Psychological Test and Assessment Modeling, Volume 53, 2011 (1), 3-34

Can a multidimensional hierarchy of skills generate data conforming to the Rasch model? A comparison of methods

Wolfgang Schoppek[1] *& Andreas Landgraf*[2]

Abstract

The question of the dimensionality of intelligent performance has kept researchers occupied for decades. We investigate this question in the context of learning elementary arithmetic. Our assumption of a polyhierarchy of skills in arithmetic (HiSkA) predicts a multidimensional structure of test data. This seems to contradict findings that data collected to validate the HiSkA conformed to the Rasch model. To resolve this seeming contradiction, we analysed test data from two samples of third graders with a number of methods ranging from factor analysis and Rasch analysis to multidimensional item response theory (MIRT). Additionally we simulated data sets based on different unidimensional and multidimensional models and compared the results of some of the analyses that were also applied to the empirical data. Results show that a multidimensional generating structure can produce data conforming to the Rasch model under certain conditions, that a general factor explains a substantial amount of variance in the empirical data, but that the HiSkA is capable of explaining much of the residual variance.

Key words: Arithmetic skills; learning hierarchy; simulation; multidimensional item-response theory; knowledge space theory

[1] *Correspondence concerning this article should be addressed to:* Wolfgang Schoppek, PhD, University of Bayreuth, Universitätsstr. 30, 95447 Bayreuth, Germany; email: wolfgang.schoppek@uni-bayreuth.de

[2] University of Bayreuth, Germany

Introduction

Since the days of Spearman and Thurstone researchers have been debating if intelligent performance can be explained by one general ability factor (Spearman, 1904, 1927) or if it is the result of a complex interplay of different independent abilities (Thurstone & Thurstone, 1941). In the present paper, we analyse the dimensionality question in the context of learning elementary arithmetic. As a basis for selecting practice problems in our software "Merlin's Math Mill" (MMM) we had developed a hierarchy of basic arithmetic skills (HiSkA). In several experiments we tested the progress pupils made when using the software with a test consisting of diverse problems, and found most of the test results to be consistent with the Rasch model (Schoppek & Tulis, 2010, Schoppek, subm.). Since the HiSkA is a polyhierarchic structure (a directed acyclic graph) containing independent substructures, the question arises, if this finding contradicts the assumptions of the HiSkA. More generally, we want to answer the question if a multidimensional hierarchy of skills can generate data that are consistent with the unidimensional Rasch model.

A look at current discussions of such questions shows that they still cannot be answered unambiguously. Since models are always simplifications of reality and data never fit models perfectly, it depends on research goals how results are interpreted. We can observe this in the recent debate between Rindermann (2006) and representatives of the German PISA consortium about what international student assessment studies really measure. Rindermann (2006) focuses on the high intercorrelations of the scales for reading, math, and science and found strong first factors in principle component analyses of reanalysed data. He concludes that international student assessment studies measure mainly general cognitive abilities, and corroborates his conclusion with content analyses of test items, showing that similar skills are required in items of all subscales. Baumert, Brunner, Lüdtke, & Trautwein (2007) respond that finding a strong *g*-factor is not sufficient for identifying it with intelligence, and that Rindermann (2006) used obsolete methods. They claim that international student assessment studies measure the results of cumulative processes of knowledge acquisition, conceding that intelligence plays an important role in these processes. To substantiate their position, they report a comparison between two structural equation models, a simple g-factor model, and a nested factor model that assumes additional domain specific factors besides a g-factor. Although the latter model fits the data better than the former and shows that the specific factors account for significant proportions of variance, the results can also be interpreted as supporting Rindermann's view, because the g-factor is still clearly the strongest one! This dispute shows that the decision if an achievement is based on a single dimension or on multiple dimensions cannot be decided simply by using the correct method. In our view, the results of different methods should be compared.

As regards the general question if items that are sensitive to differences in many dimensions can be fitted by a unidimensional model, Reckase and Stout (1995) have shown that this is the case when the items are sensitive to the same composite of skills and knowledge (see also Reckase, 2009). This occurs when the skills are highly correlated with each other. However, when the skills are elementary and hard to measure separately, the question about their correlation cannot be answered empirically. This is true for elementary arithmetic problems, where even simple equations draw on more than one skill. For example, the skill of crossing the tens boundary cannot be assessed without concurrently assessing counting skills or the knowledge of arithmetic facts about the decomposition of numbers. Therefore, we embedded simulations in our research strategy. This allowed us to analyse simulated data whose generating structure is exactly known with the same methods as the empirical data.

A research strategy of comparing simulated with real data is only possible when an exact theory is at hand. A good example for the value of theories in the context of the dimensionality question of arithmetic skills is presented by Arendasy, Sommer, and Ponocny (2005), who tested the predictions of four theories about solving different arithmetic word problems of the "compare" type (Riley, Greeno, & Heller, 1983). Compare word problems in general are the most difficult of all basic word problems involving addition and subtraction (Riley et al., 1983; Stern, 1994). However, there are subtle differences in difficulty between various subtypes, which have been explained differently by a number of theories. Arendasy et al. (2005) reviewed these theories and derived specific predictions regarding person homogeneity and item homogeneity of a set of different compare word problems. Person homogeneity means that an IRT model (such as the Rasch model) estimates the same difficulty parameters of the items, regardless of the subsample the estimation is based on. Item homogeneity is given when the estimated person parameters do not depend on the subset of items the estimation is based on. For example, the construction integration theory by Kintsch (1988; Kintsch & Lewis, 1993) predicts no item homogeneity when comparing two subtypes of problems, because different cognitive processes are assumed to be required for solving each subtype. At the same time, person homogeneity (based on different sample partitioning criteria) is predicted, because the differences between the subtypes of problems are assumed to be the same for all subjects. For testing these predictions, Arendasy et al. (2005) used nonparametric goodness-of-fit tests (*T*-statistics) that were introduced by Ponocny (2001). In a study with n=100 second graders the application of *T*-statistics revealed person homogeneity with respect to nine criteria for splitting the sample, but rejected item homogeneity between two specific subtypes of compare word problems. Without the theories, the finding would be qualified as ambiguous. However, the construction integration theory by Kintsch (1988; Kintsch & Lewis, 1993) predicts exactly this pattern of results. The study also demonstrates that it is inappropriate to rely on a single test for proving model validity.

The present analyses are structured and guided by the HiSkA, a fine-grained theory how achievement in different computation and word problems can be explained. Hence we primarily follow a deductive approach. Inductive methods are included to complement the deductive conclusions. In our attempt to test the HiSkA, we compare the results of a broad range of methods. This includes simulation of data sets based on the HiSkA and competing models and comparison of analyses between simulated and collected data. We perceive our work as a contribution to strengthen the contact between content oriented research and methodological research.

In the following, we introduce the hypothetical hierarchy of arithmetic skills, report some data collected for its empirical validation, and present the results of our analyses in order to explore the question if a multidimensional hierarchy of skills can generate data conforming to the Rasch model.

A hierarchy of skills for arithmetic – HiSkA

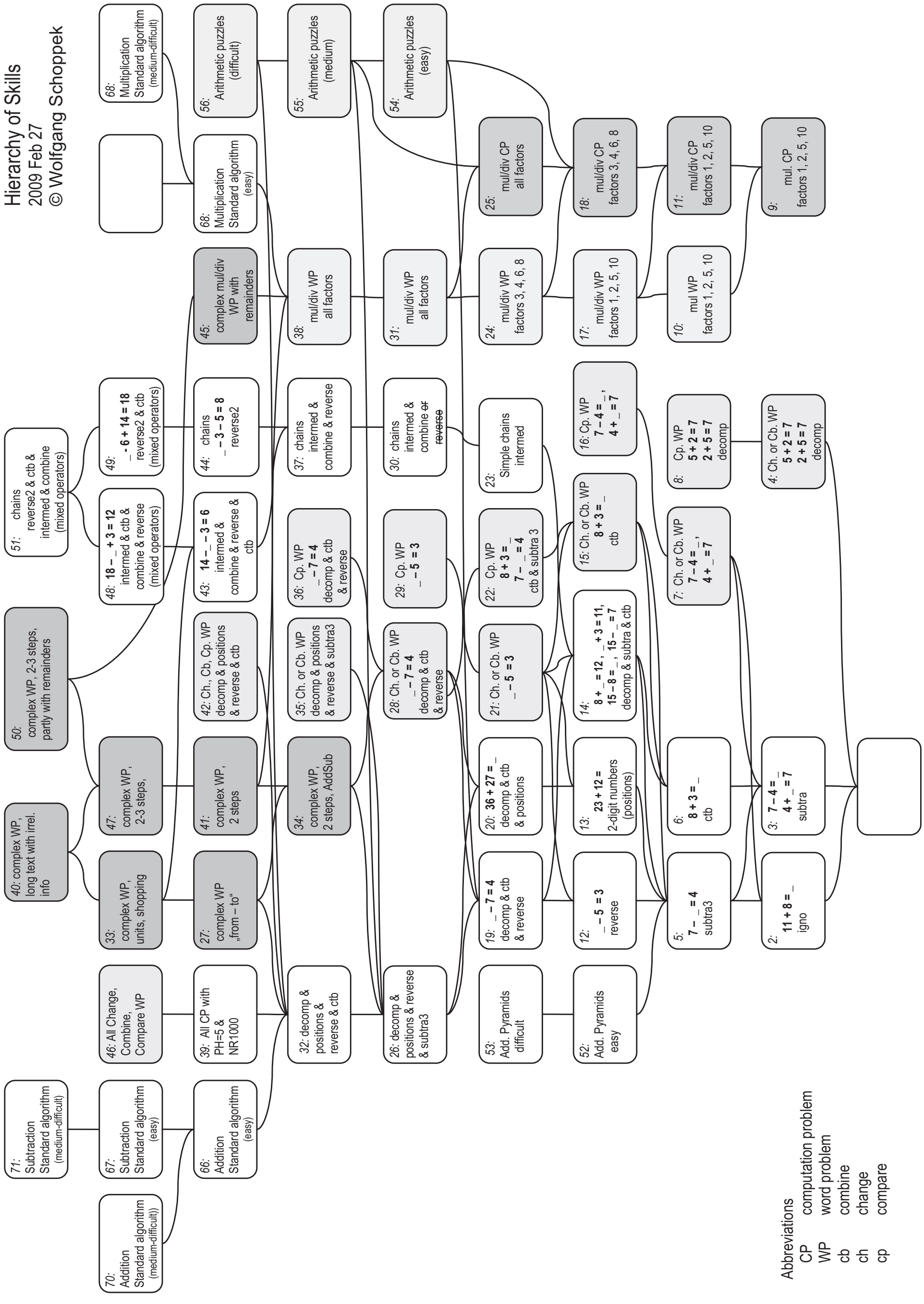
For basic arithmetic it is possible to specify hierarchies of skills that are derived from the domain structure. These hierarchies are useful for planning instruction, specific diagnosis of knowledge, and selection of appropriate practice problems. The underlying idea of decomposing complex skills was introduced by Gagne (1962), and has been validated in a number of successful applications in the 1960s and 1970s (e.g. White, 1976). The idea has recently again become subject of debate in the "math wars" (Anderson, Reder, & Simon, 2000). We have developed a hierarchy of skills for arithmetic (HiSkA) as a basis for the problem selection algorithm in the adaptive training software "Merlin's Math Mill" (MMM, Schoppek & Tulis, 2010). MMM automatically selects practice problems that are adequate to the current skill level of the user.

