Z Erziehungswiss (2015) 18:781–801
DOI 10.1007/s11618-015-0645-3

ALLGEMEINER TEIL

ZfE

Reading skills of students in different school tracks: Systematic (dis)advantages based on item formats in large scale assessments

Franziska Schwabe · Nele McElvany · Matthias Trendtel

Published online: 21 July 2015

Abstract Contrary to the broad documentation of substantial absolute differences in mean reading achievement between academic and non-academic track students, little is known about specific strengths and weaknesses of these groups of students in reading tests. Therefore, in this study we investigated Differential Item Functioning of 100 PISA 2009 reading items comparing $N=3824$ students attending academic and nonacademic school tracks in Germany. Significant interaction effects between school track and item format were found. Students of academic tracks showed specific strengths in responding to open-ended items compared to equally skilled students in non-academic tracks. Furthermore, the effects were stable even when differences on individual and social dimensions between the groups of students as well as compositional aspects of their school environments were controlled for. Institutional and compositional aspects of school tracking and their effects on reading performance were discussed.

Dr. F. Schwabe (✉) · Prof. Dr. N. McElvany
Institut für Schulentwicklungsforschung, Technische Universität Dortmund,
Vogelpothsweg 78,
44227 Dortmund, Germany
e-mail: schwabe@ifs.tu-dortmund.de

Prof. Dr. N. McElvany
e-mail: Office.mcelvany@fk12.tu-dortmund.de

M. Trendtel
Bundesinstitut für Bildungsforschung,
Innovation und Entwicklung des österreichischen Schulwesens,
Alpenstraße 121,
5020 Salzburg, Austria
e-mail: m.trendtel@bifie.at

Keywords Differential learning environments · Differential item functioning · Large-scale assessments · School tracking · Reading competence

Lesekompetenzen von Jugendlichen in unterschiedlichen Schulformen: Systematische Vor- und Nachteile bei verschiedenen Antwortformaten in Large-Scale Assessments

Zusammenfassung Im Gegensatz zur umfangreichen Dokumentation substantieller absoluter Leseleistungsunterschiede zwischen Schulformen mit und ohne Abituroption, ist in Bezug auf spezifische Stärken und Schwächen dieser Schülergruppen in Lesetestaufgaben unabhängig von ihrem Lesekompetenzniveau wenig bekannt. Vor diesem Hintergrund untersucht der vorliegende Beitrag die relativen Leistungen von $N = 3824$ Jugendlichen, die entweder eine Schule mit oder ohne Abituroption besuchen, mit Differential Item Functioning Analysen der 100 Lesetestaufgaben aus PISA 2009. Es zeigten sich signifikante Interaktionen zwischen besuchter Schulform und Kompetenzen bei der Bearbeitung unterschiedlicher Antwortformate. Schülerinnen und Schüler an Schulformen mit Abituroption besaßen spezifische Vorteile bei der Bearbeitung von open-ended Aufgaben im Vergleich zu fähigkeitsgleichen Jugendlichen an Schulen ohne Abituroption. Diese Unterschiede waren stabil auch unter Kontrolle individueller und sozialer Unterschiede zwischen den Gruppen sowie unter Kontrolle kompositioneller Aspekte ihrer Lernumwelten. Institutionelle Aspekte von Schulformeffekten und deren Zusammenhang mit Lesekompetenz werden diskutiert.

Schlüsselwörter Differenzielle Lernumwelten · Schulformunterschiede · Lesekompetenzen · Large-scale Assessment · Differential Item Functioning

1 Introduction

Reading is an essential skill for academic achievement as well as for involvement in modern society (Snow et al. 1998; Connor et al. 2011; Pfost et al. 2012). Thus, proficiency in reading, especially in reading comprehension, is of high importance to all students. The finding that average reading achievement differs considerably between school tracks (academic vs. non-academic) is a crucial issue in educational contexts (Maaz et al. 2008; Retelsdorf et al. 2012). Moreover, studies concerned with reading competence of particular subgroups of students indicated that specific strengths and weaknesses between students who are equally skilled but attend different school tracks might exist in addition to absolute differences and independently of the students' level of reading proficiency (overview for language tests: Ferne and Rupp 2007).

On the item level, proper methodologies to detect specific strengths and weaknesses of particular subgroups are analyses by *Differential Item Functioning* ([DIF]; Angoff 1993). DIF analyses refer to a psychometric difference in how an item functions across groups. In concrete terms an item shows DIF when examinees who belong to different groups, e.g., male-female or young-old, have different probabilities of correct responses to this item—despite the fact that they are equally skilled in the construct the item is supposed to assess (cf. Ferne and Rupp 2007). There is a substantial body of evidence for the existence of DIF in reading test items for subgroups of students who differ by gender, language background, ethnicity, and/or social status (e.g., Schwippert et al. 2004; Haag et al. 2013). However, studies which report on DIF for students with different academic backgrounds are rare, even though they could provide relevant information regarding both (a) test construction and interpretation as well as (b) specific competencies of students in different tracks. The existent studies have focused mostly on item content as a source of DIF (Pae 2004; Haberkorn et al. 2012). However, item format—as another item feature—has also been shown to be highly relevant in the context of DIF in reading tests (Rodriguez 2002; Lafontaine and Monseur 2009; Taylor and Lee 2012; Schwabe et al. 2015). Regarding school tracking, literature discusses mainly three potential sources of differences in mean reading achievement, which might also relate to the emergence of specific strengths and weaknesses of students attending different tracks: (a) individual and social differences in students' preconditions at school entry (Luyten et al. 2003), (b) institutional, and/or (c) compositional aspects of the different school tracks (e.g., Becker et al. 2012). This study considers DIF caused by item format in reading test items for students in different school tracks following a Generalized Linear Mixed Models (GLMM) approach (De Boeck et al. 2011). Additionally controlling for individual and social factors which might influence specific strengths and weaknesses, the study aims to gain new insights into the effects of institutional and compositional factors of school tracking on reading test performance.

2 Theoretical background

Reading comprehension is a major goal of education. In addition, it is one of the fundamental prerequisites for academic success, because proficient reading is of utmost importance for content learning in all educational subjects (Snow et al. 1998; Retelsdorf et al. 2012). Children begin learning to read systematically in elementary school, and thus, the acquisition of basic reading skills mainly takes place during the first years of schooling (for Germany: Klicpera and Gasteiger-Klicpera 1993; Pfost et al. 2012). Later students develop more advanced reading skills such as text comprehension (Snow 2002). However, substantial parts of students reveal only moderate reading competence at the age of 15, although they are at the end of compulsory school in Germany as well as in many other countries (internationally: OECD 2010a; for Germany: Klieme et al. 2010). Thus, reading competence is still a crucial issue in secondary schools.

