Language proficiency and mathematics achievement –
Empirical study of language-induced obstacles in a high stakes test, the central exam ZP10

Susanne Prediger · Nadine Wilhelm · Andreas Büchter · Erkan Gürsoy · Claudia Benholz

This document provides an English translation of the article Prediger, Susanne, Wilhelm, Nadine, Büchter, Andreas, Gürsoy, Erkan & Benholz, Claudia (2015). Sprachkompetenz und Mathematikleistung – Empirische Untersuchung sprachlich bedingter Hürden in den Zentralen Prüfungen 10. Journal für Mathematik-Didaktik, 36(1), 77-104. DOI: 10.1007/s13138-015-0074-0. English DOI 10.1007/s13138-018-0126-3.

Abstract By analysing of (co-)variance in high stakes test items (of the central exam ZP 10 written in the grade 10 classrooms in North Rhine-Westphalia, the study shows that language proficiency is the background factor with the strongest connection to mathematics achievement among all social and linguistic background factors. A differential functional analysis, an analysis of written products created by students and clinical interviews provide deeper insights into this connection and therefore a contribution to the empirical investigation of language-induced obstacles in high stakes tests. They also refer to the epistemic function of language beyond reading obstacles.

Key words Mathematics achievement · Language Proficiency · Language obstacles · DIF-Analysis

Zusammenfassung Anhand der Zentralen Prüfung 10 Mathematik in Nordrhein-Westfalen wird durch Varianz-, Kovarianz- und Regressionsanalysen gezeigt, dass sprachliche Kompetenz unter allen sozialen und sprachlichen Faktoren den stärksten Zusammenhang zur Mathematikleistung hat. DIF-Analysen sowie Analysen schriftlicher Bearbeitungen auf Basis von Interviewanalysen ermöglichen ein tieferes Verständnis des starken Zusammenhangs und damit einen Beitrag zur empirischen Untersuchung von sprachlich bedingten Hürden in den abschlussrelevanten Prüfungen. Sie verweisen insbesondere auch auf die kognitive Funktion von Sprache jenseits von Lesehürden.

Schlüsselwörter Mathematikleistung · Sprachkompetenz · Sprachlich bedingte Hürden · DIF-Analyse

Mathematics Subject Classification (2010) 97D60 · 97D70

1. Introduction

The starting point for this study originated from the empirical finding in international comparative tests showing that school achievement is connected to students' family background, in Germany more strongly than in other countries (Baumert & Schümer, 2001). Whereas most empirical studies and official statistics in Germany focus solely on social background factors and factors of language biography, the international findings suggest a more profound consideration of the factor language proficiency beyond reading proficiency (Secada, 1992; Abedi, 2006).

This article presents a study that investigates which factors of family and language have the strongest connection to mathematics achievement in a high stakes central exam called

ZP10-Mathematik (the central exam for the medium track in the German state North-Rhine-Westphalia for grade 10 in the year 2012).

In the first step, the mathematics achievement scores of 1495 students were investigated with respect to the data about immigrant status, socio-economic status, multilingualism and age of first exposure to German, as well as language proficiency and reading proficiency.

In the second step, the analysis of items enabled the investigation of the statistically detected connection between mathematics achievement and language proficiency using a profound specification of language-induced obstacles in the items. Language-induced obstacles are obstacles which emerge in the test situations on a linguistic level, e.g. reading difficulties. Furthermore, obstacles which students with low language proficiency are more likely to encounter are also subsumed under language-induced obstacles. The empirical analysis shows that language-induced obstacles can be classified as conceptual or processual and that they must be traced back to students' difficulties in earlier processes in which conceptual understanding or processual competencies are acquired. For explaining in which way these obstacles can be subsumed as language-induced, the article draws upon the epistemic function of language.

2. Theoretical and empirical backgrounds: Relevance of family and language factors

2.1 Background factors for capturing underprivileged students

Several empirical studies show that in Germany, underprivileged students achieve substantially lower scores in tests than their peers; this applies especially for mathematics (Baumert & Schümer, 2001). Previous studies use various family and language factors to operationalize the underprivileged status and relate it to mathematics achievement:

- nationality (e.g. Autorengruppe Bildungsberichterstattung, 2012; Mikrozensus, 2011),
- immigrant status operationalized by the countries of birth of the students and their parents (e.g. OECD, 2007; Autorengruppe Bildungsberichterstattung, 2012; Tarelli et al., 2012; and many others),
- multilingualism and missing correspondence between family language and language of instruction (e.g. OECD, 2007, p. 120; Heinze et al., 2007; Burns & Shadoian-Gersing, 2010; Ufer et al., 2013; Haag et al., 2013),
- socio-economic status (e.g. Bos et al., 2003; Ehmke et al., 2004; Werning et al., 2008), and
- reading proficiency (Rindermann, 2006; Leutner et al., 2004, p. 167ff.; Knoche & Lind, 2004, p. 206; Bos et al., 2012, p. 237ff.), which is also discussed as a mediator for other background factors (Walzebug, 2014).

Whereas most German studies focus on family background or reading proficiency, findings from other countries suggest that *language proficiency in a wider sense* might have a bigger impact on mathematics achievement (Pimm, 1987; Secada, 1992, S. 638; Abedi, 2006). For investigating this possible connection, the relationship between language and mathematics must be conceptualized more precisely.

2.2 Language proficiency and language-induced obstacles

In mathematics education research there is a long tradition of investigating the role of language in mathematics classrooms (Pimm, 1987; Ellerton & Clarkson, 1996). With respect to disparities according to family backgrounds, a new dimension must be established as outlined in the following (cf. Prediger & Özdil, 2011 for an overview).

The connection between family background and language proficiency has been theoretically conceptualized mostly by the distinction between everyday language and school academic language, or by BICS (Basic Interpersonal Communication Skills) and CALP (Cognitive Academic Language Proficiency, Cummins, 2000): Many students with immigrant status or from socially underprivileged families develop solid everyday language skills, but often not enough academic language proficiency which is crucial in school contexts (Schleppegrell, 2004; Gogolin, 2009; Morek & Heller, 2012). The school academic language register differs from the everyday register with regard to specific lexical demands (in German, e.g., prefix verbs, compound words, standardized technical concepts and a high lexical density) as well as numerous grammatical-syntactical and discourse features which allow high condensation and decontextualization (ibid.). The competencies related to these specific demands are not automatically acquired in everyday communication.

Even though the construct of school academic language still requires further elaboration, which is currently an issue in linguistic, psychological and educational research (actual survey in Redder & Weinert, 2013), some empirical studies already show how limitations in academic language proficiency influence mathematics learning and mathematics achievement (Kaiser & Schwarz, 2003; Heinze et al., 2007; Gellert, 2011; Rösch & Paetsch, 2011; Ufer et al., 2013; Prediger, 2013; Prediger & Wessel, 2013). These studies provide an important foundation for the current study as they show that conceptualizations of language proficiency beyond reading proficiency are required. Its leading idea is the duality of *communicative and epistemic function of language* (Maier & Schweiger, 1999, S. 18; Morek & Heller, 2012, Pimm, 1987).

Until now, the connection between language proficiency and mathematics achievement has been discussed for tests mainly with respect to the communicative function of language: Some students encounter more difficulties when decoding the test items, which may cause that they cannot show their mathematical competencies. In this case, language-induced obstacles are reading obstacles which are mostly considered (especially in the US-American discourse, Abedi, 2006; Brown, 2005; Wolf & Leon, 2009; Martiniello, 2009) as language biases. Language biases are indications of missing validity of the tests which assess reading proficiency instead of mathematics achievement. From this perspective, it makes sense to conceptualize language proficiency as reading competence and identify reading obstacles in order to eliminate them in a fair test which only focusses on mathematical competencies (Abedi, 2006).

Besides its communicative function, language also has an epistemic function, i.e., it is also a cognitive tool in thinking and learning processes (Maier & Schweiger, 1999, p. 18, Pimm, 1987). For explaining the epistemic function theoretically, linguistic models were developed to conceptualize the relationship between thinking and speaking in different ways (Morek & Heller, 2012). In particular, the epistemic function is emphasized in school academic language register, because the specific language features of condensation and decontextualisation allow more complex cognitive and epistemic processes (Halliday, 1993; Schleppegrell, 2004; Thürmann, Vollmer & Pieper, 2010; Morek & Heller, 2012).

With respect to the epistemic function of language, language-induced obstacles can emerge in mathematics tests for students with low LP, when the task itself is understood by the learners but they cannot cope with the cognitive demands. Hence, language-induced obstacles are not only reading obstacles but also include other aspects which act as obstacles for students with low language proficiency. Certain obstacles can be traced back to longer-term, language-induced limitations during thinking and learning processes, as this present study shows.