Starting point of the development was a systematic classification of addition/subtraction problems according to their attributes. Beginning with the simplest problems, we identified different procedures for solving these. Thereby we found a close correspondence between problem attributes and skills that are sufficient to solve them and defined prerequisite relations between skills and classes of problems. In classifying problems and skills we could resort to a large body of literature about the difficulty of elementary problems (Parkman & Groen, 1971; Kornmann & Wagner, 1990), about addition strategies (e.g. Torbeyns, Verschaffel, & Ghesquiere, 2005), about the difficulties of subtraction (e.g. Seyler, Kirk, & Ashcraft, 2003), and about recognising and utilising inverse relations between operations (Canobi, 2005; Gilmore & Bryant, 2008). The complete hierarchy is pictured in Figure 1.

It turned out, that some of the simple subskills occurred repeatedly in more difficult problems, whereas other subskills seem to be replaced by more sophisticated ones (as described by Bergan, Stone & Field, 1984). For example, crossing the tens boundary initially involves the decomposition of numbers, such as 7+5 = 7+3+2. Once the solutions can be retrieved from memory, the decomposition is no longer necessary.

The hierarchy consists of nodes representing classes of problems that can be solved with the same set of skills. The nodes are connected through links representing surmise relations (Doignon & Falmagne, 1999), meaning that if a person masters a node then he/she also masters the nodes connected downwards. In the instructional context, this also means that the problems represented by a node should only be tackled when the predecessor nodes are mastered.

Figure 1


Hierarchy of Skills
2009 Feb 27
© Wolfgang Schoppek
70: Addition Standard algorithm (medium-difficult))
71: Subtraction Standard algorithm (medium-difficult)
67: Subtraction Standard algorithm (easy)
66: Addition Standard algorithm (easy)
46: All Change, Combine, Compare WP
39: All CP with PH=5 & NR1000
32: decomp & positions & reverse & ctb
26: decomp & positions & reverse & subtra3
53: Add. Pyramids difficult
52: Add. Pyramids easy
40: complex WP, long text with irrel. info
33: complex WP, units, shopping
27: complex WP „from – to"
19: _ – 7 = 4 decomp & ctb & reverse
12: _ – 5 = 3 reverse
5: 7 – _ = 4 subtra3
2: 11 + 8 = _ igno
50: complex WP, 2-3 steps, partly with remainders
47: complex WP, 2-3 steps,
41: complex WP, 2 steps
34: complex WP, 2 steps, AddSub
20: 36 + 27 = _ decomp & ctb & positions
13: 23 + 12 = 2-digit numbers (positions)
6: 8 + 3 = _ ctb
3: 7 – 4 = _ 4 + _ = 7 subtra
42: Ch., Cb, Cp. WP decomp & positions & reverse & ctb
35: Ch. or Cb. WP decomp & positions & reverse & subtra3
28: Ch. or Cb. WP _ – 7 = 4 decomp & ctb & reverse
21: Ch. or Cb. WP _ – 5 = 3
14: 8 + _ = 12 , _ + 3 = 11, 15 – 8 = _ , 15 – _ = 7 decomp & subtra & ctb
51: chains reverse2 & ctb & intermed & combine (mixed operators)
48: 18 – _ + 3 = 12 intermed & ctb & combine & reverse (mixed operators)
43: 14 – _ – 3 = 6 intermed & combine & reverse & ctb
36: Cp. WP _ – 7 = 4 decomp & ctb & reverse
29: Cp. WP _ – 5 = 3
22: Cp. WP 8 + 3 = _ 7 – _ = 4 ctb & subtra 3
15: Ch. or Cb. WP 8 + 3 = _ ctb
7: Ch. or Cb. WP 7 – 4 = _ , 4 + _ = 7
49: _ – 6 + 14 = 18 reverse2 & ctb (mixed operators)
44: chains _ – 3 – 5 = 8 reverse2
37: chains intermed & combine & reverse
30: chains intermed & combine ~~or reverse~~
23: Simple chains intermed
16: Cp. WP 7 – 4 = _ , 4 + _ = 7
8: Cp. WP 5 + 2 = 7 2 + 5 = 7 decomp
4: Ch. or Cb. WP 5 + 2 = 7 2 + 5 = 7 decomp
45: complex mul/div WP with remainders
38: mul/div WP all factors
31: mul/div WP all factors
24: mul/div WP factors 3, 4, 6, 8
17: mul/div WP factors 1, 2, 5, 10
10: mul WP factors 1, 2, 5, 10
68: Multiplication Standard algorithm (easy)
25: mul/div CP all factors
18: mul/div CP factors 3, 4, 6, 8
11: mul/div CP factors 1, 2, 5, 10
9: mul. CP factors 1, 2, 5, 10
68: Multiplication Standard algorithm (medium-difficult)
56: Arithmetic puzzles (difficult)
55: Arithmetic puzzles (medium)
54: Arithmetic puzzles (easy)
Abbreviations
CP computation problem
WP word problem
cb combine
ch change
cp compare

The nodes are grouped in levels. A new level is established by the following criteria:

A new subskill is necessary for solving a problem type[3].

Several subskills are newly combined.

Problems are classified according to the following attributes:

Length (L):	5 or 7 elements (including operator and equal signs)
Number range (NR):	up to 10, up to 20, up to 100, up to 1000, more than 1000
Operator (OP):	plus, minus (problems with length 7: same or mixed)
Placeholder (PH):	Position 1, 3, 5, 7 (positions of operator and equal signs are also counted)
Crossing the tens boundary (CT):	with or without
N of digits of smallest number:	1, 2, (3)

The following description contains the rationale for establishing a new level and sample problems for the first three levels. The complete description can be found in the appendix. Attributes of the problems that can be solved on the respective level are given in parentheses.

Level 1

New: skill **decomp**	Decomposition of numbers between 2 and 10
Special case	Decompositions of 10
useful concept	Understand that 2 + 5 = 5 + 2

Level 2

New: skill **igno**	Ignore the tens position (when adding up to 20 without CT) 11 + 8 = _
New: skill **subtra**:	Subtraction / completion (based on **decomp**) 9 – 4 = _ , 4 + _ = 9

[3] On introduction of a new subskill, other requirements are kept as simple as possible. For example, when introducing the subskill „reversal of operators", using subtraction problems with placeholder at position 1, only problems without crossing the tens boundary are provided, even though the subskill "crossing the tens boundary" was introduced before. This has the effect that some problems on higher levels turn out to be easier than other problems on lower levels.

Level 3

New: skill **subtra3** Application of decomposition on subtraction with PH 3
8 – _ = 5

New: skill **ctb** Crossing the tens boundary (based on **decomp)**
(addition, PH 5)
7 + 6 = _

Level 1 is defined by the skill of decomposing the numbers between 2 and 10 (shorthand: "decomp"). At this level, it does not matter if students perform the skill by counting or by retrieval. Most children, however, quickly learn to retrieve these facts from memory (Geary, Hamson, & Hoard, 2000). The decompositions of 10 are particularly important for crossing the tens boundary. We believe that it is favourable to gain an understanding of commutativity already at this level.

The new skills "ignore the tens position" and "subtraction" make up Level 2. Note that these skills are not yet combined at this level. The subtraction skill is based on the understanding that the subtraction c-a=? and the completion a+?=c are both based on the same triple of numbers (a, b, c, where a+b=c), which in turn depends on the decomp skill.

