https://doi.org/10.20378/irbo-51699

Chapter 8

8 Reading Literacy Development in Secondary School and the Effect of Differential Institutional Learning Environments

Maximilian Pfost and Cordula Artelt

Summary

The German secondary school system is characterized by a relatively early separation of students into different types of schools or school tracks that provide different types of curricula in accordance with the prerequisites of the learners. The stratification of the students into the different school tracks is based mainly on student achievement in elementary school, but is also influenced by other factors such as the socioeconomic status or immigration background of the family. As upper academic track schools should provide more favorable developmental conditions with regard to the students' cognitive competencies due to institutional characteristics and school composition effects, pre-existing differences in reading comprehension and vocabulary between the students in the different school tracks should further increase over the course of secondary school. In tracing the development of reading comprehension and vocabulary between Grade 5 and

Author Note

Maximilian Pfost,
Department of Educational Research, University of Bamberg, Germany.

Cordula Artelt,
Department of Educational Research, University of Bamberg, Germany.

This research was supported by grant WE 1478/4-1 & AR 301/9-1 from the German Research Foundation (DFG).

We would like to thank Benjamin Nagengast (University of Tübingen) for his insightful comments on a draft of this chapter.

Correspondence concerning this chapter should be addressed to Maximilian Pfost, Department of Educational Research, University of Bamberg, Markusplatz 3, 96045, Bamberg, Germany. E-mail: maximilian.pfost@uni-bamberg.de

Grade 7 in the current study, results indicated a widening gap between upper, middle, and lower academic track school students' reading comprehension, whereas stable achievement differences in vocabulary were found. A second analysis investigated the effect of attending the different school tracks while controlling for selectivity into the different secondary schools. Results indicated substantial positive effects of attending an upper academic track school in comparison to the lower and middle academic track schools in terms of effect sizes for reading comprehension and vocabulary, though not all results reached statistical significance. Taken together, favorable learning environments seem to support reading literacy development, but the reported findings should be generalized cautiously.

In most German states, students enroll in secondary school when they reach the age of 10 after 4 years of primary education (Cortina, Baumert, Leschinsky, Mayer, & Trommer, 2008; Faust, 2006). The secondary school system in Germany, in contrast to the primary education system, is marked by a strict institutional stratification of students into different types of schools or tracks that go along with distinct school leaving certificates and that provide different learning opportunities to their students. With regard to reading literacy, the transition from primary to secondary school is also marked by different conceptions of schooling and the function of reading. Whereas during primary school, instruction focuses on teaching children to read, over the course of secondary school, students increasingly read to learn (Burns & Kidd, 2010; Chall, 1983). Nevertheless, although explicit instruction in reading is rare and the process of acquiring further reading skills becomes increasingly incidental in the course of secondary school, there is still a generally positive trend in the development of students' reading literacy until students leave school (Hill, Bloom, Black, & Lipsey, 2008; Klicpera, Schabmann, & Gasteiger-Klicpera, 1993). Therefore, it is of critical importance to investigate the role of schools in a secondary school system that is characterized by an explicit between-school tracking for the development of reading literacy.

As mentioned, the German secondary school system separates their students by different types of schools or tracks that provide different types of curricula in accordance with the competencies and prerequisites of the learners. We call this form of organizational differentiation *between-school tracking* or *curricular differentiation by school type* (LeTendre, Hofer, & Shimizu, 2003) in contrast to forms of tracking that take place within schools (e.g., differentiating by courses or streams that can often be found in U.S. high schools). Thereby, the assignment of students to the different types of schools depends primarily on an interplay between decisions made by the primary schools and by the parents (Cortina & Trommer, 2005; Faust, 2005). Over the course of the last year in primary school, the school provides a recommendation for the educational career of the student. This recommendation is primarily based on the student's aptitudes, but also takes into account other prognostic factors (e.g., familial support of the child). The bindingness of this recommendation varies between the federal states, providing different scopes for parents' decision making with regard to the educational careers of their children. In the end, this procedure leads to a separation of the students between the different types of schools according to the students' cognitive abilities but also according to their social and familial backgrounds (Baumert & Köller, 2005; Baumert & Schümer, 2001; Ditton & Krüsken, 2006; Ditton, Krüsken, & Schauenberg, 2005). The rationale behind this institutional separation of students, which Gamoran and Mare (1989) call the *Positive View of Tracking*, is "that students differ in their academic goals and in the environments in which they learn best. Ideally, a system of academic tracking matches students' aptitudes with the objectives and learning environments to which they are best suited" (Gamoran & Mare, 1989, p. 1148). Therefore, a homogenization of the group of students with regard to their ability level should ideally enhance learning for all students (Baumert, 2006). Nevertheless, empirical support for this assumption has been mixed (cf. Ariga & Brunello, 2007; Slavin, 1990).

However, focusing exclusively on the question of the productivity of tracking practices in comparison to nontracking practices on students' learning neglects a second outcome dimension: individual differences or performance inequality between students who attend different tracks. Separating students into different school tracks might, for example, be very effective for students in higher academic tracks, whereas it

might have detrimental effects for students in lower academic tracks. Of course, the opposite could also be true. Students in lower academic tracks might receive the instruction they need to catch up to the achievement level of the higher track students. Therefore, the following two questions require further analysis: How do the cognitive competencies of students who were separated into different academic tracks develop and how would these competencies have developed if the students who were assigned to a certain school track would have been assigned to another track?

Type of School and Causes of Individual Differences in Competence Development

In most German states, the secondary school system is comprised of at least three types of schools or tracks (Cortina, et al., 2008): a *lower academic track* ("Hauptschule") that provides 5 years of basic secondary education, generally preparing students for vocational training; a *middle academic track* ("Realschule"), comprising 6 years of secondary education; and a *higher academic track* ("Gymnasium") that comprises 8/9 years of secondary education and qualifies students for university admission. In addition, some German states run comprehensive secondary schools, offering all three types of school leaving certificates. As different types of schools pursue different academic goals and students are selected into these types of schools primarily according to their cognitive abilities and academic achievement, different learning environments are the result. These school-type-specific environments provide differential developmental possibilities for students based on differential distributional processes of economic, social, and cultural resources; differential institutional working and learning conditions; as well as differential school-type-specific educational and curricular traditions (Baumert, 2006; Baumert, Köller, & Schnabel, 1999; Baumert & Schümer, 2001; Gamoran & Berends, 1987). For example, whereas in lower academic track schools, it is still common to have a form teacher who teaches several or almost all subjects (Leschinsky, 2008a), teachers in middle or upper academic track schools are usually specialized to teach only two or three subjects (Leschinsky, 2008b; Trautwein & Neumann, 2008). In addition, upper academic track teachers tend to have higher levels of content knowledge as well as pedagogical content knowledge (Baumert, et al., 2010). Furthermore, comparing the cultures of instruction, relatively

clear-cut differences between tracks are apparent: In the upper academic track schools, lessons are usually characterized by a high level of cognitive activation and a low level of teacher support, whereas in lower academic track schools, lessons are usually characterized by a high level of teacher support and a low level of cognitive activation (Kunter, et al., 2005). Finally, instruction in lower tracks often seems to proceed more slowly and is conceptually simplified, thereby providing only restricted access to knowledge for students who attend this track (Gamoran & Berends, 1987).

In addition to the thus-far described institutional differences in instruction, the student composition itself might support or handicap learning processes (Baumert, Stanat, & Watermann, 2006; Harker & Tymms, 2004; Pfost, 2011; Zimmer & Toma, 2000). This means that differences in the development of cognitive competencies might be attributable not only to institutional differences in the learning environments, but might also reflect differences in the characteristics of the students within these schools. For example, it has been shown that the proportion of students with an immigration background is negatively linked to the development of the students' reading competence (Pfost, 2011; Stanat, 2006; Walter & Stanat, 2008). Further studies have shown a positive relation between the mean level of achievement and individual reading development (Baumert, et al., 2006; Dreeben & Barr, 1988; Lehmann, 2006) or mathematics (Lehmann, 2006; Opdenakker, van Damme, de Fraine, van Landeghem, & Onghena, 2002; Zimmer & Toma, 2000). Finally, evidence exists for a positive effect of the aggregated mean socioeconomic status on students' academic achievement (Dumay & Dupriez, 2007; Ma & Klinger, 2000; van Ewijk & Sleegers, 2010). As the access to different school tracks is highly selective, institutional differences in the composition of students within schools is the result and may reinforce existing institutional differences in the learning opportunities that are offered. Consequently, different learning rates between students attending different school tracks in secondary school should be expected.

When reviewing differences in the development of cognitive competencies, a third cause of individual differences needs to be taken into account: differential learning rates due to individual characteristics or traits of the students themselves. Therefore, differences in competence development between different school tracks might be attributable to observed and unobserved characteristics that govern the selectivity of

students into the different types of schools. A well-supported fact is that in primary school, students already differ in their school performances, familiar and social backgrounds, as well as expectations concerning future school achievement (Ditton & Krüsken, 2006; Gamoran & Mare, 1989; Maaz, Hausen, McElvany, & Baumert, 2006; Schneider & Stefanek, 2004). For example, parents from different economic and educational backgrounds might apply different strategies such as the utilization of paid private tutoring to realize their educational aspirations and therefore might try to actively influence the selection process into secondary school (Dang & Rogers, 2008; Schneider, 2004). Furthermore, students differ in their prior knowledge when entering secondary school, which might directly result in different learning rates (Renkl, 1996). Within the domain of reading, Stanovich (1986, 2000) describes a model of increasing interindividual differences in reading literacy; he named this the Matthew effect model. Thereby, the cumulative advantages of good readers or the cumulative disadvantages of bad readers are the result of reciprocal self-reinforcing causal processes: "The very children who are reading well and who have good vocabularies will read more, learn more word meanings, and hence read even better. Children with inadequate vocabularies – who read slowly and without enjoyment – read less, and as a result have slower development of vocabulary knowledge, which inhibits further growth in reading ability" (Stanovich, 1986, p. 381). However, empirical studies that have investigated the Matthew effect model in reading have produced mixed results. On the one hand, there is much empirical support from longitudinal studies concerning the reciprocal relation of reading ability, reading motivation, and reading behavior (McElvany, Kortenbruck, & Becker, 2008; Morgan & Fuchs, 2007; Pfost, Dörfler, & Artelt, 2010). On the other hand, studies that have focused on the development of the competence gap between good and poor readers have not yet accumulated convincing evidence which clearly supports a pattern of increasing or a pattern of decreasing differences in reading achievement over time (e.g. Aarnoutse, van Leeuwe, Voeten, & Oud, 2001; Bast & Reitsma, 1998; Kempe, Eriksson-Gustavsson, & Samuelsson, 2011; Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005; Pfost, Dörfler, & Artelt, 2012).

In sum, differences in learning rates between students attending lower, middle, and upper academic track schools are the result of an interplay between individual,

institutional, and school composition factors that may add up, reinforce, or compensate each other over the course of students' individual development.

Achievement Differences and Achievement Growth in Secondary School – Empirical Findings

Cross-sectional studies, especially the four PISA studies run by the OECD between 2000 and 2009 (Baumert, et al., 2001; Klieme, et al., 2010; Prenzel, et al., 2007; Prenzel, et al., 2005), have reported large differences in cognitive competencies between the students who attend different school tracks in Germany. In the most recent PISA study, 15-year-old students attending upper academic track schools on average achieved a reading comprehension score that was more than one and a half standard deviations above the average reading comprehension score of students attending lower academic track schools. Students attending middle academic track schools as well as comprehensive schools reached an average reading comprehension score in between these other two types of schools (Naumann, Artelt, Schneider, & Stanat, 2010). Comparable results have been reported for mathematics and science (Frey, Heinze, Mildner, Hochweber, & Asseburg, 2010; Rönnebeck, Schöps, Prenzel, Mildner, & Hochweber, 2010). Intuitively, we might conclude that these differences are the result of achievement differences prior to secondary school plus different learning rates between school tracks, but cross-sectional studies such as PISA cannot determine the time in the course of development at which differential learning rates appear. Thus, the hypothesis of a widening achievement gap between the different academic tracks needs to be analyzed longitudinally.

