Studies in Educational Evaluation 35 (2009) 64–70

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models

Olga Kunina-Habenicht [a,*], André A. Rupp [b], Oliver Wilhelm [a]

[a] Institute for Educational Progress (IQB), Humboldt University Berlin, Unter den Linden 6, D-10099 Berlin, Germany
[b] Department of Measurement, Statistics, and Evaluation (EDMS), University of Maryland, 1230A Benjamin Building, College Park, MD 20742, USA

ARTICLE INFO

Keywords:
Student evaluation
Evaluation methods
Structural equation models
Cognitive diagnosis model
Mathematics achievement
Elementary school students

ABSTRACT

In recent years there has been an increasing international interest in fine-grained diagnostic inferences on multiple skills for formative purposes. A successful provision of such inferences that support meaningful instructional decision-making requires (a) careful diagnostic assessment design coupled with (b) empirical support for the structure of the assessment grounded in multidimensional scaling models. This paper investigates the degree to which multidimensional skills profiles of children can be reliably estimated with confirmatory factor analysis models, which result in continuous skill profiles, and diagnostic classification models, which result in discrete skill profiles. The data come from a newly developed diagnostic assessment of arithmetic skills in elementary school that was specifically designed to tap multiple skills at different levels of definitional grain size.

In large-scale educational assessments, unidimensional latent variable models from item response theory (IRT) (e.g., de Ayala, 2009; Embretson & Reise, 2000) with continuous proficiency scales are typically used to model response data from a single domain such as mathematics, science, or reading. These models are statistically parsimonious and can produce reliable rank-ordering information about individual children or groups of children at aggregate levels (see, e.g., the technical reports of National Assessment of Educational Progress or Programme for International Student Assessment).

Yet, if a more differentiated understanding of the basic cognitive skills underlying assessment performance is intended, this approach can be limiting. Since most constructs in educational assessment require multiple cognitive skills for children to successfully answer questions on a diagnostic assessment, a multidimensional skills profiling approach can be more suitable if separate statistical information is desired for each component skill.

Indeed, the practical need for assessments that provide such multidimensional skill profiles has recently been clearly articulated (Huff & Goodman, 2007; see also Leighton & Gierl, 2007 more generally). If the design of the educational assessment allows for sufficient information about each of the latent skills of the children to be culled from the response data – and this is a big "if" in practice – then *multidimensional latent variable models* would consequently be more suitable models for data analysis.

Diagnostic classification models (DCMs) can yield such multidimensional diagnostic profiles based on statistically-driven multivariate classifications of children (for an overview of these models see, e.g., DiBello, Roussos, & Stout, 2007; Junker, 1999; Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010). Many DCMs are discrete alternatives to traditional multidimensional latent variable models from confirmatory factor analysis (CFA) (e.g., McDonald, 1999) or IRT (e.g., Ackerman, Gierl, & Walker, 2003).

Despite their theoretical potential in the educational measurement literature, the number of successful practical applications of these models has, so far, remained relatively small. Most applications of DCMs are found as half-hearted add-ons to simulation studies (e.g., de la Torre & Douglas, 2004). Yet, typically neither a particular added value for practical decision-making is articulated (for an exception see Templin & Henson, 2006) nor an investigation of the robustness of the conclusions is conducted (for an exception see Anozie & Junker, 2007).

Furthermore, DCMs are often retrofitted in studies that utilize data that come from assessments that were originally developed for unidimensional scaling purposes such as the TOEFL (von Davier, 2005a) or NAEP (Xu & von Davier, 2006, 2008). As a result, estimation problems frequently occur. These may include problems of non-convergence, the estimation of a number of latent dimensions that are highly correlated, and/or low reliabilities of the constituent dimensions (see also Haberman, 2008; Rupp, 2008).

To begin to fill in some of these gaps in the literature, this paper illustrates how multidimensional skill profiles of children can be developed with traditional multidimensional CFA models and state-of-the-art multidimensional DCMs. The data for the paper

* Corresponding author. Tel.: +49 30 27879089; fax: +49 30 20935336.
E-mail addresses: olga.kunina@iqb.hu-berlin.de (O. Kunina-Habenicht), ruppandr@umd.edu (A.A. Rupp), oliver.wilhelm@rz.hu-berlin.de (O. Wilhelm).