At Level 3, the new skills "subtraction with placeholder at position 3" (subtra3) and "crossing the tens boundary" (ctb) are introduced. The subtra3 skill is an extension of the subtra skill from Level 2 to a new problem type. The ctb skill draws heavily on the decomp skill as two decompositions have to be performed to solve the corresponding problems (e.g. for solving 7+5, we assume the decomposition of 10=7+3 and the decomposition 5=3+2).

Inclusion of other problem types

Other problem types, such as word problems or multiplication and division problems were included in the addition-subtraction backbone according to the following principles:

Addition / subtraction word problems: We defined mastery of the underlying computation problem as a condition for presenting the respective word problems (see Fuchs, Fuchs, Compton, Powell, Seethaler, Capizzi, Schatschneider, & Fletcher, 2006 for a rationale of this decision). Most nodes in the hierarchy separate the easier change and combine problems from the compare problems.

Multiplication / division problems: We ordered the nodes according to the multiplicands, starting with 1, 2, 5, and 10, continuing with 3, 4, 6, and 8, and finishing with 7 and 9.

Multiplication / division word problems: We defined mastery of the underlying computation problem as a condition for presenting the respective word problems. Subsequently, word problems with remainders are presented.

Arithmetic puzzles: Problems typically starting with "I think of a number" were linked into the hierarchy at places that represent mastery of the underlying operations.

Complex word problems: These problems, which involve more than one calculation step, were also linked into the hierarchy at places that represent mastery of the underlying operations. Additionally, they were ordered according to difficulties that originate from the number of calculation steps or from specific features, such as the occurrence of remainders or the peculiarities of pagination (e.g. reading from page 5 to 10 means reading 6 pages).

Theoretical foundations of the applied methods

In this section we briefly introduce the theoretical foundations of some of the methods we applied to our data. We do not discuss reliability analysis and principal component analysis because we assume that most readers are familiar with these methods.

The Rasch model

The Rasch model (RM, Rasch, 1960) is probably the most fundamental in a family of models that is characterised by assuming explicit relations between the ability of a person and the probability of solving specific test items. This family of models is known as item-response theory. The Rasch model assumes an unidimensional person parameter θ_v, representing his/her ability, and an item parameter ε_i, characterising the difficulty of the item on the same scale. The probability that a person v solves item i is specified by the following logistic equation:

$$P(\text{"+"} \mid \theta_v, \varepsilon_i) = \frac{\exp(\theta_v - \varepsilon_i)}{1 + \exp(\theta_v - \varepsilon_i)}$$

If the RM is true for a set of items, item parameters are estimated to the same values regardless of the sample this estimation is based on. Also, estimates of person parameters are largely independent of the subset of items used to estimate them. This property – referred to as “specific objectivity” – has been utilised for tests of model conformity (see Fischer, 2007 for an overview). A commonly used one is Andersen’s likelihood-ratio (LR) test (Andersen, 1973), which evaluates the differences between the CML estimates of the item parameters in different subgroups. The test statistic has an asymptotic χ^2-distribution. When the LR test indicates significance, the H_0 of assuming RM conformity must be rejected. The other aspect of specific objectivity is picked up in the Martin-Löf test (Martin-Löf, 1973), which assesses item homogeneity by comparing the likelihoods of two subsets of items with the likelihood of the complete test.

It is common practice to perform model checks based on these methods to show that a set of items conforms to the RM, which means that a single dimension is sufficient to explain individual differences in test results (e.g. Fischer, 1974). However, these tests assume the validity of the model as H_0, resulting in a high type-II-risk (accepting H_0 while in the population the H_1 holds) when the type-I-risk level is set to the usual $p < .05$ (Fischer, 2007). Therefore, power analyses should be performed, which consider both types of risk as well as the sample size, to get a more differentiated answer to the question about unidimensionality. Because according to Kubinger (2005), there is no table relating power to sample size for Andersen's LR test, we performed simulations to explore its power to detect deviations from unidimensionality under different conditions.

Multidimensional item-response theory – MIRT

Since the HiSkA defines hypotheses about the skills required to solve specific classes of items, and about the independence between some of these skills, a multidimensional model of the relationship between the skills and the probability of solution can be specified and tested directly. The theory underlying the methods for testing such models is known as "multidimensional item-response theory" – MIRT (Reckase, 2007). All MIRT models assume a space made up by a number of dimensions that represent latent traits of persons. A person's position in that space is specified by the vector of his/her values on each of the dimensions. Like the RM, most MIRT models assume logistic functions relating probability of solution to the position on the latent traits. There are compensatory models, assuming that high ability in some skills can compensate for lower ability in other skills, and non compensatory models. We applied a compensatory model because according to Reckase (2007) both models predict similar probabilities (in the regions of most interest) while the former are mathematically more tractable.

For our multidimensional analysis we used the NOHARM program by Fraser and McDonald (2003). NOHARM approximates the logistic model with the normal ogive function for estimating the parameters of a compensatory MIRT model. Omitting the guessing parameter, the model equation can be written as:

$$P(y_j = 1 | \underline{\theta}\} = N[f_{j0} + \underline{f_j '}\underline{\theta}],$$

where N[.] is the normal distribution function, $\underline{\theta}$ is a vector of latent traits, $f_{j0} = -a_j b_j$ (where a_j is the discrimination parameter and b_j is the difficulty parameter), and $\underline{f_j}$ is a vector of coefficients. The elements of $\underline{f_j}$ can be interpreted as loadings of item j on the assumed dimensions. The procedure for estimation of parameters is a variant of the one described by McDonald (1982).

For testing a model, an item × skill matrix, the so called Q-matrix, must be provided a priori, which states what skills are required for the solution of each item. This matrix is similar to the structure matrix that is specified in linear logistic test models (LLTM; Embretson & Daniel, 2008; Fischer, 1973). However, as LLTMs can be viewed as specifications of unidimensional models, they are not truly multidimensional. Another advantage of the MIRT model estimated in NOHARM is the option to perform exploratory analyses. The results of a MIRT analysis are similar to those of a factor analysis: A table of loadings of items on each dimension. NOHARM also calculates an index of goodness of fit (GFI). The Tanaka GFI (Tanaka, 1993) relates the sample covariance matrix and the residual covariance matrix. A GFI of 1 indicates perfect fit. GFI values of greater than 0.90 are considered acceptable; values above 0.95 indicate good fit (McDonald, 1999).

Knowledge space theory

In the HiSkA we assume that complex skills can be decomposed into subskills, and mastery of the subskills is a condition for performing the complex skill. It follows that a person who solves a problem requiring complex skill c, which can be decomposed into the subskills a and b, should also be capable of solving problems requiring only skill a or skill b. Although such considerations were formative for constructing the HiSkA, its nodes are actually not skills but classes of items. A formal theory about such dependences between test items, the knowledge space theory, was developed by Doignon and Falmagne (1999). The most fundamental concept in this theory is the *knowledge state*, which is defined as the subset of items (from a domain of items) a person can solve correctly in ideal conditions. Since not all possible knowledge states are plausible (e.g. states where difficult items are solved and easy items are failed), the authors assume *knowledge structures*, which are sets of plausible knowledge states. When a knowledge structure is closed under union, i.e. when the union of any two knowledge states of the structure is also an element of this structure, the knowledge structure is called a *knowledge space*. Dependences between items (as described above) are called *surmise relations*, which are commonly represented as Hasse diagrams. Figure 2 shows this for a subset of items from the HiSkA. For example, when a person solves Item 34, it can be concluded on the basis of the surmise relations that she also solves Item 26 and Item 14 (also meaning that {14, 26, 34} is a knowledge state in the corresponding knowledge space). Empirically, the knowledge state of a person corresponds with his or her response pattern for a set of items. It is important to note that the knowledge state can only be estimated as a theoretical construct, because empirically, item responses are subject to error.

It is obvious that a completely different approach for modelling the relation between skills and performance is followed in the knowledge space theory than in (M)IRT. Whereas the latter assume functional relationships between latent traits and solution probabilities, the former makes qualitative assumptions (which are expressed in the language of set theory) about the relations between items that are deterministic in nature.

A number of methods have been developed within knowledge space theory that are particularly suitable for comparing the fit of data to polyhierarchical models. However, all these methods struggle with the problem of handling the probabilistic nature of error. A thorough discussion of these problems and a suggested solution is provided by Weber (2004). For our analyses we have applied the inductive item tree analysis as it is described by Sargin and Uenlue (2009).

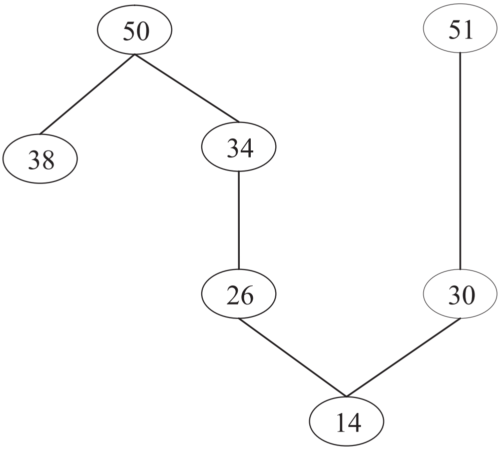


Figure 2:
Hasse diagram of the surmise relations between seven items from the math test used in our studies. The numbers refer to nodes in the HiSkA; the corresponding problems are listed in Table 1

Method

The data reported here were collected in two studies. The first sample, referred to as Sample 1, was tested as part of a larger experiment with Merlin's Math Mill (Schoppek, subm.). Sample 2 was drawn from six third classes from Bayreuth and vicinity. Data from this sample have been collected exclusively for the present work and have not been reported elsewhere.