A first indication for a longitudinal, longer-term relevance of the epistemic function of language was given by Heinze et al. (2007) in their study about the development of achievements from Grade 1 to Grade 2. This study shows that socio-economic status and multilingualism have the strongest correlation with mathematics achievement in Grade 1. In contrast, the longitudinal development is most closely connected to cognitive abilities and to language proficiency (in this particular study operationalized as listening comprehension and individual lexicon, Ufer et al. 2013), which particularly applies for conceptual understanding. The authors interpret this result as an indication that "language deficits [can] negatively impact the learning gains in subject matters [like mathematics] in a cumulative way" (Herwartz-Emden, 2003, p. 692).

Due to the fact that the epistemic function cannot only be considered in terms of reading proficiency (Duarte et al., 2011, p. 39), we conclude that language proficiency must be conceptualized to a wider extent to include lexical-semantical (related to the lexicon and its meanings) as well as grammatical skills in language receptive and language productive skills, which are tightly connected.

In the last 30 years, research in linguistics and language acquisition was dedicated to providing theoretical foundations and operationalizations of how complex combinations of the above-mentioned skills can be assessed in a simple way. The C-Test (Grotjahn, 1992) provides an example of an operationalization by using a cloze test which is constructed from texts systematically. Even though such an operationalization does not provide a deep insight into all complexities, it considers the lexical-semantical and grammatical features of school academic language when based on texts in school academic language (Daller, 1999, for information in Redder & Weinert, 2013). The C-Test enables the assessment of language proficiency in a linguistically acknowledged way and takes into account those skills which are crucial for mathematics learning. That is why this study uses the C-Test for roughly measuring language proficiency. Then language-induced obstacles are identified with means of the item analysis as well as a video-based analysis of students' processing procedure.

2.3 Language-induced obstacles in literacy-based tests

According to previous empirical findings, the connection between reading / language proficiency and mathematics achievement should be detectable especially using tests which do not assess context-free calculation skills, but – within the PISA-framework of *mathematical literacy* (OECD, 2007; Neubrand, 2001) – the mindful and flexible application of mathematic, by emphasizing strategies for solving inner-mathematical and everyday problems and conceptual understanding. Büchter and Pallack (2012, p. 63) consider the central exams in Grade 10 in North Rhine Westfalia to belong to such literacy-based tests, which is the focus of the current study. Similar to the study of Brown (2005), which shows the existence of specific obstacles for students with low LP in context-based and text-based tests in the US, this study investigates the existence of similar obstacles for students with low language proficiency in the German central exam.

With respect to *mathematical literacy*, Kaiser and Schwarz (2003) claim not to consider reading obstacles merely as biases endangering the validity, because decoding and sense making from texts is a crucial element of a mindful application of mathematics. Hence it might be beneficial (also in light of the current curriculum) to identify the reading obstacles and other obstacles for students with low LP. In future research, this could evoke the design of learning opportunities for supporting students with low LP to overcome such obstacles for improving mathematical literacy.

2.4 Research questions

Based on the current state of research, the connection between mathematics achievement and language proficiency is to be investigated in comparison to other family background factors. The study focuses on the high stakes test "Zentrale Prüfungen ZP10-Mathematik auf dem Niveau des Mittleren Schulabschlusses" in North Rhine-Westphalia, i.e. the central exam in grade 10 for the medium track at the end of their compulsory schooling. Different obstacles are identified which can be traced back to the communicative and epistemic function of language. The study pursues the following research questions:

Q1. Which social and language background factors have the strongest connection to mathematics achievement in the high stakes test ZP10?

Q2. Which items do students find difficult to accomplish? Which items are difficult for students with low LP?

Q3. Which obstacles can be reconstructed in items which are difficult for students, which one of these items is relatively more difficult for students with low LP?

Research question Q2 is an auxiliary question serving for investigating research question Q3.

3. Research design and methods

In a mixed methods design, the test scores from the high stakes test ZP10 were analyzed in relation to the background factors. These statistical analyses were complemented by a DIF analysis (for detecting differential item functioning) for the most influential factor, language proficiency, and finally by an investigation of written student solutions and observations from students' videotaped working processes. A pragmatic choice of methods supported a deeper explanation of statistical effects.

3.1 Methods and measures for the quantitative study

In this present study, mathematics achievement is treated as a dependent variable, capturing institutional success in a literacy-based test. The data analysis is based on the evaluations of the teachers in the ZP10 (written in 2012 for the medium track "Mittlerer Schulabschluss"). The particular data corpus was collected under normal field conditions of a high stakes exams. This implies a high extrinsic motivation of the students (reported by many teachers) but also a missing empirical control of interrater reliability. However, reliability was controlled in kindred studies of the same high stakes test (Büchter & Pallack, 2012).

The teachers evaluated each task by scores from 1 to 5 and filled an evaluation form for determining the exam grades. For this present study, one tasks with different demands were split into two items, so that the test consisted of 27 items. The researchers dichotomized the evaluation scheme and decided which score to expect for the item to be "mainly successfully solved" (cf. Büchter & Pallack, 2012 for legitimizing the criteria of dichotomizing) based on didactical considerations. The *item score of a person* was operationalized as the number of mainly successfully solved items.

This foundation allowed to scale the test data in a one-dimensional IRT scale and use the WLE (weighted likelihood estimates) of each person as metrically scaled measures for the mathematics achievement in the further analysis (cf. Section 1.3).

The independent variables were chosen with respect to the state of research (outlined in Section 2.1): A first set of variables comprised social background factors, captured by a self-report questionnaire before the exam. Besides age and gender, the questionnaire addressed the following variables (all captured on a three-step ordinal scale):

- *immigrant status,* operationalized by the countries of birth of the students and their parents (as usual e.g. in PISA in OECD, 2007); three-step ordinal scale: first generation – second generation – third generation or no immigrant status).
- *socio-economic status* (SES), operationalized by the visualized book-at-home-index (Paulus, 2009 showed retest-reliabilty of r = .80; also used in TIMSS by Schnabel & Schwippert, 2000, p. 269); here summarized in a three-step ordinal scale of low – medium – high).
- *age of first exposure to German language,* as an operationalization of family languages and language acquisition type (De Houwer, 2009); three-step ordinal scale: only German in the family – German and another language learned in the family before Kindergarten – German learned in Kindergarten or later)

For the second set of independent variables, the students' language competencies were conceptualized in two ways. Both tests measure related but not identical constructs a metrical scale properties was pragmatically assumed for both tests:

- *reading proficiency,* operationalized by 14 items on reading understanding in the parallel German exam (medium track). The test has a limited internal consistency (α = 0.54 in a sample of n = 1066). One reason for this is that the development of central exams must focus on a wide spectrum of demands rather than a homogeneous construct in a test-theoretical perspective. However, the items were considered for first approximations, operationalizing the reading proficiency score by the number of correctly solved items. Due to these statistical and content-related reasons, the more detailed analysis focused on the more reliable C-Test.
- *German language proficiency* was operationalized by a C-Test which is often chosen for a time-economic and standardized assessment of a complex construct of language proficiency without reduction to isolated sub-skills and sufficient reliability (Grotjahn, 1992). The administered C-Test (Baur & Spettmann, 2010) consisted of five texts in a high demanding school academic language. The text with mathematically relevant contexts had similar difficulties as well as connection measures compared to the texts without mathematically relevant contexts. Across the five texts, the C-Test had a good internal consistency (α = 0.86 for a sample of n = 698). The language proficiency score used for the regression and covariance analyses was operationalized by the number of clozes filled in correctly in the five texts. Based on this score, three equally sized groups were formed (abbreviated thirds with low – medium – high language proficiency), and two equally sized groups (median split for half with low – high language proficiency) were formed for an easier interpretation of the DIF-analysis.

3.2 Sampling

The empirical investigation is based on a sample of 1495 students in Grade 10 of 19 comprehensive schools and 67 mathematics courses of the medium track (aiming at the formal exam for the medium track "Mittlerer Schulabschluss, cf. Table 1).

