LOOK ACADEMIC PUBLISHERS
OPEN ACCESS

IEJME — MATHEMATICS EDUCATION
2017, VOL. 12, NO. 7, 667-687

Are Word Problems Really More Difficult for Students with Low Language Proficiency? Investigating Percent Items in Different Formats and Types

Birte Pöhler[a], Ann Cathrice George[b], Susanne Prediger[a], Henrike Weinert[a]

[a]Institute for Development and Research in Mathematics Education, TU Dortmund University, GERMANY;
[b]Institute for Educational Research, Innovation and Development of the Austrian School System (BIFIE), AUSTRIA.

ABSTRACT

Achievement gaps between students with low and high language proficiency appear for word problems, but is this due to their text format or their conceptual challenges? A test with percent problems of different types and in pure, text and visual format was conducted with N=308 seventh graders. Students' scores were analyzed statistically by a cognitive diagnosis model. Unlike expected, the probability for students with low language proficiency to solve items in text format is not lower than in pure format. These results are interpreted as indication that conceptual challenges might impact stronger than reading challenges.

KEYWORDS
Percentages, word problems, cognitive diagnosis model, DINA, visual models, language proficiency

ARTICLE HISTORY
Received 05 September 2017
Revised 11 October 2017
Accepted 13 October 2017

Introduction

Large-scale assessment studies have repeatedly documented achievement gaps for language minority students (Martiniello, 2008; Abedi, 2006; Haag et al., 2013) or socially disadvantaged students with low language proficiency, even if speaking the majority language (Prediger et al., 2013; Walzebug, 2014). Although it seems likely to trace these language gaps back to word problems and their language demands (Duarte et al., 2011), little is known whether it is really the text format of an item which disadvantages the language learners or its inherent conceptual demands which is of cause higher than for a procedural

CORRESPONDENCE Susanne Prediger ✉ prediger@math.uni-dortmund.de

innermathematical item. As most studies disentangling obstacles for language learners investigate complete assessments with innermathematical procedural as well as context items, there is a risk of confounding language demands and conceptual demands (e.g. in Martiniello, 2008; Wolf & Leon, 2009).

That is why the study presented constructed a test with items of comparable conceptual demands, but in different formats. If test items with similar conceptual demands are posed in pure format, text format or visual format, do students of low language proficiency really have more difficulties with the text format?

In Section 1, the theoretical background of the study is outlined, on demands for language learners and on the exemplary topic in view, percentages in grade 7. Section 2 refines the re-search questions, and Section 3 presents the research design and methods, Section 4 the empirical results, which are discussed in Section 5.

Theoretical Background and State of Research

1.1 Language gaps and word problems

Various empirical studies show that secondary students' academic language proficiency is a crucial factor for their performance in mathematics tests, this applies for students with minority home languages (Abedi, 2006; OECD, 2007; Martiniello, 2008; Haag et al. 2013) as well as for socially underprivileged students with the majority languages (Hirsch, 2003; Prediger et al., 2013, Walzebug, 2014). Already in 1992, Secada resumed in his literature review based on several Anglo-American studies that 'these studies indicate a relationship between how proficient someone is in a language and performance of mathematics achievement' (Secada, 1992, p. 638). In a German high stakes test ZP10, proficiency in the language of assessment turned out to be even more relevant for the mathematics achievement than other background factors like home languages, socio-economic and immigrant status (Prediger et al., 2013).

Especially in the context of US high stakes testing, poor performances of students with low language proficiency (in brief: low LP) are often explained through difficulties in understanding the wording of mathematical problems: 'Solving math word problems […] presents a double challenge for students whose language proficiency is limited' (Abedi, 2004, p. 31). Therefore, researchers underline that students with low language proficiency have specific *reading comprehension difficulties with word problems* (e.g. Duarte et al., 2011, for an overview), whereas test items with short texts are often assumed to be 'language fairer'. As a consequence several authors plead for reducing the linguistic complexity of test items and investigate also the effects of other accommodations for low language proficient students, including provision of extra time or a glossary (Abedi, 2006).

The focus on word problems posing specific challenges has been fueled by several studies which conducted differential item analysis for disentangling the specific challenges for (monolingual or multilingual) students with low language proficiency (Martiniello, 2008; Wolf & Leon, 2009; Walzebug, 2014; Prediger et al. 2015; Haag, Heppt, Stanat, Kuhl, Pant, 2015). These studies identified that linguistically complex test items tend to be disproportionally more difficult for language learners than items without complex language. But as these studies

analyzed complete assessments with innermathematical procedural as well as context items, there is a risk of confounding language demands and conceptual demands when considering word problems. Or they compared only word problems and their linguistically simplified parallel items (Abedi & Lord, 2001; Haag, Heppt, Roppelt, & Stanat, 2015), that means, they compared *within* the text format, but not items in text format with visual format or nearly language relieved by technical terms.

The doubt whether the difficulties must partly be traced back to conceptual challenges is nurtured by a recent study on the German high stakes test ZP10 NRW in which reading comprehension difficulties could only partially explain the poorer performance of students with low language proficiency (Prediger et al., 2015). Contrary to expectations, items, which posed high difficulties to students with low LP could *not* be characterized by reading challenges, but rather by conceptual or process-oriented challenges, which was confirmed in interview studies. These findings suggest investigating the role of the text format in comparison to other formats but the same conceptual challenges.

1.2 Students' performance with regard to different problem formats

Research results on comparing between text, visual and pure format are much older than the current the research on language gaps. Already in the 1980ies, word problems were shown to be more difficult than pure items on the same mathematical topic, and independent of students' language proficiency (e.g. Kouba et al., 1988). For example, Carpenter et al. (1980, p. 12f) reported on performance levels of word problems for basic arithmetic operations and fractions which were about 10% to 30% lower than those of corresponding innermathematical procedural items. However, also they compared items with different conceptual demands as the procedural items did not require any mathematization process. Since then, the assumption that word problems are more difficult for students is often repeated in literature, but rarely shown empirically for items with equivalent conceptual demands which only differ in their format. This also applies for the conducted differential analyses with respect to language proficiency (see above).