Dividing secondary school student populations into more or less ability-homogenous groups of learners is a common feature of many educational systems. A huge body of literature claims that *tracking* affects students' educational success and achievement as well as their emotional well-being (e.g., Oakes 1982; Baumert et al. 2006; Ariga and Brunello 2007; Becker et al. 2012). Similar to the differences between German states in regard to the transition regulations, there are considerable differences from state to state in the available range of school tracks in Germany (KMK 2014). Three main types of schools can be distinguished: (a) one exclusive academic or non-academic track, (b) combinations of non-academic—intermediate and vocational—tracks (e.g., Mittelschule [Saxony], Erweiterte Realschule [Saarland], Sekundarschule [Bremen, Saxony-Anhalt]), and (c) combinations of academic and non-academic tracks (e.g., Integrierte Sekundarschule [Berlin], Oberschule [Bremen, Lower Saxony], Stadtteilschule [Hamburg]; compare for a description in detail KMK 2014). Independent of the labels of the different school tracks tracking is discussed as a factor which affects reading performance and development to the advantage of students attending academic tracks (Becker et al. 2012; Retelsdorf et al. 2012, compare also Sect. 2.3). Beyond absolute differences in reading achievement tracking might also be correlated with specific strengths or weaknesses of the student groups which also impact reading test results.

2.1 Specific strengths and weaknesses of particular subgroups of students

Research, which was conducted in the framework of DIF, has shown that particular subgroups of students may possess differential profiles of competences even if their level of reading achievement is comparable (e.g., Schwippert et al. 2004; Haag et al. 2013). While there is a huge body of research on DIF in regard to individual (e.g., gender: Schwippert et al. 2004; Schwabe et al., 2015) and social (e.g., socioeconomic status: McElvany and Schwabe 2013; Walzebug 2014) characteristics of students, studies on DIF focusing on differences in academic background factors are comparatively rare.

Pae (2004) investigated DIF on the English subtest of the 1998 Korean National Entrance Exam for Colleges and Universities for examinees with different academic backgrounds (i.e., Humanities vs. Sciences), using the Item Response Theory Likelihood Ratio approach. This study focused on item content. Findings from the preliminary content analysis suggested that items dealing with science-related topics, data analysis, and number counting were differentially easier for the Science students, whereas items about human relationships were differentially easier for the Humanities students (Pae 2004). In conclusion, DIF was detected for students with different academic background depending on the item feature ‘item content’.

In a more recent study, DIF was investigated for students from different German school tracks based on data from the National Educational Panel Study ([NEPS], Haberkorn et al. 2012). 4887 subjects (35.2 %) who took the reading test attended “Gymnasium” and 9010 (64.8 %) did not. Overall, two items of the 31 items studied showed strong DIF and four items medium DIF. The authors did not find any item content related indicators which might have caused DIF (Haberkorn et al. 2012).

In conclusion, research indicates that there might be specific strengths and weaknesses of students in different tracks, but the number of studies which explicitly focused on these learners and their performance in reading tests is severely limited. Moreover, existent studies do not consider item format as an item feature, which is known to cause differences between groups of test takers in reading tests.

2.2 Reading test item formats

The most common formats of test items are multiple-choice (MC) and open-ended (OE) items (Haladyna and Rodriguez 2013). The distinguishing feature of all variations of MC items is their requirement to select an answer from a set of options. In contrast, OE items require the creation of an answer. OE formats vary greatly, ranging from simple fill-in items to complex essay writing. When used in test situations, both formats have advantages and disadvantages. The main reason for the use of OE items is the assumption that they measure a 'deeper understanding' and are more closely related to the demands of school lessons (e.g., Bacon 2003). Nevertheless, one crucial restriction of OE items pertains to the scoring of given answers, which has several disadvantages compared to the scoring of MC items. It tends to be more complex and subjective, and therefore, the reliability of MC items is often higher (Wainer and Thissen 1993; for a discussion: Rodriguez 2002).

Considering the differences between OE items and MC items, there are reasons for the assumption that they require different skills. First, the demand of OE items to write answers might still be an obstacle for some students even in secondary schools, independently of their reading competence level (Guthrie and Wigfield 2005), especially for those who lack productive language skills. Second, OE items require the respondents to actively formulate an answer (Solheim 2011) and therefore, the items might be perceived as more challenging. These dissimilarities in requirements have been found to interact with characteristics of the test takers such as country of birth (Grisay and Monseur 2007), gender (Lafontaine and Monseur 2009), and level of proficiency (Routitsky and Turner 2003).

There are several explanations for this observation, for example different tendencies to guess between the groups studied, or the verbal superiority of one specific group (e.g., Simkin and Kuechler 2005). Moreover, research has shown that students with lower verbal skills perform relatively worse on OE items compared to MC items (Routitsky and Turner 2003; Lafontaine and Monseur 2006, 2009; Rauch and Hartig 2010). There is some empirical evidence for the assumption that academic track students are more likely to possess the skills required by OE items (e.g., DESI Konsortium 2008) compared to non-academic track students. Considering a ninth grade sample of students in Germany Steinert et al. (2008) reported that academic track students outperformed students who attended non-academic tracks in German text production. Moreover, while only small amounts of variance of German language proficiency were explained by differences within (37 %) and between schools (11 %), more than 50 % of the variance could be traced back to differences between school tracks. Therefore, OEs' specific requirements might keep non-academic track students from fully demonstrating their reading competence in responding to these items.

Summarizing research in regard to different item formats, it is obvious that OE items require different skills compared to MC items. Moreover, academic and non-academic track students differ in their levels of proficiency in the specific requirements of OE items independently of their reading competence. Consequentially, OE items might be relatively easier for students attending an academic track compared to students attending a non-academic school track. While the relationship between the performance in responding to different item formats and students attending different school tracks has not been studied yet, there are findings in regard to effects of tracking on reading performance. These could provide insights into possible issues, which should be considered when addressing specific differences in terms of new research questions. Research in the context of tracking and its relationship to differences in reading achievement mainly discusses institutional and compositional aspects of the different tracks.

2.3 Institutional and compositional aspects of school tracking and its effects on reading competence

In the case of Germany—as an example of between-school tracking at the secondary level of schooling—variation between school tracks (academic vs. non-academic) has been demonstrated in regard to several institutional factors (overview: Baumert et al. 2006). First, prospective teachers are trained differently depending on the school track in which they will be teaching. It has been shown that these differences result in specific pedagogical knowledge and skills as well as content knowledge (Kleickmann et al. 2013). Second, non-academic and academic school tracks differ in curricula (DESI Konsortium 2008). Third, teachers in non-academic tracks are assumed to use repetitive teaching methods more often, while teaching in academic tracks is for the most part determined by cognitively activating instruction (Kunter et al. 2013). Consequently, it is possible that the differences in reading achievement between students attending different school tracks arise partly from the institutional factors mentioned.