Table 1 Overview on descriptive data of the whole sample and the subsamples

Variable	Grouping	Distribution
Whole sample	19 comprehensive schools, 67 mathematics courses	N = 1495
Age (n=1489)	17 years and older 16 years 15 years	311 (21 %) 984 (66 %) 194 (13 %)
Gender (n=1487)	Male Female	774 (52 %) 713 (48 %)
Immigrant status (n=1480)	1st generation (student immigrated) 2nd generation (parents immigrated) no / 3rd generation	152 (10 %) 623 (42 %) 705 (48 %)
Socio-economic	Low SES ("no" or "very few" books at home)	509 (34 %)

status	Medium SES ("enough books for one board")	488 (33 %)
(n=1493)	High SES ("three boards" or "complete wall of books"	496 (33 %)
Age of first exposure	Multilingual, German in Kindergarten	289 (19 %)
to German language	Multilingual, German before Kindergarten	538 (36 %)
(n=1486)	Monolingual, only German	659 (44 %)
Subsamples		
Reading Proficiency Test	Only students in German course on medium level	n = 1066
Language proficiency C-Test	Representative sample of schools	n = 698

This sample is representative for comprehensive schools in the metropole region of Ruhrgebiet with respect to social and family factors but also to achievement in the central exams. The representativeness of the test scores was checked by specialists in the Ministry of Education of North Rhine-Westphalis based on the internal data. Gymnasien (the schools for the higher track) were not included in the sample as they do not participate in the high stakes exam.

1066 out of the 1495 students of the whole sample participated also in the German Course on the medium track (a.k.a. Deutsch-Erweiterungskurs) and sat for the final exam on the medium track. The other 429 students participated in German courses on lower tracks and hence sat for another exam ("Hauptschulabschluss Klasse 10"). As the reading proficiency items in the lower tracked exam are different than in the medium tracked exam, the data is not comparable on a joint scale.

Due to the above reasons, reading proficiency can only be considered for the larger subsample of students in the medium tracked German courses: on the basis of its composition, it has slightly higher scores in the mathematics exam than the whole sample (M = 11.5; SD = 4.5 compared to M = 10.9; SD = 4.7 in the whole sample, cf. Table 2 below). The C-Test was written by 698 students, as only some schools agreed to dedicate additional time for these tests. The mathematics test scores of this subsample (M = 11.1; SD = 4.7, cf. Table 2) do not significantly differ from the mathematics test scores of the whole sample.

3.3 Modeling mathematics achievement and statistical procedures for data analysis

The raw scores in the mathematics test ZP10 was scaled in a one-dimensional dichotomic Rasch-Model (Rost, 2004, p. 115ff.). This kind of scaling was proven as adequate in previous studies for modelling raw scores from high stakes tests (Büchter and Pallack, 2012). By using the Rasch-Model, the achievement data of students and the item difficulties can be captured on a common metric scale with the Weighted Likelihood Estimated (WLE) as estimates for the person parameters (being standardized for an average item difficulty of 0). For each item, the item difficulty is measured on the common Rasch scale together with the WLE for the individual students and represented as logit.

The characteristic values of the Rasch scale show that the one-dimensional dichotomic Rasch model is adequate for the dichotomized data. For all items, the Item-Fits Weighted Mean Square (MNSQ: Mean square) range from 0.93 to 1.12, thus all values are within the interval [0.80; 1.20] considered as acceptable range for these measures in the PISA-study (cf. OECD, 2009, p. 355). The WLE reliability is satisfactory with 0.79. The item difficulties vary between -2.50 for the easiest item and 2.70 for the most difficult item (with 0 for the average item difficulty).

Besides the data from the Rasch scale, this article will also take into consideration the raw scores as they provide the statistics with the higher relevance for grades and graduating, hence they have direct relevance in the context of the high stakes test.

For analyzing the connections between background factors and mathematics achievement (research question Q1), analyses of variance, regression and covariance were conducted. For this purpose, the Rasch-scaled data were considered as dependent variable and the factors of family background, reading and language proficiency as independent variables.

- Firstly, separate models were determined for finding isolated effects of the separately treated independent variables and the respective percentages of explained variance (η^2 or R^2, resp.) were compared. For the categorically or ordinally scaled background factors, a one-factor analysis of variance (ANOVA) was applied, and linear regressions for the data on reading and language proficiency.
- Secondly, an analysis of covariance (ANCOVA) was conducted for the statistically adequate and content-valid independent variables (*SES*, age of first exposure to German and language proficiency, see above).

The second research question (Q2) focuses on the *absolute and relative item difficulties*, which were investigated by DIF analyses (Differential Item Functioning), which detected that items of the ZP10 which were "statistically unexpectedly difficult" (in the frame of the one-dimensional dichotomic Rasch model) for the half of the students with low language proficiency (median split due to C-Test, in the following abbreviated by low-LP-half). For this group, the DIF analysis determines the expected frequency of solution for each item by relating the medium WLE of the group to the difficulty of each item. These theoretically expectable item difficulties are then compared to the observable item difficulties of the group, captured by the so-called DIF-value of the item (similarly done by Abedi, 2006; Haag et al., 2013).

3.4 Methods for analyzing the individual solution processes

In order to specify the language-induced obstacles which made some items difficult or relatively difficult for students with low language proficiency (in brief: low LP) (research question Q3), the items identified as relatively difficult were further investigated in a subsample of 195 written tests (representative with respect to achievement and background factors). For this purpose, the students' written solution pathways were coded with respect to steps in the solution process of each item. This was based on item-specific coding-schemes which captured the relevant steps of the solution process. For contrasting the halves of students with high and low language proficiency, dropout rates were determined to capture the diverse mastering of obstacles. The analysis of the dropout rates allowed to determine the obstacles for the different language groups.

Furthermore, 47 solution processes of one or two students each were videorecorded in clinical interview settings. They were videotaped and qualitatively analyzed with respect to the appearing obstacles (in sum, 47 x 30-45 min. video material). The results of these qualitative analyses are presented in detail in other publications (Prediger et al., 2013; Gürsoy et al., 2013; Wilhelm, 2016). Here, only selected results are cited for backing up some aspects and for explaining them.

By comparing the different obstacles in the written solutions and videotaped solution processes, four categories of obstacles were developed:

(1) the category of *reading obstacles* includes obstacles in the steps of understanding the texts of the items;

(2) the category of *processual obstacles* denotes obstacles appearing in the cognitively demanding steps of the processes,

(3) the category of *conceptual obstacles* comprises obstacles which are identified in steps of the solution process which demand conceptual understanding (e.g. basic mental models of mathematical concepts);

(4) the category of *calculatory obstacles* includes obstacles identified only in steps of the inner-mathematical calculation or treatment.

Usually, processual and caclulatory obstacles emerge in later steps of the solution process than pure reading. In contrast, conceptual obstacles are often connected to reading obstacles because a pre-understanding of the mathematical structure is necessary for decoding of the text. Although calculatory obstacles appeared, they were not more frequent among students with low LP than among students with high LP, thus these obstacles were not considered in this present study.

4. Results of the analyses

4.1 Connections between background factors and mathematics achievement

Table 2 shows group differences in mathematics achievement. The groups were formed by the different above-mentioned background factors as well as the above-mentioned subsamples. Mathematics achievement was assessed by the average WLE of the groups as well as by the raw scores with their immediate relevance for grades. In order to ensure comparability, three groups were formed if possible in a meaningful way. In the medium achievement levels, differences in WLE of 0.2 were interpreted as approximately one solved item more or less. For the raw scores, a score difference of 11-12 correspond to one grade level: the grade "sufficient" was assigned to a score of 38-49. The first column of the table list the considered background factors, the second column shows the constructed groups for which the distribution is given in the third column.

The results show a highly significant difference between the strong and the weak group (shown in a Post hoc Scheffé test with $p < 0.001$ after a one-factorial ANOVA with significant F-test) for each factor under consideration of mathematics achievement.