Koedinger and Nathan (2004) draw a more differentiated picture: for algebra word problems, they show that differences in external representations (here called formats) "can affect performance when one representation is easier to comprehend than another" (ibid, p. 129), and also text formats can be more accessible. Also other studies show a possible positive effect of text formats: van den Heuvel-Panhuizen (2005) emphasizes that using contexts in problems (with texts or pictures) can also support students' performance. For example, sixth graders could subtract fractions in word problems with 30% more success than for the same problem in pure format; subtraction word problems were solved 20% better than in pure format, and multiplications with decimal numbers 40% better. She explains these differences by two sources: a context can enhance the accessibility of a problem and the underlying mathematical concepts (van den Heuvel-Panhuizen, 2005, p. 2) and a context problem 'provides students with the opportunity to solve problems by using informal strategies that are linked to contexts' (ibid., p. 7). Thus, the role of text and context is discussed incoherently in mathematics education research as texts can pose linguistic demands and

also increase the accessibility. The advantages and disadvantages of text formats for students with low language proficiency is to be investigated more systematically.

With respect to visual models (like diagrams or percent bars), empirical indications exist on their potential to facilitate the accessibility of a test item. For example, Walkington et al. (2013) show that for seventh graders, solving percent word problems percent bars provide an important support. One could even assume that students with low language proficiency do equally well as their more language proficient peers in visually presented items if there were no problems in conceptual understanding, only in text comprehension.

These different considerations motivated the research interest on comparing difficulties in *different problem formats,* i.e. *text format* (offering the main information and relations in written language), *visual format* (offering the main information and relations in graphical representations), and the so-called *pure format* (with mainly technical and symbolic language). This comparison between formats is treated systematically for the exemplary mathematical topic of percentages.

1.3 Conceptual demands posed by different problem types – The case of percentages

The mathematical topic percentages is chosen due to its major role in middle school mathematics and its importance in many everyday contexts. Furthermore, several empirical studies have shown students' difficulties with percentages and that percent problems in assessments bear various difficulties for students (e. g., Behr, et al., 1992; Kouba et al., 1988; Parker & Leinhardt, 1995). Nonetheless compared to other areas of arithmetic and proportions, relatively few recent studies exist that explore students' competencies and difficulties, (historical exceptions are named in Parker & Leinhardt, 1995; recent exceptions are Dole et al., 1997; Jitendra & Star, 2012 and Walkington et al., 2013).

In their research survey, Parker and Leinhardt (1995, p. 472f) resume the following four reasons for students' difficulties that help to determine the topic-specific conceptual demands:

(1) The complexity of the mathematical content (covering the coordination of percent amount and base as core concepts).

(2) The diversity of different relations which can be described by percentages (parts of wholes, comparisons, changes, …).

(3) The fact that the different relations (see the second aspect) are – except for part of whole – often not explicitly treated in the curricula.

(4) The use of 'an extremely concise linguistic form' (ibid. p. 473), which result in the fact that the relevant mathematical relations are often invisible in the language.

Whereas the first three reasons focus on conceptual understanding, the fourth point includes the problem of challenges in cracking word problems with percentages: The mathematizing process in the students' mastery of percent problems is typically characterized through one core step, the *identification of*

the problem type (Dole et al., 1997). Usually, *three elementary problem types* are distinguished (ibid., with different names): 'find the amount (if rate and base are given)', 'find the rate (if amount and base are given)', and 'find the base (if amount and rate are given)'. Several empirical studies show different success rates for different problem types (e.g. Kouba et al., 1988, p. 17). Some studies found that '[n]ot surprisingly, students were more successful in calculating a percent of a number than in solving other types of percent problems' (ibid., p. 17). As for many students the problem type 'find the amount' is easier than the two others, this type often being overgeneralized to 'find the base'.

Beyond these three elementary problem types, *more complex problem types* exist, for example 'percentage growth', 'percentage comparison' or 'find the base after reduction (if discount and reduced amount are given)' (Parker & Leinhardt, 1995, p. 439). These complex problem types pose even bigger conceptual demands for students, and perhaps also reading challenges. Kouba et al. (1988) assume that students have less experience with these more complex problem types or they 'are careless in their reading or are unable to comprehend [such] a nonroutine situation" (p. 18). The problem type 'find the base after reduction' is therefore suitable for systematically varying the conceptual demands in the test design (cf. Section 3).

Existing empirical studies have compared students' performances on percent problems mainly with respect *to problem types* (e.g. Kouba et al., 1988; Dole et al., 1997). In contrast, the comparison of *problem formats* have been less considered (see above Walkington et al., 2013 as an exception). Furthermore, little is known on difficulties with percent problems of students with varying language proficiency. Especially the more complex problem types seem to pose additional comprehension challenges that are worth being considered in more detail.

Research Questions

The resumed state of research with the constructs of problem types and problem formats provide the base for formulating the following refined research questions for capturing the role of language proficiency for different formats and types:

(Q1) How do students perform in parallel test items on percentages with text format, visual format and pure format?

(Q2) How does students' mastery of different problem formats differ between the groups of students with high and low language proficiency?

(Q3) How do the performance gaps depend on the problem types?

The research design served especially to validate or refute the following assertions from the literature:

(A1) Problems *in text format* are more difficult than *in pure format* due to comprehension difficulties for word problems (e.g. Kouba et al., 1988).

(A1*) Problems *in text format* are easier than in *pure format* since contexts can enhance students' accessibility of the problem (as resumed for elementary arithmetic problems by van den Heuvel-Panhuizen, 2005).

(A2) *Problems in visual format* are easier than in text and pure format since visual models can enhance the accessibility of the problem (Walkington et al., 2013).

(A3) *Students with low language proficiency* have difficulties with *other* problem formats than students with high language proficiency; especially they have specific difficulties with problems in *text format* (Duarte et al., 2011).

The results of a preliminary study with a similar design (Pöhler, Prediger, & Weinert, 2016) could not completely confirm these assertions derived from the literature review, as the performance gap between students with high and low language proficiency are similar for all problem formats. These facts motivate us to repeat the investigation of the research questions with a new sample and another method for the data analysis.

Research Design and Method

The study presented in this paper was conducted with a paper and pencil test on percent problems for N = 308 students in grade 7 (usually 12 to 13 years old).

3.1 Measures

3.1.1 Construction of the main instrument: The Percent-Cross-Test

The Percent-Cross-Test is constructed as paper-and pencil test with 17 items, in which three problem types are systematically crossed with three problem formats (cf. Table 1 for exemplary items, partly taken from Hafner, 2012, more items are shown in Pöhler et al., 2016). The items were coded dichotomously, the maximum score was 17 points.

For covering differing conceptual demands, the following three problem types are selected: 'find the amount', 'find the base' and 'find the base after reduction'. Based on the literature review in Section1.3, the problem type 'find the base after reduction' is expected to have the highest conceptual demand, whereas 'find the amount' is expected to have the lowest demand. The problem type 'find the rate' is omitted as it is the easiest to distinguish from the others by merely considering the involved units. Each problem type is presented in three formats (cf. Table 1):

For the 'pure format', exercises were given together with the technical terms (hence the decision of problem type is already explicit, it is of course not pure in the sense of "no language").