Participants

Sample 1 consisted of *n*=264 third graders – 128 girls and 136 boys – from seven elementary schools in the region of Bayreuth, Germany. The mean age of the participants was 9;1 (SD: 4.9). Sample 2 consisted of *n*=140 third graders – 71 girls and 69 boys – from three elementary schools in the region of Bayreuth. The mean age of these children was 8;9 (SD: 5.0).

Measures

To validate the hierarchy of arithmetic skills, we developed a test made up from items that could be mapped to the nodes of the hierarchy. As there are more nodes in the hierarchy than the number of items that can be administered in a test for children of age 8, a subset of nodes had to be selected for the test. We decided to exclude nodes from the first three levels, because the corresponding problems would be too easy. We also excluded multiplication and division computation problems, because in earlier studies performance on these problems depended markedly on how recently the multiplication tables have been practised (Schoppek & Laue, 2005). We selected 17 addition and subtraction computation problems with two or three operands and the unknown at different positions. The test also included six (seven) word problems, ranging from a compare word problem to a complex word problem requiring three calculation steps, including multiplication.

The application of methods based on the knowledge space theory requires that the number of subjects is greater than the number of possible response patterns. Thus with our samples of $n \geq 140$ we can analyse a subset of seven items (2^7=128). We selected seven Items that are particularly representative for the underlying structure of skills (see Figure 2). The items are listed in Table 1.

Table 1:
Problems of the reduced HiSkA tests

Node number	**Problem**	**Mean 2006**	**Mean 2008**	**Skills**
Node 14	___ + 8 = 11	.97	.96	decomp, subtra, ctb
Node 26	54 – ___ = 41	.86	.91	decomp, positions, reverse, subtra3
Node 30	14 + 3 – 16 = ___	.87	.74	intermed, combine or reverse
Node 51	93 – 62 – ___ = 17	.44	.34	intermed, reverse2, ctb, combine
Node 34	“There are 25 children in a class. All children shall present their favourite book. 4 boys and 5 girls haven’t read a book yet. How many books will then be presented?”	.41	.84[4]	complex WP, 2 steps, AddSub
Node 38	“Sabrina is building ducks with Lego. She needs 9 bricks per duck. How many ducks can she build when she has 63 suitable bricks?”	.53	.49	MulDiv WP
Node 50	“The Circus Pellegrini needs 60 fishes everyday to feed the animals. Each of the 4 ice bears eats 8 fishes. The remaining fishes are given to the 7 seals. How many fishes does each seal get?”	.23	.24	complex WP, 3 steps, AddSub, MulDiv

[4] There is a marked difference in difficulty between 2006 and 2008 because two different problems were selected as instances for Node 34. The text displayed in Table 2 is from the 2008 test. In 2006 the problem involved adding times on a bike tour.

The structure depicted in Figure 2 combines three strands of skills. The strand from Node 14 up to Node 50 requires more and more sophisticated addition and subtraction skills. Those skills are also required in the strand from Node 14 to Node 51; but this strand additionally requires handling three operands. The third strand from Node 38 to Node 50 involves multiplication. Thus, besides much dependence between the nodes, there are also regions in the structure that are expected to be quite independent: Node 38 should be independent from Nodes 14, 26, 30, 34 and 51. Also, the greater the distance of the successors of Node 14 from the root, the more independent they should be (e.g. Nodes 34 and 51).

Procedure

Data from Sample 1 were collected in March 2007 as a follow-up test in a training experiment with MMM. The experiment had started with a pretest in October 2006, followed by a eight week intervention with MMM (or regular math instruction in the control classes), and a posttest in December 2006. The test was administered with a time limit of 45 minutes, which is sufficient for solving all 23 problems. Each correctly answered equation problem scored one point. For word problems, one point was scored only if the calculation, the correct result, and the unit stated in the answer were correct.

Data from Sample 2 were collected in October 2008, about four weeks after the beginning of the school year. Time and scoring scheme were the same as in Sample 1. At the end of both studies, teachers were handed out the results of the tests and were told to debrief the pupils.

Data analysis

As mentioned above, we analysed our data with a broad range of methods in order to compare results. At this point, we will give an overview of the methods we have applied. Some more details are provided in the appropriate sections. In a first step, we applied traditional methods based on classical test theory: internal consistency (Cronbach's alpha) and factor analysis. We continued with testing our data for RM conformity, applying Andersen's LR test and the Martin-Löf test. To test the multidimensional structure of the data directly, we conducted multidimensional IRT analyses and analyses based on the knowledge space theory. Finally, we simulated data sets on the basis of three competing models and compared the results of some of the methods that we had applied to the empirical data.

Results

Methods based on classical test theory

The most common and traditional methods for testing item homogeneity are the analysis of internal consistency (e.g. Cronbach's α) and factor analysis. Both are associated with classical test theory (Lord & Novick, 1968). Because factor analysis may create spurious factors when applied to Pearson correlations of dichotomous data (Kubinger, 2003), we based our analyses on tetrachoric correlation matrices. The analyses were conducted using the R-package “psych” (Revelle, in prep.).

Reliability analysis

The analysis of the complete tests resulted in satisfactory internal consistencies of $\alpha = .84$ (2006) and $\alpha = .81$ (2008). For a test of 23 items, this is not particularly high and leaves room for the interpretation that despite one main ability factor, there are other factors influencing the responses. This becomes even more obvious when analysing the reduced test consisting of seven items, resulting in consistencies of $\alpha = .56$ (2006) and $\alpha = .47$ (2008). Of course, these results are not conclusive as to whether they are due to measurement error or to a multidimensional generating structure.

Factor analysis

The results of principal component analyses parallel those of the reliability analyses. For the complete tests we found one main factor with a sharp bend in the scree-plot of eigenvalues (*e*) (2006: $e_1 = 8.73$, $e_2 = 2.61$; 2008: $e_1 = 8.12$, $e_2 = 2.76$). However, this factor accounts for only 38 % (2006) and 34 % (2008) of variance. Again paralleling the results of the reliability analyses, the dominance of a main factor is less obvious in factor analyses of the reduced tests. Because we present results of a MIRT version of factor analysis in more detail below, we don’t want to elaborate further on traditional methods.

To summarise, traditional methods converge in the interpretation that a substantial share of variance can be explained by a fairly general ability factor. This does not preclude the existence of other specific factors and is thus in line with the assumptions of the HiSkA.

Rasch model analyses

One advantage of methods based on item response theory (IRT) is that many of them were developed specifically for dichotomous data. Probably the most fundamental of all IRT models – the Rasch model – has the additional advantage of allowing real tests of the predictions of the model and not merely goodness-of-fit tests (Kubinger, 2007).

We conducted Rasch analyses using the R software system with the eRm package (Mair & Hatzinger, 2007). Specifically, we used two common likelihood-ratio (LR) tests to estimate the consistency of our data with the unidimensional Rasch model: Andersen’s LR test (Andersen, 1973), which evaluates the differences between the CML estimates of the item parameters in different subgroups, and the Martin-Löf test, which assesses item homogeneity by comparing the likelihoods of two subsets of items with the likelihood of the complete test. For Andersen’s LR test we formed the subgroups by a median split of the total scores; for the Martin-Löf test we separated the word problems from the computation problems.

For the complete test data from 2006, Andersen's LR test indicates a deviation from the Rasch model (*LR* = 38.14, *df* = 22, $p < .05$). The Martin-Löf test, however, indicates no significant deviation from item homogeneity (*LR* = 67.93, *df* = 101, $p > .05$). For the complete test data from 2008 none of the LR tests suggests rejection of the H_0 (RM conformity) (Andersen's *LR* = 29.20, *df* = 23, *p* = .17, Martin-Löf *LR* = 100.86, *df* = 118, $p > .05$).

We conducted the same analyses with the reduced tests. For the 2006 data not all items could be used in Andersen's LR test when splitting at the median because the easiest item was answered correctly by all participants of the group with raw scores above the median. The remaining six items indicate no significant deviation from person homogeneity (Andersen's *LR* = 8.04, *df* = 5, *p* = .154) and item homogeneity (Martin-Löf *LR* = 5.21, *df* = 11, $p > .05$). Similar results were found for the 2008 data, where also one item was dropped from the analysis when splitting at the median, resulting in a likelihood ratio that indicates no deviation from the RM (*LR* = 3.778, *df* = 5, *p* = .582). All items could be used for the Martin-Löf test, which also indicated no significant deviation (Martin-Löf *LR* = 12.62, *df* = 11, $p > .05$).

The overall picture can be interpreted as evidence for the RM conformity of all tests – the full tests as well as the reduced tests. Given the assumed multidimensional generating structure these results are astonishing. However, they are in line with earlier findings of our group where similar tests consisting of different types of problems proved to conform to the RM (Schoppek & Tulis, 2010). But failure to reject the H_0 does not mean its confirmation. Hence these results again leave space for the assumption of additional factors.

Analyses based on multidimensional IRT (MIRT)

The methods applied hitherto are commonly used to demonstrate unidimensionality. But are the results of these tests sufficient to reject our assumption of the HiSkA as generating structure? We tried to answer this question by testing the assumed multidimensional structure of our data in a confirmatory manner using the NOHARM software (Fraser & McDonald, 2003), which performs parameter estimations for compensatory MIRT models.