Table 2 Group differences in mathematics achievement for groups with regard to different background factors (two or three groups for each factor)

Background Factors	Established groups	Distribution of all students to the groups	Math Achievement by average raw score in M-Test (max. 85) m (SD)	Math Achievement by Average WLE m (SD)
Whole sample	19 school	N=1495	43.5 (13.6)	-0.62 (1.07)
Gender	male	774 (52.1 %)	45.3 (14.0)	-0.44 (1.09)
(n=1487)	female	713 (47.9 %)	41.3 (12.8)	-0.81 (1.02)
Immigrant	1st generation	152 (10.3 %)	41.3 (13.6)	-0.77 (1.13)
status	2nd generation	623 (42.1 %)	40.9 (13.5)	-0.81 (1.06)
(n=1480)	no / from 3rd generation	705 (47.6 %)	46.2 (13.0)	-0.41 (1.03)
Socio economic	low SES	509 (34.1 %)	41.9 (14.0)	-0.74 (1.13)
status	medium SES	488 (32.7 %)	42.9 (12.9)	-0.67 (1.04)
(n=1493)	high SES	496 (33.2 %)	45.7 (13.4)	-0.43 (1.01)
Moment of	Multilingual, German after 3	289 (19.4 %)	39.5 (13.7)	-0.91 (1.09)
German acquisition	Multilingual, German before 3	538 (36.2 %)	42.2 (13.5)	-0.71 (1.07)
(n=1486)	Monolingual, only German	659 (44.3 %)	46.3 (13.0)	-0.40 (1.02)
Language	low LP	235 (33.7 %)	37.3 (13.4)	-1.04 (1.08)
proficiency	medium LP	233 (33.4 %)	44.2 (12.6)	-0.53 (1.01)
(C-Test,	high LP	230 (33.0 %)	50.3 (11.4)	-0.11 (0.94)
n=698)	*Whole subsample C-Test*	*698 (100 %)*	*43.9 (13.6)*	*-0.56 (1.08)*
Reading proficiency	low RP	365 (34.2 %)	40.3 (12.9)	-0.85 (1.02)
(German-Course on	medium RP	405 (38.0 %)	46.6 (12.6)	-0.39 (0.95)
medium level,	high RP	296 (27.8 %)	50.0 (12.5)	-0.14 (0.99)
n=1066)	*Whole subsample reading*	*1066 (100 %)*	*45.4 (13.3)*	*-0.48 (1.02)*

However, the group differences vary, as the 4th and 5th column of Table 2 show: The differences in social factors score less than 7, hence half a grade level, gender differences score 4, the immigrant status 5.3, SES 3.8 and the age of first exposure to German 6.8. In contrast, the score differences for the language factors are higher: the (slightly more homogeneous) reading subsample shows scores differences of 9.7. The subsample of C test participants demonstrates the largest score difference of 13 between the thirds with low and high LP. The average of the third with low LP reaches the average grade between "sufficient" and "fail", whereas the third with high LP obtains a "satisfactory". Thus, the score difference corresponds to more than one grade level.

In a more systematic statistical way, the different effects of social and language background factors are captured by the analysis of variance and the regression analysis showing the percentage of explained variance (Table 3). Whereas the social factors explain between 1 % and 3 % of variance, the language factors have substantially more explaining potential with 10 % and 14 %.

Table 3 Percentages of explained variance (for mathematics achievement, WLE) for different background factors in the analysis of variance and regression analysis

Background factor	Procedure	Explained variance	df	F	Significance
Socio economic status (n = 1493)	ANOVA	$\eta^2 = 0.02$	2	11.44	$p < 0.01$
Immigrant status (n = 1480)	ANOVA	$\eta^2 = 0.03$	2	25.64	$p < 0.01$
Age of first exposure to German (n = 1486)	ANOVA	$\eta^2 = 0.03$	2	26.24	$p < 0.01$
Reading proficiency* (n = 1066)	linear regression	$R^2 = 0.10$	1	111.08	$p < 0.01$
Language proficiency (C-Test) (n = 698)	linear regression	$R^2 = 0.14$	1	124.44	$p < 0.01$

* This test has a critical internal consistency of $\alpha = 0.54$ (see above); It was integrated into the table for nevertheless allowing heuristical comparisons.

The isolated consideration of independent variables shows that the language proficiency explains a higher percentage of the variance of WLE than the reading proficiency (whose reliability was critical). The observation that language proficiency is presumably more mighty for explaining variance is plausible, as the construct differences also takes into account language skills beyond reading skills which might explain the different achievements (cf. Section 4.2). That is why the further analyses rely on the factor language proficiency.

For the social factors, the immigrant status and the age of first exposure to German explain equally sized percentages of variance. Both factors are closely connected, but the age of first exposure to German seems to have a stronger interpretation for the connection to mathematics achievement. Hence, the analysis of covariance (in Table 4) only takes into account the age of first exposure to German when determining the influence of the social factors under consideration of the covariate of language proficiency.

The analysis of covariance shows that the socio-economic status has no separate contribution to account for the variance of WLE when the language proficiency is controlled ($F(2, 682) = 1.38$; $p = 0.25$). Considering sociolinguist findings, this result can be interpreted in a way that the impact of the SES is mainly mediated by differences in LP. In contrast, the age of first exposure to German shows a significant effect on the WLE, even after controlling of LP ($F(2, 682) = 9.29$, $p < 0.01$). One possible interpretation (which must be further studied) for this finding might be in the epistemic function of language for mathematics learning: If students have the same LP at the time of testing (Grade 10) but have started to learn the language of instruction afterwards, then the lower LP in earlier grades might have led to an insufficient or partial acquisition of mathematical competencies.

Table 4 Analysis of covariance
(Mathematics achievement by WLE with the covariate language proficiency; n = 692)

Source of Variance	Sum of squares	df	F	Signifikance	partial η^2
Corrected model	133.79	9	15.02	$p < 0.01$	0.17
Language proficiency (C-Test)	49.02	1	50.12	$p < 0.01$	0.07
Socio economic status (SES)	2.70	2	1.38	$p = 0.25$	0.00
Age of first exposure to German	18.16	2	9.29	$p < 0.01$	0.03
SES * Age of first exposure to German	0.78	4	0.20	$p = 0.94$	0.00
Error	666.98	682			
Corrected total variation	800.77	691			

Summing up, the first statistical analysis shows that language proficiency has the strongest connection to mathematics achievement among all social and language factors by the highest percentage of explained variance. This finding resonates with the international research results (cf. Section 2.2) and justifies why all further analyses focus on the language proficiency as operationalized by C tests.

4.2 Analysis of relative and absolute difficulties of items

For treating research question Q2, Table 5 lists all relative and absolute difficulties of items, each in two representations: The frequency of solutions in the whole sample shows which items are difficult for all students, the frequencies of solution of the half of students with low or high LP show the group differences and can be related to the total scores in Table 2. The next columns show item difficulties on the Rasch scale and the DIF-values as the main measure for relative difficulties for students with low LP, which are explained in more detail below. For ensuring best possible interpretability, two same-sized groups were formed for the DIF analysis: one low-LP-half and a high-LP-half. This means that both halves have DIF-values of the same absolute value, the minus or plus indicating the direction of difficulty shift.

As expected, the general tendency is that the third of students with low LP reaches lower frequencies of solution than the third of students with high LP. The DIF analysis enables the identification of those items which are, beyond that, “statistically unexpectedly difficult” for the lower half of students (in relation to the expectable item difficulty, this is abbreviated by “relatively more difficult”). In the last column of Table 5, the DIF-value of each item is given, i.e. the group-specific shifts of the item difficulties (observed item difficulty in relation to the expectable item difficulty) for the low-LP-half of students. The Rasch scale captures the item difficulties and WLE on a metric joint scale. For example, for Item 1a, one can find out that the item was more difficult by 0.177 units on the Rasch scale for the low-LP-half than for the whole sample (and more easier y 0.177 units on the Rasch scale for the high-LP-half).

Table 5 Overview on absolute and relative difficulties of items

Items	Brief description of item content	Frequency of solution in the whole sample	in the halves with low | high LP	Item difficulty (in Rasch modell)	DIF-value (for half with low LP)
1 Basic skills					
1a	Estimating the tower of coins (Fermi)	31 %	23 % | 44 %	0.311	0.177*
1b	Growth of bacteria (finding alg. expressions)	46 %	38 % | 49 %	-0.424	-0.087
1c	Party hat (calculating cones)	24 %	23 % | 34 %	0.720	-0.023
1d11	Tables (calculating the length of a rectangle)	74 %	72 % | 75 %	-1.841	-0.284*
1d12	Tables (calculating the circumference)	84 %	85 % | 83 %	-2.539	-0.444*
1d2	Spread sheet (finding formula)	12 %	8 % | 13 %	1.716	-0.063
1e	Internet use (reading off tables)	69 %	64 % | 73 %	-1.558	-0.138
2 Fuel consumption (Quadratic function)					
2a1	Reading off an x-value for a given y-value	68 %	66 % | 70 %	-1.508	-0.246*
2a2	Percentage comparison of two values	12 %	9 % | 17 %	1.646	0.071
2b1	Proportional consumption (division)	68 %	62 % | 78 %	-1.526	0.083
2b2	Calculating the liters (multiplication)	33 %	28 % | 43 %	0.207	0.034
2c1	Finding y in the quadratic equation	50 %	46 % | 55 %	-0.612	-0.146
2c2	Finding x in the quadratic equation	11 %	10 % | 13 %	1.804	-0.146
3 Octopus Paul (Probabilities)					
3a	Simulation	49 %	37 % | 60 %	-0.558	0.209*
3b	Two step tree diagrams	59 %	56 % | 76 %	-1.063	0.178*
3c	Two step probability	68 %	64 % | 82 %	-1.486	0.158
3d	Determining the complementary probability	17 %	14 % | 25 %	1.276	0.069
3e1	Probability for more steps	53 %	45 % | 64 %	-0.736	0.107
3e2	Drawing the values in a graph	37 %	29 % | 45 %	0.030	0.079
3e3	Why the graph must not be continuous	5 %	4 % | 11 %	2.701	0.261
3f	Formula for W(n)	9 %	5 % | 11 %	1.997	0.180
4 Stairs (Algebra)					
4a1	2.60 : 14 (approaching the context)	39 %	36 % | 52 %	-0.068	0.029
4a2	Calculating the slope in %	23 %	20 % | 27 %	0.841	-0.083
4a3	Calculating the angle of the slope (Tangent)	34 %	32 % | 41 %	0.161	-0.118
4b1	Checking a rule of thumb	52 %	46 % | 55 %	-0.701	-0.132
4b2	Determining values for the rule of thumb	51 %	41 % | 59 %	-0.685	0.074
4b3	Find values by trial and error	10 %	10 % | 10 %	1.897	0.204