Moreover, it should be noted that school tracks differ not only in institutional factors but likewise in compositional aspects (e.g., Rjosk et al. 2014). Considering these differences within the student body and reflecting the work of Baumert et al. (2006), Becker et al. (2014) distinguished five crucial dimensions of heterogeneity on the compositional level: (a) achievement, (b) socio-cultural background, (c) psychosocial risk factors, (d) ethnic-cultural background, and (e) the learning biography. Baumert et al. (2006) demonstrated that academic and non-academic tracks have nearly laterally reversed profiles with respect to these dimensions. Moreover, there is empirical evidence indicating that students' achievement in several domains is related to compositional aspects of their learning environment (e.g., socio-economic status: Van Ewijk and Sleegers 2010; race and ethnicity: Brenner and Crosnoe 2011). In a recent investigation, Rjosk et al. (2014) have found the effect of the socioeconomic composition on achievement to be mediated partially by aspects of instructional quality. These aspects of instructional quality were in turn related to institutional factors. In consequence, it seems reasonable to assume that absolute differences as well as specific strengths and weaknesses in reading achievement arise partly from a complex interaction between institutional and compositional aspects of school tracking.

Empirical research on the effects of tracking on reading achievement is ambiguous. Investigations based on cross-country data, which analyzed the influence of differences in institutional factors between countries on international test scores have used regression analyses and have revealed mixed results. Applying a differences-in-differences approach, Hanushek and Woessman (2006) studied the impact of early tracking on standardized reading test scores. They demonstrated that early tracking does not affect average test performance, but increases social inequalities. Contradicting these results, Ariga and Brunello (2007) observed a positive effect of tracking on literacy achievement. They investigated the effect of the duration spent in a tracked system on the performance in a standardized cognitive test. Conflicting results also appear for the intra-national effects of tracking in Germany (overview: Becker 2009). Comparisons of different school tracks have revealed parallel or even convex average growth rates in reading achievement (e.g., Retelsdorf et al. 2012; Köller et al. 2013). Beyond investigating average growth rates, some researchers have analyzed distinct positive effects of a system of different school tracks on reading achievement, controlling for individual differences in prerequisites of students (for a description of the method, see Becker et al. 2012). Findings suggest that reading competence increases significantly in the academic track when individual differences are controlled (Bos and Gröhlich 2010; Bos and Scharenberg 2010; Pfost et al. 2010). However, investigating sub-components of reading separately, Retelsdorf et al. (2012) have revealed mixed results: they did not find any variation in the increase of reading comprehension between academic and non-academic tracks, but a more positive development for students in academic tracks in terms of decoding speed. Considering literacy development between grades 3 and 4 as well as between grades 5 and 6, Becker et al. (2014) reported no advantages for academic track students in reading comprehension, vocabulary, or decoding ability, controlling for inter-individual differences in intra-individual change.

Summarizing these findings, there is some empirical evidence supporting the assumption of effects of different school tracks on reading achievement and development. The academic track seems to foster reading comprehension greater compared to non-academic tracks, even after controlling for individual differences in pre-conditions. However, the findings are not straightforward as some studies also suggest no differences in reading development. These mixed results with respect to absolute differences in reading test performance might arise due to specific competencies in regard to different item formats of students attending different tracks, which in turn might be caused by differences between the performance of the different students groups on OE and MC items. Research suggests that academic track students might possess specific advantages in responding to OE items compared to students, who attend non-academic tracks. Moreover, these specific strengths of academic track students in responding to OE items might be correlated with (a) individual differences on the student-level, and/or (b) institutional and compositional differences on the school-level. In order to gain new insights into the relationship between school track attendance and specific strengths in different item formats these factors should be taken into account.

3 Research questions and hypotheses

The aim of the current study is to analyze specific strengths and weaknesses of students who attend different school tracks in Germany (academic vs. non-academic) in regard to different item formats which are used in large scale assessments of reading competence (OE vs. MC). Moreover, the study aims to investigate the effect of institutional and compositional factors on these specific strengths. In detail, the study explores the following research questions:

Hypothesis 1: Is there a specific strength of students who attend the academic track with respect to OE items compared to students who attend a non-academic track?

Taking into account research on the specific requirements of OE items as well as the theoretical considerations and empirical findings concerning students with different academic backgrounds, we expect a significant interaction between item format and school track. Regarding the level of superiority, we assume that students attending an academic school track relatively outperform students attending a non-academic school track in OE items, because of the assumed differences in regard to requirements of OE items between the groups studied (hypothesis 1).

Hypothesis 2: Does the expected specific strength of students who attend the academic track with respect to OE items in comparison to students who attend a non-academic track remain when individual and social factors are controlled for?

Considering prior research on effects of institutional and compositional factors of school tracking, which demonstrated that absolute differences in reading competence exist even when individual and social differences were controlled for, we expect the interaction between item format and students' attended school track to remain significant when controlling for these variables (hypothesis 2).

Hypothesis 3: Does the expected specific strength of students who attend the academic track with respect to OE items compared to students who attend a non-academic track remain when in addition to individual and social factors also compositional aspects of their school environment are controlled for?

Because of the assumed stand-alone effect of institutional factors and by analogy with the arguments concerning the second research question, we expect the interaction between item format and students' attended school track to remain significant when compositional aspects as well as individual and social factors are controlled for (hypothesis 3).

4 Method

4.1 Sample

Analyses were based on data from the *Programme for International Student Assessment* (PISA) 2009, which was conducted in Germany in April and May 2009 by a national consortium coordinated by the *German Institute for International Educational Research* (DIPF), Frankfurt. In the German 2009 PISA study a total of 3824 students aged 15 took part. They attended either a non-academic track ($N=2202$)

or an academic track ($N=1622$). There were no group-differences with respect to the percentages of females ($\chi^2(1)=4.254$, $p=0.10$). However, the subsamples differed by language background, as 1621 (80.2 %) of the non-academic track students stated that they spoke the test language at home, while 1383 (85.3 %) of the academic track students made the same statement ($\chi^2(1)=59.633$, $p<0.001$). Moreover, the subsamples differed by the number of books at home as an indicator of the social background: 1332 (65.88 %) of the non-academic track students stated that their family possessed less than 100 books, while only 489 (30.2 %) of the academic track students made the same statement ($\chi^2(1)=385.000$, $p<0.001$). Also, in regard to self-reported reading motivation the subsamples showed significant differences: 1229 (60.8 %) of the non-academic track students demonstrated low levels of reading motivation. This was true for only 499 (30.8 %) of the students who attended an academic track ($\chi^2(1)=255.495$, $p<0.001$).