We calculated a series of analyses varying the number of extracted dimensions and compared the Tanaka GFI values. Exploratory analyses produced results that could be easily interpreted for the data from 2008. In a next step, we performed the same confirmatory analyses on both data sets. Starting from the seven items introduced earlier, we added some more items from the complete tests that resemble the seven items in their requirements, ending up with 12 and 13 items for 2006 and 2008, respectively. In the confirmatory analyses, we assumed one "general skill" dimension (g) for the solution of all items, a "multiplication/division" dimension ("MulDiv"), and a third dimension representing skills of dealing with more than two operands ("Chains"). The assumption of a "general skill" dimension is justified on the basis of the results reported above that indicate homogeneity of our items.

The results of the MIRT analyses are displayed in Table 2. In the “Comments” column we list the interpreted names of the factors for exploratory analyses; for confirmatory analyses we also list the name of the factors, but also the number of factor loadings greater than 0.4 (after varimax rotation), and in parentheses the number of items that were expected to load on the factor. We could not find traces of the skill “crossing the tens boundary” and of problem size in our data.

The Tanaka goodness-of-fit index indicates that all models displayed in Table 2 fit the data well. It is not surprising that in the exploratory analyses more factors result in a better fit. In both data sets, the assumed factors “g” and “MulDiv” appear in most analyses. This is not the case for the factor “Chains” (representing the skill of handling more than two operands). Clear indications for this factor are only found in the confirmatory analysis with three factors of the 2008 data. Our interpretation is that handling more than two operands is not a central skill, probably because there are so many strategies supporting the skill.

The results of the MIRT analyses connect well to those of the unidimensional analyses: Again we find a strong main factor explaining much of the variation in the data. As the plain addition/subtraction problems have the highest loadings on this factor (and most problems in the analysed set require these operations), it seems to represent the corre-

Table 2:
Results of the MIRT analyses of 13 (12) items using NOHARM. In the Comments column, the numbers of loadings > .4 are listed; theoretically expected numbers are given in parentheses

Model Description	Tanaka GFI	RMS of residuals	Comments
2008			
4 factors exploratory	0.991	0.005	"g", MulDiv, 2 item specific factors
3 factors exploratory	0.985	0.007	plain AddSub, MulDiv, 1 item specific factor
2 factors exploratory	0.977	0.008	plain AddSub, MulDiv
3 factors: g, MulDiv, Chains	0.971	0.009	g: 7 (13), MulDiv: 3 (4), Chains: 3 (3)
1 factor	0.958	0.011	g: 6 (13)
2006			
3 factors: g, MulDiv, Chains	0.990	0.003	g: 10 (12), MulDiv: 2 (3), Chains: 1 (3)
1 factor	0.986	0.008	g: 9 (12)

sponding skill. The requirement to detect and perform multiplication and division in word problems shows up reliably as a second factor. However, not all subtleties of the HiSkA, such as the skill for crossing the tens boundary, are reflected in the data.

Analyses based on knowledge space theory

The knowledge space theory defines a knowledge state of a person as the particular subset of questions that this person is capable of solving (Falmagne, 1989). This means that a subjects' pattern of responses to a set of items corresponds to his/her knowledge state. Of course, it is plausible to assume certain probabilities of errors that blur the correct states. The selected seven items with the surmise relations shown in the Hasse diagram of Figure 2 form a knowledge space of 23 states. We refer to response patterns that are elements of this knowledge space as "regular patterns".

Distribution of response patterns

The 264 cases from 2006 produced 45 different response patterns (out of 128 possible patterns). Their distribution is depicted in Figure 3. Two hundred and six (78 %) of the response patterns equalled regular knowledge states. Most of the non-regular patterns occurred only once or twice. The most frequent non-regular pattern occurred seven times. Only three of the 23 regular patterns did not occur at all.

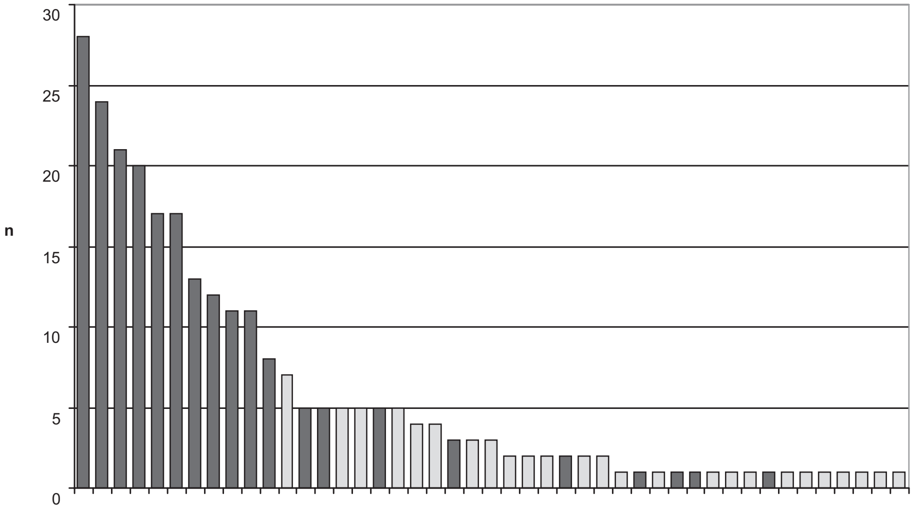


Figure 3:
Distribution of response patterns from 2006. The dark columns represent regular response patterns

We can compare these results with the linear structure that results from a unidimensional model, assuming that the solution of an item implies the solution of all easier items (so the items would form a Guttman scale). The corresponding knowledge space for our seven items then consists of eight knowledge states (0000000, 1000000, 1100000, …, 1111111). Only 128 of the 264 (48 %) response patterns were regular patterns of this linear model.

From this comparison we might conclude, that the HiSkA model is more appropriate than the linear model. But these analyses suffer from the problem that the knowledge space of the linear model is a subset of the knowledge space of the HiSkA: All regular patterns of the unidimensional model are also regular patterns of the multidimensional model, and there are more regular patterns in the latter model. Therefore, the probability of occurrence of an irregular pattern is greater for the unidimensional model.

The data from 2008 yield similar results. The 140 cases produced 34 different response patterns. Seventy-nine per cent of the response patterns were regular knowledge states of the HiSkA. The most frequent non-regular pattern occurred only three times. The unidimensional model with its eight knowledge states covers only 51 per cent of the response patterns.

To summarize, these analyses show that there is a broad overlap between the empirical response patterns and those expected in the knowledge space derived from the HiSkA. As the linear model has fewer knowledge states, the comparison of the two models with respect to overlap is unfair.

Item tree analysis

In knowledge space theory, the surmise relation states for pairs of items (*i, j*) from an item set *I* that mastering item *j* implies mastering item *i*. These implications form a quasi-order on *I*. The empirical cases that contradict the quasi-order can be represented in a matrix of the numbers of counterexamples. This matrix is the starting-point of item tree analyses. Schrepp (2003) introduced an algorithm that allows to extract inductively a set of competing quasi-orders from data. He also proposed a measure *diff* as the mean quadratic difference between the observed counterexamples to j→i and the expected number of counterexamples for each quasi-order. The best fitting quasi order is the one with the minimal *diff* value. Sargin and Uenlue (2008) proposed some corrections and improvements to Schrepp's ideas that are implemented in the *R* package DAKS (Sargin & Uenlue, 2009).

In the following analyses, we compared three knowledge structures, using the minimized corrected algorithm from the DAKS package: (a) the structure induced by the inductive item tree analysis, (b) the structure derived from the HiSkA, and (c) the linear structure, which assumes that solution of an item implies the solution of all easier items. The results are shown in Table 3. For both data sets, the HiSkA structure fits the data better than the linear structure. In the 2008 data, the fit for this structure is as good as the induced structure.

Table 3:
diff values from item tree analyses

	Induced structure	HiSkA structure	Linear structure
2006	61.32	87.31	194.37
2008	10.03	9.97	35.89

Although the induced structures are theoretically unproductive (they contain surmise relations between the most difficult items and most items of medium difficulty, and no strands of surmise relations spanning more than 3 items), they serve as a lower bound estimation of the *diff* values for the present data. Again, these results clearly favour the HiSkA structure over the linear structure. Moreover, for the 2008 data the *diff* value indicates that the HiSkA structure fits the data even better than the induced structure.

Simulations

In the previous sections, we have assumed tacitly that the HiSkA as a polyhierarchic structure should generate multidimensional data. Now we want to discuss this assumption more closely. We have already mentioned the conclusion of Reckase and Stout (1995) that multidimensional structures produce unidimensional data when the items are sensitive to the same composite of skills. We want to postpone the discussion if this is plausible in our case. Rather, we want to explore if the HiSkA might generate unidimensional data on account of its very properties. After all, the HiSkA predicts (a) different difficulties for classes of problems and (b) positive correlations between many of the pairs of items. Qualitatively, this can be illustrated with the expected relationship between Node 34 and Node 51 (see Figure 2). Although both nodes have Node 14 as an ancestor, we would not expect a high correlation in the solution rates, because beside the skills for Node 14, they require completely different skills (Node 51: Handling more than two operators with varying placeholders; Node 34: transforming text into a mathematical model). On the other hand, both nodes represent moderately advanced problems, which beginners aren't expected to solve. Node 38 is the only one that is largely independent of the other nodes. But again it represents a rather advanced problem type.

It is impossible to predict exactly the results we can expect when analysing data generated by the HiSkA. Even with only seven nodes the structure is too complex. This question can better be answered with simulations. When the domain is complex, simulations sometimes yield surprising results. In our case, it might be that despite the polyhierarchic structure of the HiSkA, it generates data that can well be explained by a single dimension.

