British Journal of Educational Psychology (2012), 82, 647–671
© 2011 The British Psychological Society

The British Psychological Society

www.wileyonlinelibrary.com

Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching

Jan Retelsdorf[1]*, Michael Becker[2,3], Olaf Köller[1] and Jens Möller[4]

[1]Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[2]University of Potsdam, Germany
[3]Max Planck Institute for Human Development, Germany
[4]Christian-Albrechts-University of Kiel, Germany

Background. Assigning students to different school tracks on the basis of their achievement levels is a widely used strategy that aims at giving students the best possible learning opportunity. There is, however, a growing body of literature that questions such positive effects of tracking.

Aims. This study compared the developmental trajectories of reading comprehension and decoding speed between students at academic track schools that typically prepare students for university entrance and students at non-academic track schools that usually prepare students for vocational education.

Sample. In a longitudinal design with three occasions of data collection, the authors drew on a sample of $N = 1{,}508$ 5th graders (age at T1 about 11 years, age at T3 about 14 years) from 60 schools in Germany. The academic track sample comprised $n = 568$ students; the non-academic track sample comprised $n = 940$ students.

Method. Achievement measures were obtained by standardized tests of reading comprehension and decoding speed. Students at the different tracks were closely matched using propensity scores. To compare students' growth trajectories between the different school tracks, we applied multi-group latent growth curve models.

Results. Comparable results were recorded for the complete (unmatched) sample and for the matched pairs. In all cases, students at the different tracks displayed a similar growth in reading comprehension, whereas larger growth rates for students at academic track schools were recorded for decoding speed.

Conclusions. Our findings contribute to an increasing body of literature suggesting that tracking might have undesired side effects.

*Correspondence should be addressed to Jan Retelsdorf, Leibniz Institute for Science and Mathematics Education, Kiel, Olshausenstraße 62, 24118 Kiel, Germany (e-mail: jretelsdorf@ipn.uni-kiel.de).

DOI:10.1111/j.2044-8279.2011.02051.x

Reading skills are mainly developed during preschool and elementary school years. Even though these fundamentals affect later comprehension in adolescence (Cunningham & Stanovich, 1997), reading development does not come to an end on finishing elementary school. In fact, fostering reading comprehension is still an important task of secondary schools, since many students lack sufficient proficiency as readers even at the age of 15 (Kirsch *et al.*, 2002). During secondary school, tracking – the assignment of students to different school tracks on the basis of their achievement levels – is a particular feature of school in many countries, which might affect the development of reading skills during these years. However, research on reading skill development in a tracked school system is scarce. This study aimed at comparing the developmental trajectories of reading comprehension and decoding speed between students at different school tracks. First, we will review some literature on tracking before we present the aims of our research in detail.

Tracking

Although there are certain differences, the educational systems of most industrialized countries use ability grouping in one way or another (LeTendre, Hofer, & Shimizu, 2003). The most important rationale underlying grouping students with regard to their achievement level is that all students are supposed to learn best, when they are in a homogeneous group of students with comparable abilities (e.g., Oakes, 1987; Pallas, Entwisle, Alexander, & Stluka, 1994). According to this assumption, teaching commensurate with individual requirements is said to be much easier and more effective in groups displaying homogenous capabilities and teachers can more easily provide appropriate learning opportunities (e.g., Kulik & Kulik, 1992). Thus, the aim of tracking is to give students the best possible learning opportunity. Apart from this, tracking has been criticized because students in lower track schools are at a disadvantage to those in higher track schools with a resultant lowering of achievement and motivation (e.g., Lucas, 1999; Oakes, 1987) or even an anti-school culture (Van de Gaer, Pustjens, Van Damme, & De Munter, 2006). However, there is also some indication that tracking enhances students' achievement at all tracks (Mulkey, Catsambis, Steelman, & Crain, 2005) or that students at lower track schools might even benefit from tracking with regard to a positive development of the self-concept (Ireson, Hallam, & Plewis, 2001; Liu, Wang, & Parkins, 2005; Marsh, 1987) and lower amounts of burnout (Salmela-Aro, Kiuru, & Nurmi, 2008). Since there are many different forms of tracking, following we will describe the particular kind of tracking that is relevant in our study before we will discuss its possible effects on achievement development.