4.2 Measures

4.2.1 Reading competence and motivation

In PISA 2009 a total of 100 items were administered in 13 booklets in a multi-matrix design (OECD 2010a). The booklets consisted of continuous, non-continuous, mixed, and multiple texts (OECD 2010a). Five different item formats were administered. Table 1 gives an overview of the item types. The different formats were combined into the superordinate categories MC (Complex Multiple Choice; Multiple Choice) and OE (Open Constructed Response; Short Response; Closed Constructed Response). Percentages of correct responses show that OE and MC items do not in general differ by difficulty. On the whole, 47 items in an MC format and 53 in an OE format were administered and item format was used as a predictor on the item level. In order to obtain descriptive statistics, reading items were scaled according to a 1PL model (compare analyses), and revealed an EAP reliability of .81.

Reading motivation was assessed by eleven items. Six of them were positively phrased (e.g., “Reading is one of my favorite hobbies”), and five were negatively

Table 1 Item types, percentages of correct responses, omitting rates, and not reached rates

Item type	*N*	Entire sample			Academic track students			Non-academic track students		
		Cor. res. (%)	Omi. (%)	Not rea. (%)	Cor. res. (%)	Omi. (%)	Not rea. (%)	Cor. res. (%)	Omi. (%)	Not rea. (%)
Complex multiple choice	8	45.7	1.7	1.0	59.4	0.7	0.5	35.7	2.5	1.1
Multiple choice	39	67.4	2.2	1.3	78.8	1.1	0.8	58.1	3.0	1.7
Open constructed response	36	56.0	13.7	1.2	71.1	5.3	5.3	44.9	20.0	1.3
Short response	7	58.1	8.0	1.2	68.3	2.5	0.8	50.6	12.0	1.5
Closed constructed response	10	72.4	3.9	1.2	82.8	0.9	0.5	64.7	6.2	1.7

*Cor. res.*correct response, *Omi.*omitted, *Not rea.*not reached phrased (e.g., “I only read if I have to”). Students rated their agreement with the statements on a 4-point Likert scale (1 = *disagree completely* to 4 = *agree completely*). The scale had a high reliability (α = 0.92).

4.2.2 Students' background

Students’ background was assessed by student questionnaires. Variables on the student level were operationalized as follows: In regard to language background, statements on the use of the language of the test at home were evaluated. This indicator was chosen because of its’ high importance in the context of reading assessment and students with an immigrant background (Stanat et al. 2010). The scale comprised three possible answers with respect to the use of the test language at home (“always or almost always”, “sometimes”, and “never”). A dummy-variable was created which assigned a multilingual language background to students who either sometimes or never speak the language of the test in their family (0 = *monolingual language background*; 1 = *multilingual language background*). Social background was operationalized through the number of books possessed by the parents (0 = *less than 100 books*; 1 = *more than 100 books*). This indicator was chosen because it simultaneously incorporates cultural and economic capital and is of high importance in the context of reading (Bos et al. 2007). Also grade repetition was measured by the student questionnaires and dichotomized (0 = *no repetition*; 1 = *at least one repetition*).

In addition to student level variables, variables on the school level were computed as follows: To include the school track a dummy-variable was created which assigned *academic track* to students who attended an academic track (0 = *non-academic track*; 1 = *academic track*). The other variables on the school level—percentage of students with multilingual language background, percentage of students with high socioeconomic status, percentage of grade repeaters, and percentage of highly motivated students—were included in the models (compare following section) as continuous variables.

4.3 Analyses

For preliminary analyses of the data *t*-tests were conducted. Interpretation of mean differences was done following Cohen’s *d*. The clustered samples resulting from the survey design of PISA might have led to biased standard errors and significance tests. In order to avoid such bias, standard errors for mean comparison were computed using a jackknife procedure (OECD 2012).

In order to investigate the research questions, a GLMM approach proposed by De Boeck et al. (2011) was applied to the data, and analyses of uniform *Differential Item Functioning* (DIF) were conducted. The analyses were computed in the program R (R Core Team 2013) with the packages MCMCglmm (Hadfield 2010) and TAM (Kiefer et al. 2013).

The one-parametric-logistic (1PL) Item Response Theory (IRT) model can be formulated for test taker *p* and item *i* as follows (cf. De Boeck et al. 2011):

$$\eta_{pi} = \theta_p + \beta_i$$

with η_{pi} describing the logit of the probability of a correct response, $\theta_p \sim N(0,\sigma^2_\theta)$ the ability of the test taker, and $\beta_i \sim N(0,\sigma^2\beta)$ the easiness of the item as random effects. According to De Boeck et al. (2011, pp. 18) "a DIF model can be formulated as follows:

$$\eta_{pi} = \theta_p + \beta_i + \zeta_{focal} Z_{(p,i)focal} + \sum_{(h=1)^H} \omega_h W_{(p,i)h}$$

with θ_p and β_i as above and ζ_{focal} the global effect of the focal group in comparison with the reference group: with $Z_{(p,i)focal} = 1$ for the focal group, and 0 for the reference group; with $W_{(p,i)h}$ as the person-by-item covariate h, in such a way that $W_{(p,i)h} = 1$ if both $Z_{(p,i)focal} = 1$ and the considered person is part of the focal group (item subset DIF), and $W_{(p,i)h} = 0$ otherwise; and ω_h as the corresponding DIF parameter."

In order to control for individual and social factors on the student level (research question 2) as well as for compositional aspects on the school level (research question 3), we specified four cross-classified multilevel models (compare for a similar approach: Bakker et al. 2014). We extended the DIF model above by a school-level random effect $\gamma_s \sim N(0,\sigma^2\beta)$ for school s.

Inclusion of predictor groups was done hierarchically (no predictors [Model 0], item level predictors [Model 1], student level predictors [Model 2], and student and school level predictors [Model 3]). Furthermore, we included the interaction effect of item format and school track for all models except of the empty model. All models were compared using variance explanation measures (R^2). Regression coefficients were tested on difference from zero at the 1 % significance level.

4.4 Missing data

As a result of the PISA design, missing achievement data appeared due to (a) the test design (not all items were administered to all students), (b) omission of administered items, and/or (c) not reaching of administered items. Following OECD (2012), missing data of type (a) were coded as missing in any case, and those of type (b) were coded as an incorrect answer in any case. Type (c) missing data were coded as missing for item calibration and for estimating cross-classified models and as an incorrect answer for the estimation of person ability for mean comparison. Moreover, missing data appeared in the background information variables. Multiple imputations are one recommended solution to handle this problem (e.g., van Buuren and Groothuis-Oudhoorn 2011). They were done with the package mice (van Buuren and Groothuis-Oudhoorn 2011). Twenty complete data sets were generated and analyzed separately. Results were conflated following Rubin (1987).