Within the domain of mathematics, the assumption of a widening achievement gap has been investigated and verified several times (Becker, Lüdtke, Trautwein, & Baumert, 2006; Köller & Baumert, 2001) with the exception of Schneider and Stefanek (2004), who reported stable mathematics achievement differences between Grade 2 and Grade 11. The reported results from Germany converge well with studies that have investigated the effect of taking advanced courses in U.S. high schools (Gamoran & Mare, 1989; Schmidt, 2009).

Within the domain of reading, however, studies have been less frequent and the results have been more controversial. This might, at least partially, be attributable to differences in the learning opportunities that underlie the development of different cognitive skills (cf., Köller & Baumert, 2008). Whereas for the development of mathematical skills, schools play almost a monopolistic role in the transfer of knowledge, within the domain of reading, further learning opportunities such as leisure time reading (e.g., Pfost, Dörfler, et al., 2010; Spear-Swerling, Brucker, & Alfano, 2010) are of high relevance. Consequently, it might be reasonable to expect that differences in school learning environments might be more related to the development of mathematics than to the development of reading literacy. Retelsdorf and Möller (2008), in analyzing data from the LISA study, reported small but nonsignificant differences in the development of reading literacy from Grade 5 to Grade 6 between lower (d = 0.59), middle (d = 0.62), and upper academic track schools (d = 0.82). Initial differences in reading literacy in Grade 5, when students enter secondary school, however, were already relatively large, with students in the upper academic track scoring on average more than one standard deviation (d = 1.22) above students from the middle academic track and even more than two standard deviations (d = 2.30) above students from the lower academic track. Similar results were presented by Gröhlich, Bonsen, and Bos (2009): In analyzing data from more than 10,000 students from the Hamburg KESS study, the authors reported the highest growth in reading literacy between the end of Grade 4 and Grade 6 for students who attended comprehensive schools (d = 0.47), followed by students who attended lower and middle academic track schools (d = 0.45). The lowest average growth was reported for upper academic track students (d = 0.42). The results confirm the findings from the antecedent LAU study (Lehmann, Peek, Gänsfuß, & Hußfeldt, 1998). Taken together, the results in the domain of reading have been less stringent and have not confirmed the assumption of a widening gap over the course of secondary school.

The question of whether a privileged school learning environment is linked to an increased learning rate was also addressed by the Berlin ELEMENT study (Lehmann & Lenkeit, 2008), which was subsequently reanalyzed by Baumert, Becker, Neumann, and Nikoleva (2009). In the state of Berlin, students have the opportunity to switch to some upper academic track schools ("grundständiges Gymnasium") after Grade 4 or to

stay in a prolonged elementary school and change to secondary school after Grade 6. Students who chose to attend early upper academic track schools after Grade 4 had, in comparison to the students who remained in elementary school, better marks, better reading, and mathematics competencies and came from families with a higher socioeconomic status. Results describing the competence development between Grade 4 and Grade 6 showed, beyond initial differences in reading literacy, a comparable learning rate for students in the two types of schools. With regard to mathematics, students in the early upper academic track school showed an increased learning rate in comparison to the elementary school students. The reanalysis of the data by Baumert et al. (2009), however, focusing on the role of the learning environment on the development of reading and mathematics, did not demonstrate a more favorable learning rate in reading or in mathematics for students in the early upper academic track schools after students' individual characteristics, driving the transition from elementary to early upper academic track school, had been taken into account. Therefore, the hypothesis that a privileged learning environment leads to higher learning rates was not confirmed by this study. Finally, using data from the BiKS study, Pfost, Karing, Lorenz, and Artelt (2010) report a widening achievement gap or fan-spread effect between students attending the lower academic track and the middle as well as upper academic track for reading comprehension, but not vocabulary, between Grade 5 and Grade 6. In addition, a fan-spread effect between students attending different secondary schools was already traceable when students still attended primary school.

Taken together, whereas in the domain of mathematics, fan-spread effects have been demonstrated several times, within the domain of reading, results have been less stringent and have mostly indicated relatively stable achievement differences between different types of schools across the course of secondary school. However, due to the assumption of different learning environments, also fan-spread effects in the domain of reading can be expected.

Research Questions

The current study focused on the following two questions: First, can differences in the development of reading literacy by type of school/school track be found? With regard

to the assumption that upper academic track schools provide a favorable learning environment due to institutional and compositional factors and that students attending upper academic track schools on average have higher cognitive abilities, which should additionally promote further learning, different learning rates in favor of students in upper academic track schools were expected. Furthermore, as lower academic track schools should provide the least favorable learning conditions, the lowest learning rates were expected within this school type. Second, it seemed important to ask whether an effect of attending different types of schools on reading achievement measures could be verified independent of students' characteristics that govern the selectivity into the different secondary school tracks. Again, we expected a favorable effect of attending upper academic track schools in comparison to middle and lower academic track schools, after controlling for important covariates that go along with the choice of a certain track. Due to sample-size restrictions, students from middle and lower academic track schools were grouped together. Therefore, only the effect of attending upper academic track schools in comparison to attending an alternative type of school (middle and lower academic tracks) was estimated.

The current paper extends the findings reported by Pfost, et al. (2010) in at least two ways: at first, data up to Grade 7 was available. Second, the role of covariate selection for the estimation of effects of different institutional learning environments was addressed in more detail.

Method

Design and Participants

All analyses were based on data from the BiKS-8-14 panel study. At the first point of measurement, in the second term of Grade 3, $N = 2{,}395$ students were assessed. After the transition from primary into secondary school, a subsample of $n = 922$ students (38.5% of the original sample) was further followed across secondary school ($n = 268$ in the lower, 188 in the middle, and 466 in the upper academic tracks). Students were selected for further participation in the BiKS-8-14 panel study when they agreed to participate further, when they chose a school within the BiKS inquiry region that had at least one class with at least three participants, and when the school was not

characterized by comprehensive or remedial instruction (cf., Schmidt, Schmitt, & Smidt, 2009). Furthermore, n = 879 secondary school students (n = 102 in the lower, 135 in the middle, and 642 in the upper academic tracks) were additionally recruited in Grade 5 for participation in the BiKS panel study, resulting in a total sample of N = 1,801 secondary school students. Whereas in primary school, data collection took place every half year (Measurement Waves 1, 2, and 3), in secondary school, data were collected annually at the end of each academic year (Measurement Waves 4, 5, and 6). The following analyses focused on the development of measures of reading comprehension and vocabulary between Grade 5 and Grade 7. Additional data from the elementary school years were taken into account for the second set of analyses. The average age of the students was 11.4 years (SD = 0.5) in Grade 5. Furthermore, in our sample, 13.8% of the students lived in households with immigration backgrounds. The gender of the students was almost equally distributed; 47.8% of the students were male and 52.2% were female.

Measures

Students, teachers, and parents were tested on a wide range of measures. In the following section, the measures that were used in the current analysis are presented. At first, the two measures of reading comprehension and vocabulary used in secondary school (Grade 5 to 7) are depicted. Developmental differences between school tracks on these two variables are of major interest in our analyses. Therefore, these two variables are presented in detail. Subsequently, the variables/covariates that were used in the second analysis, in order to control for the selectivity into the different school tracks, are depicted. All covariates were assessed in primary school.

Reading comprehension. In Grade 5, reading comprehension was measured by a sample of six short texts with a total of 43 multiple-choice items developed by the BiKS research group. For the reading comprehension test, the students had to read a given text, search relevant information, and generate more or less high inferences from the text to answer the given items. In Grade 6, three texts with a total of 31 multiple-choice items were used. Finally, in Grade 7, again, three texts with a total of 26 multiple-choice items were used. For the three waves of measurement, a common item design with a nonequivalent groups/anchor-item test design was applied (Holland, Dorans, &

Peterson, 2007; Kolen & Brennan, 2004), allowing the estimation of students' reading comprehension on a common metric within an IRT framework. Therefore, for all reading comprehension test items, the item difficulty parameters were estimated with a three-dimensional 1-parameter Rasch model by using the ConQuest software package (Wu, Adams, Wilson, & Haldane, 2007). A design matrix was specified and the item difficulty parameters of the three waves of measurement were estimated in a single simultaneous run (concurrent estimation). Item difficulty parameters for the same items across different waves of measurement were set equal. Subsequently, individual students' abilities were estimated in a second run by weighted likelihood estimates (WLEs) for every wave of measurement using the item difficulty parameters of the concurrent estimation. Missing responses were treated as incorrect during the item calibration stage as well as during the estimation of the person parameters. The estimated individual ability scores were conclusively T-standardized ($M = 50$, $SD = 10$) in Grade 5. The reliabilities (WLE-reliability) of the reading comprehension measures were satisfactory for all waves of measurement ($\text{Reliability}_{\text{Grade 5}} = .78$, $\text{Reliability}_{\text{Grade 6}} = .77$, $\text{Reliability}_{\text{Grade 7}} = .76$).

Vocabulary. Students' vocabulary was measured by a set of 35 items from the subscale V1 (Vocabulary) of the *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision* (KFT 4-12 + R; Heller & Perleth, 2000). Additional vocabulary items that were used in Grade 7 were disregarded in the present analysis in order to keep the metric constant. Ceiling effects were negligible as still in Grade 7 the maximum test score was reached by just one student of the sample. For every item, a target word as well as a selection of four additional words was presented for reading. Students had to indicate the word whose definition best matched the presented target word. Students' vocabulary was estimated by summing the number of correct answers. For ease of interpretation, students' vocabulary scores were also T-standardized ($M = 50$, $SD = 10$) in Grade 5 by a linear transformation. The internal consistency (Cronbach's α) of the vocabulary test was satisfactory for the three waves of measurement ($\alpha_{\text{Grade 5}} = .78$, $\alpha_{\text{Grade 6}} = .80$, $\alpha_{\text{Grade 7}} = .78$).

Covariates. Socioeconomic and ethnic-cultural backgrounds. Data concerning students' socioeconomic and ethnic-cultural backgrounds were collected in a highly standardized telephone interview in the first and third waves of measurement in Grade

3 and Grade 4 of elementary school. In order to determine students' immigration backgrounds, parents were asked questions concerning their cultural origin. Students were classified as having an immigration background when at least one parent was born in a foreign country. Furthermore, the parents were asked questions concerning their familial, educational, as well as occupational status. With this information, the highest ISEI (International Socio-economic Index of Occupational Status; Ganzeboom, De Graaf, & Treiman, 1992) and educational level of the parents was determined.

Cultural capital. Parents were asked to specify the number of books they had at home. The responses were categorized by the interviewers. Categories ranged from 1 (*not one*) to 7 (*more than 500*).

Extracurricular reading behavior. Students' habitual extracurricular reading behavior was assessed by a single item ("Does [the name of the child] read for pleasure?") in the parental telephone interview in Grade 4. Parents rated the frequency of their children's reading behavior on a 4-point Likert-type scale with the response options 1 (*almost never or never*), 2 (*rarely*), 3 (*yes, several times a week*), and 4 (*yes, everyday*).

Reading self-concept. Students' reading self-concept was assessed by a single item ("How good are you in school in... reading?") in the students' questionnaire in Grade 4. Students rated their reading self-concept on a 4-point Likert-type scale ranging from 1 (*bad*) to 4 (*very good*).

Vocabulary. In Grade 4, students' vocabulary was measured by a set of 30 items from the supplementary vocabulary test of the culture fair intelligence test (CFT 20, german version: Weiß, 1987).

Mathematics competence. Students' mathematics competence in Grade 4 was measured by a selection of 19 items from the DEMAT 4 (Gölitz, Roick, & Hasselhorn, 2005).

Spelling. Spelling was measured in Grade 4 by using 21 items from the DRT 4 (Grund, Haug, & Naumann, 2003).

General cognitive abilities. Students' general cognitive abilities were assessed in Grade 4 with a set of 15 items from the matrices subtest of the culture fair intelligence test (CFT 20-R, german version: Weiß, 2006).

Reading comprehension. In Grade 4, reading comprehension was measured by a sample of 13 short texts with 20 multiple-choice items from the subscale text comprehension of the ELFE 1-6 (Lenhard & Schneider, 2005). The test was prolonged by adding three new texts with six multiple-choice items developed by the authors to avoid ceiling effects.