A minus sign in a DIF-value means that the item difficulty decreases, thus this item is relatively easy for the subsample compared to the expectable difficulties. By means of the non-printed standard error, the shifts of item difficulties were tested for statistical significance[1]. Those items whose DIF-value is significant on the 5 %-level are marked with by *.

Three items were identified to be relatively easy for the students with low LP (1d11, 1d12 Spread sheet and 2a1 Reading off fuel consumption in a diagram). Only three items (1a Estimating the tower of coins, 3a and 3b Octopus Paul) were identified as statistically relatively difficult. One more item (2a2 Percentage comparison of two values for fuel

[1] The significance of the DIF-value depends on the difference on the metric Rasch scale, but also from the item difficulty: Very easy and very difficult items require larger shifts for reaching significant results than items of medium difficulty.

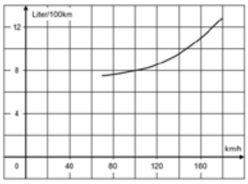
consumption) did not reveal a significant DIF-value, but it induced language obstacles for all students, which made it particularly interesting for investigating language-induced obstacles.

Item 1a: Estimating the tower of coins (relatively more difficult for students with low LP)
Estimate how many kilometers high would a tower of 2.4 billion 1-cent-coins be. Give an explanation for your solution.

Items 1d12 und 1d12: Spread Sheet
(relatively easier for students with low LP)
Malak explores a rectangle with an area of 144 cm^2 by means of a spreadsheet. He calculates the circumference for different lengths of the sides a and b.
(1) Calculate the missing values for the cells B7 and C10.
(2) Find a formula for each of the cells B5 and C5.

	A	B	C
1	Gegebener Flächeninhalt des Rechtecks: 144 cm²		
2			
3			
4	Seite *a* [in cm]	Seite *b* [in cm]	Umfang [in cm]
5	1,5	96,0	195,0
6	3,0	48,0	102,0
7	4,5		73,0
8	6,0	24,0	60,0
9	7,5	19,2	53,4
10	9,0	16,0	
11	10,5	13,7	48,4
12	12,0	12,0	48,0
13	13,5	10,7	48,3

Item 2a: Fuel consumption (a1 relatively easier for students with low LP; a2 difficult for everybody)
The fuel consumption for vehicles is specified by the consumption in liters (l) for a distance of 100 km. The fuel consumption of a car depends on the speed. The diagram shows the fuel consumption for a car that drives in the highest gear. That is why the graph starts at 70 km/h.



(1) What speed does the car have on average, when it consumes 11 l for 100 km?
(2) How much (in percent) lies the consumption for 180 km/h over the consumption for 100 km/h? Write down your calculation. [literal translation from German]

Items 3a und 3b: Octopus Paul (relatively more difficult for students with low LP)
During the World Championship of Soccer in 2010, octopus Paul became famous all over the world. Before every game […], Paul chose one of its feeding dishes. The media interpreted its choice as a "prediction" for the winner of the game. […] Mathematically, the "predictions" are random experiments with two equally probable outcomes.
a) Explain how this random experiment can be simulated by means of a dice.
b) Draw a tree diagram that determines the probability for two predictions.

Figure 1 Items identified as relatively easy or difficult for students with low language proficiency in the central exam ZP10

These seven identified items are printed in Figure 1. The relatively easier items (which had a negative DIF-value for students with low LP) are only treated briefly here, the relatively more difficult are discussed in more detail in the next section.

Item 1d11 and *1d12* demand simple knowledge about spreadsheets, which all students seem to have acquired in their classes (notwithstanding their language proficiency). Thus, they are not more difficult for students with low LP than for students with high LP. For *Item 2a1* (Reading off fuel consumption in a diagram), an analysis of corresponding videotaped student processes showed that most students with a superficial standard strategy can read the value without having understood the functional relationship. This provided a hypothesis for explaining the relative low difficulty for students with low LP.

4.3 Reconstructed obstacles in the identified items

The analysis of 195 written tests and 47 videotaped solution processes for the items identified as difficult or relatively difficult (Items 1a, 2a2, 3a, and 3b) allowed the researchers to specify the repetitive obstacles which could be classified in four types (cf. Section 3.4): reading obstacles, processual obstacles, conceptual obstacles, and calculatory obstacles. The former ones were not only encountered by students with low LP, hence they are not considered as language-induced. The other three obstacles will be illustrated exemplarily below.

Reading obstacles on the sentence level – exemplified for Item 2a2 (Fuel consumption)

Item 2a2 (Percentage comparison of two values for fuel consumption, represented in Figure 1) was a difficult item for all students (cf. Table 5). It was solved by 12 % of all students, only 9 % of the half with low LP.

An analysis of the videotaped solution processes revealed repetitive reading obstacles on a sentence level: It is not the single unknown word which led to the obstacle, but the complex sentence structure created by nested prepositional phrases ("How much (in percent) lies the consumption for 180 km/h over the consumption for 100 km/h?" literally translated from German). It substitutes several main clauses and condenses the complex relations into a short phrase.

This reading obstacle becomes evident in the transcript of Berna's solution process, a 16 year-old girl with Turkish as her family language who belongs – according to her C test – to the lower third of students with low LP:

5	B	[*reads the problem silently to herself, 14 sec*] Well, here, this is, I believe, then the problem, that we shall then find, how many percent are, eh, 180 of 100 kilometers.
…		
19	B	[*2 sec break*] Well, the – we have this fuel consumption, eh 100 km/h.
20	I	Mhm.
21	B	Ok. And when someone drives his car, then he drives 100- eh 180 km/h. And we shall find out, how much in percent this lies over the normal mileage.
...		[*evaluates the percent formular by 180 and 100*]
33	B	[*3 sec break*] Eh, the fuel consumption is at 55 percent [*2 sec break*] over the consumption at 100 km/h [*laughs*] Ehm. [*4 sec break*] I would not have an answer, now.

Berna simplifies the question

How much (in %) lies the consumption for 180 km/h over the consumption for 100 km/h?
to
How many percent is 180 of 100?
(in turn 21) and calculates (in turn 33)
How many percent is 100 of 180?

Berna does not recognize the shortened prepositional phrase "consumption for 100 km/h" as the "functional value of the consumption function for the speed of 100", but associates the 100 with the consumption itself (turns 19, 21). Although she repeats the phrase "How much lies this over" (in turn 21), she calculates a simple proportion (between turn 21 and 33). By this, she has not correctly identified both relevant relations in the sentence, hence her underlying mathematizing process is not successful. She realizes that she cannot interpret her result and articulates this by the phrase, “I would not have an answer, now.” (turn 33). In contrast, students with high LP focus on the relations which enables them to find an adequate mathematization (cf. Gürsoy et al., 2013 and Wilhelm, 2016 for further analyses).

In order to examine how far Berna’s difficulties for overcoming the reading obstacles reoccur for other students, 195 written tests were coded with respect to mastering the necessary steps of the solution process. The contrast of the dropout rates for each solution step was interpreted as following: A step was coded as not mastered if it is not completed successfully or left out. The dropout rate was operationalized as conditioned relative frequency: for each step, the dropout rate is related to the whole number? of those tests in which the preceding steps would have allowed to master it. This operationalization of the dropout rate allows to draw conclusions about the relevant obstacles.