5 Results

5.1 Descriptives

The achievement gap, which had already been reported for PISA 2009, was replicated on the basis of the operationalization chosen. Students attending an academic school track ($N = 1622$; $M_\theta = 0.76$; $SD_\theta = 0.03$; Correct answers 74.6 %; Omitted 10.0 %; Not reached 1.0 %) outperformed students who attended a non-academic track ($N = 2202$; $M_\theta = -0.65$; $SD_\theta = 0.03$; Correct answers 51.7 %; Omitted 2.6 %; Not reached 0.1 %) in absolute reading achievement ($t(3,822) = -268.84$, $p < 0.001$). The difference reveals a large effect size (Cohen's $d > 0.8$).

5.2 Specific strength of students in an academic track in OE items

In order to answer the first research question on the interaction between students' attended school track and item format, we specified among the set of four cross-classified models (see Table 2) a DIF model (Model 1, Table 2). Model 1 included the main effect of item format and an interaction effect between students' attended school track and item format. Item subset DIF was studied.

About 26 % of the total variance goes back to variance between schools as representation of the different school tracks (Model 0). In Model 1 the interaction predictor explained 11 % of this school level variance ($R^2 = 0.11$). The estimated coefficient of the main effect ($\zeta_{\mathrm{OE}} = -0.37$) did not differ significantly from zero ($p = 0.20$), which can be explained by the fact that OE and MC items did not differ in difficulty in general (compare instruments section). Due to this fact the proportion of variance even had a negative sign and was therefore set to zero. R^2 was slightly negative but far away from a warning magnitude of 0.05 (Snijders and Bosker 2012).

The interaction between students' attended school track and OE item format had a significant, positive effect ($\omega = 0.28$; $p < 0.001$). The positive sign of the interaction effect indicated a specific strength of students in an academic school track in OE items compared to students in a non-academic track. In conclusion, hypothesis 1 is supported by the data:

Academic track students have a specific strength in OE reading items compared to non-academic track students.

5.3 Impact of individual and social factors on the interaction

In order to answer the second research question, namely to analyze the interaction between students' attended school track and item format while controlling for individual and social differences between the students, a second DIF model was specified (Model 2, Table 2).

Model 2 included student level predictors as well as an interaction effect between students' attended school track and item format. Moreover, interaction effects between item format and the specified main effects were included (see Table 2). Introducing student level predictors in Model 2 explained 14 % of the student level and 69 % of the school level variance (see Table 2). The amount of explained school level variance may result from the fact that students with different backgrounds tend to attend different schools. All estimated coefficients of the main effects on student level in Model 2 ($\zeta_{mot} = 0.32$, $\zeta_{books} = 0.20$, $\zeta_{lang} = -0.28$) as well as the track affiliation ($\zeta_{aca} = 1.16$) were significantly different from zero ($p < 0.001$) with the exception of the effect of being female ($\zeta_{fem} = 0.06$, $p = 0.12$). Also, the interaction between students' attended school track and item format still had a significant, posi-

Table 2 Cross-classified multilevel models

Random effects		Model 0: No predictors			Model 1: Item level and interaction			Model 2: Item and student level and interactions			Model 3: All predictors		
		Var	SE	%Var	Var	SE	R^2	Var	SE	R^2	Var	SE	R^2
Item		2.03	0.10	55.1	2.03	0.11	0^x	2.03	0.10	0.00	2.04	0.10	0^x
Student		0.69	0.01	18.8	0.69	0.01	0^x	0.60	0.01	0.14	0.60	0.01	0.14
School		0.96	0.06	26.1	0.86	0.05	0.11	0.30	0.04	0.69	0.17	0.03	0.82
Total		3.68		100	3.58		0.03	2.93		0.21	2.80		0.24
Predictor level					Par.	SE	*p*	Par.	SE	*p*	Par.	SE	*p*
Item	Item format[a]				−0.37	0.28	0.20	−0.55	0.29	0.06	−0.56	0.29	0.06
Student	Gender[b]							0.06	0.04	0.12	0.06	0.04	0.13
	Motivation							0.32	0.03	<0.001	0.31	0.03	<0.001
	Books at home[c]							0.20	0.04	<0.001	0.17	0.04	<0.001
	Language background[d]							−0.28	0.06	<0.001	−0.24	0.06	<0.001
School	School track[e]							1.16	0.09	<0.001	0.37	0.12	<0.001
	Mean motivation										0.30	0.20	0.13
	Mean books at home										1.11	0.29	<0.001
	Mean language background										−1.31	0.33	<0.001
	Mean grade repetition										−0.82	0.29	<0.001
Interaction	School tracke × item format[a]				0.28	0.03	<0.001	0.22	0.03	<0.001	0.22	0.03	<0.001
	Gender[b] × item format[a]							0.09	0.03	<0.001	0.09	0.03	0.01
	Motivation × item format[a]							0.06	0.02	0.01	0.06	0.02	0.01
	Books at home[c] × item format[a]							0.01	0.04	0.81	0.01	0.04	0.79
	Language background[d] × item format[a]							−0.02	0.05	0.70	−0.02	0.05	0.70

Var variance, R^2 proportion of explained variance, ^{x} The value was negative and therefore set to 0

[a] Category (Ca): open-ended

[b] Ca: female

[c] Ca: more than 100

[d] Ca: multilingual

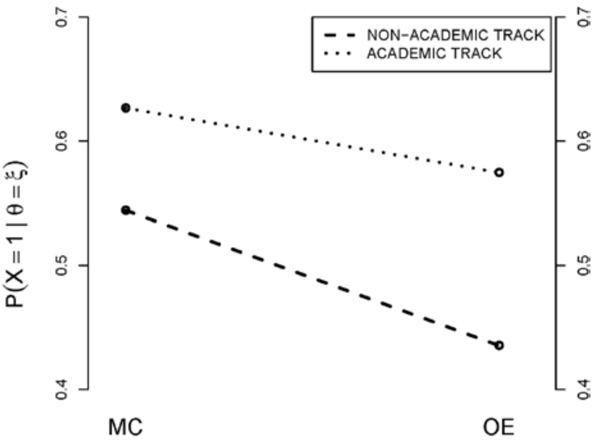
[e] Ca: academic

tive effect (ω= 0.22; $p < 0.001$), indicating a specific strength of academic track students in OE items. In addition, the interactions between gender and item format and between motivation and item format had significant positive effects ($\omega_{fem \times item} = 0.09$; $p < 0.001$; $\omega_{mot \times item} = 0.06$; $p < 0.05$). In conclusion, hypothesis 2 is supported by data, as the interaction between students' attended school track and item format remained significant when individual and social differences on student level were controlled for.