The items in the 'visual format' alsways used the bar model, an established visual model for percentages (van den Heuvel-Panhuizen, 2003), here contextualized in download bars which stem from a familiar everyday context for teenagers. The vidual model was kept constant in order to keep comparability.

Three or two items for each problem type are constructed in 'text format' with varying language challenges. Thus, the items in text format encompassed different levels of linguistic complexity which were constructed to correspond to the usual complexity of textbooks and exams in that age group.

Although not all items can be shown here, it is important to mention that the items across the three formats contained the same structural cores (same number sets) in each format. Table 1 shows items from different number sets.

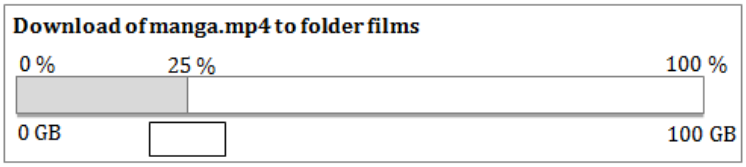
The limitation to 1-3 items per problem type and problem format results from research pragmatic reasons of the field, namely restricted time of student's concentration. Although previous studies have shown the validity of the items (Pöhler et al., 2016), in particular the results relating to the few items in visual format may only be interpreted with caution.

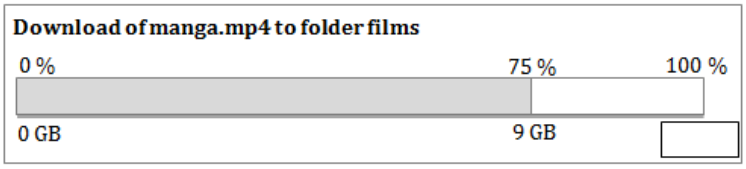
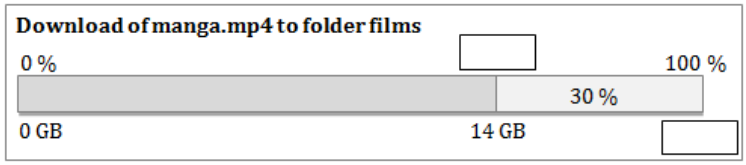
3.1.2 Other measures

Questionnaire for background variables. A students' questionnaire was administered to ask for gender, age, immigrant status (operationalized by students' and parents' countries of birth, at least one parent born abroad), and multilingualism (operationalized by languages spoken in the family and with friends).

The *socio economic status* was measured in students' self-report by the book scale with graphical illustrations that is widely used and has shown good retest scores (mean r = 0.80, cf. Paulus, 2009; 5-ary scale with 1 being low status, 5 being high).

Table 1. Item sets in three different formats for three problem types (translated)

Item set for problem type "Find amount"	
Pure format (2 Items)	What is 75% of 1000 g? Find the amount.
Visual format (1 Item)	How many gigabyte (GB) have already been downloaded? Find the missing value. 
Text formats (3 Items)	Potatoes consist of 75% water. How much water (in g) is contained in 1000 g potatoes?
Item set for problem type "Find base"	
Pure format (2 Items)	5% are 250 €. Find the base.
Visual format (1 Item)	What is unknown here? Find the missing value.

	
Text formats (3 Items)	When buying a new kitchen, Family Mays receives a discount of 250 €, that was 5% of the regular price. What is the normal price of the kitchen?
Item set for problem type "Find base after reduction"	
Pure format (2 Items)	Calculate the former price (base). New price: 30 € Discount: 40%
Visual format (1 Item)	What is unknown here? Find the missing values. 
Text formats (2 Items)	Mrs. Schmidt pays 30 € for a dress in the summer sale. The dress was reduced by 40%. How much did the dress cost before?

Language proficiency. Students' language proficiency was operationalized by the BISPRA-test (Redder & Wagner, 2015 adapted from Uesseler et al., 2013) which assesses receptive syntactical and semantical dimensions of composed verbs from the school academic register in typical contexts of percentages. With Cronbach's □□= 0.791, it shows a good internal consistency in the sample of N = 308 students and □ = 0.813 for a larger sample in German schools (N = 1124). The correlation of 0.18 between BISPRA scores and SES measure shows that both constructs are not identical.

3.2 Sampling and subsampling

The sample consisted of students of 25 classes in seven urban schools, in sum N = 308 students who were taught percentages in the recent months. The sample is representative for schools medium and low privileged milieus. For investigating differences in students' achievements with varied language proficiency, the sample of students was split into two subsamples according to their scores in the BISPRA-test: Students with a BISPRA score lower than the median in the whole sample were assigned to a low language proficient group, those with a higher BISPRA score into a high language proficient group. Table 2 gives an overview on descriptive characteristics of the whole sample and the two subsamples.

Table 2. Characteristics of the whole sample and the subsamples

	Whole Sample	Subsample students with low language	Subsample students with high language proficiency

		proficiency	
Number of students	308	174	134
Gender: Share of Boys / Girls	49% / 51%	52% / 48%	45% / 55%
Immigrant Status: Share yes / no	53% / 47%	64% / 36%	39% / 61%
Multilingual: Share yes / no	46% / 54%	66% / 33%	37% / 63%
Age M (SD)	12.69 (0.68)	12.76 (0.69)	12.60 (0.66)
SES M (SD)	3.07 (1.19)	2.94 (1.16)	3.24 (1.21)
Language Proficiency: BISPRA Score M (SD)	20.03 (4.95)	16.59 (3.64)	24.49 (2.00)
Percent-Cross-Test Total Raw Score M (SD)	5.91 (4.76)	4.30 (3.64)	7.99 (5.22)

3.3 Data Analysis

For the statistical data analysis, a probabilistic model is applied which accounts for the specific multidimensional structure of the Percent-Cross-Test, a member of the family of so-called Cognitive Diagnosis Models (CDMs). As this statistical method cannot be assumed to be known by all readers, the following sections account for its statistical background. However, the results can also be understood intuitively without these details.