0191-491X/$ – see front matter © 2009 Published by Elsevier Ltd.
doi:10.1016/j.stueduc.2009.10.003

come from a newly developed diagnostic assessment of arithmetic abilities for elementary school in grades 3 and 4. Specifically, the study was guided by the following three research questions:

1. How well can pilot data from a newly developed diagnostic assessment of arithmetic ability for elementary school be scaled with multidimensional CFA models and DCMs?
2. How do global and local model fit and resulting parameter interpretations differ across the CFA models and the DCMs?
3. What lessons can be learned about the benefits and limitations of DCMs vis-à-vis CFA models?

This paper has three main sections. In the next section of the paper, the theoretical framework for developing the diagnostic assessment for this study is presented. The section after that discusses the data structure and the CFA and DCMs that were fitted. The final section describes the results from the data analyses. The paper concludes with a synthesis of how the three research questions could be answered with these data and what future research directions should be.

Theoretical framework

Instrument development

The diagnostic mathematics assessment (DMA) for elementary school children for this study aims to provide a differentiated profile of basic arithmetic and modelling skills for children in grades 3 and 4. In the first step of the item development process, relevant skills in mathematics in elementary school were identified based on the review of the international didactic literature (see, e.g., Carpenter, Fennema, Franke, Levi, & Empson, 1999; NCTM, 2000). In addition, reports of typical conceptual mistakes that children make on basic arithmetic items were used to guide this process. There are two components to the DMA, one consisting of context-free arithmetic items, so-called *computation items*, and one consisting of contextualized arithmetic items, so-called *word items*. Fig. 1 shows sample problems used on the DMA.

(a)

	+	22	=	27
	+	45	=	80
25	+		=	73
	+	6	=	104
113	+		=	204

(b)

T_AS5

Marlen has 55 songs on her MP3-Player. Leon has 15 songs less than Marlen. When they go through their song lists, they realize that the lists have 10 songs in common.

How many different songs do Marlen und Leon have altogether?

Marlen und Leon have ☐ songs altogether.

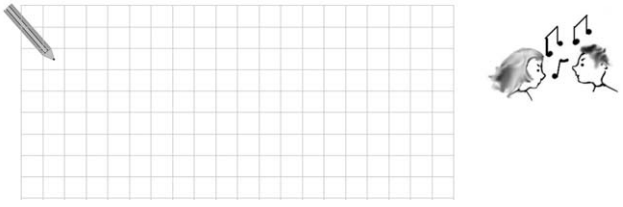

Fig. 1. (a) Example of simple computation items. (b) Example of a word item for addition and subtraction T_AS5.

Table 1
Sample Q-matrix for simple computation items and word item T_AS5.

Item	Addition	Subtraction	Multiplication	Division	Modelling
$X+22=27$	1	0	0	0	0
$113+X=204$	1	0	0	0	0
$95-X=60$	0	1	0	0	0
$76-45=X$	0	1	0	0	0
$14\times X=28$	0	0	1	0	0
$X\times 6=72$	0	0	1	0	0
$24:X=8$	0	0	0	1	0
$72:4=X$	0	0	0	1	0
T_AS5	1	1	0	0	1

The following assessment design was used. At the coarsest level, the objective was to develop at least a reliable unidimensional "basic arithmetic skills" scale. At the next finer level, the objective was to develop five reliably distinct, albeit correlated, proficiency scales for "addition", "subtraction", "multiplication", "division" and "modelling". Table 1 shows a classification for some simple computation problems and one word item according to the skills they measure for illustration purposes.

Note that this table is also known as the loading- or *Q-matrix* (e.g., Tatsuoka, 1983) for these items. It can contain single entries per item reflecting simple skill structure or multiple entries per item reflecting complex skill structure, which is also known as *within-item multidimensionality* (see Adams, Wilson, & Wang, 1997). The word items were classified in the same manner except that they were all classified as measuring the "modelling" skill in addition to the arithmetic skills.

Overall, there are a total of 52 computation items and a total of 35 word items on the DMA; all answers were dichotomously scored (i.e., "correct" vs. "incorrect") for this study. For the computation items, this resulted in a total of 14 simple addition items, 14 simple subtraction items, 12 simple multiplication items, and 12 simple division items.