Grades. Information concerning the students' grades after the first term of Grade 4 was provided by the class teachers. In Germany, grades range from 1 (*excellent*) to 6 (*insufficient*).

Analytic Strategy

The first set of analyses addressed the question of whether differences in the development of reading comprehension and vocabulary between students attending different types of schools could be demonstrated. In order to test for developmental differences, difference scores for reading comprehension and vocabulary, using models of true intraindividual change (cf. Geiser, 2010; Steyer, Eid, & Schwenkmezger, 1997), were computed (Figure 1). The type of school was used as a grouping variable. As there was only one indicator of reading comprehension or vocabulary available for each wave of measurement, a latent achievement indicator was not estimated. Consequently, the measurement error of the manifest variables was set to zero. The initial unconstrained model was just identified, fitting the data perfectly. To test for differences between groups, mean change scores between different types of schools were set to be equal and compared to the model without this constraint. All multigroup models of difference scores were estimated with M*plus* 6.1 (Muthén & Muthén, 1998-2010). In order to take the nested data structure into account, the *type is complex* option was used. Although an MLR estimator was used, the chi-square value for testing the constrained model against the alternative, unconstrained (just-identified) model was not corrected as there was not yet a routine within M*plus* for doing this when missing data were replaced by multiple imputation.[1] The analyses were run two times. In the first analysis, students were grouped according to the type of school that these students attended in Grade 5. Changes in the school type between Grade 5 and Grade 7 that

[1] cf. M*plus* Discussion board, posting by Linda K. Muthén on 16th June 2006 on http://www.statmodel.com/discussion/messages/22/381.html [17th March 2012].

may have occurred were ignored. In the same way, students who had to repeat a class between Grade 5 and Grade 7 were treated as though they had advanced in the normal manner. In these two cases (change of school and grade repetition), test information of Grade 6 and/or Grade 7 was almost never available, and the achievement scores were imputed. To support the interpretation of our results, the models describing differences in reading comprehension and vocabulary development were reanalyzed in a second set of analyses, considering only students who were still actively participating in the study in Grade 7, who did not change their type of school, and who did not repeat a class during the time period under investigation.

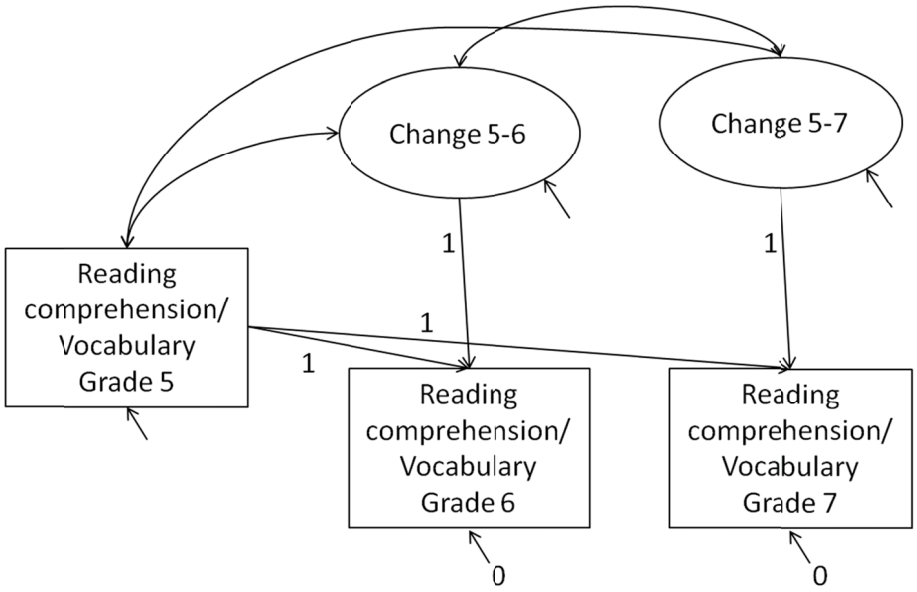

Figure 1. Specified difference score model for reading comprehension/vocabulary. The model is specified as baseline model.

The second set of analyses addressed the question of whether an effect of attending different types of schools or school tracks on the development of reading comprehension and vocabulary could be verified independent of individual characteristics influencing the selectivity into the secondary school system. To address this research question, we used the reduced subsample of students in secondary school for whom test information, including inter alia measures of reading comprehension and vocabulary from the elementary school years, was available. In order to disentangle

institutional from individual effects, interindividual differences between students prior to their secondary school attendance needed to be adequately controlled. One of the most efficient tools for estimating treatment effects (e.g., the effect of attending different types of schools) in nonexperimental studies is Propensity-Score-Matching (PSM). In general, matching methods within observational studies aim to equate a distribution of covariates in treatment and control groups by drawing students from both groups who are similar on a set of observed covariates (Rosenbaum & Rubin, 1985; Stuart, 2010). Matching methods often come into operation when causal inferences about treatment effects in observational designs are of particular interest (c.f. Morgan & Winship, 2007; Rubin, 1997; West & Thoemmes, 2010). PSM traditionally comprises two analytical steps: First, for every student, the probability of being in either the treatment (TG) or the control group (CG) is calculated on the basis of the covariates that are taken into account. In the present analysis, attending an upper academic track school comprised the treatment condition and lower or middle academic track schools the control condition. In the current analysis, the following covariates were considered: the state where the school was located (dummy coded: 0 = Hesse, 1 = Bavaria), students' age and sex (dummy coded: 0 = female, 1 = male), parents' education (dummy coded: 0 = parents did not reach university entrance qualification, 1 = parents reached university entrance qualification), students' immigration background (dummy coded: 0 = no immigration background, 1 = students have an immigration background), parents' HISEI, cultural capital of the parents (the categories were dummy coded), students' time spent in extracurricular reading (the categories were dummy coded), students' reading self-concept (the categories were dummy coded), and Grade 4 achievement measures of vocabulary, mathematics, spelling, general cognitive abilities, and reading comprehension. Only linear effects of the covariates were considered. In the second matching analysis, in addition to the already denoted variables, students' grades after the first term of Grade 4 in mathematics and German were taken into account. As denoted, students' grades from the first term of Grade 4 were directly linked to the choice of school track. However, school grades are often not comparable to each other due to different applied reference scales (Maaz, et al., 2008; Trautwein, Lüdtke, Becker, Neumann, & Nagy, 2008; Treutlein & Schöler, 2009) and should therefore be treated and interpreted with caution.

On the basis of these variables, a probit score which indicates a student's probability of attending the upper academic track school (TG) given that student's covariates was estimated. Then, students in the two groups were matched to each other on the basis of the calculated probit score using radius matching (see Dehejia & Wahba, 2002; Morgan & Winship, 2007). Therefore, for each treatment case control cases were selected that were located within a particular distance – the radius – of the calculated propensity score. In cases in which more than one control student was located within the maximum acceptable distance around the treatment group student, the selected control cases were given equal weights. The radius was set at $\delta = 0.005$. Treatment cases that did not have a possible counterpart within the control cases were said to be off the support and were not considered for further analysis. The same was true for control cases without possible counterparts from the treatment cases. Therefore, the interpretability of the treatment effect was limited to those for whom possible counterparts existed (common-support treatment effect for the treated). In other words, the estimated average effect of attending an upper academic track school (TG), in comparison to attending lower or middle academic track schools (CG), on the development of reading comprehension and vocabulary is only informative with regard to those students who typically attend an upper academic track school and for whom comparable counterparts who attend lower and middle academic track schools exist. As mentioned, students attending lower and middle academic track schools were grouped together because of their small sample size. After the matching procedure, balance with respect to the incorporated covariates and the overlap between the two groups was checked. Therefore, the standardized differences of the covariates between the two treatment groups before and after the matching procedure were computed. In the final step, the analysis of the outcomes, differences in reading comprehension and vocabulary in Grade 7 between the matched groups were tested. Propensity-Score-Matching was done with STATA 11 using the psmatch2 routine (Leuven & Sianesi, 2003).

Missing data. Missing data is a typical problem of research in the social sciences, especially in longitudinal studies. In the current study, missing data may have occurred on the one hand because parents did not give consent for their child to participate in the study. What is known from the literature is that active informed

parental consent is related to factors such as the degree of deviant behavior of the students, students' scholastic performance, and the social and ethnic backgrounds of families (Courser, Shamblen, Lavrakas, Collins, & Ditterline, 2009; Esbensen, et al., 1996; Esbensen, Hughes Miller, Taylor, He, & Freng, 1999; Unger, et al., 2004). On the other hand, parents may have given their informed consent but students might not have been present on the testing day, might not have correctly answered the questions, or may have left the study after a certain amount of participation (dropout). Study dropout in particular may be a sign of educational problems such as repeating a year or changing school type, and therefore needs to be treated cautiously (van de Grift, 2009). In other words, treatment-related attrition may be a serious threat to the internal validity of the estimated results (West & Thoemmes, 2010). In the first analysis, the data of all secondary school students in schools in which competence measurement took place and for whom parental consent was present were included in the analysis. Missing data on measures of reading comprehension and vocabulary were replaced by multiple imputation (m = 5) using a broad set of auxiliary variables. Multiple imputation was implemented by using an R script by Robitzsch (personal communication, March 18, 2011) controlling the imputation with Partial Least Squares regression within MICE (van Buuren & Oudshoorn, 2000). In order to verify the results of the first descriptive analysis, a second descriptive analysis was run by which, again, a dataset to which multiple imputation was applied was used, but the analysis was restricted to students who were still actively participating in the study in Grade 7, who did not change their type of school, and who did not repeat a class during the time period under investigation. We will denote this reduced sample as the "active sample" as students were still actively participating in the study in Grade 7. Finally, an EM algorithm that applied single imputation was used on the covariates that were used in the Propensity-Score-Matching. Although single imputation does not seem to be an adequate strategy in outcome analyses, it seems to be a sufficient and effective approach in the context of Propensity-Score-Matching (Stuart, 2010). The propensity score matching analysis was run exclusively using the active subsample of n = 658 students, for whom data from the primary school years were available and who were still active participants in the BiKS-8-14 longitudinal study in Grade 7.

Results

Developmental Differences in Reading Comprehension and Vocabulary

In order to trace interindividual differences in the development of reading comprehension and vocabulary, difference scores based on models of true intraindividual change were computed. The models were specified as baseline models, allowing for the analysis of differences in changes in reading comprehension and vocabulary between Grade 5 and Grade 6 (Change 6-5) as well as Grade 5 and Grade 7 (Change 7-5). A graphical illustration of the development of reading comprehension and vocabulary by type of school for the entire sample of secondary school students is depicted in Figures 2 and 3. The corresponding estimated results are presented in Table 1.

Table 1. Reading Comprehension and Vocabulary Development by School Track

	Grade 5 *M (SD)*	Grade 6 *M (SD)*	Grade 7 *M (SD)*	Change 5-6 *M (SD)*	Change 5-7 *M (SD)*
	Reading comprehension				
Lower academic track	40.47 (8.47)	41.98 (9.16)	43.80 (11.31)	1.51 (10.30)	3.33 (11.25)
Middle academic track	47.60 (7.77)	50.49 (9.41)	50.93 (11.80)	2.90 (8.88)	3.34 (11.26)
Upper academic track	53.90 (8.58)	58.21 (11.36)	60.26 (13.97)	4.32 (10.61)	6.36 (12.83)
Full sample	50.01 (10.00)	53.49 (12.45)	55.20 (14.74)	3.49 (10.32)	5.20 (12.34)
Test of significance[a]	$p < .01$[b]			$p < .01$	$p < .01$
	Vocabulary				
Lower academic track	40.84 (8.81)	45.13 (9.98)	50.22 (8.83)	4.29 (8.65)	9.38 (8.96)
Middle academic track	47.03 (7.92)	52.20 (9.53)	54.93 (9.10)	5.16 (8.27)	7.89 (8.95)
Upper academic track	53.92 (8.50)	58.54 (8.20)	61.09 (7.29)	4.62 (7.47)	7.17 (8.15)
Full sample	50.00 (10.00)	54.65 (10.35)	57.75 (9.14)	4.65 (7.88)	7.75 (8.52)
Test of significance[a]	$p < .01$[b]			*ns*	$p < .01$

Note. Sample size was $n = 370$ students in lower academic track schools, $n = 323$ in middle academic track schools, and $n = 1{,}108$ students in upper academic track schools.
[a]It was tested whether estimates were equal between students attending lower, middle and upper academic track schools.
[b]Mplus Type is General was used as Grade 5 reading comprehension/vocabulary was treated as manifest.