Pursuing the research questions Q1 to Q3 in a methodologically sound and deep way puts some specific requirements on the statistical analysis and its empirical model to be applied: (1) A multi-dimensional approach is required for allowing a differentiated analysis of the test results based on the diverse problem types and problem formats. It should be possible to evaluate the students' ability for each problem type and each problem format individually as well as for combination of format and type. (2) The model should support a pre-assignment of the test items to the problem types and problem format. The test design yields a well-grounded theory of items belonging to specific problem types and formats. Thus this assignment should not be deduced empirically. (3) It is assumed that students' have to master both, the assigned problem type and the problem format, for effectively solving the respective item, i.e., a lack in either of the two parts cannot be compensated by a surplus in the other one. In terms of empirical models, that demand accounts for a non-compensatory model. (4) Because of the fine-grained definition of the skills, i.e. the three problem types and the three formats, it may be useful to characterize these skills as either present or absent from the students. Note that 'skill' is a technical term in the language of CDMs, we relate it here to abilities to cope with different types or formats, all referring toc conceptual demands.

The family of cognitive diagnosis models (CDMs; DiBello et al., 2007) allows for a realization of these targets. Specifically, the so called Deterministic

Input Noisy "And" Gate model (DINA; Haertel, 1989) was chosen because of its simplicity and non-compensatory assumption.

3.3.1 Background of Cognitive Diagnosis Models

The Deterministic Input Noisy "And" Gate model (DINA; Haertel, 1989) is a parsimonious and thus easily interpretable non-compensatory variant of a cognitive diagnosis model (CDM). Roughly spoken, CDMs yield a classification of the students with respect to a set of predefined skills or attributes (i.e. basic sub-competencies underlying a more coarse competence). More precisely the students' classifications can be split into three main outcomes:

(1) The skill distribution evaluates the percentages $P(\alpha_k)$ of students mastering each of the K individual skills. Skill is used here as technical term for sub-competency, not in the reduced sense of procedural routine skills.

(2) Because students can either master or not master each of the K skills, $L=2^K$ possible combinations of skill possession, the so-called skill classes, arise. The percentage of students $P(\boldsymbol{\alpha}_l)$ belonging to these classes is the second main outcome of a CDM, named the skill class distribution.

(3) A CDM model allows estimating each individual student's possession of the K skills, the student's dichotomous skill profile $\boldsymbol{\alpha}_i$.

For a more detailed introduction to CDMs and especially the DINA model, see for example George & Robitzsch (2015). For a discussion about the need of CDMs (compared to traditional educational assessments rooted in item response theory or classical test theory), their goals and the considerations involved in the development process, see de la Torre and Minchen (2014).

For applying a CDM, two components are to be prepared: Firstly, each substantial skill is to be assigned to a so-called latent categorical skill variable, termed α_k (cf. Table 3: e.g. α_A corresponds to the first skill 'find the amount').

It is assumed that each student possesses a subset out of a total of K skills $\alpha_A,\ldots,\alpha_{VF}$. In the case of the Percent-Cross-Test, there exist K=6 skills, i.e. the three problem types 'find the amount' α_A, 'find the base' α_B and 'find the base after reduction' α_{BR} and the three problem formats 'pure format' α_{PF}, 'text format' α_{TF} and 'visual format' α_{VF}. The problem types (cf. Section 1.3) can be considered as skills, since they require different mathematizations by the students. Even though the students are able to solve one of the elementary problem types 'find the amount' or 'find the base', it is not obvious that they could deal successfully with the other one since one may not assume 'that if students are able to do task in one direction they can automatically do it in the logically opposite manner' (Carpenter et al. 1980, p. 10). The problem type 'find the base after reduction' is more complex as it requires a further step (Parker & Leinhardt, 1995, p. 439). Even for solving items in different formats, students need specific skills. This shows especially the debate (cf. Section 1.2) about specific difficulties with word problems (e.g. Duarte et al., 2011, for an overview). The students' mastery of six skills is measured in a dichotomous format, i.e. for example $\alpha_A = 1$ corresponds to mastery of the skill 'find the amount' and $\alpha_A = 0$ to non-possession.

Table 3. Latent skill variables for the CDM model and substantial skills

Latent skill variable	Description	Domain
α_A	Find the amount	Problem type
α_B	Find the base	Problem type
α_{BR}	Find the base after reduction	Problem type
α_{PF}	Pure format	Problem format
α_{TF}	Text format	Problem format
α_{VF}	Visual format	Problem format

The second component to be prepared for an application of a CDM is the combination of skills that students should require for mastering a specific item j, $j = 1, \dots, J$. Educational experts define this assignment in a K-length dichotomous vector q_j. If the k^{th} skill is required to solve item j, $q_{jk} = 1$, or otherwise $q_{jk} = 0$. Collecting all vectors q_j for a test of length J results in a $J \times K$ weight matrix $\boldsymbol{Q}$, the so called Q-matrix (Fischer & Molenaar, 1995; Tatsuoka, 1984). In the case of the Percent-Cross-test, the information which skills are required to solve the items is already given through the item construction process, in which each item was meant to cover exactly one problem type and one problem format (cf. Table 1). Thus, the Q-matrix inherits a substantial theory about basic attributes (i.e. the skills) which are needed to effectively solve the test (cf. Section 3.3.3). Based on the Q-matrix and the dichotomous $I \times J$ response matrix $\mathbf{X}$ (containing the empirical responses of I students to J items), the CDM classifies the students with respect to the K skills.

One simple approach for estimating the percentages of students mastering a skill could be first calculating the percentage of students solving each item belonging to the respective skill and then building the average of these percentages. This approach differs from the main idea of the models' procedure, in that the model does not evaluate the responses to each item separately, but instead coherently estimates the classification based on the responses of a student to all J items.

3.3.2 The CDM DINA Model

This section introduces some basic notation to represent the mathematical form of the CDM DINA model, discuss its characteristics and introduce its application in the present study. As already mentioned above, the DINA model is non-compensatory, as a lack in one skill cannot be compensated by a surplus in another skill. For reasons of simplicity, let us assume for a moment that the skills possessed by an individual student i are known, i.e. the students' skill profile $\boldsymbol{\alpha}_i = [\alpha_{iA}, \dots, \alpha_{iVF}]$ is given. As $\boldsymbol{\alpha}_i$ is defined to be dichotomous, student i possesses skill k, k=$A,\dots,VF$, in case of $\alpha_{ik} = 1$ and lacks skill k in case of $\alpha_{ik} = 0$. Given the individual student's dichotomous skill profile $\boldsymbol{\alpha}_i$ the non-compensatory assumption of the DINA model is reflected in the construction of the student's expected latent response

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

If $\xi_{ij} = 1$, the student is expected to master item j, otherwise not (i.e. $\xi_{ij} = 0$). Actually, a student may master an item with probability g_j (the so-called guessing parameter) in spite of not being expected to. In the same line, a student may slip an item with probability s_j (the so-called slipping parameter) even though being expected to master it. If the DINA model holds, we can express the response probability of a student i with skill profile $\boldsymbol{\alpha}_i$ as

$$P\left(X_{ij} = 1 \mid \boldsymbol{\alpha}_i\right) = P\left(X_{ij} = 1 \mid \xi_{ij}, g_j, s_j\right) = (1 - s_j)^{\xi_{ij}}\, g_j{}^{1-\xi_{ij}}.$$

Note that in real life applications, the students' skill profiles $\boldsymbol{\alpha}_i$ are unknown. It is one of the model's goals to estimate them. To achieve that marginal maximum likelihood estimation and expectation maximization algorithm are deployed (de la Torre, 2009).