We simulated data sets using the knowledge space theory. It is straightforward to specify all knowledge states that are consistent with a specific model – the so called regular patterns. Furthermore, the distribution of the knowledge states has to be determined a priori. In some simulations we based the distribution of the knowledge states on the empirical distribution; in others we assumed uniform distribution of the knowledge states.

We simulated data according to three models: (1) The HiSkA model with the empirical distribution of knowledge states (emp model), (2) the HiSkA model with a uniform distribution of knowledge states (unif model), and (3) a linear model assuming that the items conform to a Guttman scale (see above). All simulated data sets were created in two steps. First we drew cases from the set of regular patterns with probabilities that were derived from the aforementioned distributions. In a second step each entry of the response pattern was randomly switched according to two parameters: The probability for guessing, p_g, and the probability for a slip, p_s. The guessing parameter was set to $p_g = 0.05$, because results for our problems can be selected from a plausible range of about 20 numbers. The slip parameter was set to $p_s = 0.1$. Preliminary analyses have shown that these settings produced distributions of response patterns that were similar to those in the empirical sample.

To see what relationships between items can be expected, we simulated two samples of $n = 10000$ cases each, one based on the empirical distribution, and one based on the uniform distribution of knowledge states. The phi-coefficients between pairs of items in the dataset from 2008 (columns r) and the two simulated samples (columns e for emp model, u for unif model) are depicted in Table 4. Correlations between .10 and .19 are marked in light grey; correlations greater than .19 are dark grey (Note that the maximum possible phi-coefficient is less than 1.0 for many of the marginal distributions in these samples). It is very interesting that despite the relative independence of Node 38 the simulations not only predict a strong correlation with Node 50, which is also present in the empirical data, but also correlations with the nodes 34, 30, and 51. Overall, the HiSkA predicts correlations between unrelated items – predominantly, but not only when using the empirical distribution.

Table 4:
Correlations (phi) between items in the empirical sample from 2008 (r) and two simulated samples (e – emp model, u – unif model). ▨: .10≤ phi<.20, ■: .20≤ phi

	26			34			30			51			38			50		
	r	e	u	r	e	u	r	e	u	r	e	u	r	e	u	r	e	u
14			■			▨			■			▨				▨		
26				■	■	■			▨	▨						▨		▨
34													▨	▨	▨	▨	▨	■
30										▨	■	■	▨	▨		▨	▨	
51													▨	▨		▨	▨	
38																■	■	■

The next question we want to answer by means of simulation is how well the item tree analysis reproduces the generating structure. In the previous chapter, we have reported *diff* values for the discrepancy between the observed and expected numbers of instances (pairs of items) contradicting the assumed structure. One problem with the *diff* values is that their distribution is not known, preventing significance testing. To circumvent this, one can use a bootstrapping method: We simulated n = 1000 datasets (of n = 140 cases each) generated by the HiSkA as described above, and calculated the corresponding *diff* values. The empirical *diff* value can then be compared with the distribution. We did this with the empirical distributions of response patterns of both datasets. Both resulting *diff*-distributions resemble a χ^2-distribution with a median of 54 and an interquartile range of 20 for 2006, and a median of 9 and an interquartile range of 5 for 2008.

For 2006 we found that only 2.2 % of the simulated *diff* values were greater than the empirical value of *diff* = 87.31, indicating a significant deviation of the empirical value from what is expected assuming validity of the HiSkA. However, all simulated *diff*-values were less than *diff* = 194.37, the value resulting from the assumption of a Guttman model. For 2008, the location of the empirical *diff* value indicated a good fit of the HiSkA: 40.6 % of the simulated *diff* values were greater than the empirical value of *diff* = 9.97. This pattern of results is in line with the findings from the MIRT analyses, that the data from 2008 are more consistent with the HiSkA than the data from 2006.

Now we turn to the central question if the HiSkA can generate unidimensional data. For each of the three models (emp, unif, linear) we simulated 1000 samples of n = 240 and noted rejection of the H_0 of RM conformity assessed with the Martin-Löf test and Andersen's LR test. For the Martin-Löf test we divided the items in computation problems and word problems. For all tests, the common significance level of $p < .05$ was applied.

We expected more deviations from the Rasch model for the uniform data sets because they contain more response patterns that contradict the unidimensional model (recall that these patterns were rare in the empirical distribution). As a matter of course, we predicted more deviations from the Rasch model for the multidimensional data sets than for the unidimensional sets. Results of the simulations are shown in Table 5. The most obvious result is that we obtained completely different results depending on the split criterion. When we split the samples randomly for the LR test, rejection rates are very low.

Table 5:
Martin-Löf-test and Andersen's LR-test for simulated data (for each model, N=1000 samples of n=240 cases were simulated).

	Rejection rates		
	Martin-Löf test	LR test median split	LR test random split
Guttmann model (unif.dist.)	0.954	0.980	0.068
HiSkA model (emp. dist.)	0.775	0.573	0.057
HiSkA model (unif. dist.)	0.988	0.528	0.057

For stricter split criteria rejection rates are much higher. Our expectations for the HiSkA-models are approximately met by the results of the Martin-Löf test: The uniform model is almost always rejected, whereas the empirical model passes the test in about 23 % of the cases. The rejection rate of Andersen's LR test does not differ much between the HiSkA-models based on different distributions and is less sensitive to deviations from the RM. Unexpectedly, the Guttman model is also rejected in most cases when applying strict split criteria. According to Kubinger (2005, 2007) this is typical for Guttman scales for the following reason: In the subsample scoring above the median, almost everyone solves the easy items, leading to erroneous parameter estimations for these items. In the subsample scoring below the median, this happens with the difficult items because almost nobody in that subsample solves them. Thus, deviations between accurately and erroneously estimated parameters on either end of the scale are likely, leading to high LRs and rejection of the RM. We think that similar effects are responsible for the high rejection rates of the HiSkA models. Nevertheless, our simulations show that even with these effects in place, between 23 % and 43 % of the simulated samples in the “empirical distribution” condition have been classified as consistent with the RM. It is not implausible that our empirical data were similar to those samples. So we can answer our central question that in most cases, a multidimensional structure like the HiSkA does not generate data conforming to the RM, but that the probability of obtaining results that would be interpreted as indicating RM conformity is not negligible, particularly when using Andersen's LR test.

A possible explanation for the moderate rejection rates reported in Table 5 is that Andersen's LR test does not have enough power with sample sizes around 200. To investigate this we performed simulations varying sample size. Thousand samples were simulated for each sample size ranging from $n = 100$ to $n = 1000$ with increments of 50. Split criterion was the median of the raw scores. The results are displayed in Figure 4. We found that for all models the rejection rate approaches an asymptote of greater than 0.95. This indicates that our empirical sample sizes of $n = 264$ and $n = 140$ were not large enough to reject the H_0 of being consistent with the RM. The slope differences between the three models are due to the effects discussed above for Guttman scaled data. The knowledge space that conforms to the HiSkA deviates from that conforming to the Guttman model in important aspects. Please refer to Figure 2 for the following explanations. Considering Item 51 and Item 26, the HiSkA allows all four combinations of solving vs. not solving these items, although Item 51 is much more difficult than Item 26. The same is true for Item 50 and Item 30. This independence between items results in more accurate parameter estimation and hence to a lower probability of H_0 rejection in the LR test. In the empirical distribution, the combinations of Item 51 solved and Item 26 failed or Item 50 solved and Item 30 failed are rare, making the situation more similar to the Guttman model than the uniform HiSkA model, where these combinations occur with the same frequency as any other regular combination.

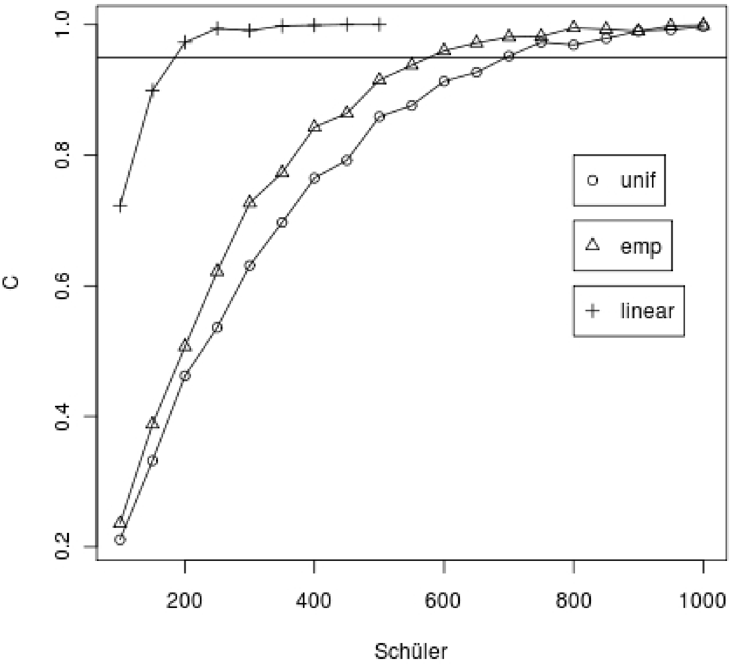


Figure 4:
Rejection rates (C) of Andersen's LR test applied to simulated data of varying sample size (unif: HiSkA model with uniform distribution of knowledge states; emp: HiSkA model with empirical distribution; linear: Guttman scaled data)

To summarise, the simulations have shown that (a) the HiSkA predicts correlations between item classes where at first glance one would expect none, that (b) the bootstrap method for simulating distributions of *diff* values demonstrated that the 2008 data fit the HiSkA well, and that (c) the HiSkA generates data that are not unlikely to be classified as being consistent with the RM when using sample sizes around *n* = 200.