For the word items, this resulted in a total of 3 problems that require just addition skills, 3 problems that require just subtraction skills, 3 problems that require just multiplication skills, 3 problems that require just division skills and 22 problems that require a combination of two skills. While the computation items were identical across both grades, a multi-matrix design for the word items was developed due to ceiling effects for word items in grade 3 grounded in curricular differences and children's cognitive development across grades.

This paper focuses on the analyses that concern the four basic arithmetic skills and the modelling skill. Attributes analyzed in the current study were a subset of a larger number of attributes.

Selecting a suitable DCM

Due to the fact that the data for this paper come from a pilot study that provides a limited number of responses for each item, it was decided to utilize a relatively simple DCM in terms of the number of model parameters. The DCM that was selected is defined as a member of the *general diagnostic model* (*GDM*) family (e.g., von Davier, 2005a). The GDM family subsumes a wide variety of compensatory DCMs (i.e., models that reflect the assumption that a lack of proficiency on one skill can be compensated for by a surplus on another skill).

A two-parameter GDM with two ability levels was employed for this study. This model is a discrete version of a compensatory two-parameter multidimensional IRT model. Due to the close structural relationship between this IRT model and a compensatory CFA model with unrestricted loadings and error variances – despite some estimation differences – the GDM is also structurally comparable to the latter (e.g., McDonald, 1999; Thissen & Wainer, 2001; see also Takane & de Leeuw, 1987). Consequently, it would

be expected that statistical information provided from such a CFA model and the GDM would be very similar resulting in very similar conclusions about the children. Mathematically, the two-parameter version of the GDM for dichotomous response data can be represented as follows:

$$P(X_{ij}=1)=\frac{\exp(\beta_{0j}+\sum_{k=1}^{K}\beta_{1jk}\alpha_{ik}q_{jk})}{1+\exp(\beta_{0j}+\sum_{k=1}^{K}\beta_{1jk}\alpha_{ik}q_{jk})}.$$

Symbolically, X_{ij} is the observed response of child i to item j, β_{0j} is an intercept parameter that can be viewed as the difficulty parameter for item j, β_{1jk} is a slope parameter that can be viewed as the discrimination parameter for each item on each skill dimension, α_{ik} is the discrete latent variable for child i on skill k indicating whether child i has mastered skill k ($\alpha_{ik}=1$) or not ($\alpha_{ik}=0$), and q_{jk} is the binary entry in the Q-matrix indicating whether skill k is measured by item j ($q_{jk}=1$) or not ($q_{jk}=0$). Thus, the response probability for item j is computed as a function of a baseline probability that depends on β_{0j} and increments that depend on β_{1jk} for mastered skills relevant to responding to the item.

Methods

Data collection

Data for $N=464$ elementary school children (238 boys, 224 girls, 2 unspecified) from 10 classes in the 3rd grade ($n=241$) and 10 classes in the 4th grade ($n=223$) were used in this study; the data were collected in April 2008. The mean ages in 3rd grade and 4th grade were 8.72 years (SD = 0.52 years) and 9.86 years (SD = 0.55 years), respectively. The 10 classes were selected from six different schools located in different federal states of Germany. The DMA was administered under standardized conditions at the particular schools by seven trained graduate students. Additionally, school grades in mathematics as well as demographic information about sex, age and migrant status were collected from all children. The testing session consisted of two school periods of 45 min with a 10 min break in between.

Item exclusion for latent variable analyses

In order to calibrate the latent variable models from the two different measurement frameworks, poorly functioning items had to be excluded to alleviate convergence problems due to severe item misfit. The first criterion for item exclusion was that items with an empirical item difficulty greater than .90 or smaller than .10 were removed from the analyses; item difficulty was estimated by the percent-correct or p-value from classical test theory (see, e.g., Thissen & Wainer, 2001, chapter 2). Similarly, items with low discrimination values, as measured by point-biserial correlation indices below .20, were excluded. Finally, items with low factor loadings based on a stepwise exploratory factor analysis solution for tetrachoric correlation matrices (see Kano & Harada, 2000) were excluded as well; the analyses consisted of fitting one-factor models for each basic skill (e.g., addition, subtraction) separately. For the 3rd grade, six addition, four subtraction, two multiplication, four division and six modelling items were eliminated due to the criteria mentioned above. For the 4th grade, seven addition, four subtraction, two multiplication, one division, and two modelling items were excluded from the analysis due to the same criteria.