After estimating, the DINA model yields values for the following parameters: the item parameters g_j (guessing) and s_j (slipping); the skill possession parameters $P(\alpha_k)$ (i.e. the percentage of students possessing each skill) and skill class parameters $P(\boldsymbol{\alpha}_l)$ (i.e. the percentage of students possessing a specific combination of skills). Based on the item parameters we may interpret $\omega_{1j} = 1 - g_j - s_j$ as item discrimination, where values close to or greater than 1 indicate a good separation of examinees with low abilities from examinees with high abilities (George & Robitzsch, 2015). For statistical inference of the parameters (standard errors and test statistics) the analysis were recalculated 50 times with randomized jackknife zones. A criterion for evaluating the absolute model fit is the standardized root mean square residual (SRMSR; Maydeu-Olivares, 2013). Maydeu-Olivares suggests that SRMSR values smaller than 0.05 indicate well-fitting models.

To compare the extent to which groups (i.e. students with high versus low language proficiency) differ in their skill possession $P(\alpha_k)$ a multiple group DINA model is established (e.g. Johnson et al., 2013). For avoiding a biased estimation of the group differences, invariant item parameters are chosen as identification condition (cf. de la Torre & Lee, 2010; Johnson et al., 2013; Xu & von Davier, 2008).

3.3.3 Q-matrix of present model and derived parameters

Directly implementing the skill to item assignment in a Q-matrix would end in a matrix having for each item exactly one 1 in the first three columns (i.e. the assignment of one of the three problem types) and one 1 in the last three columns (i.e. the assignment of one of the three problem formats). As George and Robitzsch (2014) showed such matrices would leave the skills unidentified, all combinations are established between one problem type and one problem format as skills in the CDM model for the Percent-Cross-Test. This leads to a model with $K^* = 9$ skills as outlined in Table 4. As this matrix only includes assignments of items to one single skill, the non-compensatory model of the DINA corresponds to any other compensatory model.

Table 4. Q-matrix for assignment of nine combinations between problem type and problem format skills to test items

	$\alpha_{A;PF}$	$\alpha_{A;TF}$	$\alpha_{A;VF}$	$\alpha_{B;PF}$	$\alpha_{B;TF}$	$\alpha_{B;VF}$	$\alpha_{BR;PF}$	$\alpha_{BR;TF}$	$\alpha_{BR;VF}$
Item 1	1	0	0	0	0	0	0	0	0
Item 2	0	0	0	1	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Item 10	0	0	0	0	0	0	0	1	0
Item 11	0	0	0	0	1	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Item 16	0	0	0	0	0	0	0	0	1

$\alpha_{A:PF}$ find the amount / pure format, $\alpha_{A:TF}$ find the amount / text format, $\alpha_{A:VF}$ find the amount / visual format, $\alpha_{B:PF}$ find the base / pure format, $\alpha_{B:TF}$ find the base / text format, $\alpha_{B:VF}$ find the base / visual format, $\alpha_{BR:PF}$ find the base after reduction / pure format, $\alpha_{BR:TF}$ find the base after reduction / text format, $\alpha_{BR:VF}$ find the base after reduction / visual format.

The CDM allows for determining the percentages $P(\alpha_{A:PF}),\ldots,P(\alpha_{BR:VF})$ of students possessing these nine skills. Based on percentages of students possessing the $K^* = 9$ combinations of one problem type and one problem format, the percentages $P(\alpha_k)$ of students possessing the original $K = 6$ sub-competencies can be derived. Therefore, $P(\alpha_k)$ is defined as mean values of the three percentages including skill α_k at a time. For example, the percentage of students possessing α_A (find the amount) by $P(\alpha_A) = [P(\alpha_{A;PF}) + P(\alpha_{A;TF}) + P(\alpha_{A;VF})]/3$.

As the CDM model provides a ratio scale of measurement, the differences between the subsamples of students with high and low language proficiency are meaningful, e.g.

$$\Delta P(\alpha_{A;PF}) = P(\alpha_{A;PF} \mid \text{high language}) - P(\alpha_{A;PF} \mid \text{low language})$$

and analogically the differences of possessing the (six) original skills. Positive values in both measures indicate advantages for students with high language proficiency. This allows to evaluate if the students' mastery of nine skill combinations and six skills differs significantly between both subsamples.

The calculations were conducted with the statistical programming software R (R Core Team, 2015) and especially the R package CDM (George, Robitzsch, Kiefer, Gross, Ünlü, 2016; Robitzsch; Kiefer, George, Ünlü, 2016). All significance tests referred to the significance level of .05 unless specified differently.

3.3.4 Model fit and item related parameters

The fitted DINA model with the nine skills described in Section 3.3.3 has an overall model fit of SRMSR=0.074 and thus the model is accurate.

Table 5 shows a summary of the item related parameters. The item p-values describe the percentage of students solving the items: 9% of the students

solved the most difficult item, whereas 62% solved the easiest item. On average 29% of the students solved the items, which makes the test relatively hard. The item guessing parameters range from .01 to 0.46 (SD=.11) and the item slipping parameters range from .01 to 0.65 (SD=.21). The item discriminations ω_{1j} range from .34 to .91 with a mean value of .67 (SD=.21). These values indicate that the items separate well between students possessing the relevant skills and those who do not, and thus the items can separate between students who rule the skills from those who do not. The difficulty of the items based on the empirical solution frequency is satisfying.

Table 5. Model summary - Item related parameters

	Minimum	Maximum	Mean	Standard Deviation
Item p-values	.09	.63	.31	.17
Guessing parameter g_j	.01	.47	.06	.11
Slipping parameter s_j	.01	.62	.25	.20
Item discrimination ω_{1j}	.37	.90	.68	.20
Correlations between nine skills	.39	.96	.67	.17

Furthermore, the correlations between the nine skills were inspected. They have a mean of .67 and range between .39 and .96 (SD=.17). The correlations indicate relations between the mastery of the problem types and formats. This is what we expect since students mastering e.g. find the base after reduction in pure format might also master find the base in pure format. However, it is no perfect correlation, meaning that students mastering one problem type and format (i.e. one skill) do not automatically master all the other skills, too.