Discussion

What do all these analyses contribute to the question whether a multidimensional generating structure can produce data conforming to the RM and to the question of the validity of the HiSkA? In our context of extensive model testing it should not be forgotten that there is no strictly objective procedure that tells if a data set conforms to a specific model or not. Eventually the researchers are responsible for a decision. Also, a data set may match more than one model and it depends on the researchers' goals what model is preferred. A second caveat pertinent to our interpretations concerns the sizes of the empirical samples. To arrive at sound conclusions about the structure of arithmetic skills, the results need to be replicated with larger samples. For some of the analyses reported, our sample sizes were very small.

Our simulations have demonstrated that a multidimensional structure like the HiSkA produces data that are not unlikely to be qualified as consistent with the RM when applying common criteria to sample sizes around n = 200. However, this does not preclude that multidimensional models might fit the data as well. Firstly, Andersen’s LR test does not have much power at these sample sizes[5] and secondly, the “confirmation” of the H_0 (unidimensionality) with the type-I-error set to $p < 0.05$ is associated with a large type-II-error. Compared to Andersen's LR test, the Martin-Löf test was more sensitive to deviations of our simulated data from the RM, resulting in many cases where one test indicated person homogeneity and the other test rejected item homogeneity in the same simulated data set. This is not unusual: The same pattern of results has been found in human data from solving word problems by Arendasy et al. (2005), who applied Ponocny's (2001) non-parametric T statistics rather than the traditional likelihood ratio tests. Preliminary analyses of our empirical data with Ponocny's (2001) *T2*-statistic have shown that multiplication/division vs. addition/subtraction as partitioning criterion suggested rejection of the H_0 of item homogeneity whereas random partitioning did not.

So regarding the question if whether multidimensional generating structure can produce unidimensional data we can conclude that depending on the methods used, deviations from homogeneity may or may not be detected. Global tests such as Andersen's LR test with median split may be appropriate in the context of test development. However, when the focus is on validating cognitive models, more specific tests such as the Martin-Löf test with theoretically grounded partitioning criteria, or Ponocny's (2001) T-statistics provide better chances of detecting deviations.

A final remark about our simulation results: They contradict findings of an early simulation study by Stelzl (1979), who found likelihood tests to be seriously insensitive to specific violations of homogeneity even with sample sizes of n = 1000. We attribute this discrepancy to the fact that our simulated samples were more realistic than some of the scenarios simulated by Stelzl (1979), because they were carefully derived from plausible knowledge spaces.

The discussion of the second question – for the validity of the HiSkA – is structured along four possible interpretations of our results:

a) One main factor accounts for a substantial amount of variance; the rest is error variance.

b) The test draws on multiple skills some of which are highly correlated with each other.

c) The data conform well to the HiSkA, but the HiSkA is not as multidimensional as expected.

d) The data are multidimensional, but the generating structure is different from the HiSkA.

[5] For recent developments of RM tests with predictable power see Kubinger, Rasch & Yanagida (2009).

For each interpretation we will check how well it is supported by the various analyses. Interpretation a gets some support from the results of the factor analyses and the reliability analyses. If our goal was to develop a homogeneous arithmetic test, our items would constitute a good basis from which this goal could be attained by item selection and item reformulation. But even these analyses indicate that a one factor solution is unsatisfactory. Particularly the subset of seven items does not form a homogeneous scale. The MIRT analyses clearly contradict Interpretation a, because in the sample of 2008 two additional factors could be confirmed: One representing multiplication/division skills, and one representing the skill of handling more than two operators. In the sample of 2006 only the former factor could be confirmed. Hence, variance not explained by the first factor can be explained by specific other skills. The item tree analyses yielded worse fits for the unidimensional models than for the HiSkA models, which speaks against Hypothesis a. But as the knowledge space derived from the Guttman model has much fewer states than the one derived from the HiSkA, the linear models were clearly disadvantaged in the item tree analysis. Mirroring the observation that data conforming to a Guttman model cannot conform to the RM (Kubinger, 2007), our assumptions for modelling a linear structure might be too restrictive for item tree analyses as well.

At first glance, the MIRT results militate against Interpretation b, because the solutions with two or three independent factors fit the data better than one-factor solutions. However, it is not clear what exactly is represented by the first factor. Some indicators point to the possibility that the strong first factor found in most analyses (FA, MIRT) is a compound of skills: Firstly, we found correlations between items that should be unrelated according to the HiSkA, for example between items from Node 34 and Node 38. (However, the fact that the simulations also predicted unexpected correlations attenuates this argument). Secondly, there are reasons why theoretically independent skills might be anyway correlated. Let us illustrate this with the following example. Although in principle the skill of handling more than two operators in a computation problem (Node 51) has nothing to do with the skill of identifying the requirement of performing a division in a word problem (Node 50), it is likely that a person who masters one skill has some experience in the other skill, too, just because both are subject of mathematics instruction at about the same time and the person has reached a certain level of development. In the sample of 2008, items loading high on the first factor require various levels of addition and subtraction skills. In the sample of 2006 more diverse items have loadings on the first factor and only one additional factor (mul/div) could be identified. For this sample, Interpretation b (correlated skills) is probably adequate. All these considerations are leading to the question, whether general intelligence (*g*) might establish the first factor. Our interpretation of the first factors in the MIRT analyses as reflecting the level of development fits well into the conception of van der Maas, Dolan, Grasmann, Wicherts, Huizenga, and Raijmakers (2006). These authors have demonstrated that *g* can be reconstructed as emerging by positive beneficial interactions between cognitive processes during development.

Interpretation c can be viewed as a variation of Interpretation b – at least they are not mutually exclusive. The HiSkA as a directed acyclic graph is not per se multidimensional in the sense of orthogonal dimensions defining a space. It only opens the possibility of independent development on different branches, which does not necessarily occur in reality. For 2008, all relevant analyses (MIRT, ITA, simulations) indicate that in spite of the existence of a strong first factor, the data conform well to the HiSkA. Hence for this data set we favour Interpretation c as the best explanation: The data conform well to the HiSkA, but the HiSkA is not as multidimensional as expected. This is not so obvious for 2006. But taking into account that the data in this sample has been collected later in the school year, it could well be that theoretically independent skills have assimilated through the unifying influence of training and instruction. Baumert et al. (2007) have pointed out that when the treatment is constant (curricula, similar instruction for all students), differences in academic performance necessarily vary with general cognitive abilities (prior knowledge, faster comprehension, etc.). This would explain why the 2006 data are closer to unidimensionality.

Finally, turning to Interpretation d (that the data are multidimensional, but based on a different generating structure), only the inductive item tree analysis of the 2006 data points to that. However, the induced structure can hardly be called multidimensional, because it has Nodes 30, 38, 50, and 51 at the upper level, Nodes 26 and 34 at the intermediate level, and Node 14 as a root node, with the levels almost completely linked.

Although the analyses converge to Interpretation c, we should keep in mind that many of them are restricted to a subset of 7 items. This limitation is due to the fact that our moderate sample sizes did not allow item tree analyses of larger item sets. So we could provide some arguments for the validity of the HiSkA, but did not validate it as a whole. As the HiSkA was developed for application in training software, it needed to be very comprehensive, so a complete validation is probably neither feasible nor desirable.

Regardless of the size of the item set, some lessons about the structure of elementary arithmetic skills have been learned: Firstly, prerequisite relations between computation problems and word problems as well as among word problems postulated in the HiSkA have been confirmed on a fine grained level. These analyses complement analogous coarse grained results obtained by Fuchs et al. (2006). Secondly, identifying and performing multiplication in word problems has been identified as a relatively independent skill. This finding is interesting in a research culture where most authors focus either on addition/subtraction or on multiplication/division, but rarely on arithmetic in its entirety. Thirdly, we found indications that handling more than two operators is another independent skill. Because handling more than two operators involves subskills that are relevant to algebra (e.g. the reversal of operators), we believe that practising this problem type is a good way for preparing young students for this important area of mathematics. Many authors agree that "an algebraic strand should be integrated with arithmetic from the earliest years" (Freiman and Lee, 2004 cited after Verschaffel, Greer, & Torbeyns, 2006).

References

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

Andersen, E. B. (1973). A goodness of fit test for the Rasch-Model. *Psychometrika*, *38*, 123-140.

Anderson, J. R., Reder, L. M., & Simon, H. A. (2000). Applications and misapplications of cognitive psychology to mathematics education. *Texas Educational Review*.

Arendasy, M., Sommer, M., & Ponocny, I. (2005). Psychometric approaches help resolve competing cognitive models: When less is more than it seems. *Cognition and Instruction*, *23*, 503-521.

Baumert, J., Brunner, M., Lüdtke, O., & Trautwein, U. (2007). Was messen internationale Schulleistungsstudien? – Resultate kumulativer Wissenserwerbsprozesse. [What do international student assessment studies measure? – Results of cumulative knowledge acquisition.] *Psychologische Rundschau*, *58*, 118-145.

Bergan, J. R., Stone, C. A., & Feld, J. K. (1984). Rule replacement in the development of basic number skills. *Journal of Educational Psychology*, *76*, 289-299.

Canobi, K. H. (2005). Children's profiles of addition and subtraction understanding. *Journal of Experimental Child Psychology*, *92*, 220-246.

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

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly*, *50*, 328-344.

Falmagne, J. C. (1989). A latent trait theory via a stochastic learning theory for a knowledge space. *Psychometrika*, *54*, 283-303.

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

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction in psychological test theory]. Bern: Huber.