Model estimation

Multidimensional CFA models were estimated with *Mplus 5* (Muthén & Muthén, 1998–2008) using the mean- and variance-corrected weighted least-squares (WLSMV) estimator for dichotomous indicators based on tetrachoric correlations. The GDM was estimated with the MDLTM software (von Davier, 2005b) that was made available to the authors as a research license.

A sequence of nine models was fit within the CFA framework corresponding to different underlying latent skill structures that reflected reasonable hypotheses about skill separability based on the DMA design as well as empirical fit results from some models. All models were fit to data from grades 3 and 4 separately. It was decided to differentiate model fit for a larger number of CFA models first and then to fit the GDM that corresponded to the best-fitting CFA model. This was done because assessing global model fit, local item fit, as well as the fit of nested and non-nested models is better understood and documented within the CFA literature at this point.

The first eight models postulate a simple loading structure, meaning that variance in each item is explained by only one latent dimension. Specifically, in model M1 only one latent variable representing basic arithmetic skills was estimated. In model M2-a, four correlated latent variables were estimated corresponding to addition, subtraction, multiplication, and division skills. In model M2-b, one higher-order latent variable was additionally estimated. In model M3-a, five correlated latent variables were estimated corresponding to addition, subtraction, multiplication, division, and modelling skills. In model M3-b, an additional higher-order latent variable was estimated. In model M4-a, two correlated latent variables were estimated, one representing addition and subtraction skills and one representing multiplication and division skills. In model M4-b, an additional correlated latent variable was estimated representing modelling skills, which is a model that is statistically equivalent to including a higher-order latent variable that is not shown. In the last model M5, a complex loading structure for computation items and word items was estimated. That model included the same latent variables as model M3-a. While for each computational item only one regression parameter for each relevant computational skill (addition, subtraction, multiplication or division) was postulated, one loading for the latent factor for modelling as well as one or more additional loadings for each required computational skills were estimated for the word items.

Results

Model fit, selection, and interpretation for CFA models

Table 2 shows the model fit indices for all nine models. Using cut-off values of the *Comparative fit index* (*CFI*) >.95 and the *Root mean square error of approximation* (*RMSEA*) <.06 as a rough general guideline (Hu & Bentler, 1999; Kaplan, 2000), model fit indices indicate an unacceptable model fit of model 1 for grades 3 and 4. The model fit of models with five correlated factors (3-a and

Table 2
Fit indices for competing CFA models.

Model	Grade 3				Grade 4			
	χ^2	d.f.	CFI	RMSEA	χ^2	d.f.	CFI	RMSEA
1	572.7	79	.864	.161	375.8	79	.945	.130
2-a	233.2	87	.960	.083	197.2	91	.980	.072
2-b	234.1	87	.959	.084	196.7	91	.980	.072
3-a	247.1	127	.951	.063	224.2	132	.981	.056
3-b	245.7	127	.951	.062	227.7	132	.980	.057
4-a	256.0	85	.953	.091	213.0	89	.977	.079
4-b	258.1	126	.946	.066	235.7	131	.978	.060
5	270.7	123	.940	.071	226.1	130	.980	.058

Notes. d.f.: degrees of freedom; CFI: comparative fit index; and RMSEA: root mean square error of approximation.

Table 3
Factor loadings of the final CFA model (model 3-a) for the 3rd and 4th grade.