Results

4.1 Students' performances in different problem formats and problem types

The overall average performance of students on the Percent-Cross-Test clearly indicates that students, both with high and low language proficiency, lack conceptual understanding to solve items involving percentages, but of course with differences. With regard to the first research question Q1 on students' performances in different problem formats, the empirical results in Table 6 document nearly equivalent mastery probabilities in the three problem formats, with slightly positive trends for the text and especially the visual format. Students have a general probability of 33.7% to solve items in pure format, 35.0% for items in text format and 38.1% for items in visual format. This means that problems in text format are not potentially significant more difficult than items in pure or visual format.

Thus, the assertion (A1) that problems in text format are more difficult than in pure format due to comprehension difficulties for word problems (Kouba et al., 1988) couldn't be confirmed by these results. Otherwise, tasks in pure format seem to be slightly more difficult than in the other two formats. Hence the claim (A1*) that problems in text format are easier than in pure format slightly tend to apply, since their contexts can enhance students' accessibility of the problem (van den Heuvel-Panhuizen, 2005).

Table 6. Mastery probabilities of nine combinations of problem types and problem formats (with standard errors) and mastery probabilities of problem types and problem formats

Problem format **Problem type**	Pure format a_{PF}	Text format a_{TF}	Visual a_{VF}	Mastery probability of problem types
Find the amount a_A	.353 (.066)	.425 (.054)	.617 (.030)	.461 (.038)
Find the base a_B	.498 (.052)	.434 (.055)	.305 (.029)	.413 (.034)
Find the base after reduction a_{BR}	.161 (.029)	.202 (.042)	.221 (.023)	.195 (.025)
Mastery probability of problem formats	.337 (.036)	.350 (.045)	.381 (.023)	

However, these results on general probabilities must be differentiated with respect to the different problem types, as the other lines in Table 6 show:

- For the problem type 'find the amount', the probabilities for each format vary substantially: students have a 35.3% probability to solve items in pure format, 42.5% for items in text format and 61.7% for items in visual format. For this problem type, assertion (A1*) that text formats can enhance accessibility, is more powerful.

- In contrast, for the problem type 'find the base', the assertion (A1) that problems in text format are more difficult than in pure format (Kouba et al., 1988) tends at least not to contradict the results (49.8% for items in pure format, 43.4% for items in text format).

- For the more complex problem type 'find the base after reduction' the picture changes again, the probability for the pure format is with 16.1% little lower than for the text format with 20.2%.

The assertion (A2) was that visual formats are easier than other formats since they can enhance the accessibility of a problem (Walkington et al. 2013). According to the results in Table 6, this mainly applies for 'find the amount', but also but to a lesser extent for the more unknown problem type 'find the base after reduction' where the visual format received a probability of 22.1%, compared to 16.1% for the pure and 20.2% for the text format.

4.2 Differences between the subsamples with high and low language proficiency

The second research question Q2 asked for group differences in students' mastery of problem formats. For this purpose, Table 7 provides the separated data for mastery probabilities in both subsamples, in the upper part for students with high language proficiency, in the middle part for students with low language proficiency and in the lower part the differences between both subsamples.

Table 7 shows that for students with low language proficiency the format of an item seems to have a greater influence than for students with high language proficiency. For the second group there are only minimal differences between the mastery probabilities of the problem formats, ranging from 46.0% for items in pure format, 48.0% for items in visual format to 49.2% for items in text format. The students with low language proficiency have lower mastery probabilities for all problem types and all problem formats as their more language proficient peers. Furthermore, they show a preference for items in visual format (mastery probability of 30.5%) and have slightly higher mastery probabilities for tasks in text format (mastery probability of 24.5%) than for items in pure format (mastery probability of 22.1%). However, the performance gaps between the groups are similar for the pure and the text format (Δ = 23.9% for pure format, Δ = 24.7% for text format) and smaller for the visual format (Δ = 16.6%). Thus, the students with low language proficiency don't have a considerable bigger disadvantage with the text format.

Research questions Q3 asks for differences in the performance gaps between students of high and low language proficiency within the problem types. The general pattern of difficulties between problem types seem to be parallel for both subsamples. The problem type 'find the amount' is solved with the highest probabilities, i.e. in the subgroup with high language proficiency with 60.2% and with 33.3% significantly worse in the group with low language proficiency.

Table 7. Skill mastery probabilities of nine combinations of problem types and problem formats (with standard errors)
for students with high and low language proficiency and their differences
(with differences being significant at the level of .05 printed in **bold numbers**).

Students with high language proficiency	Pure format α_{PF}	Text format α_{TF}	Visual α_{VF}	Skill mastery of problem types
Find the amount α_{A}	.485 (.044)	.582 (.044)	.739 (.047)	.602 (.034)
Find the base α_{B}	.643 (.044)	.592 (.052)	.388 (.044)	.541 (.037)
Find the base after reduction α_{BR}	.254 (.040)	.303 (.042)	.313 (.036)	.290 (.033)
Skill mastery of problem formats	.460 (.035)	.492 (.039)	.480 (.035)	
Students with low language proficiency	Pure format α_{PF}	Text format α_{TF}	Visual α_{VF}	Skill mastery of problem types

Find the amount α_A	.197 (0.027)	.281 (0.033)	.523 (0.03)	.333 (0.024)
Find the base α_B	.384 (0.045)	.314 (0.036)	.241 (0.034)	.313 (0.031)
Find the base after reduction α_{BR}	.083 (0.021)	.140 (0.028)	.149 (0.027)	.124 (0.022)
Skill mastery of problem formats	.221 (0.022)	.245 (0.025)	.305 (0.027)	
Differences between groups of low and high language proficiency	Pure format α_{PF}	Text format α_{TF}	Visual α_{VF}	Skill mastery of problem types
Find the amount α_A	Δ = **.288** (.048)	Δ = **.301** (.051)	Δ = **.216** (.054)	Δ = **.268** (.037)
Find the base α_B	Δ = **.259** (.056)	Δ = **.278** (.060)	Δ = **.164** (.042)	Δ = **.228** (.047)
Find the base after reduction α_{BR}	Δ = **.171** (.046)	Δ = **.163** (.049)	Δ = **.171** (.046)	Δ = **.166** (.038)
Skill mastery of problem formats	Δ = **.239** (.039)	Δ = **.247** (.042)	Δ = **.166** (.036)	

Slightly lower probabilities can be found for 'find the base', 54.1% for the high language proficiency group and a significant lower probability of 31.3% for their less language proficient peers. For both groups the probabilities for solving the more complex problem type 'find the base after reduction' are far lower than for the other problem types: For students with high language proficiency the probability to solve such an item is with 29.0% only half and in the subgroup of low language proficiency with 12.4% only about a third of the respective size. The differences between the groups are significant and greater for the basic problem types 'find the amount' and 'find the base' (Δ = 26.8% and Δ = 22.8%) than for the more complex problem type which reaches Δ = 16.6% more for the group of high language proficiency.