First, results indicated large differences in reading comprehension in Grade 5 between students in the different school tracks. Students attending upper academic track schools on average achieved the highest reading comprehension score, whereas students in the lower academic track schools achieved the lowest. Furthermore,

significant differences in the development of reading comprehension between different school tracks were found: Between Grade 5 and Grade 6, students in the upper academic track schools showed the largest increase in reading comprehension, followed by students attending middle academic track schools. The smallest increase was measured in the group of lower academic track students.[2] A model constraint representing equal average reading comprehension development between the three type of schools was significant ($\Delta\chi^2 = 12.212$, $df = 2$, $p < .01$), indicating that developmental differences between school tracks are of statistical relevance. Regarding the development of reading comprehension for the full 2-year period between Grade 5 and Grade 7, we still found a clear statistically significant difference between students in the different school tracks ($\Delta\chi^2 = 22.458$, $df = 2$, $p < .01$). Again, students attending upper academic track schools showed the highest learning rate in comparison to lower and middle academic track students. The average learning rate of students attending lower academic track schools was comparable in size to the learning rate of the middle academic track students.

[2] Due to the application of a different scaling and imputation procedure as well as the usage of different analytic models, the reported growth rates may slightly vary from the results reported by Pfost, Karing, Lorenz, and Artelt (2010).

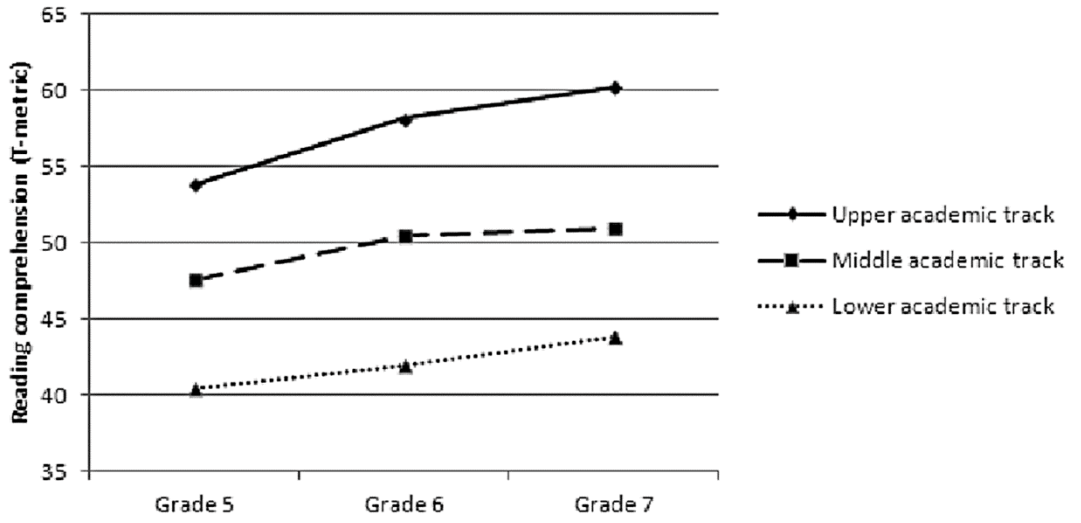

Figure 2. Development of reading comprehension by type of school. Estimates are based on the full sample of secondary school students (cf. Table 1 for corresponding data).

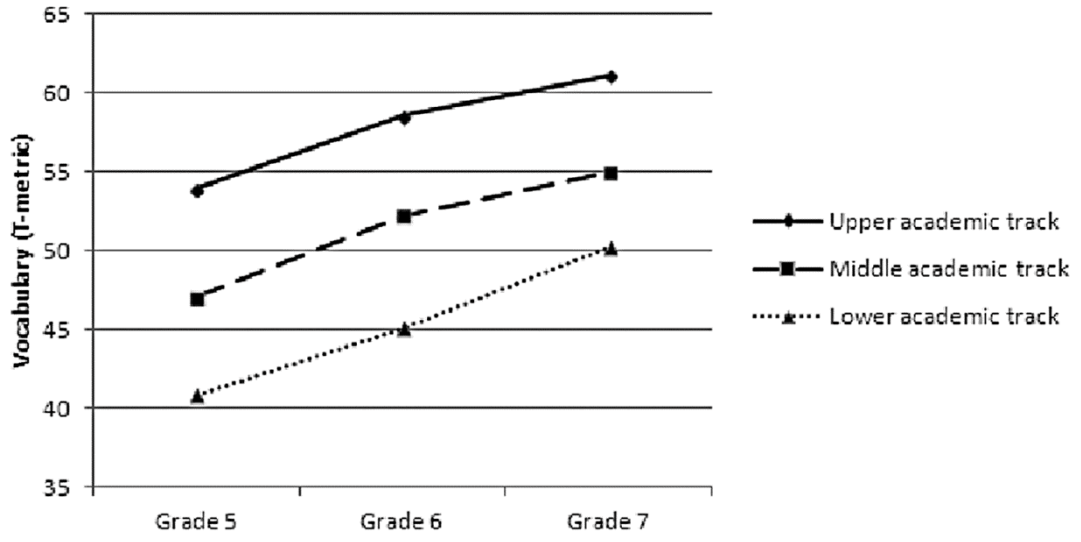

Figure 3. Development of vocabulary by type of school. Estimates are based on the full sample of secondary school students (cf. Table 1 for corresponding data).

Regarding vocabulary, again, strong interindividual differences in Grade 5 between students attending the different types of schools were present. When tracing the development of vocabulary between Grade 5 and Grade 6, no differences in the learning rate between students attending different types of schools were found ($\Delta\chi^2 = 1.220$, $df = 2$, ns). However, when analyzing the long-term development of vocabulary between Grade 5 and Grade 7, significant differences occurred ($\Delta\chi^2 = 10.144$, $df = 2$, $p < .01$). Interestingly, the developmental pattern was different from the one found for reading comprehension. Whereas for reading comprehension, the highest learning rate was found for students attending upper academic track schools; for vocabulary, the highest learning rate was found for students attending lower academic track schools. This means that lower academic track students caught up to the performance of the better performing middle and upper academic track students who were comparable in their learning rates.

In summary, results based on the full sample of secondary school students provide evidence for a widening gap or fan-spread effect for reading comprehension between students attending different school tracks, whereas with regard to the development of vocabulary, the opposite seems true: On average, students attending lower academic track schools showed the largest gains in vocabulary, whereas the smallest gains were found for upper academic track students.

Then, the same two difference score models for reading comprehension and vocabulary were estimated, but analyses were restricted to the sample of students who were still actively participating in the BiKS study in Grade 7, who did not change their type of school, and who did not have to repeat a class. This restriction reduced the sample size by $n = 443$ (24.6%) students, leading to an effective sample size of $n = 1{,}358$ (75.4% of the full sample) students. The reduced or active sample was composed of $n = 196$ (formerly $n = 370$; 53.0%) lower academic track students, $n = 267$ (formerly $n = 323$; 82.7%) middle academic track students, and $n = 895$ (formerly $n = 1{,}108$, 80.8%) upper academic track students. The estimated model results for the active sample are presented in Table 2.

Table 2. Reading Comprehension and Vocabulary Development by School Track (Active Sample)

	Grade 5 *M (SD)*	Grade 6 *M (SD)*	Grade 7 *M (SD)*	Change 5-6 *M (SD)*	Change 5-7 *M (SD)*
	Reading comprehension				
Lower academic track	40.47 (8.70)	42.37 (8.61)	43.75 (11.42)	1.90 (9.79)	3.28 (10.96)
Middle academic track	48.06 (7.57)	50.94 (9.09)	51.61 (11.65)	2.88 (8.78)	3.56 (11.38)
Upper academic track	54.57 (8.52)	59.51 (11.12)	61.76 (13.69)	4.94 (10.55)	7.19 (12.76)
Full sample	51.25 (9.80)	55.35 (12.17)	57.17 (14.66)	4.10 (10.19)	5.91 (12.38)
Test of significance [a]	$p < .01$ [b]			$p < .01$	$p < .01$
	Vocabulary				
Lower academic track	40.87 (8.78)	45.33 (10.08)	49.67 (8.95)	4.46 (8.50)	8.81 (7.92)
Middle academic track	47.53 (7.76)	52.67 (9.15)	55.30 (9.10)	5.14 (8.10)	7.77 (8.67)
Upper academic track	54.85 (7.99)	59.65 (7.59)	62.06 (6.74)	4.80 (7.21)	7.21 (7.68)
Full sample	51.39 (9.58)	56.21 (9.82)	58.94 (8.90)	4.82 (7.59)	7.55 (7.94)
Test of significanc [a] e	$p < .01$ [b]			*ns*	*ns*

Note. The estimates refer to students who were still actively participating in the BiKS study in Grade 7, who did not change their type of school, and who had not repeated a class during the time period under investigation (active sample). Sample size was $n = 196$ students in lower academic track schools, $n = 267$ in middle academic track schools, and $n = 895$ students in upper academic track schools.
[a]It was tested whether estimates were equal between students attending lower, middle and upper academic track schools.
[b]Mplus Type is General was used as Grade 5 reading comprehension/vocabulary was treated as manifest.

In comparison to the estimated results for the full sample (cf. Table 1), the estimations for the active sample (cf. Table 2) differed in two ways: First, the overall reading comprehension and vocabulary levels were about one tenth of a standard deviation higher in the reduced, active sample than in the full sample. This may be due to two causes. On the one hand, dropout was higher in lower academic track schools than in middle and upper academic track schools. On the other hand, especially within the upper academic track schools, students with lower achievement levels tended to drop out more often. Second, whereas in the first set of analyses, significant differences in the development of vocabulary between Grade 5 and Grade 7 between school tracks were found, analyses based on the active sample did not confirm this result ($\Delta\chi^2 = 3.543$, $df = 2$, *ns*). This difference might be attributable at least in part to a lower estimated vocabulary gain between Grade 5 and Grade 7 for students attending lower academic track schools in the active sample in comparison to the complete sample that included student dropouts. With regard to the development of reading comprehension, significant developmental differences in favor of students attending upper academic

track schools were found, confirming the results of the first analysis that was based on the data of all secondary school students.

The Effect of Institutional Differences in Learning Environment on the Development of Reading Comprehension and Vocabulary

In order to test whether differences in the development of reading comprehension and vocabulary could be attributed to institutional differences in the learning environment, the selectivity of the students into the different school types had to be taken into account. Analyses were restricted to a subsample of n = 658 students, for whom information – inter alia test data – from the elementary school years was available and who were still active study participants in Grade 7 (active subsample). The developmental trends for reading comprehension and vocabulary for this longitudinal subsample of active secondary school students were comparable to the developmental trends for the full sample of active secondary school students (the full sample comprised also students that were not tested in primary school; cf. Tables 2 and 3).