Addition			Subtraction			Multiplication			Division			Modelling		
#	λ^3	λ^4	#	λ^3	λ^4	#	λ^3	λ^4	#	λ^3	λ^4	#	λ^3	λ^4
A1	.61	.77	*S1*	.84	.83	*M1*	.90	.94	*D1*	.91	.94	*MD1*	.51	–
A2	.72	.77	*S2*	.94	.95	*M2*	.82	.85	*D2*	.96	.94	*MD2*	.68	.60
A3	.73	.83	*S3*	.84	.87	*M3*	.92	.84	*D3*	.97	.96	*MD3*	.59	–
A4	.74	.82	*S4*	.92	.97	*M4*	.80	.96	*D4*	.82	.92	*MD4*	.60	.53
A5	.87	.53	*S5*	.89	.97	*M5*	.79	.79	*D5*	.89	.81	*MD5*	.69	.48
A6	.75	.81	*S6*	.74	.77	*M6*	.86	.86	*D6*	.85	.94	*MD6*	.68	.75
A7	.83	–	*S7*	.82	.69	*M7*	.86	.88	*D7*	.92	.89	*MD7*	.84	.65
A8	.84	.65	*S8*	.70	.75	*M8*	.88	.80	*D8*	–	.99	*MD8*	.61	–
			S9	.69	.75	*M9*	.82	.92	*D9*	–	.96	*MD9*	.51	.58
			S10	.67	.55	*M10*	.78	.93	*D10*	.90	.91	*MD10*	.57	.66
									D11	–	.91	*MD11*	.58	–
												MD12	.62	–
												MD13	.63	.67
												MD14	.86	–
												MD15	.66	.74
												MD16	.67	.71
												MD17	–	.73
												MD18	–	.86
												MD19	–	.67
												MD20	–	.45
												MD21	–	.81
												MD22	–	.87
												MD23	–	.82
												MD24	–	.77
												MD25	–	.76
												MD26	–	.68
												MD27	–	.72

Notes. λ^3: factor loading in the 3rd grade; and λ^4: factor loading in the 4th grade.

3-b) is slightly better than the model fit of models with three correlated factors (4-a and 4-b) due to higher complexity. As expected, the model fit indices for models with higher-order latent variables are very similar to those of their non-hierarchical counterparts. Model 5 with the complex loading structure had to be rejected due to several non-significant loadings for items with multiple loadings.

Based on these results, model 3-a with a simple loading structure was selected as the final CFA model and was used for the computation of latent skill scores as well as for comparison with the GDM. All factors loadings in this model were significant at $\alpha = .05$ and ranged from $\lambda = .51$ to $\lambda = .97$ for the 3rd grade and from $\lambda = .45$ to $\lambda = .99$ for the 4th grade, respectively. Table 3 provides detailed information on the factor loadings for the final model for the 3rd and 4th grade. Table 4 presents the correlations between the latent skill variables for this model for both grades.

An inspection of the correlation pattern for both grades reveals that there exist high latent correlations between addition and subtraction in both grades ($\varphi = .83$ and $\varphi = .75$ for grades 3 and 4, respectively) and high correlations for multiplication and division ($\varphi = .88$ and $\varphi = .95$ for grades 3 and 4, respectively). Put simply, with the current form of the DMA it is not possible to achieve a reliable empirical separation between these pairs of skills; each latent skill dimension contributes little unique information over and above the other. Additionally, modelling skills correlate substantively with all basic arithmetic skills in grade 4 showing that an empirical separation of this skill is similarly difficult in this grade.

Table 4
Correlations between latent skill variables in the best-fitting CFA model.

Item	Addition	Subtraction	Multiplication	Division	Modelling
Addition	.92/.90	.75	.64	.66	.78
Subtraction	.83	.95/.95	.68	.67	.76
Multiplication	.42	.42	.96/.97	.95	.83
Division	.44	.47	.88	.97/.98	.85
Modelling	.78	.68	.44	.39	.92/.95

Notes. Correlations for 3rd grade are shown in the lower off-diagonal while correlations for 4th grade are shown in the upper off-diagonal. Reliability estimates for the latent skill variables for both grades (3rd/4th) are shown in the diagonal; these are measured by the ω coefficient (e.g., McDonald, 1999).

Model fit and parameter interpretation of GDM

For the purposes of this pilot study, the GDM corresponding to CFA model M3-a was fit to the data. The MDLTM software provides for each child the probabilities of latent class membership for all of the $2^5 = 32$ theoretically possible latent classes as well as a marginal distribution of all these latent classes in the sample. Person parameter estimates for a latent skill variable depend on the percentage of correct answers given to items that measure this skill in the entire sample as well as on the responses of the individual child. The sufficient statistic for the ability estimates is the weighted sum of item scores that are supposed to measure that skill calculated as the maximum a posteriori estimate (von Davier, 2005a). For such skills, several points along the continuum are defined to represent different ability levels and the crossing of the levels across latent dimensions results in the total number of latent classes that are estimated by the GDM.