The performance gap between the groups with respect to the problem formats vary with the problem types. The differences vary between Δ = 16.4% for 'find the base' in visual format and Δ = 30.1% for 'find the amount' in text format. However, the text format does reach with only small differences to the pure format the highest performance gap for the problem types-'find the base' (Δ = 27.8% and Δ = 25.9%) and 'find the amount' (Δ = 30.1% and Δ = 28.8%).

In total, the assertion A3 must be differentiated in many ways, and the text format does not turn out to considerably disadvantage students with low language proficiency the most.

Discussion

Are word problems really more difficult for students with low language proficiency? Although this assertion is often stated in literature, there exist only very old empirical evidences (Carpenter et al. 1980, Kouba et al., 1988) which seem to have not been replicated since then, whenever the problem format is thoroughly disentangled from the cognitive demand (conceptual or procedural demands?). Although a lot of evidence exists that within the text format, students' success depend on the language proficiency (Walzebug 2014; Haag et al. 2013 and many others), there was a need to construct a test with contestant cognitive demands and numbers sets for comparing really only the difficulty of problem formats.

In our study, the often claimed assertion that the performance in cracking percent problems depends to a great extent on the problem format is not confirmed by the presented test with N = 308 students. Thus there rarely are differences between the probabilities for solving problems in pure, visual and text format for the whole sample. For the group of students with low language proficiency, the probabilities of all considered abilities and ability combinations are lower than for the more language proficient learners, which was expectable. Thereby, the disparities between the two groups are much lower for problems in visual format than in the other two formats. According to that, the problem format has a slightly greater influence on the probability to solve an item for students with low than for high language proficiency. With respect to the fact that the percent bar is not commonly seen in conventional German textbooks, this result is interpreted as an indicator for its intuitive accessibility. This could serve as an argument for its use as visual representation to introduce percentages (see approaches of van den Heuvel-Panhuizen, 2003; Pöhler & Prediger, 2015). As a restriction, it must be noted repeatedly that only a limited number of items in the visual format are considered.

With regard to the three considered problem types, the mastery probabilities follow the same general pattern: As expected, they are considerably higher in each case for the basic problem types 'find the amount' and 'find the base' than for the more complex problem type 'find the base after reduction'. For students with high language proficiency, the text format seems to enhance the accessibility for the more unknown problem types a bit more than the visual models, whereas their low language proficient peers seem to be a bit more supported by the visual format.

In this way, the here presented study conducted by means of a DINA-model confirmed the central result of a similar study with another sample (Pöhler et al., 2016). Accordingly, the students with high language proficiency outperformed the low language proficient students in all items and not only with regard to a particular problem format. Based on the results from the investigation of percent items in different problem formats and problem types, the question in the title "Are word problems really more difficult for students with low language proficiency" has to be denied. This suggests that not the restricted language proficiency alone is responsible for the disadvantages, especially for word problems, of the low language proficient group. Other studies which combine the tests with qualitative investigations (e.g. Prediger et al., 2015) provide empirical indications that the main issue for students with low

language proficiency is their lacking conceptual understanding, not only reading word problems.

As a practical consequence for classrooms, the findings provide quantitative evidence for the relevance of Moschkovich's (2013) practical recommendation not to restrict support for low language proficient learners in mathematics to word problems alone. Instead of that, a consequent intertwinement of language learning with the development of conceptual understanding is required (Moschkovich, 2013). A developed learning arrangement for percentages which consequently integrate lexical and conceptual learning, attempt to meet these demands (Pöhler & Prediger, 2015) and has shown significant efficacy for students' learning (Pöhler, Prediger, & Neugebauer, 2017).

The claim of validity is content-related. It is limited by focusing only the exemplary mathematical content area percentages and furthermore considering merely three of at least five possible problem types within the design of the test. This selection can be pragmatically justified, such as the small number of items per problem format and type, which act as limitation on the methodological level, with the time-effective conduction of the test in the educationally context.

For future research, an extension of the findings should be planned with deeper qualitative insights into students' processes while solving the items in an interview study, as well as a transfer of the study to other content areas.

Acknowledgement

The data for this study was gathered in the context of the project MuM-Multi, funded by the German ministry BMBF (grant 01JM1403A to Susanne Prediger). We thank our partners Angelika Redder and Jonas Wagner for providing the BISPRA-Test and Alexander Schüler-Meyer and Lena Wessel for managing the test administration.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Birte Pöhler - Institute for Development and Research in Mathematics Education, TU Dortmund University, Germany.

Ann Cathrice George - Institute for Educational Research, Innovation and Development of the Austrian School System (BIFIE), Salzburg, Austria.

Susanne Prediger - Institute for Development and Research in Mathematics Education, TU Dortmund University, Germany.

Henrike Weinert - Institute for Development and Research in Mathematics Education, TU Dortmund University, Germany.

References

Abedi, J. (2004). Will You Explain the Question?. *Principal Leadership*, *4*(7), 27-31.

Abedi, J. (2006). Language issues in item-development. In S. M. Downing & T. M. Haldyna (Eds.), *Handbook of test development* (pp. 377-398). Mahwah: Erlbaum.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.

Behr, M. J., Harel, G., Post, T. R., & Lesh, R. (1992). Rational number, ratio, and proportion. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 296–332). New York: Macmilian.

Carpenter, T. P., Corbitt, M. K., Kepner, H. S., Lindquist, M. M., & Reys, R. E. (1980). NAEP note: Problem solving. *The Mathematics Teacher*, *73*(6), 427-433.

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

de La Torre, J. & Minchen, N. (2014). Cognitively Diagnostic Assessments and the Cognitive Diagnosis Model Framework. *Psicología Educativa*, 20(2), 89–97.