5.4 Impact of institutional and compositional factors on the interaction

In order to answer the third research question, namely to analyze the interaction between students' attended school track and item format while controlling for individual and social differences on the student level as well as compositional dissimilarities on the school level, a third DIF model was specified (Model 3, Table 2). Model 3 included all predictors and explained 82 % of the school level variance. The estimated coefficients of the main effect of being female on the student level ($\zeta_{fem} = 0.06$, $p = 0.13$) and the level of motivation on school level ($\zeta_{mot\text{-}school} = 0.30$, $p = 0.13$) did not differ from zero. All other estimated coefficients of the main effects on the student level in Model 3 ($\zeta_{mot} = 0.31$, $\zeta_{books} = 0.17$, $\zeta_{lang} = -0.24$) as well as coefficients on the school level ($\zeta_{aca} = 0.37$, $\zeta_{books\text{-}school} = 1.11$, $\zeta_{lang\text{-}school} = -1.31$, $\zeta_{g.rep\text{-}school} = -0.82$) were significantly different from zero ($p < 0.001$). Also, the interaction between students' attended school track and item format had a significant, positive effect (ω= 0.22; $p < 0.001$), indicating again a specific strength of academic track students with respect to OE items. Figure 1 illustrates this differential advantage showing probabilities of correct answers controlled for person ability and background variables on student and school level. In conclusion, hypothesis 3 is supported by the data, as the interaction between students' attended school track and item format remained significant when individual and social differences on the student level as well as compositional dissimilarities on the school level were controlled for.

Fig. 1 Probabilities of correct answers controlled for person abilty and background variables on student and school level



5.5 Discussion and conclusion

This study investigated relative strengths and weaknesses in reading test items for students who attended either a non-academic or an academic track in German secondary schools. In contrast to previous studies, we focused on item format as a source of these specific differences in performance. Moreover, we addressed the question whether or not controlling for differences in individual and social aspects as well as for compositional factors of different school tracks had an effect on the assumed interaction between students' attended school track and item format. The findings show that there is a considerable interaction between students' attended school track and item format in the 2009 PISA reading test items: OE items are relatively easier for students attending an academic track compared to students attending a non-academic track. In addition, the observed interaction between students' attended school track and item format remained stable even when individual and social aspects of the students as well as compositional factors of school environment were controlled for.

Comparable findings in regard to specific strengths of particular subgroups of students in OE reading items have already been reported by several studies (e.g., gender comparisons: Routitsky and Turner 2003; Lafontaine and Monseur 2009). Thus, this study is in line with previous research, and, furthermore, enlarges findings on DIF due to item format. Concerning the interaction between gender and DIF, different reasons have been discussed as potential explanations for the finding of specific strengths in OE items, which might also apply to the case at hand: (a) differences in productive language skills, (b) different test taking behaviors, and/or (c) even other aspects like confounding between OE items and item content or item difficulty (Lafontaine and Monseur 2009; Rauch and Hartig 2010).

As far as analyses of students who attend different school tracks are concerned, the first explanation is supported by findings that these students differ considerably by productive language skills (DESI Konsortium 2008), which are a key requirement of OE items. In addition, different levels of reading motivation of the groups studied might result in varying test taking strategies (for a discussion in detail: Schwabe et al. 2015). Poorly motivated students, who are overrepresented within the non-academic tracks, might try less hard when answering OE items or even omit these items. From our point of view even higher rates of omission are an indication of a differential validity of these items. In regard to the correlation between item format and item content or item difficulty, we showed that the formats do not differ by difficulty in general. Nevertheless, our study does not take into account differences in item content. In PISA OE items often assess higher order reading skills, but there are also OE items measuring basic skills. Furthermore, we have no indication of the cognitive demands of the different item formats. Descriptive comparisons of difficulties yield no systematic difference between OE and MC PISA items.

The result that the interaction between students' attended school track and item format remains stable even when individual differences like different levels of motivation are controlled for, indicates that institutional and/or compositional differences between the school tracks studied might influence reading performance substantially (Maaz et al. 2008; Becker et al. 2012; Retelesdorf et al. 2012). The observed specific strength of academic track students compared to non-academic track students in answering OE items might be the result of more intensive training especially of productive language skills. Moreover, amounts of instruction on creating an answer in contrast to selecting an answer might differ between academic and non-academic tracks. Differences between academic and non-academic tracks in instructional quality are caused by differences in teacher training, specific curricula, and dissimilar forms of instruction in German secondary schools (Baumert et al. 2006). For example, curricula of the different school tracks vary in regard to the explicit promotion of productive language skills (DESI Konsortium 2008). Furthermore, previous research on the effects of institutional aspects of tracking suggests that these effects appear in a complex interplay together with effects of compositional aspects of tracking (Baumert et al. 2006; Becker 2009).

The generalizability of the results discussed is limited by several factors. First, the investigation of reading competence was done on the basis of a specific set of items. In order to reveal more differences between the two studied groups, more items and item types could be considered. Information, e.g. in regard to the cognitive demand of the different item formats could allow for even more detailed results. Item types, which could be considered, are especially the newly constructed formats, which were developed in the PISA context for the assessment of complex problem solving competencies or the assessment of digital reading competence. Moreover, students' background information was dichotomized in all cases except for motivation. The dichotomization of constructs such as books possessed may be a threat to the validity of the scales. Second, the categorization of the different tracks was only sketchily. Schools which are classified as belonging to the same track might differ considerably. Moreover, students' background variables were chosen according to frequently used indicators. Nevertheless, results might have varied, if different predictors for example the students' country of birth or parents' profession had been incorporated. Third, in order to identify DIF just one method was employed. As different DIF detection methods might provide somewhat different results, it is not known how the identified DIF effects would change if some other method were used for DIF detection. Finally, the criterion to match students with the same ability inherent to the test under investigation may have already been contaminated by DIF effects, and so it may not be an accurate indicator for the overall proficiency.

Despite the restrictions mentioned, the results of the current study have important implications for standardized testing, particularly in regard to large scale assessments and the promotion of productive language skills in non-academic school tracks. As far as assessment is concerned, the relatively better performance of students in academic track schools in OE items as compared to students in a non-academic track can be seen as a limitation of the validity of the used instruments. To date, a high percentage of assessments includes the application of OE items, and crucial decisions in educational contexts are often made based on the revealed test results. Therefore, increasing knowledge about the nature of the competencies which are assessed by OE items is of great importance. Furthermore, the degree to which these competencies are relevant and valid with regard to what is being assessed should be investigated in future research.