The MDLTM software provides an estimate of the reliability that is similar to reliability indices in IRT (e.g., Adams, 2006). It also provides two information-based fit indices for relative model fit comparisons, the Akaike's information criterion (AIC) (Akaike, 1974) and a Bayesian information criterion (BIC) (Schwarz, 1978). It also provides an item fit statistic (*Item-fit RMSEA*), which essentially compares the model-predicted item response probabilities for a correct response for respondents in different latent classes with the observed proportions of correct responses by the responses weighted by the proportion of respondents in each latent class. This fit statistic should not be confused with the popular omnibus test for model fit in the CFA literature; its formula is given in Appendix A (M. von Davier, personal communication, January 20, 2009).

Table 5
Correlations between latent skill variables in the GDM.

	Addition	Subtraction	Multiplication	Division	Modelling
Addition	.77/.76	.60	.61	.58	.64
Subtraction	.71	.87/.88	.55	.57	.64
Multiplication	.39	.39	.85/.93	.95	.74
Division	.32	.31	.85	.84/.93	.73
Modelling	.54	.54	.30	.24	.82/.88

Notes. Correlations for 3rd grade are shown in the lower off-diagonal while correlations for 4th grade are shown in the upper off-diagonal. Reliability estimates for both grades (3rd/4th) are shown in the diagonal.

Item fit

The item fit indices for the GDM showed that only 2% of the items showed good fit (RMSEA $< .05$), 50% of the items showed moderate fit (RMSEA $< .10$), and 48% of the items showed poor fit (RMSEA $> .10$). Even though the impact of such item misfit on subsequent inferences about respondents and items has not been studied in detail for the GDM at this point, it is not advisable to use the model for high-stakes inferences in this context. Additional analyses using a larger data set from the field trial coupled with more extensive model fit comparisons using a larger class of DCMs will, hopefully, present a different picture. Nevertheless, a few additional descriptive details based on this GDM are presented in the following to make some relevant general points about the statistical similarity of output and conclusions between the CFA model and the GDM.

Correlation patterns

The latent correlations between the discrete latent skill variables estimated by the MDLTM software are shown in Table 5.

As expected from the pattern in Table 4 for the CFA model, the correlation pattern of the discrete individual skill estimates for the GDM is similar. That is, addition and subtraction are moderately correlated in both grades ($\varphi = .71$ and $\varphi = .60$ for grades 3 and 4, respectively) while multiplication and division are highly correlated in both grades ($\varphi = .85$ and $\varphi = .95$ for grades 3 and 4, respectively). Furthermore, the modelling skill is correlated moderately with the four basic arithmetic skills in grades 3 and 4. The absolute magnitude of these correlations is lower in the DCM than in the CFA model, which is essentially a result of the fact that the former model uses discrete latent variables while the latter uses continuous latent variables.

Latent class distribution

Even though there are a total of $2^5 = 32$ latent classes that can be theoretically distinguished without postulating any conditional relationships among the latent skill variables, fewer latent classes could be empirically distinguished for the DMA data due to the design of the instrument. Table 6 shows the *mixing proportions* (i.e., the proportions of children in each latent class) for the seven most frequently populated classes in grade 3 and their counterparts in grade 4 as well as the mastery rates for each skill for both grades. These mastery rates are based on the estimated empirical ability parameters.

The most prevalent latent class membership in 3rd grade is observed for the latent class where none of the skills were mastered (23.5%), followed by a latent class where all five skills were mastered (21.0%). Another 14.2% of the children succeeded in addition, subtraction and modelling skills, but had difficulties in multiplication and division. Similarly, in 4th grade the most commonly populated latent class is the one where all five skills were mastered (36.5%) followed by a latent class where none of the skills was mastered (27.7%). This pattern is typical for many real-data analyses of DCMs. Many respondents are often classified into the two latent classes that represent complete non-mastery of all skills and complete mastery of all skills. The challenge of fitting DCMs is, thus, to differentiate reliably between the classes with mixed mastery patterns of skills.

An inspection of the skill patterns across latent classes in this data set shows that the skills form a divergent hierarchy with a linear component and two branches that split from it. That is, the analyses suggest that the first skills to be mastered are addition and subtraction, which is the linear component. Subsequently, multiplication and division are mastered alongside with modelling, which are the two diverging branches. This hierarchy represents the curricular sequencing of these skills in German elementary schools well.

