56	.42	.36	.43	.42	.36
SE	.03	.04	.03	.03	.03	.04	.03	.04	.03

Note. See Table 15 for the variable descriptions.

Table 17 ■ Differences in skill possession between boys and girls. The differences' standard errors SE are used for determining if boy and girls differ significantly in skill possession (see row significance, significance level .05).

	α_{DA}	α_{DK}	α_{DR}	α_{GA}	α_{GK}	α_{GR}	α_{NA}	α_{NK}	α_{NR}
$P(\alpha_k\|\text{boys}) - P(\alpha_k\|\text{girls})$	−.01	.03	−.06	.03	.03	.07	.09	−.02	.11
SE	.05	.05	.04	.04	.05	.05	.05	.05	.05
significance	-	-	-	-	-	-	-	-	*

Note. See Table 15 for the variable descriptions.

example. The simple example offers a good starting point for such a discussion.

The present article referred to basic CDMs embedded in two educational examples using dichotomous responses and skills. More general CDMs allow for polytomous items and polytomous skills as well (Chen & de la Torre, 2013; von Davier, 2005). Furthermore, we presented all models on the assumption that all 2^K skill classes are represented in the analysis. However there are situations in which we only estimate a number of $L \leq 2^K$ skill classes: Firstly, one may have prior information, that not all skill classes are plausible. Secondly, the user wants to avoid ambiguous skill classes (for details Groß & George, 2014), or thirdly, the analysis involves too many skills for estimating all skill classes given the sample size (George & Robitzsch, 2014; Pan & Thompson, 2007).

More advanced methods for CDMs are available and are implemented in the **CDM** package of `R`. For example, based on the results of the ECPE example, one could define CDM models including skill hierarchies (Templin & Bradshaw, 2014) supported by the method `skillspace.hierarchy`. Or, if the whole TIMSS data should be analyzed analogously to the second example methods of skill space reduction (Xu & von Davier, 2008) as implemented in `skillspace.reduction` may prove useful.

Beyond the application of CDMs, readers may be interested in linking CDMs to other latent variable models: As CDMs rely on both discrete manifest and discrete latent variables, they can be considered as restricted latent class models (Formann, 1985). If manifest indicators are discrete and latent variables are continuous, one should apply item response models (for an introduction see e. g. de Ayala, 2009); if, in contrast, both manifest indicators and latent variable are continuous, structural equation models (for an introduction see e. g. Beaujean, 2014) are the method of choice.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). VA: American Psychiatric Publishing. doi:10.1176/appi.books.9780890425596

Beaujean, A. A. (2014). *Latent variable modeling using R: a step-by-step guide*. New York: Roudledge.

BIFIE; (2015). *BIFIEsurvey*. R package Version 1.4-0, https://cran.r-project.org/package=BIFIEsurvey.

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.

Buck, G. & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157. doi:10.1191/026553298667688289

Chen, J. & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*, 419–437. doi:10.1177/0146621613479818

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123–140. doi:10.1111/j.1745-3984.2012.00185.x

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

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*(3), 163–183. doi:10.1177/0146621608320523

de la Torre, J. (2009b). DINA model parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130. doi:10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199. doi:10.1007/s11336-011-9207-7

de la Torre, J. & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353. doi:10.1007/BF02295640

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. doi:10.1007/s11336-008-9063-2

de la Torre, J., van der Ark, L. A., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Develpment, online first*. doi:10.1177/0748175615569110

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

DiBello, L., 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, Psychometrics* (pp. 979–1030). Amsterdam: Elsevier. doi:10.1016/S0169-7161(06)26031-0

Feasel, K., Henson, R., & Jones, L. (2004). *Analysis of the Gambling Research Instrument (GRI)*.

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology, 38*, 87–111.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Berlin: Springer.

George, A. C., Oberwimmer, K., & Itzlinger-Bruneforth, U. (in press). Sampling. In C. Schreiner & S. Breit (Eds.), *Large-Scale Assessment mit R: methodische Grundlagen der österreichischen Bildungsstandardüberprüfung [Methods in large-scale assessments with R]*. Wien: UTB.

George, A. C. & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test and Assessment Modeling, 56*(4), 405–432.

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (in press). The R package CDM for cognitive diagnosis modeling. *Journal of Statistical Software*.

Groß, J. & George, A. C. (2014). On permissible attribute classes in noncompensatory cognitive diagnosis models. *Methodology: European Journal of Research Methods for Behavioral Sciences, 10*(3), 100–107.

Groß, J., Robitzsch, A., & George, A. C. (2015). Cognitive diagnosis models for baseline testing of educational standards in math. *Journal of Applied Statistics, online first*. DOI: 10.1080/02664763.2014.1000841.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–323. doi:10.1111/j.1745-3984.1989.tb00336.x

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

Hélie, S. (2006). An introduction to model selection: tools and algorithms. *The Quantitative Methods for Psychology, 2*(2), 1–10.

Henson, R. & Templin, J. (2007, April). *Large-scale language assessment using cognitive diagnosis models*. Paper presented at the Annual meeting of the National Council for Measurement in Education, Chicago, IL.

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. doi:10.1007/s11336-008-9089-5

Jaeger, J., Tatsuoka, C., Berns, S. M., & Varadi, F. (2006). Distinguishing neurocognitive functions in schizophrenia using partially ordered classification models. *Schizophrenia bulletin, 32*(4), 679–691. doi:10.1093/schbul/sbj038

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. doi:10.1007/s12564-011-9196-3

Li, H., Hunter, C. V., & Lei, P.-W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing, online first*. doi:10.1177/0265532215590848

Martin, M. O. & Mullis, I. V. (Eds.). (2013). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Maydeu-Olivares. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Inter-*

disciplinary Research and Perspectives, *11*, 71–137. doi:10.1080/15366367.2013.831680

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

Pan, J. & Thompson, R. (2007). Quasi-monte carlo estimation in generalized linear mixed models. *Computational Statistics & Data Analysis*, *51*(12), 5765–5775. doi:10.1016/j.csda.2006.10.003

R Core Team. (2015). *R: a language and environment for statistical computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Pædagogiske Institut.

Ravand, H. & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation*, *20*(11), 1–12.

Rupp, A. & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, *6*, 219–262. doi:10.1080/15366360802490866

Tatsuoka, K. (1984). *Analysis of errors in fraction addition and subtraction problems*. University of Illinois, Urbana-Champaign.

Templin, J. & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*, 317–339. doi:10.1007/s11336-013-9362-0

Templin, J. & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. doi:10.1037/1082-989X.11.3.287

Templin, J. & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50. doi:10.1111/emip.12010

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

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

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)*. Educational Testing Service.

Citation

George, A. C., & Robitzsch, A. (2015) Cognitive Diagnosis Models in R: A Didactic. *The Quantitative Methods for Psychology*, *11*(*3*), 189-205.

Received: 12/08/2015 ~ Accepted: 18/09/2015