Focusing on short and medium term implications of the findings of this study in the context of large scale reading assessment, a central issue is creating consciousness for the observed specific strength of students attending academic tracks in Germany. In recent debates on test and assessment practices of reading competence rather little attention is given to the aspect of item features (Schroeder and Tiffin-Richards 2014). Our study indicates that although the methodological standards of large scale assessments are outstanding high, there are still aspects, which could be optimized. Being aware of subgroup-dependent differential item difficulties can serve as the starting point for a discussion, which aims at a continuous advancement in terms of measuring competencies. A crucial step in this process is a decision about whether differences depicted by our analyses were intended by the test developers in terms of construct representation. In regard to gender the OECD (2010b) states that the advantage of girls in responding to OE items is intended because their advantage was there since the first PISA wave and minimizing this advantage would result in limiting possibilities for trend analyses. Moreover, OE items are needed to measure higher order reading skills such as reflecting on the entire text and stating an own opinion, which could be hardly assessed by MC items (OECD 2010b).

Our findings implicate an offensive report on differences of subgroups in regard to different item formats. In the long term, the development of new item formats or the optimization of current formats should incorporate knowledge gained about specific strengths of particular subgroups in assessments. From our perspective a great challenge in the context of test construction is the development of adequate MC items, which depict higher order reading skills. This optimization process should take into account efforts that have been already made in the assessment of digital reading in PISA 2012. Improved MC items might offer possibilities for a more fair comparison of the results of academic and non-academic track students in large scale reading assessments. One possibility for handling the observed differences in terms of test interpretation might also be to report two different scores by using multidimensional (at least two-dimensional) models, which has already been suggested by Rauch and Hartig (2010). The authors emphasized that if multidimensional models are intended to be applied, the proportion of OE items should not be too small.

Considering the promotion of productive language skills across all school tracks, our findings provide empirical evidence for the notion that instruction should definitely focus on these kind of language skills in order to make sure that all students get a chance to develop them. Moreover, promotion of productive language skills should start at an early age and continue up to higher grades, as our results indicate deficits of students at the age of 15. Future research is necessary to develop, implement, and evaluate training programs which satisfy the needs of students in all tracks in a convincing manner.

In order to draw a conclusion, it can be said that this study extended previous research findings on specific strengths in reading tests of students attending different school tracks. Considering the implications both concerning standardized assessment and regarding the promotion of productive language skills across all tracks, the study gained scientific knowledge relevant for educational and political decisions and for teaching activities in schools and classrooms.

References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale: Lawrence Erlbaum Associates.

Ariga, K., & Brunello, G. (2007). *Does secondary school tracking affect performance?* Evidence from IALS, IZA Discussion Papers, No. 2643. http://nbn-resolving.de/urn:nbn:de:101:1–20080416197. Accessed: 4th April 2015.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple choice and short answer questions in a marketing context. *Journal of Marketing Education, 25*, 31–36.

Bakker, M., van den Heuvel-Panhuizen, M., & Robitzsch, A. (2014). First-graders' knowledge of multiplicative reasoning before formal instruction in this domain. *Contemporary Educational Psychology, 39*(1), 59–73.

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus [School structure and the genesis of differential environments of learning and development]. In J. Baumert, P. Stanat & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit. Vertiefende Analysen im Rahmen von PISA 2000* (pp. 99–188). Wiesbaden: Springer VS.

Becker, M. (2009). *Kognitive Leistungsentwicklung in differenziellen Lernumwelten: Effekte des gegliederten Sekundarschulsystems in Deutschland [Development of cognitive performance in differential environments of learning: Effects of the tracked system of secondary schools in Germany]*. Berlin: Max-Planck-Institut für Bildungsforschung.

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology, 104*(3), 682–699.

Becker, M., McElvany, N., Lüdtke, O., & Trautwein, U. (2014). Lesekompetenzen und schulische Lernumwelten: Besondere Fördereffekte des Frühübergangs in Gymnasien? [Reading skills and learning environment. *Are there specific effects of early transition to academic tracks?]. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 46*(1), 35–50.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., & Tuerlinckx, F. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1–28.

Bos, W., & Gröhlich, C. (eds.) (2010). *KESS 8: Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen am Ende der Jahrgangsstufe 8 [KESS 8: Skills and attitudes of students in Hamburg's schools at the end of grade 8]*. Münster: Waxmann.

Bos, W., & Scharenberg, K. (2010). Lernentwicklung in leistungshomogenen und -heterogenen Schulklassen [Development of learning in classes with a homogenous and classes with a heterogenous achhievement level]. In W. Bos, E. Klieme, & O. Köller (eds.), *Schulische Lerngelegenheiten und Kompetenzentwicklung: Festschrift für Jürgen Baumert* (pp. 173–194). Münster: Waxmann.

Bos, W., Schwippert, K., & Stubbe, T. C. (2007). Die Koppelung von sozialer Herkunft und Schülerleistung im internationalen Vergleich [The connection between social background and achievement in international comparison]. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, K. Schwippert, & R. Valtin (eds.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 225–247). Münster: Waxmann.

Brenner, A. D., & Crosnoe, R. (2011). The racial/ethnic composition of elementary schools and young children's academic and socioemotional functioning. *American Educational Research Journal, 48*(3), 621–646.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67.

Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., & Underwood, P. S. (2011). Testing the impact of child characteristics × instruction: Interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly, 46*(3), 189–221.

DESI-Konsortium (eds.) (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie [Teaching and competency acquisition in German and English. Results of the DESI study]*. Weinheim: Beltz.

van Ewijk, R., & Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review, 5*, 134–150.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*(2), 113–148.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33*, 69–86.

Guthrie, J. T., & Wigfield, A. (2005). Roles of motivation and engagement in reading comprehension assessment. In S. G. Paris & S. A. Stahl (eds.), *Children's reading comprehension and assessment* (pp. 187–214). Mahwah: Lawrence Erlbaum.

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, 28*, 24–34.

Haberkorn, K., Pohl, S., Hardt, K., & Wiegand, E. (2012). *NEPS technical report for reading—Scaling results of starting cohort 4 in ninth Grade (NEPS Working Paper No. 16)*. Bamberg: Otto-Friedrich-Universität (Nationales Bildungspanel).

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software, 33*(2), 1–22.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York: Routledge.

Hanushek, E. A., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal, 116*, C63–C76.

Kiefer, T., Robitzsch, A., & Wu, M. (2013). *TAM: test analysis modules*. R package version 0.7–35.

Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., & Krauss, S. (2013). Teachers' content knowledge and pedagogical content knowledge: The role of structural differences in teacher education. *Journal of Teacher Education, 64*(1), 90–106.

Klicpera, C., & Gasteiger-Klicpera, B. (1993). *Lesen und Schreiben: Entwicklung und Schwierigkeiten [Reading and writing: Development and difficulties]*. Bern: Huber.

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., & Stanat, P. (eds.) (2010). *PISA 2009. Bilanz nach einem Jahrzehnt [PISA 2009: Result after one decade]*. Münster: Waxmann.

KMK (2014) = Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2014). *Basic structure of the education system in the Federal Republic of Germany*. http://www.kmk.org/fileadmin/pdf/Bildung/AllgBildung/Schema-Bildungsgaenge_und_Schularten-Stand_2014.pdf. Accessed: 4th April 2015.

Köller, O., Schütte, K., Zimmermann, F., Retelsdorf, J., & Leucht, M. (2013). Starke Klasse, hohe Leistungen? Die Rolle der Leistungsstärke der Klasse für die individuellen Mathematik- und Leseleistungen in der Sekundarstufe I [Strong class. *high performances? The role of class performance with regard to individual mathematic and reading performances at secondary level I]. Psychologie in Erziehung und Unterricht, 60*, 184–197.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805–820.

Lafontaine, D., & Monseur, C. (2006). *Impact of test characteristics on gender equity indicators in the assessment of reading comprehension*. Liège: University of Liège.

Lafontaine, D., & Monseur, C. (2009). Gender gap in comparative studies of reading comprehension: To what extent do the test characteristics make a difference? *European Educational Research Journal, 8*(1), 69–79.

Luyten, H., Cremers-van Wees, L. M., & Bosker, R. J. (2003). The Matthew effect in Dutch primary education: Differences between schools, cohorts, and pupils. *Research Papers in Education, 18*, 167–195.

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcome. *Child Development Perspectives, 2*(2), 99–106.

McElvany, N., & Schwabe, F. (2013). Fairness von Lesetestaufgaben für Kinder aus Familien mit unterschiedlichem sozioökonomischem Status bei Large-Scale-Studien [Fariness of large scale reading test items for children with different social backgrounds]. In N. McElvany & H. G. Holtappels (eds.), *Empirische Bildungsforschung—Theorien, Methoden, Befunde und Perspektiven. Festschrift für Wilfried Bos* (pp. 219–234). Münster: Waxmann.

Oakes, J. (1982). The reproduction of inequity: The content of secondary school tracking. *The Urban Review, 14*(2), 107–114.

OECD (2010a) = (Organization for Economic Co-operation and Development). (2010a). *PISA 2009 results: What students know and can do—student performance in reading, mathematics and science (Volume I)*. http://dx.doi.org/ö.pl010.1787/9789264091450-en. Accessed: 4th April 2015.

OECD (2010b) = (Organization for Economic Co-operation and Development). (2010b). *PISA 2009 Assessment framework: Key competencies in reading, mathematics and science*. Pisa: OECD Publishing.

OECD (2012) = (Organization for Economic Co-operation and Development). (2012). *PISA 2009 Technical Report*. PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264167872-en. Accessed: 4th April 2015.

Pae, T.-I. (2004). DIF for examinees with different academic backgrounds. *Language Testing, 21*(1), 53–73.

Pfost, M., Karing, C., Lorenz, C., & Artelt, C. (2010). Schereneffekte im ein- und mehrgliedrigen Schulsystem: Differentielle Entwicklung sprachlicher Kompetenzen am Übergang von der Grund- in die weiterführende Schule [Fan spread effects in a tracked and a nontracked school system. *Is there evidence for differential linguistic competence development at the transition from primary to secondary school?]. Zeitschrift für Pädagogische Psychologie, 24*, 259–273.

Pfost, M., Dörfler, T., & Artelt, C. (2012). Reading competence development of poor readers in a German elementary school sample: An empirical examination of the Matthew effect model. *Journal of Research in Reading, 35*(4), 411–426.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/. Accessed: 4th April 2015.

Rauch, D., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling, 52*(4), 354–379.

Retelsdorf, J., Becker, M., Köller, O., & Möller, J. (2012). Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching. *British Journal of Educational Psychology, 82*, 647–671.

Rjosk, C., Richter, D., Hochweber, J., Lüdtke, O., Klieme, E., & Stanat, P. (2014). Socioeconomic and language minority classroom composition and individual reading achievement: The mediating role of instructional quality. *Learning and Instruction, 32*, 63–72.

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (eds.), *Large scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213–231). Mahwah: Lawrence Erlbaum.

Routitsky, A., & Turner, R. (2003). *Item format types and their influences on cross-national comparisons of student performance*. Chicago: Paper presented at the annual meeting of the American Educational Research Association.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schroeder, S., & Tiffin-Richards, S. P. (2014). Kognitive Verarbeitung von Leseverständnisitems mit und ohne Text [Cognitive handling of reading comprehension test items with and without text]. *Zeitschrift für Pädagogische Psychologie, 28*(1–2), 21–30.

Schwabe, F., McElvany, N., & Trendtel, M. (2015). The school age gender gap in reading achievement: Examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly*, published online first, 1–14.

Schwippert, K., Bos, W., & Lankes, E. M. (2004). Lesen Mädchen anders? Vertiefende Analysen zu Geschlechtsdifferenzen auf Basis der Internationalen Grundschul-Lese-Untersuchung IGLU [Do girls read differently? In-depth analyses of gender differences based on the Progress in International Reading Literacy Study PIRLS]. *Zeitschrift für Erziehungswissenschaft, 7*(2), 219–234.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education, 3*(1), 73–97.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage Publishers.

Snow, C. E. (2002). *Reading for understanding: Toward a research and development program in reading comprehension*. Santa Monica: RAND.

Snow, C. E., Burns, M. S., & Griffin, P. (eds.) (1998). *Preventing reading difficulties in young children*. Washington: National Academy Press.

Solheim, O. J. (2011). Impact of reading self-efficacy and task value on reading comprehension scores in different item formats. *Reading Psychology, 32*, 1–27.

Stanat, P., Rauch, D., & Segeritz, M. (2010). Schülerinnen und Schüler mit Migrationshintergrund [Students with immigrant backgrounds]. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, & P. Stanat (eds.), *PISA 2009. Bilanz nach einem Jahrzehnt* (pp. 200–230). Münster: Waxmann.

Steinert, B., Hartig, J., & Klieme, E. (2008). Institutionelle Bedingungen der Sprachkompetenz [Institutional factors impact on language proficiency]. In E. Klieme (ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch* (pp. 411–448). Weinheim: Beltz.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25*(3), 246–280.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103–118.

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