The marginal mastery rates for the different skills support the face validity of these findings, because they show that many children, especially in the 3rd but also in the 4th grade, have great difficulties with multiplication and division. Based on the school curricula one would expect that children that master multiplication and division skills should master addition and subtraction skills as well. Yet 16% of children in the 3rd grade and 13% in the 4th grade are classified into latent classes that contradict such an acquisition pattern, which is probably a reflection of model-data misfit. Note that differences in the skill profiles are not directly comparable across grades because partly different items were considered in the corresponding models.

Despite some concerns about such misfit these findings generally show that a DCM analysis provides information that is in alignment with information from traditional multidimensional scaling models within the CFA framework. Finally, due to the aforementioned similarities between the GDM and CFA models that were chosen for these analyses one would expect that the skill scores on the latent variables are reasonably consistent across the two models. Indeed, that is what is found with correlation estimates between factors scores from CFA and person parameters

Table 6
Latent class distribution and mastery proportions in GDM.

Skills					Class membership	
Addition	Subtraction	Multiplication	Division	Modelling	3rd grade	4th grade
0	0	0	0	0	23.5%	27.7%
1	1	1	1	1	21.0%	36.5%
1	1	0	0	1	14.2%	4.5%
1	1	1	1	0	6.6%	1.7%
1	1	0	0	0	6.3%	5.8%
1	0	0	0	0	6.0%	8.0%
1	1	1	0	1	4.6%	<1.0%
69.2%	54.5%	44.8%	37.2%	48.5%	Mastery 3rd grade	
63.4%	54.7%	46.1%	46.9%	46.3%	Mastery 4th grade	

Notes. The differences in mastery probabilities are due to the fact that test versions for the 3rd and 4th grade differed in terms of items that were kept for the current analyses.

Table 7
Correlation of latent skill scores between final CFA model and GDM.

Skill	Grade 3 (n = 241)	Grade 4 (n = 223)
Addition	.85	.89
Subtraction	.87	.86
Multiplication	.86	.83
Division	.88	.83
Modelling	.87	.85

estimated in the GDM ranging from r = .83 to r = .89 for the five latent skill variables (see Table 7).

Discussion

This study has provided some preliminary evidence that has helped to answer the three research questions posed at the outset. With regards to the first research question, a CFA model provided a better fit to the data than the GDM that was chosen as a discrete counterpart to it. Descriptively, the GDM led to plausible interpretations of latent class membership probabilities, mastery probabilities, and latent skill correlations. Given the fact that there is evidence of item misfit for a large number of items, however, it is not advisable to use this model to scale the current data for high-stakes purposes. Moreover, CFA models and the GDM struggle to accommodate a complex loading structure for the DMA data.

With regards to the second research question, the correlation patterns of the continuous latent skill variables in the CFA model and the discrete latent skill variables in the GDM, viewed separately or jointly, show that both latent variable models provide similar outcomes. These results empirically illustrate the degree to which the theoretical relationship between different multidimensional latent variable models is reflected in practice.

With regards to the third research question, the results show that it is essentially impossible to empirically separate addition and subtraction skills as well as multiplication and division skills under either modelling approach. It is possible to separate modelling skills from the basic arithmetic skills to some degree, which is also reflected in the acceptable fit for model 4-b under the CFA framework.

One important message of this paper is that the GDM does not provide any "new" information beyond the multidimensional CFA model per se. It provides "different" information — to wit, a direct representation of possible conditional skill relationships and a classification of children. This may represent potentially useful diagnostic information for teachers and parents.

From a psychometric point of view, a discretization of continuous proficiency scales for the purpose of classification almost always means a statistical loss of information. However, Haberman, von Davier, and Lee (2008) found that GDMs with only four or five ability levels per dimension are quite comparable in terms of model-data fit compared to models that assume a continuous ability distribution. Thus, in the end, arguing for an added value of one modelling framework over another depends on the purpose to which the constituent models are put as both frameworks have specific strengths and weaknesses.