Table 3. Reading Comprehension and Vocabulary Development by School Track (Active Elementary-Secondary-School Longitudinal Subsample)

	Grade 5 *M (SD)*	Grade 6 *M (SD)*	Grade 7 *M (SD)*	Change 5-6 *M (SD)*	Change 5-7 *M (SD)*
	Reading comprehension				
Lower academic track	40.27 (8.92)	42.20 (8.71)	42.76 (10.98)	1.92 (10.16)	2.48 (10.85)
Middle academic track	47.10 (7.42)	50.29 (9.47)	50.60 (12.12)	3.19 (9.19)	3.50 (11.67)
Upper academic track	53.71 (8.39)	58.13 (10.88)	61.34 (13.64)	4.43 (10.52)	7.64 (13.28)
Full sample	49.42 (9.88)	53.05 (11.99)	55.05 (14.89)	3.63 (10.21)	5.63 (12.67)
Test of significance [a]	$p < .01$ [b]			*ns*	$p < .01$
	Vocabulary				
Lower academic track	40.79 (8.98)	44.86 (10.43)	49.29 (9.44)	4.07 (8.34)	8.51 (8.03)
Middle academic track	47.06 (7.86)	51.92 (10.03)	54.86 (9.76)	4.86 (8.51)	7.79 (9.02)
Upper academic track	54.34 (7.70)	59.37 (7.33)	61.59 (6.99)	5.04 (7.40)	7.25 (8.01)
Full sample	49.88 (9.72)	54.68 (10.48)	57.51 (9.63)	4.80 (7.87)	7.63 (8.25)
Test of significance [a]	$p < .01$ [b]			*ns*	*ns*

Note. The estimates refer to the subsample of all secondary school students for whom data from the elementary school years were available. Furthermore, students were still actively participating in the BiKS study in Grade 7, did not change their type of school, and had not repeated a class during the time period under investigation (active sample). Sample size was n = 136 students in lower academic track schools, n = 150 in middle academic track schools, and n = 372 students in upper academic track schools.
[a]It was tested whether estimates were equal between students attending lower, middle and upper academic track schools.
[b]Mplus Type is General was used as Grade 5 reading comprehension/vocabulary was treated as manifest.

Due to unequal sample sizes of the students attending different school tracks in the current sample and the special interest in the effect of attending upper academic track schools, in which the curriculum has a strong focus on preparing students for university entrance, in comparison to lower and middle academic track schools, which both mainly focus on preparing students for vocational training, students attending the lower and middle academic track schools were combined into one comparison group. Therefore, the analyses that were conducted by using Propensity-Score-Matching (PSM) focused on the estimation of the effect of attending an upper academic track school in comparison to attending lower or middle academic track schools between Grade 5 and Grade 7 on the development of reading comprehension and vocabulary. A broad set of covariates was used in order to adequately control for the treatment assignment. Radius matching with caliper was used as the matching procedure.

The distribution of the estimated propensity scores for students attending the lower and middle academic track schools (the controls) and students attending upper academic track schools is depicted in Figure 4 (without taking mathematics and German grades into account) and Figure 5 (after additionally taking mathematics and German grades into account). A graphical inspection of Figure 4 indicates that the distribution of propensity scores for students attending the lower and middle academic track schools was highly positive or right-skewed, whereas the distribution of the propensity scores of the upper academic track students was highly negative or left-skewed. Nevertheless, the figure also indicates that in between the two peaks, a relatively large region of overlap between the two distributions was present. Therefore, we expected a satisfactory number of comparable students for the matching procedure in the two groups and a good extrapolation with regard to the interpretation of the estimated results. By contrast, regarding the distribution of the propensity scores in Figure 5, when additionally considering mathematics and German grades of the students in Grade 4, it becomes obvious that the region of overlap decreased substantially. This can be seen by the lower number of students of the two groups who fell into the middle region or region of overlap when comparing Figure 5 with Figure 4. This effect is mainly attributable to the fact that in the state of Bavaria in particular, school choice is almost directly linked to the students' grades in Grade 4. Therefore, estimations of the effect of attending an upper academic track school in comparison to

lower and middle academic track schools that take students' mathematics and German grades into account might be less affected by systematic biases due to unconsidered covariates but at the price of a lower extrapolation of the results to a larger population of students.

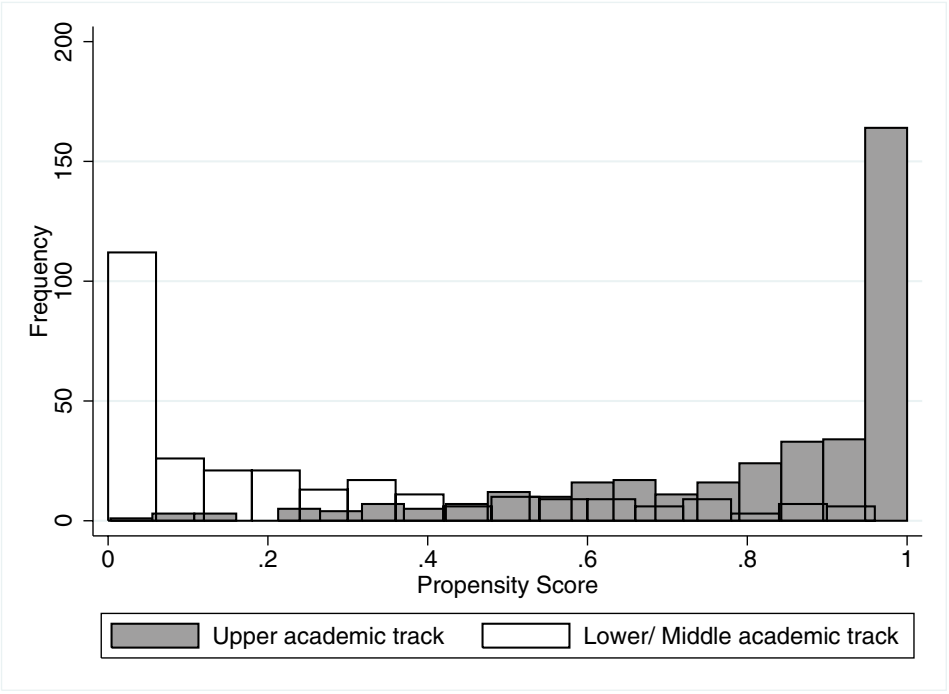

Figure 4. Distribution of propensity scores by school track without taking grades into account. Before matching, active sample: *M*(Upper academic track students) = 0.817; *M*(Lower/Middle academic track students) = 0.239; Standardized Difference = 234.1%; After radius matching: *M*(Upper academic track students) = 0.709; *M*(Lower/Middle academic track students) = 0.708; Standardized Difference = 0.1%.

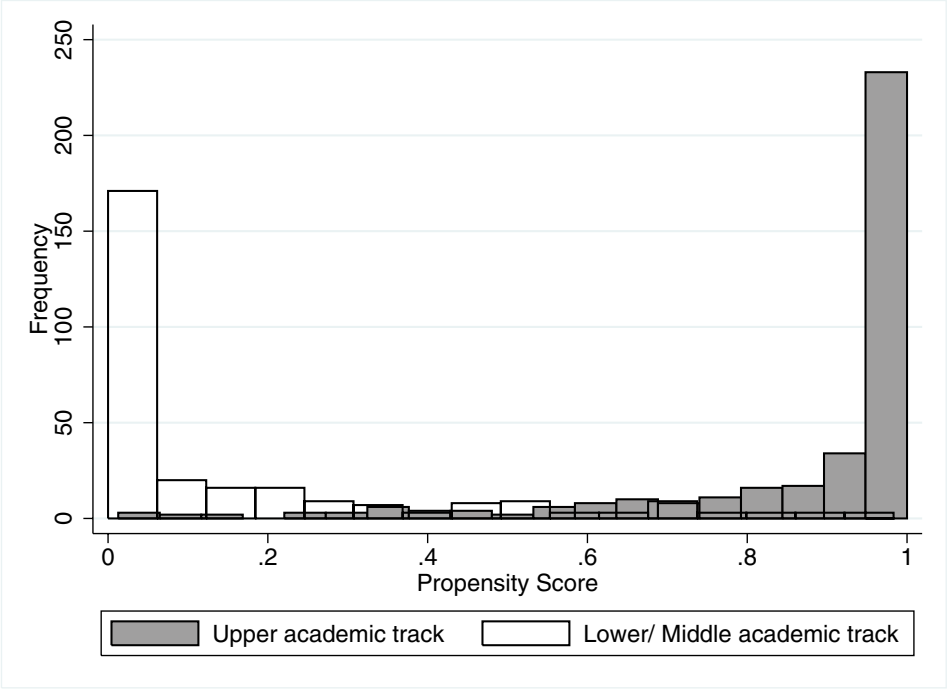

Figure 5. Distribution of propensity scores by school track after taking grades into account. Before matching, active sample: *M*(Upper academic track students) = 0.882; *M*(Lower/Middle academic track students) = 0.154; Standardized Difference = 326.9%; After radius matching: *M*(Upper academic track students) = 0.757; *M*(Lower/Middle academic track students) = 0.757; Standardized Difference = 0.0%.

In the next step, the balance with regard to the covariates between the two groups before and after the matching procedure was checked (Table 4). In the unmatched full sample, the estimates clearly indicated marked differences in the characteristics of the students who entered the upper academic track schools in comparison to the students who entered the lower and middle academic track schools (first column). Students attending upper academic track schools on average came more often from the federal state of Hesse, were younger, had better educated parents, came from families possessing more economic and cultural capital, read more in their leisure time, had a higher reading self-concept, and performed better on a wide range of achievement tests (vocabulary, mathematics, spelling, general cognitive abilities, and reading comprehension) in Grade 4 of elementary school. Finally, large differences in the German and mathematics grades in Grade 4 were present. After the first matching procedure, differences between the two groups of students were reduced substantially on most variables. However, some significant differences, especially on the categorical dummy-coded variables and the immigration background of the students remained,

reflecting problems due to the small sample size in combination with large differences on several characteristics between students attending different school tracks. Furthermore, substantial differences in the German, mathematics, and science grades in Grade 4 remained, as these three variables were not included as covariates in the matching procedure.

Table 4. Covariate Imbalance in Unmatched and Matched Samples

Factor	Before matching[1]	Matched, without grades[1]	Matched, grades included[16]
State (1 = Bavaria)[2]	-48.3**	-14.9	-27.5*
Sex (1 = male)[2]	-13.0	-2.9	2.2
Age	-41.8**	7.2	-0.2
Education Parents [23]	117.0**	6.7	-11.8
Immigration (1 = immigration background) [2]	10.7	20.6*	22.4*
HISEI	104.1**	-7.9	-16.8
Cultural capital category 3[2]	-48.4**	9.7	0.2
Cultural capital category 4[2]	-28.3**	-11.1	21.7
Cultural capital category 5[2]	-16.8*	9.3	-4.6
Cultural capital category 6[2]	28.4**	-23.3*	-22.4
Cultural capital category 7[2]	51.8**	16.6	11.4
Reading behavior category 2[24]	-17.5*	4.1	-16.8
Reading behavior category 3[24]	-25.8**	-0.4	4.3
Reading behavior category 4[24]	-25.6**	8.0	21.6*
Reading self-concept category 2[2]	-24.2**	-16.8	-10.3
Reading self-concept category 3[2]	-35.8**	22.2*	24.7
Reading self-concept category 4[2]	51.8**	-15.0	-19.4
Vocabulary	101.1**	-5.3	-15.9
Mathematics competence	87.4**	12.7	22.4
Spelling	114.2**	-15.6*	-8.6
General cognitive abilities	63.9**	1.3	-4.1
Reading comprehension	100.4**	-8.3	8.3
Mathematics grades[5]	-134.9**	-73.8**	6.5
German grades[5]	-193.8**	-96.3**	1.8
Science grades[5]	-137.0**	-56.0**	9.1
Mean value[7]	64.9	18.6	12.6

Note. Standardized differences in percent (%). Formula from Rosenbaum and Rubin (1985).
[1] In general, a positive algebraic sign indicates a higher mean value in the treatment group (= upper academic school track); Results were computed using pstest implemented in psmatch2 (Leuven & Sianesi, 2003).
[2]The variable was dummy-coded.
[3]1 = parents reached university entrance qualification.
[4]Reading behavior was negatively keyed from category 1 = yes, every day to 4 = never or almost never;
[5]In Germany, grades are negatively keyed ranging from 1 = excellent to 6 = insufficient; the negative algebraic sign therefore indicates better (= lower) grades in the treatment group (= upper academic track).
[6]German and mathematics grades were included in the PSM; Science grades were not included as this led to severe imbalances on further covariates.
[7]All differences were treated as positive values.
* $p < .05$. ** $p < .01$.