The extent of tracking varies across countries and educational systems. Referring to Trautwein, Lüdtke, Marsh, Köller, and Baumert (2006) there are three main features describing forms and intensity of tracking: the institutional level, the role of achievement, and the impact on future academic careers. First, tracking strategies differ with regard to the institutional level. Hereby, we can distinguish between forms of within-class ability grouping (mainly in the early grades), course-level grouping, which is common in secondary school, and grouping at school level. According to the latter, on the one hand there is some form of implicit tracking depending on the catchment areas of schools and, on the other hand, there is explicit tracking involving different school types characterized by achievement levels and specific curricula. The second feature, the role of achievement, deals with the decision criteria on placement in a certain track. This decision can be influenced by students' achievement (achievement grouping) or other factors such as parents' socio-economic status (SES) and educational aspirations (opt-in tracking, Trautwein *et al.*, 2006). Third, the impact of tracking on future academic careers can differ to some extent. In some educational systems, placing students in a lower track reduces their chances of attending university and obtaining a degree (e.g., Japan), whereas in other systems, this association is less rigid (e.g., United States).

In this study, we investigated the effects of tracking on reading achievement in the German educational system where we are dealing with explicit grouping at school level. After elementary school, students in Germany are assigned to different types of school that either place a focus on students' gaining qualifications that would enable them to begin a vocational apprenticeship (non-academic track schools) or prepare them for university entrance (academic track schools) so that tracking does affect future academic careers to a certain extent. Based on students' mean achievement level, teachers recommend an appropriate school track for each student at the end of elementary school. This recommendation, however, is not obligatory, and the final decision regarding which school track a child is to be placed in is ultimately that of the parents. Thus, in Germany the decision criterion for a particular track is a hybrid of achievement grouping and opt-in tracking. All in all, the German educational system is said to be the most strictly stratified school system of the Western industrialized countries (cf. Trautwein *et al.*, 2006).

Academic and non-academic tracks differ with regard to composition, such as mean achievement levels or parents' SES, and institutional factors such as the curriculum (e.g., LeTendre *et al.*, 2003). These compositional and institutional track differences can have manifold consequences that might affect students' learning and achievement. For example, the composition of students can create an environment, in which particular values and norms are predominant. Students internalize these norms and values (e.g., Barth, Dunlap, Dane, Lochman, & Wells, 2004) and thus develop attitudes towards learning that might have either positive or negative effects on achievement. Moreover, there is some evidence that teachers' beliefs, knowledge, and instructional practices differ between tracks. Hallam and Ireseon (2003) found that teachers differ in their beliefs about ability grouping. Furthermore, in a recent study Baumert *et al.* (2010) observed significant differences between teachers from academic and non-academic tracks in content knowledge and pedagogical content knowledge. These differences in teachers' beliefs and their knowledge might influence their use of instructional practices and in the long run students' achievement. Indeed, there is some research showing that teachers at higher track schools provide higher levels of problem solving and cognitive activating instruction, whereas in lower track schools exercises in class, memorization, and disciplining students are emphasized (Kunter & Baumert, 2006; Oakes, 1985; Raudenbush, Rowan, & Cheong, 1993; Retelsdorf, Butler, Streblow, & Schiefele, 2010; Van Houtte, 2004). All in all, these track differences in compositional and institutional conditions might be responsible for achievement gaps between academic and non-academic track schools.

Despite the long-term political and scientific debate on explicit tracking (e.g., Oakes, 1985), empirical studies on its particular effects on reading achievement are scarce. One of the rare longitudinal studies dealing with the effects of explicit tracking on reading achievement is from Maughan and Rutter (1987). They found that reading scores at age of 14 were higher in grammar schools than at non-selective schools when controlling for differences in intake scores. The sample size in that study, however, was rather small ($N = 160$). Another investigation of track differences in reading was a previous study with the sample of the present study, in which Retelsdorf and Möller (2008) researched the consequences of explicit school-level tracking in Germany just at the beginning of secondary school with two available waves of data. On applying latent difference score analyses, large differences in the initial level between different school types were found but no significant differences for the development of reading comprehension were recorded. The effect sizes of reading comprehension growth, however, indicated that, by trend, academic track students increased more (growth at academic track: $d = 0.82$; average growth on non-academic tracks: $d = 0.61$). In addition, academic track students displayed significantly lower decreases in reading motivation than non-academic track students. Thus, the authors speculated that the achievement gap might widen during forthcoming school years.

In contrast to research on reading development in tracked school systems, there seems to be strong empirical support for widening achievement gaps in the mathematics domain (e.g., Argys, Rees, & Brewer, 1996; Becker, Lüdtke, Trautwein, & Baumert, 2006; Hoffer, 1992; Köller & Baumert, 2001). These studies found disadvantages for students attending non-academic tracks. Moreover, Becker (2009) recently found that students in academic track schools benefited with regard to psychometric intelligence when compared to students in non-academic track schools.