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

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

Dole, S., Cooper, T.J., Baturo, A.R., & Conoplia, Z. (1997). Year 8, 9 and 10 students' understanding and access of percent knowledge. In A. Begg (Ed.), *People in mathematics education. Proceedings of 20th MERGA* (pp. 7-11). Rotorua: Merga.

Duarte, J., Gogolin, I. & Kaiser, G. (2011). Sprachlich bedingte Schwierigkeiten von mehrsprachigen Schülerinnen und Schülern bei Textaufgaben. In S. Prediger & E. Özdil (Eds.), *Mathematiklernen unter Bedingungen der Mehrsprachigkeit* (pp. 35-53). Münster: Waxmann.

Fischer, G.H. & Molenaar, I.W. (1995). *Rasch models: foundations, recent developments and applications*. New York: Springer.

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

George, A. C. & Robitzsch, A. (2015). Cognitive diagnosis models in R: A didactic. *The Quantitative Methods for Psychology*, 11(3), 189–205.

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R Package CDM for cognitive diagnosis modeling. *Journal of Statistical Software, 74*(2), 1–24.

Haag, N., Heppt, B., Roppelt, A., & Stanat, P. (2015). Linguistic simplification of mathematics items: effects for language minority students in Germany. *European Journal of Psychology of Education, 30*(2), 145-167.

Haag, N.,Heppt, B., Stanat, P.,Kuhl, P.,& Pant, H. A. (2013). Second language learners'performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction, 22*(28), 24–34.

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

Hafner, T. (2012). Proportionalität und Prozentrechnung. Wiesbaden: Vieweg + Teubner.

Hirsch, E. D. (2003). Reading Comprehension Requires Knowledge - of Words and the World. Scientific Insights into the Fourth-Grade Slump and the Nation's Stagnant Comprehension Scores. *American Educator, 4*(1), 10-44.

Jitendra, A. K., & Star, J. R. (2012). An exploratory study contrasting high- and low-achieving students' percent word problem solving. *Learning and Individual Differences, 22*(1), 151-158.

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

Koedinger, K.R. & Nathan, M. J. (2004). The real story behind the story problems. Effects of representations on quantitative reasoning. *The Journal of the Learning Sciences, 13*(2), 129-164.

Kouba, V., Brown, C., Carpenter, T., Lindquist, M., Silver, E., & Swafford, J. (1988). Results of 4th NAEP Assessment of Mathematics. *Arithmetic Teacher, 35*(8), 14-19.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*(2), 333–368.

Maydeu-Olivares. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*, 71–137.

Moschkovich, J. (2013). Principles and Guidelines for Equitable Mathematics Teaching Practices and Materials for English Language Learners. *Journal of Urban Mathematics Education, 6*(1), 45-57.

OECD (2007). *PISA 2006*. Vol. 2: Data. Paris: OECD.

Parker, M. & Leinhardt, G. (1995). Percent: A Privileged Proportion. *Review of Educational Research, 65*(4), 421-481.

Paulus (2009). *Die Bücheraufgabe zur Bestimmung des kulturellen Kapitals bei Grundschülern*. URL: http://psydok.sulb.uni-saarland.de/volltexte/2009/2368/.

Pöhler, B., & Prediger, S. (2015). Intertwining lexical and conceptual learning trajectories - A design research study on dual macro-scaffolding towards percentages. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(6), 1697-1722.

Pöhler, B., Prediger, S., & Neugebauer, P. (2017, in press). Content- and language integrated learning: A field experiment for percentages. To appear in *Proceedings of the 41st Annual Meeting of the International Group for the Psychology of Mathematics Education (PME 41).* Singapore: PME.

Pöhler, B., Prediger, S., & Weinert, H. (2016). Cracking percent problems in different formats - The role of texts and visual models for students with low and high language proficiency. In K. Krainer & N. Voundrová (Eds.), *CERME 9. Proceedings of the Ninth Congress of the European Society for Research in Mathematics Education* (pp. 331-338). Prague: Charles University / ERME.

Prediger, S., Renk, N., Büchter, A., Gürsoy, E. & Benholz, C. (2013). Family background or language disadvantages? Factors for underachievement in high stakes tests. In A. Lindmeier & A. Heinze (Eds.), *Proceedings of 37th PME* (4, 49-59). Kiel: PME.

Prediger, S., Wilhelm, N., Büchter, A., Gürsoy, E., & Benholz, C. (2015). Sprachkompetenz und Mathematikleistung–Empirische Untersuchung sprachlich bedingter Hürden in den Zentralen Prüfungen 10. *Journal für Mathematik-Didaktik*, *36*(1), 77-104.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

Redder, A. & Wagner, J. (2015): Bispra-Test. Project internal test development, adapted from Uesseler et al. 2015.

Robitzsch, A., Kiefer, T., George, A. C. & Ünlü, A. (2016). CDM: Cognitive Diagnosis Modeling. R Package version 3.1-14. Retrieved from http://CRAN.R-project.org/package=CDM.

Secada, W. G. (1992). Race, ethnicity, social class, language and achievement in mathematics. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 623–660). New York: MacMillan.

Tatsuoka, K. K. (Ed). (1984). Analysis of errors in fraction addition and subtraction problems. Final Report for Grant No. NIE-G-81-0002. Urbana, IL: University of Illinois.

Uesseler, S., Runge, A., & Redder, A. (2013). „Bildungssprache" diagnostizieren. Entwicklung eines Instruments zur Erfassung von bildungssprachlichen Fähigkeiten bei Viert- und Fünftklässlern. In A. Redder & S. Weinert (Eds.), *Sprachförderung und Sprachdiagnostik. Interdisziplinäre Perspektiven* (pp. 42-67). Münster: Waxmann.

Van den Heuvel-Panhuizen, M. (2003). The didactical use of models in realistic mathematics education. *Educational Studies in Mathematics, 54*(1), 9-35.

Van den Heuvel-Panhuizen, M. (2005). The role of contexts in assessment problems in mathematics. *For the Learning of Mathematics*, 25 (2), 2-9.

Walkington, C., Cooper, J., & Howell, E. (2013). Effects of visual representations and interest-based personalization on solving percent problems. In Martinez, M. & Castro Superfine, A. (Eds.), *Proceedings of 35th PME-NA* (pp. 533-536). Chicago: University of Illinois.

Walzebug, A. (2014). Is there a language-based social disadvantage in solving mathematical items? Learning, Culture and Social Interaction 3 (2), 159-169.

Wolf, M. K., & Leon, S. (2009). An Investigation of the Language Demands in Content Assessments for English Language Learners. *Educational Assessment, 14*(3-4), 139-159.

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