The analyses of the outcome variables for the unmatched and matched samples, without taking school grades into account, are presented in Table 5. The results indicate that even after adjusting for a broad set of covariates, significant differences remained in reading comprehension and vocabulary between students attending upper academic track schools and students attending lower or middle academic track schools. For reading comprehension, the estimated effect of attending 3 years of an upper academic track school was about $d = 0.33$ in the matched sample. With regard to the development of vocabulary, an effect of $d = 0.34$ was estimated. The effect just missed the 5% significance level, but the sample size had been substantially reduced due to the matching. However, when considering German and mathematics grades in Grade 4 as additional covariates, the results changed (Table 6). Whereas in the first matching, substantial differences in the matched groups in German, mathematics, and science grades were still present, the second analysis also achieved a satisfactory balance on these three covariates (Table 4). However, the balance on most other covariates was less satisfactory. Furthermore, as already mentioned, the number of students within the region of common support and to whom the analyses referred decreased substantially after the inclusion of the German and mathematics grades (from $n = 351$ to $n = 170$; cf. Figures 4 and 5). With regard to the outcome – the development of reading comprehension – the estimated average treatment effect for the treatment group was $d = 0.48$. For the second outcome – vocabulary – the results of the radius matching did not indicate a significant difference between school types ($d = 0.31$).

Table 5. Reading Comprehension and Vocabulary in Grade 7 by School Track Before and After Matching

Outcome	Effect	*M* (upper academic track)	*M* (lower academic track)	*Diff.*	*SE*	*Diff/ SE*	*d*	Grade 4 *d*
Reading comprehension	Unmatched	61.343	46.873	14.470	1.064	13.595**	0.97	0.91
	Matched	58.052	53.129	4.923	2.177	2.261*	0.33	-0.08
Vocabulary	Unmatched	61.588	52.211	9.378	0.718	13.052**	0.97	0.92
	Matched	60.694	57.427	3.267	1.696	1.926	0.34	-0.05

Note. Grades were not included as covariates in the matching. Sample size was $n = 658$ students in the unmatched and $n = 351$ students in the Radius matched sample. *SD*(Reading comprehension, Grade 7) = 14.902; *SD*(Vocabulary, Grade 7) = 9.640.
* $p < .05$. ** $p < .01$.

Taken together, the results of the Propensity-Score-Matching analyses indicate a substantial positive effect of attending 3 years of an upper academic track school in comparison to lower and middle academic track schools. The estimated size of this effect varied from around $d = 0.3$ to $d = 0.5$ for reading comprehension as well as vocabulary. As mentioned, the selection process of attending the upper, middle, or lower academic tracks was, at least in the regions from where the present sample stemmed, strongly determined by the Grade 4 grades. However, grades are difficult to compare across different schools and classes, so taking these measures into account as covariates in the matching process might go along with imbalances on additional unobserved variables.

Table 6. Reading Comprehension and Vocabulary in Grade 7 by School Track Before and After Matching (incl. grades as covariates)

Outcome	Effect	*M* (upper academic track)	*M* (lower academic track)	*Diff.*	*SE*	*Diff/ SE*	*d*	Grade 4 *d*
Reading comprehension	Unmatched	61.343	46.873	14.470	1.064	13.595**	0.97	0.91
	Matched	59.850	52.633	7.218	3.400	2.123*	0.48	0.07
Vocabulary	Unmatched	61.588	52.211	9.378	0.718	13.052**	0.97	0.92
	Matched	60.855	57.899	2.956	2.749	1.075	0.31	-0.14

Note. Grades were considered as covariates in the matching procedure. Sample size was $n = 658$ students in the unmatched and $n = 170$ students in the Radius matched sample. *SD*(Reading comprehension, Grade 7) = 14.902; *SD*(Vocabulary, Grade 7) = 9.640.
* $p < .05$. ** $p < .01$.

Discussion

With regard to the first research question, the question of whether differences in the development of reading comprehension and vocabulary between different types of schools or school tracks could be found, the analyses showed a widening gap between students attending upper, middle, and lower academic track schools in reading comprehension between Grade 5 and Grade 7. Furthermore, the effect of increasing differences in reading comprehension was demonstrated independently of the treatment of student dropout in the analytic model. Therefore, the developmental pattern of reading comprehension in the first years of elementary school fits well with the notion of a fan-spread effect and converges well with results that have been

reported in the domain of mathematics (Becker, et al., 2006; Köller & Baumert, 2001; Schmidt, 2009) but contrast with findings often reported in reading (Gröhlich, et al., 2009; Lehmann, et al., 1998; Retelsdorf & Möller, 2008).

In the domain of vocabulary, the findings did not support the assumption of a widening gap between different types of schools. Furthermore, results differed slightly by the different treatment of student dropout: Analyses that ignored student dropout by imputing all missing values indicated a small, although significant catch-up effect for students attending lower academic track schools, whereas analyses that excluded all students who were no longer participating in the last wave of measurement found stable differences in vocabulary between the three different school tracks. When taking a closer look at the differences between the estimated values of these two analyses, we see that the subsample of the "survivors" (students who still active participate in the study in Grade 7) in general scored higher on measures of reading comprehension and vocabulary than the full sample, indicating that lower competence is linked to an increased probability of student dropout. Furthermore, this tendency was moderated by the school track: Whereas student dropout was almost not or only slightly positively linked to achievement measures in lower academic track schools, student dropout was negatively linked to achievement differences in middle and upper academic track schools. These differences might be attributable to characteristics of the school system: Whereas in upper academic track schools, students can change only to a less demanding school type, students in lower academic track schools can additionally change to more demanding school types. Taken together, the vocabulary gap between students staying in the different school tracks (and therefore still active participating in the BiKS-study) seemed to remain stable. Slightly higher vocabulary trends however were estimated for students leaving the lower track (and therefore in most cases dropping-out of the study), indicating the need for further research dedicated to the analyses of developmental trends for students changing school track.

But why did differences in vocabulary remain more or less stable, whereas differences in reading comprehension between school tracks tend to increase with time? There are at least two explanations for this result. According to a technical explanation, differences in the development of reading comprehension and vocabulary might be an artifact of different test characteristics. Tests might differ in their sensitivity to detect

changes in the latent trait. The second explanation, an educational explanation, assumes that differences in the learning mechanisms are responsible for these developmental differences. Whereas vocabulary knowledge may be mostly acquired subconsciously by processes of incidental learning (Krashen, 1989), the fostering of reading comprehension may still be explicitly due to instruction in school. As a consequence, measures of reading comprehension should be more sensitive to between-school differences due to institutional differences in the content and quality of instruction. Nevertheless, this explanation is only partially supported by the findings of the second set of analyses, which will be discussed next.

What is the Effect of Attending an Upper Academic Track School on Learning?

Tracing interindividual differences in learning between different school tracks does not instantaneously mean that these differences are the product of different learning environments. Rather, differences in learning rates between different types of schools or school tracks might arise from the interplay of institutional characteristics with differences in the composition of the students and the individual traits and abilities of the students that already exist prior to the attendance of secondary school (Ditton & Krüsken, 2006; Pfost, Karing, et al., 2010; Schneider & Stefanek, 2004). Disentangling these different sources is of special scientific interest, but creating experimental conditions in which students can be randomly assigned to different school tracks is not feasible. The BiKS study, however, provides analytic possibilities for addressing this question because data on the students who attend different secondary school tracks are available, and these data have already been measured in elementary school (prior to the treatment exposure). To make use of this favorable circumstance in the current study, Propensity-Score-Matching as a tool for analyzing treatment effects in nonequivalent treatment groups was applied. In order to control for selectivity into the different secondary schools, a broad number of factors, including achievement measures from Grade 4, which might influence students' school choice or the outcome, were taken into account as covariates. Students' school grades in German and mathematics in the middle of Grade 4 were considered in an additional analysis, but their use went along with the loss of a broad number of matches. Furthermore, school grades are often not directly comparable beyond classes, schools, and regions because teachers are

inveigled into using different reference scales (Maaz, et al., 2008; Trautwein, et al., 2008; Treutlein & Schöler, 2009). Science grades were not included as an additional covariate. A model that included the grades of all three main subjects (German, mathematics, and science) led to a strong imbalance on most covariates and was therefore not considered. Although not included as a covariate, differences in science grades between the different school tracks were nevertheless substantially reduced by the applied Propensity-Score-Matching.

The results of the matching analyses that had not taken school grades into account as a covariate indicated a positive effect of attending an upper academic track school on the development of reading comprehension and vocabulary (the effect for vocabulary slightly missed the 5% significance level but was still substantial in terms of effect size). Regarding the magnitude of the effect on reading comprehension and vocabulary across a 3-year period, from the end of Grade 4 to Grade 7, students in upper academic track schools gained about one third of a standard deviation more than we expected that they would have learned when attending lower and middle academic track schools (the estimated counterfactual outcome). When taking grades in mathematics and German into account as further covariates, this positive significant effect of attending an upper academic track school on learning did not change substantially for reading comprehension. For vocabulary there was as strong increase in the standard error, so the effect was far away from reaching statistical significance although just marginally changing in terms of effect size. This means that although the null hypothesis of equal development between the matched pairs who attended different school tracks could not be rejected, differences in the sample that were not negligible in size remained. Comparing this cumulative 3-year effect to an empirical benchmark indicated that the emerging difference between the end of Grade 4 and Grade 7 in our sample was comparable to the normative change we would expect in the domain of reading from at least a half year of schooling (Bloom, Hill, Black, & Lipsey, 2008; Hill, et al., 2008).

So, taken together, what do the results of the matching analysis tell us? First, results need to be interpreted against the background of the assumptions underlying the analysis. As long as unobserved or unconsidered covariates that influence the treatment assignment as well as the treatment outcome and that have not been blocked by conditioning on the considered covariates are present, results may be systematically

biased. In the current study, we tried to map the process of selecting a certain school track by taking a set of prominent covariates into account. Nevertheless, it should be acknowledged that the real process of selecting a certain type of school might be much more complex than assumed in the present analyses. And second, the role of school grades as a confounding factor between school choice and competence development beyond objective achievement measures, measures of the economic, ethnic, and familial background of the students, as well as further individual characteristics of students need further investigation. Thereby, we should ask about the appropriateness of using measures such as school grades that differ in meaning between subjects due to differential context conditions.

Limitations

Analyzing the development of reading literacy in the different school tracks is a sensitive topic that needs to be treated cautiously. Analyses are sensitive to the subjects who are considered. Student dropout in longitudinal studies may occur for meaningful reasons such as a change in school type, moving to another city, the repetition of a grade, and so on (van de Grift, 2009). Therefore, in the analysis of fan-spread effects the treatment of missing values may become a central theme that has to be taken into account. In our first model, reading comprehension and vocabulary development were analyzed under the assumption that no change in the type of school occurred during the period under investigation. All missing values regardless of participation status were estimated by multiple imputation. However, we should keep in mind that student dropout was quite substantial, as only 1,358 out of 1,801 (75.4%) secondary school students participated in Grade 7 (additionally, for 120 participating students, competence measures were missing in Grade 7). Imputation of such large amounts of missing data might be critical and might explain by itself the differences found in estimated growth when compared to the students who were still actively participating. Consequently, the same analysis was run by considering only the students who were still present in Grade 7 – the active sample (N = 1,358). Nevertheless, both approaches neglected the dynamic character of the students who remained but also changed schools. Additionally, the present analyses were limited to students whose parents decided to actively participate in the BiKS study (active informed consent). Within the

BiKS study, students with an immigration background as well as students with higher (i.e., worse) grades were underrepresented in the sense that these students (i.e., their parents) more frequently actively or passively refused to participate in the study (cf. Pfost, 2011). Therefore, the current sample was not fully representative of all students from the participating schools or of all students in the federal states of Bavaria and Hesse.

Another limitation of this study concerns the measurement and scaling of reading comprehension. In the current study, reading comprehension was measured by using different items at different waves of measurement in combination with items that were presented to the students a second time (common item design with nonequivalent groups/ anchor-item test design: Holland, et al., 2007; Kolen & Brennan, 2004), and students' reading comprehension was estimated on a common metric by using a logit-link function within an IRT framework. However, equating across grade levels (vertical scaling) in particular may produce different results depending on the equating methodology used in combination with substantial equating error, particularly when assumptions of the measurement model are not met (Wu, 2010). A new presentation of identical test material, as practiced in the domain of vocabulary, does not necessarily solve scaling problems and may create additional problems such as memory effects. Thus, in summary, as long as we do not have natural metrics, research findings may be substantially biased by scaling artifacts (Embretson, 2006).