The CFA approach provides a variety of descriptive model fit indices and the likelihood ratio test for the comparison and evaluation of competing models. However, this approach cannot easily yield information about mastery levels. The GDM approach seems to be an appropriate approach to look more closely at transition through the acquisition of arithmetic skills. Certainly, item evaluation and model comparison remain challenging. In the presented analyses the CFA model seems to outperform the GDM in terms of model fit. Nevertheless, GDM can be a gainful add-on to established psychometric models in the case of the estimation of non-compensatory models or comparison of alternative Q-matrices.

Providing more empirical evidence on the robustness of the results presented in this paper is indispensable to develop defensible and interpretable multidimensional skills profiles to stakeholders. To achieve these objectives, extended analyses with the field trial data from a larger sample will be conducted. These analyses will allow for more comprehensive statements about the functioning of the DMA and the utility of DCMs vis-à-vis alternative multidimensional scaling approaches.

Acknowledgements

This research was sponsored in part by grant number RU-424/3-1 from the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) in the Priority Program 1293 "Models of Competencies for the Assessment of Individual Learning Outcomes and the Evaluation of Educational Processes". We would like to thank Dr. Matthias von Davier at the *Educational Testing Service* (*ETS*) in Princeton, NJ, USA, for making a research license of his software program MDLTM available to us for the purpose of this study as well as for his comments on an early draft of this paper. We would also like to thank Kerstin Gomm for help with data collection, entry, and clean-up. Additional thanks go out to the reviewers of the first draft of this paper whose comments were very helpful to our revision.

Appendix A. Formula for the calculation of the item-fit RMSEA index

$$\mathrm{RMSEA}_j = \sum_{k=1}^{K}\sum_{d=1}^{D}\pi(a_{kd})\left[P(X=1|a_{kd,j}) - \frac{N(X=1|a_{kd,j})}{N(a_{kd},j)}\right]^2$$

Notes. a_{kd} = skill level d for dimension k; $\pi(a_{kd})$ = proportion of students belonging to the ability group a_{kd}; $P(X=1|a_{kd,j})$ = estimated model-predicted probability to solve the item j for the students belonging to the ability group a_{kd}; $N(X=1|a_{kd,j})$ = observed number of student in ability group a_{kd} who provided a correct answer to item j; $N(a_{kd,j})$ = observed number of students in the ability group a_{kd} who provided answers to the item j overall (M. von Davier, personal communication, January, 20, 2009).

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37–53.

Adams, R. J. (2006, April). Reliability and item response modeling: Myths, observations and applications. *Presented at the 13th International Objective Measurement Workshop*.

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Anozie, N., & Junker, B. (2007, April). Investigating the utility of a conjunctive model in Q-matrix assessment using monthly child records in an online tutoring system. *Paper presented at the annual meeting of the National Council on Measurement in Education (NCME)*.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L. W., & Empson, S. B. (1999). *Children's mathematics: Cognitively guided instruction*. Heinemann: Portsmouth, NH.

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

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

DiBello, L., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol. 26, psychometrics* (pp. 979–1027). Amsterdam, Netherlands: Elsevier.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 4*, 1–55.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19–60). Cambridge: Cambridge University Press.

Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment*. Unpublished manuscript. Accessed November 28, 2006, from http://www.stat.cmu.edu/~brian/nrc/cfa.

Kano, Y., & Harada, A. (2000). Stepwise variable selection in factor analysis. *Psychometrika, 65*, 7–22.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. London: Sage.

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and practice*. Cambridge: Cambridge University Press.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Muthén, L. K., & Muthén, B. O. (1998–2008). *Mplus (Version 5.0) [Computer software]*. Los Angeles, CA: Muthén & Muthén.

NCTM – National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Rupp, A. A. (2008, April). Psychological vs. psychometric dimensionality in diagnostic reading assessment: Challenges for creating integrated assessment narratives based on multidimensional profiles. *Presented at the ETS/IEA conference entitled Assessing Reading in the 21st century: Aligning and applying advances in the reading and measurement sciences*, Philadelphia, PA.

Rupp, A. 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.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

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

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Erlbaum.

von Davier, M. (2005a). *A general diagnostic model applied to language testing data (Research Report No. RR-05-16)*. Princeton, NJ: Educational Testing Service.

von Davier, M. (2005b). *mdltm – multidimensional discrete latent trait modeling software [Computer software]*. Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data (Research Report 06-08)*. Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008). *Linking in the general diagnostic model (Research Report 08-08)*. Princeton, NJ: Educational Testing Service.