Fischer, G. H. (2007). Rasch models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics 26 - Psychometrics* (pp. 515-586). Amsterdam: Elsevier.

Fraser, C., & McDonald, R. P. (2003). NOHARM: A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, Australia: The University of New England.

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., Schatschneider, C., & Fletcher, J. M. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, *98*, 29 -43.

Gagne, R. M. (1962). The acquisition of knowledge. *Psychological Review*, *69*, 355-365.

Geary, D. C., Hamson, C. O., & Hoard, M. K. (2000). Numerical and Arithmetical Cognition: A Longitudinal Study of Process and Concept Deficits in Children with Learning Disability. *Journal of Experimental Child Psychology*, *77*, 236-263.

Gilmore, C. K., Bryant, P. (2008). Can children construct inverse relations in arithmetic? Evidence for individual differences in the development of conceptual understanding and computational skill. *British Journal of Developmental Psychology*, *26*, 301-316.

Kintsch, W. (1988). Role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*, 163-182.

Kintsch, W., & Lewis, A. B. (1993). The time course of hypothesis formation in solving arithmetic word problems. In M. Denis & G. Sabah (Eds.), *Modèles et concepts pour la science cognitive: Hommage à Jean-Francois Le Ny* (pp. 11-23). Grenobles Press Universitaires.

Kornmann, R., & Wagner, H. J. (1990). Ermittlung der Lernbasis bei einfachen Kopfrechenaufgaben im Zahlenraum 0-20. *Zeitschrift für Heilpädagogik, 41*, Beiheft 17, 211-218.

Kubinger, K. D. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science*, *45*, 106-110.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch Model – Some critical suggestions on traditional approaches. *International Journal of Testing*, *5*, 377-394.

Kubinger, K. D. (2007). Probleme bei der Testkonstruktion nach dem Rasch-Modell [Some problems in calibrating an item pool according to the Rasch model]. *Diagnostica*, *53*, 131-143.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mair, P., & Hatzinger, R. (2007). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, *49*, 26-43.

Mair, P., & Hatzinger, R. (2009). *eRm: Extended Rasch Modeling*. R package version 0.10-2. http://r-forge.r-project.org/projects/erm/

Martin-Löf, P. (1973). *Statistika modeller (Statistical models). Notes from a seminar held 1969/1970, taken by Rolf Sundberg; reprinted 1973.* Stockholm: Institutet för Försäckringsmatematik och Matematisk Satistisk vid Stockholms Universitet.

McDonald, R. P. (1982). Unidimensional and multidimensional models in item response theory. In: Proceedings of the I.R.T, C.A.T. conference, Minneapolis.

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

OECD (2005). *Pisa 2003*. Technical report. Paris: OECD.

Parkman, J. M., & Groen, G. J. (1971). Temporal aspects of simple addition and comparison. *Journal of Experimental Psychology*, *89*, 332-342.

Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika*, *66*, 437-460.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Paedagogiske Institut, Copenhagen.

Reckase, M. D. (2007). Multidimensional item response theory. In C.R. Rao & S. Sinharay (Eds.), *Handbook of Statistics 26 - Psychometrics* (pp. 607-642). Amsterdam: Elsevier.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M. D., & Stout, W. (1995). *Conditions under which items that assess multiple abilities will be fit by unidimensional IRT models.* Paper presented at the European meeting of the Psychometric Society, Leiden, The Netherlands, July.

Revelle, W. (in prep.). *An introduction to psychometric theory with applications in R*. <URL: http://personality-project.org/r/book/>

Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? [What do international student assessment studies measure?] *Psychologische Rundschau*, *57*, 69-86.

Rost, J., & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement*, *26*, 42-56.

Sargin, A., & Uenlue, A. (2008). Inductive item tree analysis: Corrections, improvements, and comparisons. University of Augsburg, Germany: Institute of Mathematics, Preprint Nr. 24/2008.

Sargin, A., & Uenlue, A. (2009). DAKS: An R package for data analysis in knowledge space theory. Manuscript submitted for publication. http://www.math.uni-augsburg.de/~uenlueal/

Schoppek, W. (subm.). Efficient computer assisted practice based on a hierarchy of arithmetic skills. *Zeitschrift für Pädagogische Psychologie*.

Schoppek, W., & Laue, C. (2005). *Individuelle Förderung von Rechenfertigkeiten in der Grundschule durch ein computergestütztes Übungsprogramm*. [Individualized computer assisted training of basic arithmetic skills in elementary school.]. URL: http://psydok.sulb.uni-saarland.de/volltexte/2008/1556/

Schoppek, W., & Tulis, M. (2010). Enhancing arithmetic and word problem solving skills efficiently by individualized computer-assisted practice. *The Journal of Educational Research*, 103, 239-252.

Schrepp, M. (2003). A method for the analysis of hierarchical dependencies between items of a questionnaire. *Methods of Psychological Research Online*, *8*, 43-79.

Seyler, D. J., Kirk, E. P., & Ashcraft, M. H. (2003). Elementary Subtraction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1339-1352.

Spearman, C. (1904). “General intelligence”, objectively determined and measured. *American Journal of Psychology*, *15*, 72-101.

Spearman, C. (1927). *The abilities of man*. New York: The Macmillan Company.

Stelzl, I. (1979). Ist der Modelltest des Rasch-Modells geeignet, Homogenitätshypothesen zu prüfen? Ein Bericht über Simulationsstudien mit inhomogenen Daten. *Zeitschrift für Experimentelle und Angewandte Psychologie*, *26*, 653-672.

Stegelmann, W. (1983). Expanding the Rasch model to a general model having more than one dimension. *Psychometrika*, *48*, 259-267.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In: *Testing structural equation models*, 10-39.

Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial Studies of Intelligence*. Chicago: University of Chicago Press.

Torbeyns, J., Verschaffel, L., & Ghesquiere, P. (2005). Simple addition strategies in a first-grade class with multiple strategy instruction. *Cognition and Instruction*, *23*, 1-21.

Verschaffel, L., Greer, B., & Torbeyns, J. (2006). Numerical Thinking. In A. Gutierrez & P. Boero (Eds.), *Handbook of research on the psychology of mathematics education: Past, present and future* (pp. 51-82). Rotterdam, Netherlands: Sense Publishers.

Weber, M. (2004). *Die Anwendbarkeit probabilistischer Modelle im Rahmen der Wissensraumtheorie.* GRIN-Verlag (http://www.grin.com/).

White, R. T. (1976). Effects of guidance, sequence, and attribute-treatment interactions on learning, retention, and transfer of hierarchically ordered skills. *Instructional Science*, *5*, 133-152.

Appendix

Level 1

New: Subskill **decomp**	Decomposition of numbers between 1 and 10
Special case	Decompositions of 10
Additional skill	Understand that 2 + 5 = 5 + 2

Level 2

New: Subskill **igno**	Ignore the tens position (when adding up to 20 without CT) 11 + 8 = _
New: Subskill **subtra**:	Subtraction / completion (based on **decomp**) 9 – 4 = _ , 4 + _ = 9

Level 3

New: Subskill **subtra3**	Application of decomposition on subtraction with PH 3 8 – _ = 5
New: Subskill **ctb**	Crossing the tens boundary (based on **decomp)** (addition, PH 5) 7 + 6 = _

Level 4

New: Subskill **reverse**	Reversal of operators (subtraction, PH 1) _ – 5 = 3 ⇒ 3 + 5 = _
New: Subskill **positions**	Calculate with two digit numbers (without CT) 22 + 17 = _ , 48 – 25 = _
New: Combination	**decomp** & **subtra** & **ctb** (NR 20 with CT, subtraction: PH 3 und 5, addition: PH 1 and PH 3) 15 – 8 = _ , 15 – _ = , 8 + _ = 12 , _ + 3 = 11

Level 5

New: Subskill **intermed** Memorise intermediate results in problems with 3 operands
(L7, NR20, PH 7)
7 + 5 + 3 = _

New: Subskill **combine** Combine arbitrary terms in problems with 3 operands
(didactically simplified: two terms cancel each other out)
8 + 11 – 8 = _

New: Combination **decomp** & **ctb** & **reverse**
(subtraction NR20, PH1)
_ – 7 = 4

New: Combination **decomp** & **ctb** & **positions**
(NR100, CT, PH5)
36 + 27 = _
23 – 17 = _

Level 6

New: Combination **decomp** & **positions** & **reverse** (& **subtra3)**
(NR100, no CT, PH 1 and PH 3)
_ + 17 = 29 , 56 – _ = 21

New: Combination **intermed** & **combine**
(L7, NR20, PH 7)
17 – 11 + 3 = _

Level 7

New: Combination **decomp** & **positions** & **reverse** & **ctb** (& **subtra3)**
(NR100, CT, PH 1 und 3)
36 + _ = 81 , _ – 37 = 26

New: Combination **intermed** & **combine** & **reverse**
(L7, NR20, PH 5, without CT)
8 – 6 + _ = 18

Level 8

New: Subskill **reverse2** gradual backward-calculation in problems with L7, PH1
_ – 3 – 5 = 8

Enlargement All types of addition/subtraction problems in NR1000

Enlargement **intermed** & **combine** & **reverse** & **ctb**
(L7, NR20, CT, PH 3)
14 – _ – 3 = 6

Level 9

New: Combination	**combine** & **reverse2** & **reverse** & **ctb** (L7, NR20, PH3) 18 – _ + 3 = 12
Enlargement	**reverse2** with mixed operators (L7, PH 1) _ – 6 + 14 = 18

Level 10

Enlargement	All problems L7 im NR100