Finally, it should be noted that Propensity-Score-Matching is only a weak alternative for the analysis of treatment effects in comparison to randomized experiments. PSM can adjust only for observed confounding covariates, whereas randomization tends to balance the distribution of all covariates, observed and unobserved (Rubin, 1997). Therefore, the estimated effects of attending an upper academic track school in comparison to lower or middle academic track schools can be interpreted only against the background of covariates that were taken into account and for which balance between the matched samples could be achieved. Furthermore, the estimated results can only be interpreted as a narrower treatment effect, the common-support treatment effect for the treated (Morgan & Winship, 2007). This means that, even if the assumption of conditional ignorability was true in the present case, the estimated effect refers only to those students who typically get the treatment, which means

students who would typically choose an upper academic track school and for whom valid counterparts in the control condition could be found. Or, in simpler terms, the estimates refer primarily to those students for whom the choice of type of school after Grade 4 was not perfectly determined by their performance, ethnic or social background, and so forth. Further discussion and assumptions concerning the causal interpretability of estimated results in observational studies are presented in Morgan and Winship (2007), Rubin (1986, 2004), Shadish (2010), and West and Thoemmes (2010). To conclude, although estimations of the effect of attending different school tracks on the development of reading comprehension and vocabulary tried to take into account a broad set of potential confounding variables that have been observed in the BiKS study in combination with up-to-date analytical methods, all estimated results should be interpreted with great caution and after reflecting upon the underlying assumptions.

Implications for Future Research

Tracing the development of cognitive competencies in different types of schools or school tracks with observational studies is a very sensitive topic. Therefore, future research should devote more resources toward further improving studies with regard to the measures used, the scaling techniques applied, and the sample selected for observation. On the other hand, estimating the effect of attending different school tracks on the development of cognitive competencies does not tell us anything about the mechanisms that mediate these effects. Therefore, beyond asking how successful schools are in promoting the cognitive development of students, we further need to ask why these differences occur. And finally, we may be interested in the question of the fit between the type of school and student characteristics. Effects of attending different school tracks may vary for different subpopulations of students, a topic that needs further attention in future research.

References

Aarnoutse, C., van Leeuwe, J., Voeten, M., & Oud, H. (2001). Development of decoding, reading comprehension, vocabulary and spelling during the elementary school years. *Reading and Writing, 14*, 61-89. doi:10.1023/A:1008128417862

Ariga, K., & Brunello, G. (2007). *Does secondary school tracking affect performance? Evidence from IALS*. Bonn: Forschungsinstitut zur Zukunft der Arbeit.

Bast, J., & Reitsma, P. (1998). Analyzing the development of individual differences in terms of matthew effects in reading: Results from a dutch longitudinal study. *Developmental Psychology, 34*, 1373-1399. doi:10.1037/0012-1649.34.6.1373

Baumert, J. (2006). Was wissen wir über die Entwicklung von Schulleistungen? *Pädagogik, 58*, 40-46.

Baumert, J., Becker, M., Neumann, M., & Nikolova, R. (2009). Frühübergang in ein grundständiges Gymnasium – Übergang in ein privilegiertes Entwicklungsmilieu? Ein Vergleich von Regressionsanalyse und Propensity Score Matching. *Zeitschrift für Erziehungswissenschaft, 12*, 189-215. doi:10.1007/s11618-009-0072-4

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, J. M., et al. (Eds.). (2001). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Baumert, J., & Köller, O. (2005). Sozialer Hintergrund, Bildungsbeteiligung und Bildungsverläufe im differenzierten Sekundarschulsystem. In V. Frederking, H. Heller & A. Scheunpflug (Eds.), *Nach PISA. Konsequenzen für Schule und Lehrerbildung nach zwei Studien* (pp. 9-21). Wiesbaden: VS Verlag für Sozialwissenschaften.

Baumert, J., Köller, O., & Schnabel, K. (1999). Schulformen als differentielle Entwicklungsmilieus – eine ungehörige Fragestellung? Erwiderung auf die Expertise "Zur Messung sozialer Motivation in der BIJU-Studie" von Georg Lind. In G. E. u. Wissenschaft (Ed.), *Messung sozialer Motivation: Eine Kontroverse* (pp. 28-69). Frankfurt am Main: Bildungs- und Förderwerk der Gewerkschaft Erziehung und Wissenschaft im DGB e. V.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*, 133-180. doi:10.3102/0002831209345157

Baumert, J., & Schümer, G. (2001). Schulformen als selektionsbedingte Lernmilieus. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann & M. Weiß (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 454-467). Opladen: Leske + Budrich.

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus. In J. Baumert, P. Stanat & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungssystem* (pp. 95-188). Wiesbaden: VS Verlag für Sozialwissenschaften.

Becker, M., Lüdtke, O., Trautwein, U., & Baumert, J. (2006). Leistungszuwachs in Mathematik. Evidenz für einen Schereneffekt im mehrgliedrigen Schulsystem? *Zeitschrift für Pädagogische Psychologie, 20*, 233-242. doi:10.1024/1010-0652.20.4.233

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*, 289-328. doi:10.1080/19345740802400072

Burns, M. S., & Kidd, J. K. (2010). Learning to read. In P. Peterson, E. Baker & B. McGaw (Eds.), *International Encyclopedia of Education* (3 ed., pp. 394-400). Oxford, United Kingdom: Elsevier.

Chall, J. S. (1983). *Stages of reading development*. New York: McGraw-Hill.

Cortina, K. S., Baumert, J., Leschinsky, A., Mayer, K. U., & Trommer, L. (Eds.). (2008). *Das Bildungswesen in der Bundesrepublik Deutschland. Strukturen und Entwicklungen im Überblick*. Reinbek bei Hamburg: Rowohlt.

Cortina, K. S., & Trommer, L. (2005). Bildungswege und Bildungsbiographien in der Sekundarstufe 1. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer & L. Trommer (Eds.), *Das Bildungswesen in der Bundesrepublik Deutschland* (pp. 342-391). Reinbek bei Hamburg: Rowohlt.

Courser, M. W., Shamblen, S. R., Lavrakas, P. J., Collins, D., & Ditterline, P. (2009). The impact of active consent procedures on nonresponse and nonresponse error in youth survey data. Evidence from a new experiment. *Evaluation Review, 33*, 370-395. doi:10.1177/0193841X09337228

Dang, H.-A., & Rogers, F. H. (2008). *How to interpret the growing phenomen of private tutoring: Human capital deepening, inequality increasing, or waste of resources?* Policy Research Working Paper 4530. The World Bank. Retrieved from http://www-wds.worldbank.org/servlet/WDSContentServer/WDSP/IB/2008/02/25/000158349_20080225153509/Rendered/PDF/wps4530.pdf

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics, 84*, 151-161.

Ditton, H., & Krüsken, J. (2006). Der Übergang von der Grundschule in die Sekundarstufe I. *Zeitschrift für Erziehungswissenschaft, 9*, 348-372. doi:10.1007/s11618-006-0055-7

Ditton, H., Krüsken, J., & Schauenberg, M. (2005). Bildungsungleichheit – der Beitrag von Familie und Schule. *Zeitschrift für Erziehungswissenschaft, 8*, 285-304. doi:10.1007/s11618-005-0138-x

Dreeben, R., & Barr, R. (1988). Classroom Composition and the design of instruction. *Sociology of Education, 61*, 129-142.

Dumay, X., & Dupriez, V. (2007). *Does the school composition effect matter? Some methodological and conceptional considerations*. Les Cahiers de Recherche en Éducation et Formation, (60). Université Catholique de Louvain, Louvain-la-Neuve.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61,* 50-55. doi:10.1037/0003-066X.61.1.50

Esbensen, F. A., Deschenes, E. P., Vogel, R. E., West, J., Arboit, K., & Harris, L. (1996). Active parental consent in school-based research. An examination of ethical and methodological issues. *Evaluation Review, 20,* 737-753. doi:10.1177/0193841X9602000605

Esbensen, F. A., Hughes Miller, M., Taylor, T. J., He, N., & Freng, A. (1999). Differential attrition rates and active parental consent. *Evaluation Review, 23,* 316-335. doi:10.1177/0193841X9902300304

Faust, G. (2005). Übergänge in den Sekundarbereich. In W. Einsiedler, M. Götz, H. Hacker, J. Kahlert, R. W. Keck & U. Sandfuchs (Eds.), *Handbuch Grundschulpädagogik und Grundschuldidaktik* (pp. 291-296). Bad Heilbrunn: Klinkhardt.

Faust, G. (2006). Zum Stand der Einschulung und der neuen Schuleingangsstufe in Deutschland. *Zeitschrift für Erziehungswissenschaft, 9,* 328-347. doi:10.1007/s11618-006-0054-8

Frey, D., Heinze, A., Mildner, D., Hochweber, J., & Asseburg, R. (2010). Mathematische Kompetenz von PISA 2003 bis PISA 2009. 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. 153-176). Münster: Waxmann.

Gamoran, A., & Berends, M. (1987). The effects of stratification in secondary schools: synthesis of survey and ethnographic research. *Review of Educational Research, 57,* 415-435. doi:10.3102/00346543057004415

Gamoran, A., & Mare, R. D. (1989). Secondary school tracking and educational inequality: compensation, reinforcement, or neutrality? *American Journal of Sociology, 94*, 1146-1183.

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A Standard International Socio-Economic Index of Occupational Status. *Social Science Research, 21*, 1-56.

Geiser, C. (2010). *Datenanalyse mit Mplus. Eine anwendungsorientierte Einführung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Gölitz, D., Roick, T., & Hasselhorn, M. (2005). Deutsche Mathematiktests für dritte und vierte Klassen (DEMAT 3+ und DEMAT 4). In M. Hasselhorn, H. Marx & W. Schneider (Eds.), *Diagnostik von Mathematikleistungen* (Vol. Band 4, pp. 167-198). Göttingen: Hogrefe.

Gröhlich, C., Bonsen, M., & Bos, W. (2009). Von KESS 4 zu KESS 7: Lernentwicklung in der Beobachtungsstufe. In W. Bos, M. Bonsen & C. Gröhlich (Eds.), *KESS 7. Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7* (pp. 91-122). Münster: Waxmann.

Grund, M., Haug, G., & Naumann, C. L. (2003). *DRT 4. Diagnostischer Rechtschreibtest für 4. Klassen*. Göttingen: Beltz.

Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School effectiveness and school improvement, 15*, 177-199. doi:10.1076/sesi.15.2.177.30432

Heller, K. A., & Perleth, C. (2000). *KFT 4-12+ R. Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision*. Göttingen: Hogrefe.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172-177. doi:10.1111/j.1750-8606.2008.00061.x

Holland, P. W., Dorans, N. J., & Peterson, N. S. (2007). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics 26. Psychometrics* (pp. 169-203). Amsterdam: Elsevier.

Kempe, C., Eriksson-Gustavsson, A.-L., & Samuelsson, S. (2011). Are there any Matthew effects in literacy and cognitive development. *Scandinavian Journal of Educational Research, 55*, 181-196. doi:10.1080/00313831.2011.554699

Klicpera, C., Schabmann, A., & Gasteiger-Klicpera, B. (1993). Lesen- und Schreibenlernen während der Pflichtschulzeit: Eine Längsschnittuntersuchung über die Häufigkeit und Stabilität von Lese- und Rechtschreibschwierigkeiten in einem Wiener Schulbezirk. *Zeitschrift für Kinder- und Jugendpsychiatrie, 21*, 214-225.

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., et al. (Eds.). (2010). PISA 2009. Bilanz nach einem Jahrzehnt. Münster: Waxmann.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking. Methods and Practices*. New York, NY: Springer.

Köller, O., & Baumert, J. (2001). Leistungsgruppierungen in der Sekundarstufe 1. Ihre Konsequenzen für die Mathematikleistung und das mathematische Selbstkonzept der Begabung. *Zeitschrift für Pädagogische Psychologie, 15*, 99-110. doi:10.1024//1010-0652.15.2.99

Köller, O., & Baumert, J. (2008). Entwicklung schulischer Leistungen. In R. Oerter & L. Montada (Eds.), *Entwicklungspsychologie* (pp. 735-768). Weinheim: Beltz.