Like Berna, 79 % of the low-LP-half of students and 63 % of the students with the high-LP-half did not recognize that the phrase “consumption for 100 km/h” demands the identification of the y-values in the diagram before continuing the solution process. Then, both values must be included in the calculation, this produces a further drop put of 64 % or 38 %, resp. (many students use non-appropriate proportional strategies instead).

Table 6 Overview on dropout rates for the solution steps in Item 2a2 (Fuel consumption) for students halves with low and high language proficiency

	Dropout rate in the mutual solution step for …	
Solution process step	... half of students with low language proficiency	... half of students with high language proficiency
Reading off the value for consumption in diagram	79 %	63 %
Transforming units (km/h, 1/100km)	31 %	25 %
Using both y-values in the calculation	64 %	38 %
Translating into a calculation	79 %	70 %
Interpreting the result	84 %	78 %

Those students who did not recognize the (typical German) phrase “*How much (in %) lies … over …*”, could not translate the data in an adequate calculation for the percentage comparison: 79 % or 70 %, resp., of the remaining students failed in this step. This step does not only entail reading obstacles, such as for Berna, but also conceptual obstacles, i.e. the conceptual understanding of a percentage comparison in contrast to determining the share. Besides the calculatory step of “transforming units”, which has low dropout rates in both student groups (31 % and 25 %), all other steps can involve reading obstacles (what is it about?) as well as conceptual obstacles (e.g., in functional relationships, two quantities must be connected). This cannot be distinguished in the students’ written solutions. In general, few differences could be identified between the two groups.

Hence, even if reading obstacles and conceptual obstacles are potentially closely linked in Item 2a2, this item still provides a good example for reading obstacles due to its syntactical complexity and the high relevance of prepositional phrases carrying the relational structure. Moreover, the high density of syntactic complexity in other items can result in reading obstacles which are difficult to overcome by students. The high density may be the cause why the feature of long sentences or long texts (which is often pointed out by teachers) is not the main source for difficulties for this age group. The relevance of prepositional constructions for capturing complex relations seems to be typical for mathematical texts, as emphasized by Jorgensen: “It is difficult to think of teaching mathematics without the use of prepositions” (Jorgensen, 2011, S. 324). Also for German, this is a typical difficulty with a high linguistic relevance (Grießhaber, 1999), especially for mathematics.

Processual obstacles - exemplified for the Item 1a (Estimating the tower of coins)

Item 1a had a solution frequency of 31 % in total, thus it was quite difficult for all students. However, this item was particularly more difficult for the low-LP-half of students, because only 23% of them managed to obtain solve it correctly (the DIF-value is exactly at the limit of significance of a 5%-level).

Nearly all students in the videotaped solution processes could paraphrase the text of Item 1a correctly (“Estimate, how many kilometers high a tower of 2.4 billion 1-cent-coins would be.”). This led to the conclusion that the items do not initiate reading obstacles. The second assumption was that the students with low LP might frequently fail in presenting their solution pathway. For this reason, the coding of 195 tests captured the presentation of the estimation separately.

Table 7 shows the different dropout rates for the half of students with low and high LP. Whereas “decoding the item text” had little dropouts in both groups, the second solution process step shows a larger gap: Whereas only 27 % of the high-LP-half began with a non-adequate estimation of the coin height, 56 % of the low-LP-half had problems in solving this item. The latter students estimated 8-12 mmm instead of 1-2 mm, which is an indication that they took the first length which came into their mind without constructing a complete situational model (cf. Reusser, 1989 for distinguishing decoding the text and construction a situation model). The videotaped solution processes confirmed that all students who were asked to draw the tower of coins (hence to make the situation model explicit) immediately revised their estimation for the height of the coin (Wilhelm, 2016). This observation allowed to exclude the assumption that the students explicitly intended an alternative construction of the tower on the thin sides of the coin.

The second assumption was that the written explanation, i.e. the language production, might have created further obstacles for students with low LP. This assumption holds for 36 % of the remaining students with low LP (and only 12 % of the students with high LP). Hence, it has a certain relevance but much less than the construction of the situation model. Thus, in this item, the communicative function of language does play a role, but much less than the epistemic function.

Table 7 Overview of the dropout rates for the solution steps in Item 1a (Tower of coins) for students with low and high language proficiency

	Dropout rate in the mutual solution step for …	
Solution process step	... half of students with low language proficiency	... half of students with low language proficiency
Decoding the item text	12 %	4 %
Estimating the heights of a coin	56 %	27 %
Explaining the estimation process	36 %	12 %
Choosing the operation for calculating	34 %	20 %
Multiplication with adequate place values	33 %	13 %
Transforming billions and units (mm to km)	76 %	62 %

The construction of situation models also appears in other items, such as Item 3b (Octopus Paul), as a processual obstacle which is difficult to overcome for students with low LP. The reason for the difficultly of this obstacle does not only emerge in the text, but it can be attributed to later modelling steps and their underlying cognitive processes. Other processual obstacles comprise the specification of a coherent sample space as a central step of modelling (Wilhelm, 2016).

Conceptual obstacles in Items 2a2 (fuel consumption) and Item 3b (Octopus Paul)

Item 2a2 ("*How much (in percent) lies the consumption for 180 km/h over the consumption for 100 km/h*?") did not only reveal reading obstacles, but also conceptual obstacles. They arose mainly within the solution process steps "Using both y-values in the calculation" (64 % compared to 38 % dropout in Table 6) and "Translating into a calculation". In the videotaped solution processes, students who could not activate conceptual understanding (with the adequate basic mental model) for percentage comparison and for structuring the relation could not overcome the step "Translating into a calculation" successfully. One such student is Berna who solves Item 2a2 (fuel consumption) by evaluating the percent formula with the amount 100 km/h and the base 180 km/h, resulting in 55 % (see above). As she cannot explain her way of thinking, the possibility of an alternative situation model can be excluded. In the same item, other students could not overcome the conceptual obstacle that gives meaning to the functional relationship between speed and fuel consumption, because they could not activate the basic mental model about functions linking quantities (cf. Wilhelm, 2016).

Similarly, *Items 3a and 3b* (Octopus Paul) were identified as relatively difficult for students with low LP in the DIF analysis, both items containing profound conceptual obstacles (cf. Wilhelm, 2016). This can be seen in Delia's solution process of Item 3b, a tenth grader who belongs to the third portion? with the lowest LP. Delia could not justify her decision for drawing the tree diagram (in Figure 2) conceptually:

3 D Yes, well, ehm. Because of two predictions. Well, this is just the one [*hints to the upper branch in the tree diagram*] and this is just one [*zhints tot he lower branch*]. [*2 sec breank*]
And, ehm, since there are two in the beginning, I would above 2 [*hints to the 2/6 on the upper branch*] and down simply – I do not know why I tool the 9, but – [1 sec break] because of the dice.

Similar to Delia, the analysis of videotaped solution processes revealed substantial deficits in the conceptual understanding of several students with low LP, in particular with respect to stochastic concepts, like multistep random experiments or simulation. However, the reformulation of the item text posed no problems for the same students (Wilhelm, 2016).

In total, the processual and conceptual obstacles turned out to be most focal, especially for explaining the relative difficulties of identified items. In contrast, reading obstacles played only a minor role in the items with significant DIF-value.

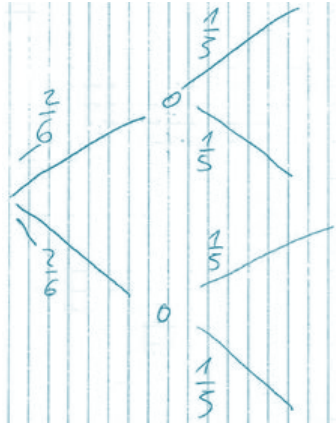


Figure 2 Delia's tree diagram

5. Discussion of results

This present study only addressed one specific exam, namely ZP 10 mathematics for the medium track in North Rhine-Westphalia in 2012. Therefore, the transferability of results to other assessments must be investigated in future research, especially for exams which focus less on the literacy-approach. However, at present the analyses reveal interesting findings to the leading research questions.

5.1 Connections between background factors and mathematics achievement

Q1. Which social and language background factors have the strongest connection to mathematics achievement in the high stakes test ZP10?

By analysis of variance and regression analysis as well as analysis of covariance, existing findings about social disparities can be differentiated: The analyses show that language proficiency has a stronger connection to mathematics achievement in the ZP10 than social factors (SES, immigrant status, age of first exposure to German). In our study, language proficiency was measured by C tests which comprise receptive, productive, lexical as well as grammatical components. These components turned out to be more relevant (cf. Table 2) than the reading proficiency (for which reliability was constrained) in this present study. Whereas the third of students with low LP (measured by C-test) reaches a mean score of 37.3 with a grade either "failed" or "sufficient", the third with highest LP reaches a mean score of 50.3 and a "satisfactory", hence more than one grade level difference.

Therefore, the high relevance of language proficiency for mathematics achievement, which was often outlined in US research (Abedi, 2006; Secada, 1992), also applies for the German language context. This finding is not only valid (as so far usual) for language biographic variables or reading proficiency, but also for a more comprehensive construct of language proficiency which is oriented to academic language. This finding suggests language proficiency should be included in all large-scale assessments and the government data on equity as it seems to mediate social disparities.

Nevertheless, social equity issues are still relevant, since according to many sociolinguistic findings, language proficiency is closely linked to learning opportunities in family., thus language proficiency is also a socially determined construct (Cook-Gumperz, 1973, p. 1). However, whereas the relevance of SES and immigrant status can mainly have global consequences on the policy level, the relevance of language proficiency initiates activities within classrooms, as it is a *didactically highly crucial starting point for reducing social disparities*. Since one can assume that all students can profit from a language-responsive mathematics classrooms, language-responsiveness should be a core of classroom innovation projects, even if such assumptions still need to be proven.

However, the limitations of the present study must be taken into account: The sample is only representative for the medium track and does not consider the higher and lower achieving students. The reliability of teachers' assessments, the very rough measure for SES by the book scale and the missing control of general cognitive abilities, form further constraints which have to be considered when interpreting the results. A future study should be administered outside the high stakes conditions in order to overcome these limitations. Moreover, a more refined operationalization of language proficiency should be desirable, even if some linguists doubt its feasibility.

5.2 Identification of language-induced obstacles

For more concrete consequences for classrooms, the analyses on an item level were necessary in order to specify the language-induced obstacles in a refined way. The statistical correlations cannot solely explain how language proficiency influences mathematics achievement or if the connection must be traced back to other common factors, such as general cognitive abilities which were not investigated in this present study.

In contrast, the analyses of written solutions and videotaped solution processes of striking items enabled the deeper understanding of the connection between language proficiency and achievement. Even if the identification and categorization of obstacles is not finalized by the current study and not yet generalizable to other types of exams, it provides insightful first categories which appear to be beneficial for further investigation in future studies: Three types of obstacles for students with low language proficiency were reconstructed:

- *reading obstacles* for decoding the item text, especially by complex sentence structures and morphological obstacles with a high relevance of prepositions (Gürsoy, 2013; Grießhaber, 1999). Reading obstacles do not only refer to language *biases* (Abedi, 2006) which should be eliminated from exams (even if the project has revealed some of these unnecessary biases, cf. Gürsoy et al., 2013). Some of these obstacles also belong to adequate reading demands for which the students are not yet well prepared; hence, suitable learning opportunities for overcoming reading difficulties must be developed and investigated.
- *processual obstacles* in cognitively demanding processes, e.g. when building a situation model or specifying one's own definitions (Wilhelm, 2016; similarly in Duarte et al., 2011) and

- *conceptual obstacles* in the conceptual understanding of the core mathematical concepts (Wilhelm, 2016; similarly in Ufer et al., 2013).

The outlined results from the in-depth analysis show that students with low language proficiency do not only encounter not reading obstacles in the test situations, thus belonging to the communicative function of language. Instead, a lot of such students fail during further steps of the solution process. Similar results were found in DIF analyses for the test VERA 3 for third graders (Haag et al., 2013).

These obstacles do not only refer to short-term problems in the test situation, but also to longer-term accumulation of deficits for overcoming processual and conceptual obstacles. They provide first indications for the cumulatively growing difficulties of students with low LP (Herwartz-Emden, 2003, p. 692), which seem to be traced back more to the epistemic function of language. An in-depth analysis of learning processes provides further hints for this hypothesis (Prediger, 2013; Zindel, 2015).

Especially the findings on processual and conceptual obstacles seem to be crucial for the academic field, which calls for the urgent necessity for investigating language limitations in learning processes. Further research in mathematics and language education on this practically highly relevant topic is required during the next years in order to delineate different obstacles, whether reading, processual, or conceptual, and understand their interplay. Such studies should be followed by design research studies for developing learning opportunities (cf. Prediger & Özdil, 2011 for the research and design needs and Prediger & Wessel, 2013 for a first realization).

Acknowledgment. The study MuM-ZP was financially supported by the Ministry of Education of the state North Rhine-Westphalia. The responsibility for the content of this article is solely within the authors. We would like to thank all involved schools, teachers and students for their participation. We would also like to thank the reviewers and the journal editor for their careful reading. Their comments have substantially contributed to develop the rigor and accessibility of this paper.

References

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

Autorengruppe Bildungsberichterstattung (2012). *Bildung in Deutschland. Ein indikatorengestützter Bericht mit einer Analyse zu Perspektiven des Bildungswesens im demografischen Wandel*. Bielefeld: Bertelsmann.

Baumert, J., & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb. In Deutsches PISA-Konsortium (Ed.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 323-410). Opladen: Leske + Budrich.

Baur, R. S., & Spettmann, M. (2010). Lesefertigkeiten testen und fördern. In C. Benholz, G. Kniffka, & E. Winters-Ohle (Eds.), *Fachliche und sprachliche Förderung von Schülern mit Migrationsgeschichte* (pp. 95-114). Münster: Waxmann.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Walther, G., & Valtin, R. (Eds.). (2003). *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich*. Münster: Waxmann.

Bos, W., Wendt, H., Ünlü, A., Valtin, R., Euen, B., Kasper, D., & Tarelli, I. (2012). Leistungsprofile von Viertklässlerinnen und Viertklässlern in Deutschland. In W. Bos, H. Wendt, O. Köller, & C. Selter (Eds.), *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 269-301). Münster: Waxmann.

Brown, C. L. (2005). Equity of Literacy Based Math Performance Assessments for English Language learners. *Bilingual Research Journal, 29*(2), 337-363.

Büchter, A., & Pallack, A. (2012). Methodische Überlegungen und empirische Analysen zur impliziten Standardsetzung durch zentrale Prüfungen. *Journal für Mathematik-Didaktik*, *33*(1), 59-85.

Burns, T., & Shadoian-Gersing, V. (2010). The importance of effective teacher education for diversity. In OECD (Ed.), *Educating Teachers for Diversity. Meeting the Challenge* (pp. 19-40). Paris: OECD.

Cook-Gumperz, J. (1973). *Social control and socialization: A study of class differences in the language of maternal control*. London: Routledge & Kegan Paul.

Cummins, J. (2000). *Language, power and pedagogy*. Clevedon: Multi lingual Matters.

Daller, H. (1999). *Migration und Mehrsprachigkeit*. Frankfurt et al.: Peter Lang.

De Houwer, A. (2009). *An Introduction to Bilingual Development*. Bristol: Multilingual Matters.

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. 35-54). Münster: Waxmann.

Ehmke, T., Hohensee, F., Heidemeier, H., & Prenzel, M. (2004). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, J. Rost, & U. Schiefele (Eds.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (pp. 225-253). Münster: Waxmann.

Ellerton, N., & Clarkson, P. (1996). Language Factors in Mathematics Teaching and Learning. In A. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International Handbook of Mathematics Education* (Vol. 4, S. 987-1033). Dordrecht: Kluwer.

Gellert, U. (2011). Mediale Mündlichkeit und Dekontextualisierung. Zur Bedeutung und Spezifik von Bildungssprache im Mathematikunterricht der Grundschule. In S. Prediger, & E. Özdil (Eds.), *Mathematiklernen unter Bedingungen der Mehrsprachigkeit* (pp. 97-116). Münster: Waxmann.

Gogolin, I. (2009). Zweisprachigkeit und die Entwicklung bildungssprachlicher Fähigkeiten. In I. Gogolin, & U. Neumann (Eds.), *Streitfall Zweisprachigkeit – The Bilingualism Controversy* (pp. 263-280). Wiesbaden: Verlag für Sozialwissenschaften.

Grießhaber, W. (1999). Präpositionen als relationierende Prozeduren. In A. Redder, & J. Rehbein (Eds.), *Grammatik und mentale Prozesse* (pp. 241-260). Tübingen: Stauffenburg.

Grotjahn, R. (1992). *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Band 1. Bochum: Broockmeyer.