Krashen, S. (1989). We acquire vocabulary and spelling by reading: additional evidence for the input hypothesis. *The Modern Language Journal, 73*, 440-464.

Kunter, M., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum, W., et al. (2005). Der Mathematikunterricht der PISA-Schülerinnen und Schüler. *Zeitschrift für Erziehungswissenschaft, 8*, 502-520. doi:10.1007/s11618-005-0156-8

Lehmann, R. H. (2006). Zur Bedeutung der kognitiven Heterogenität von Schulklassen für den Lernstand am Ende der Klassenstufe 4. In A. Schründer-Lenzen (Ed.), *Risikofaktoren kindlicher Entwicklung. Migration, Leistungsangst und Schulübergang*. (pp. 109-121). Wiesbaden: VS Verlag für Sozialwissenschaften.

Lehmann, R. H., & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis. Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin*. Berlin: Senatsverwaltung für Bildung, Jugend und Sport.

Lehmann, R. H., Peek, R., Gänsfuß, R., & Hußfeldt, V. (1998). *LAU 7. Aspekte der Lernausgangslage und der Lernentwicklung - Klassenstufe 7. Ergebnisse einer längsschnittlichen Untersuchung in Hamburg*. Hamburg: Behörde für Bildung und Sport.

Lenhard, W., & Schneider, W. (2005). *ELFE 1-6: Ein Leseverständnistest für Erst- bis Sechstklässler*. (1. Auflage ed.). Göttingen: Hogrefe.

Leschinsky, A. (2008a). Die Hauptschule - von der Be- zur Enthauptung. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer & L. Trommer (Eds.), *Das Bildungswesen in der Bundesrepublik Deutschland* (pp. 377-406). Reinbek bei Hamburg: Rowohlt.

Leschinsky, A. (2008b). Die Realschule - ein zweischneidiger Erfolg. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer & L. Trommer (Eds.), *Das Bildungswesen in der Bundesrepublik Deutschland* (pp. 407-436). Reinbek bei Hamburg: Rowohlt.

LeTendre, G. K., Hofer, B. K., & Shimizu, H. (2003). What is tracking? Cultural expectations in the United States, Germany, and Japan. *American Educational Research Journal, 40*, 43-89. doi:10.3102/00028312040001043

Leuven, E., & Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing (Version 4.0.4). Retrieved from http://ideas.repec.org/c/boc/bocode/s432001.html

Ma, X., & Klinger, D. A. (2000). Hierarchical linear modelling of student and school effects on academic achievement. *Canadian Journal of Education, 25*, 41-55.

Maaz, K., Hausen, C., McElvany, N., & Baumert, J. (2006). Stichwort: Übergänge im Bildungssystem. Theoretische Konzepte und ihre Anwendung in der empirischen Forschung beim Übergang in die Sekundarstufe. *Zeitschrift für Erziehungswissenschaft, 9*, 299-327. doi:10.1007/s11618-006-0053-9

Maaz, K., Neumann, M., Trautwein, U., Wendt, W., Lehmann, R. H., & Baumert, J. (2008). Der Übergang von der Grundschule in die weiterführende Schule: die Rolle von Schüler- und Klassenmerkmalen beim Einschätzen der individuellen Lernkompetenz durch die Lehrkräfte. *Schweizerische Zeitschrift für Bildungswissenschaften, 30*, 519-548.

McElvany, N., Kortenbruck, M., & Becker, M. (2008). Lesekompetenz und Lesemotivation. Entwicklung und Mediation des Zusammenhangs durch Leseverhalten. *Zeitschrift für Pädagogische Psychologie, 22*, 207-219. doi:10.1024/1010-0652.22.34.207

Morgan, P. L., & Fuchs, D. (2007). Is there a bidirectional relationship between children's reading skills and reading motivation? *Exceptional children, 73*, 165-183.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference. Methods and principles for social research*. Cambridge, NY: Unversity Press.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide*. (6 ed.). Los Angeles, CA: Muthén & Muthén.

Naumann, J., Artelt, C., Schneider, W., & Stanat, P. (2010). Lesekompetenz von PISA 2000 bis PISA 2009. 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. 23-71). Münster: Waxmann.

Opdenakker, M.-C., van Damme, J., de Fraine, B., van Landeghem, G., & Onghena, P. (2002). The effect of schools and classes on mathematics achievement. *School effectiveness and school improvement, 13*, 399-427. doi:10.1076/sesi.13.4.399.10283

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J.-E., & Kirby, J. (2005). Development of individual differences in reading: results from longitudinal studies in English and Finnish. *Journal of Educational Psychology, 97*, 299-319. doi:10.1037/0022-0663.97.3.299

Pfost, M. (2011). *Klassenkompositionseffekte in der Sekundarstufe und Prüfung vermittelnder Mechanismen*. Unveröffentlichte Dissertationsschrift, Otto-Friedrich-Universität Bamberg.

Pfost, M., Dörfler, T., & Artelt, C. (2010). Der Zusammenhang zwischen außerschulischem Lesen und Lesekompetenz. Ergebnisse einer Längsschnittstudie am Übergang von der Grund- in die weiterführende Schule. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42*, 167-176. doi:10.1026/0049-8637/a000017

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*, 411-426. doi:10.1111/j.1467-9817.2010.01478.x

Pfost, M., Karing, C., Lorenz, C., & Artelt, C. (2010). Schereneffekte im ein- und mehrgliedrigen Schulsystem. Differenzielle Entwicklung sprachlicher Kompetenzen am Übergang von der Grund- in die weiterführende Schule? *Zeitschrift für Pädagogische Psychologie, 24*, 259-273. doi:10.1024/1010-0652/a000025

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., et al. (Eds.). (2007). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie.* Münster: Waxmann.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R. H., Leutner, D., Neubrand, M., et al. (Eds.). (2005). *PISA 2003. Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche?* Münster: Waxmann.

Renkl, A. (1996). Vorwissen und Schulleistung. In J. Möller & O. Köller (Eds.), *Emotionen, Kognitionen und Schulleistung* (pp. 175-190). Weinheim: Psychologie Verlags Union.

Retelsdorf, J., & Möller, J. (2008). Entwicklungen von Lesekompetenz und Lesemotivation. Schereneffekte in der Sekundarstufe? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 40*, 179-188.

Rönnebeck, S., Schöps, K., Prenzel, M., Mildner, D., & Hochweber, J. (2010). Naturwissenschaftliche Kompetenz von PISA 2006 bis PISA 2009. 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. 177-198). Münster: Waxmann.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistican, 39*, 33-38. doi:10.2307/2683903

Rubin, D. B. (1986). Which ifs have causal answers. *Journal of the American Statistical Association, 81*, 961-962.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*, 757-763.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observatinal studies. *Journal of Educational and Behavioral Statistics, 29*, 343-367. doi:10.3102/10769986029003343

Schmidt, S., Schmitt, M., & Smidt, W. (2009). *Die BiKS-Studie. Methodenbericht zur zweiten Projektphase*. Bamberg: Otto-Friedrich-Universität.

Schmidt, W. H. (2009). *Exploring the relationship between content coverage and achievement: unpacking the meaning of tracking in eight grade mathematics*. Michigan: The Education Policy Center at Michigan State University.

Schneider, T. (2004). Nachhilfe als Strategie zur Verwirklichung von Bildungszielen. Eine empirische Untersuchung mit Daten des Sozio-oekonomischen Panels (SOEP) *Discussion papers 447*. Berlin: DIW.

Schneider, W., & Stefanek, J. (2004). Entwicklungsveränderungen allgemeiner kognitiver Fähigkeiten und schulbezogener Fertigkeiten im Kindes- und Jugendalter. Evidenz für einen Schereneffekt? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 36*, 147-159. doi:10.1026/0049-8637.36.3.147

Shadish, W. R. (2010). Campbell and Rubin: a primer and comparison of their approaches to causal inference in field settings. *Psychological Methods, 15*, 3-17. doi:10.1037/a0015916

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research, 60*, 471-499. doi:10.3102/00346543060003471

Spear-Swerling, L., Brucker, P. O., & Alfano, M. P. (2010). Relationships between sixth-graders' reading comprehension and two different measures of print exposure. *Reading and Writing, 23*, 73-96. doi:10.1007/s11145-008-9152-8

Stanat, P. (2006). Schulleistungen von Jugendlichen mit Migrationshintergrund: Die Rolle der Zusammensetzung der Schülerschaft. 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. 189-219). Wiesbaden: Verlag für Sozialwissenschaften.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-407. doi:10.1598/RRQ.21.4.1

Stanovich, K. E. (2000). *Progress in understanding reading. Scientific foundations and new frontiers*. New York, NY: Guilford Press.

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true individual change: true change as a latent variable. *Methods of Psychological Research Online, 2,* 21-33.

Stuart, E. A. (2010). Matching methods for causal inference: a review and a look forward. *Statistical Science, 25*, 1-21. doi:10.1214/09-STS313

Trautwein, U., Lüdtke, O., Becker, M., Neumann, M., & Nagy, G. (2008). Die Sekundarstufe I im Spiegel der empirischen Bildungsforschung: Schulleistungsentwicklung, Kompetenzniveaus und die Aussagekraft von Schulnoten. In E. Schlemmer & H. Gerstberger (Eds.), *Ausbildungsfähigkeit im Spannungsfeld zwischen Wissenschaft, Politik und Praxis* (pp. 91-107). Wiesbaden: VS Verlag für Sozialwissenschaften.

Trautwein, U., & Neumann, M. (2008). Das Gymnasium. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer & L. Trommer (Eds.), *Das Bildungswesen in der Bundesrepublik Deutschland* (pp. 467-501). Reinbek bei Hamburg: Rowohlt.

Treutlein, A., & Schöler, H. (2009). Zum Einfluss der schulischen Lernumwelt auf die Schulleistung. In J. Roos & H. Schöler (Eds.), *Entwicklung des Schriftspracherwerbs in der Grundschule* (pp. 109-143). Wiesbaden: VS Verlag für Sozialwissenschaften.

Unger, J. B., Gallaher, P. G., Palmer, P. H., Baezconde-Garbanati, L., Trinidad, D. R., Cen, S., et al. (2004). No news is bad news: Characteristics of adolescents who provide neither parental consent nor refusal for participation in school-based survey research. *Evaluation Review, 28,* 52-63. doi:10.1177/0193841X03254421

van Buuren, S., & Oudshoorn, C. G. M. (2000). *Multivariate Imputation by Chained Equations. MICE V1.0 User's manual*. Leiden: TNO Prevention and Health.

van de Grift, W. (2009). Reliability and validity in measuring the value added of schools. *School effectiveness and school improvement, 20*, 269-285. doi:10.1080/09243450902883946

van Ewijk, R., & Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review, 5*, 134-150. doi:10.1016/j.edurev.2010.02.001

Walter, O., & Stanat, P. (2008). Der Zusammenhang des Migrantenanteils in Schulen mit der Lesekompetenz: Differenzierte Analysen der erweiterten Migrantenstichprobe von PISA 2003. *Zeitschrift für Erziehungswissenschaft, 11*, 84-105. doi:10.1007/s11618-008-0005-7

Weiß, R. H. (1987). *Wortschatztest (WS) und Zahlenfolgentest (ZF). Ergänzungstests zum Grundintelligenztest CFT 20. Handanweisung*. Göttingen: Hogrefe.

Weiß, R. H. (2006). *Grundintelligenztest Skala 2 - Revision - (CFT 20-R) mit Wortschatztest und Zahlenfolgentest - Revision (WS/ZF-R)*. Göttingen: Hogrefe.

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods, 15,* 18-37. doi:10.1037/a0015917

Wu, M. L. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice, 29,* 15-27. doi:10.1111/j.1745-3992.2010.00190.x

Wu, M. L., Adams, R. J., Wilson, M., & Haldane, S. A. (2007). ACER ConQuest version 2.0: generalised item response modelling software. Camberwell: ACER Press.

Zimmer, R. W., & Toma, E. F. (2000). Peer effects in private and public schools across Countries. *Journal of Policy Analysis and Management, 19,* 75-92. doi:10.1002/(SICI)1520-6688(200024)19:1<75::AID-PAM5>3.0.CO;2-W