Gürsoy, E. (2013). *Textkohäsion und Textkohärenz in den Zentralen Prüfungen 10 Mathematik (NRW) - Eine multiperspektivische Untersuchung zu Kohäsionsmitteln im Deutschen und Türkischen und zur Repräsentation von Kohäsion bei türkisch-deutschsprachigen Schülerinnen und Schülern*. Dissertation. Essen: Universität Duisburg-Essen.

Gürsoy, E., Benholz, C., Renk, N., Prediger, S., & Büchter, A. (2013). Erlös = Erlösung? – Sprachliche und konzeptuelle Hürden in Prüfungsaufgaben. *Deutsch als Zweitsprache, 13*(1), 14-24.

Haag, N., Hept, 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, 28*(0), 24-34.

Halliday, M.A.K. (1993). Some grammatical problems in scientific English. In M.A.K. Halliday & J.R. Martin (Eds.), *Writing science: Literacy and discursive power* (pp. 69-85). Pittsburgh: University Press.

Heinze, A., Herwartz-Emden, L., & Reiss, K. (2007). Mathematikkenntnisse und sprachliche Kompetenz bei Kindern mit Migrationshintergrund zu Beginn der Grundschulzeit. *Zeitschrift für Pädagogik*, *53*(4), 562-581.

Herwartz-Emden, L. (2003). Einwandererkinder im deutschen Bildungswesen. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer, & L. Trommer (Eds.), *Das Bildungswesen in der Bundesrepublik Deutschland. Strukturen und Entwicklungen im Überblick* (pp. 661-709). Reinbek: Rowohlt.

Jorgensen, R. (2011). Language, culture and learning mathematics: A Bourdieuian analysis of indigenous learning. In C. Wyatt-Smith, J. Elkins, & S. Gunn (Eds.), *Multiple perspectives on difficulties in learning literacy and numeracy* (pp. 315-329). Dordrecht: Springer.

Kaiser, G., & Schwarz, I. (2003). Mathematische Literalität unter einer kulturell-sprachlichen Perspektive. *Zeitschrift für Erziehungswissenschaft*, *6*(3), 357-377.

Knoche, N., & Lind, D. (2004). Bedingungsanalysen mathematischer Leistungen: Leistungen in den anderen Domänen, Interesse, Selbstkonzept und Computernutzung. In M. Neubrand (Ed.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland: Vertiefende Analysen im Rahmen von PISA 2000* (pp. 205-226). Wiesbaden: Verlag für Sozialwissenschaften.

Leutner, D., Klieme, E., Meyer, K., & Wirth, J. (2004). Problemlösen. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Eds.), *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (pp. 147-175). Münster: Waxmann.

Maier, H., & Schweiger, F. (1999). *Mathematik und Sprache. Zum Verstehen und Verwenden von Fachsprache im Unterricht*. Wien: oebv & hpt.

Martiniello, M. (2009). Linguistic complexity, schematic representations and differential item functioning for English language learners in math tests. *Educational assessment*, *14*(3-4), 160-179.

Mikrozensus (2011). *Bevölkerung nach Migrationsstatus regional.* Edited by Statistische Ämter des Bundes und der Länder. https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/ MigrationIntegration/ BevoelkerungMigrationsstatus5125203117004.pdf?_ blob=publicationFile. Last access May, 13, 2013.

Morek, M., & Heller, V. (2012). Bildungssprache – Kommunikative, epistemische, soziale und interaktive Aspekte ihres Gebrauchs. *Zeitschrift für angewandte Linguistik, 57*(1), 67-101.

Neubrand, M. (2001). PISA – „Mathematische Grundbildung" / „mathematical literacy" als Kern einer internationalen und nationalen Leistungsstudie. In G. Kaiser, N. Knoche, D. Lind, & W. Zillmer (Eds.), *Leistungsvergleiche im Mathematikunterricht. Ein Überblick über aktuelle Studien* (pp. 177-194). Hildesheim: Franzbecker.

OECD (2007). *PISA 2006. Science competencies for tomorrow's world*. Vol. 2: Data. Paris: OECD.

OECD (2009). *PISA 2006. Technical Report*. Paris: OECD.

Paulus, C. (2009). *Die „Bücheraufgabe" zur Bestimmung des kulturellen Kapitals bei Grundschülern*. http://psydok.sulb.uni-saarland.de/volltexte/2009/2368/pdf/BA_Artikel.pdf. Last Access May, 5, 2015.

Pimm, D. (1987). *Speaking mathematically. Communication in mathematics classrooms*. London, New York: Routledge.

Prediger, S. (2013). Darstellungen, Register und mentale Konstruktion von Bedeutungen und Beziehungen. In M. Becker-Mrotzek, K. Schramm, E. Thürmann, & H. J. Vollmer (Eds.), *Sprache im Fach – Sprachlichkeit und fachliches Lernen* (pp. 167-183). Münster: Waxmann.

Prediger, S., & Özdil, E. (Eds.). (2011). *Mathematiklernen unter Bedingungen der Mehrsprachigkeit – Stand und Perspektiven zu Forschung und Entwicklung*. Münster: Waxmann.

Prediger, S., & Wessel, L. (2013). Fostering second language learners' constructions of meanings for fractions – Design and effects of a language- and mathematics-integrated intervention. *Mathematics Education Research Journal*, *25*(3), 435-456.

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 the 37th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, S. 49-56). Kiel: PME.

Redder, A., & Weinert, S. (Eds.) (2013). *Sprachförderung und Sprachdiagnostik. Interdisziplinäre Perspektiven.* Münster: Waxmann.

Reusser, K. (1989). *Vom Text zur Situation zur Gleichung. Kognitive Simulation von Sprachverständnis und Mathematisierung beim Lösen von Textaufgaben*. (Habilitationsschrift). Bern: Universität Bern.

Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? Schulleistungen, Schülerfähigkeiten, kognitive Fähigkeiten, Wissen oder allgemeine Intelligenz? *Psychologische Rundschau*, *57*(2), 69-86.

Rösch, H., & Paetsch, J. (2011). Sach- und Textaufgaben im Mathematikunterricht als Herausforderung für mehrsprachige Kinder. In S. Prediger, & E. Özdil (Eds.), *Mathematiklernen unter Bedingungen der Mehrsprachigkeit* (pp. 55-76). Münster: Waxmann.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. ed.). Bern: Hans Huber.

Schnabel, K. U. & Schwippert, K. (2000). Einflüsse sozialer und ethnischer Herkunft beim Übergang in die Sekundarstufe II und den Beruf. In J. Baumert, W. Bos & R. Lehmann (Eds.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn.* Band 1 (pp. 261-281). Opladen: Leske + Budrich.

Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistics perspective*. Mahwah: Lawrence Erlbaum.

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

Tarelli, I., Schwippert, K., & Stubbe, T. C. (2012). Mathematische und naturwissenschaftliche Kompetenzen von Schülerinnen und Schülern mit Migrationshintergrund. In W. Bos, H. Wendt, O. Köller & C. Selter (Eds.), *TIMSS 2011 – Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 247-268.). Münster: Waxmann.

Thürmann, E., Vollmer, H., & Pieper, I. (2010). *Language(s) of schooling: Focusing on vulnerable learners.* (= The linguistic and educational integration of children and adolescents from migrant backgrounds. Studies and resources 2). Straßbourg: Council of Europe. http://www.coe.int/t/dg4/linguistic/Source/Source2010_ ForumGeneva/2-VulnerLLearnersThurm_EN.pdf. Last access February, 13, 2015.

Ufer, S., Reiss, K., & Mehringer, V. (2013). Sprachstand, soziale Herkunft und Bilingualität: Effekte auf Facetten mathematischer Kompetenz. In M. Becker-Mrotzek, K. Schramm, E. Thürmann & H. J. Vollmer (Eds.), *Sprache im Fach – Sprachlichkeit und fachliches Lernen* (pp. 167-184). Münster: Waxmann.

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

Werning, R., Löser, J. M., & Urban, M. (2008). Cultural and social diversity: An analysis of minority groups in German schools. *The Journal of Special Education, 42*(1), 47-54.

Wilhelm, N. (2016). *Zusammenhänge zwischen Sprachkompetenz und Bearbeitung mathematischer Textaufgaben - Quantitative und qualitative Analysen sprachlicher und konzeptueller Hürden*. Wiesbaden: Springer.

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.

Zindel, C. (2015). „Wenn ich wüsste, was davon was ist...“ – konzeptuelle und sprachliche Hürden bei funktionalen Abhängigkeiten. Erscheint in H. Linneweber-Lammerskitten (Ed.), *Beiträge zum Mathematikunterricht* (pp. 1024-1027). Münster: WTM.