As such, one might well conclude that the type of school affects students' achievement growth. Drawing such causal inferences, however, is usually inappropriate when based on data derived from observational studies that lacks additional assumptions and adjustment, respectively. Thus, we are unable to decide whether a widening achievement gap between different school tracks is a result of the type of school or if it is quite simply the result of previously existing differences among students.

The present investigation

The purpose of our study was to investigate the effect of explicitly different school tracks in Germany on students' reading development over a 3-year period. Therefore, we extended a previous study (Retelsdorf & Möller, 2008) drawing on the same sample. The previous study, however, comprised of only two occasions of data collection and did not address the problem of causal inference. The results of that previous study indicated that the reading achievement gap between different school tracks might widen as the students grow older, because a trend favouring academic track students was observed. Moreover, academic track students benefited from lower motivational decreases. Thus, the authors conjectured that the achievement gap might widen over a longer period of time. Furthermore, Retelsdorf and Möller (2008) investigated only reading comprehension as an achievement measure of the domain of reading. However, in research investigating individual Matthew effects (Stanovich, 1986) in reading – which describe reading development as a cumulative process, where those with a high initial reading level also gain higher growth than those with a lower initial level – it was observed that various reading skills might develop quite differently (Bast & Reitsma, 1998; Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005). In fact, there is some evidence that reading skills might even develop in a rather compensatory manner, that is, children with lower initial levels catch up (Aarnoutse, Leeuwe, Voeten, & Oud, 2001; Aunola, Leskinen, Onatsu-Arvilommi, & Nurmi, 2002; Baumert, Nagy, & Lehmann, in press). Describing the research on a cumulative versus compensatory individual reading development in detail is far beyond the scope of this paper. Some particular findings, however, might reveal that there is a great variability in the development of various aspects of reading, which emphasizes the importance of the investigation of different reading skills. For example, Parrila *et al.* (2005) found rather consistent support for the compensatory model across different reading measures in a Canadian sample, whereas in a Finnish sample their results were comparable to Bast and Reitsma (1998). These authors observed that a compensatory development was applicable for sentence comprehension, while a cumulative development was the result for word recognition. To test the developmental trajectories of different reading skills, an additional test covering decoding speed was included in this study to investigate exactly that. Since the development of reading skills is not necessarily linear, another limitation of Retelsdorf and Möller (2008) was that their study only comprised of two points of assessment and thus did not allow for a modelling of non-linear developmental trajectories. To address this problem and to test the authors' assumption that the achievement gap might widen in the long run, we used data from three points of measurement in the present study.

For both aspects of reading, we first applied latent growth curve models (LGCMs) for the whole sample to describe the developmental trajectories within the different school tracks. To enhance the validity of the conclusion that possible differences are due to type of school, we then repeated our analyses for a matched sample obtained by applying propensity score matching (e.g., Rosenbaum & Rubin, 1983). This approach allows for a careful drawing of causal conclusions from non-experimental data. The idea of propensity score matching is to approximate the effect of randomization by modelling the mechanism of group assignment (in our case, the assignment to different school tracks) and, thus, to eliminate the correlation between this assignment and the outcome. This procedure, if successfully applied, leads to groups of individuals with the same probability of belonging to either academic or non-academic track schools. Thus, propensity score matching is a question of modelling group assignment rather than the outcome (e.g., Schafer & Kang, 2008). Therefore, it is important to have comprehensive background information that is connected to the group assignment. In this study, we used covariates, which have been associated with the preference or choice of a particular school track in former research (Arnold, Bos, Richert, & Stubbe, 2007; Maaz, Trautwein, Lüdtke, & Baumert, 2008; Schnabel *et al.*, 2002). Our set of covariates comprised demographics (sex, age), social background indicators (Highest International Socio-Economic Index of Occupational Status [HISEI], parents' educational degree, ethnic background), student's achievement (elementary school grades and T1 achievement test scores), school career recommendation, preschool/kindergarten time, and parents' educational aspirations.

With regards to findings showing that ability grouping does matter (e.g., Becker, 2009; Ireson & Hallam, 2001; Lucas, 1999; Oakes, 1985), we expected to observe an effect of school track on reading development. With regards to the observation that – on the individual level – different reading skills develop in a different manner in terms of the cumulative versus compensatory model, it was not that clear whether comparable tracking effects would result for both reading skills. As aforementioned, Bast and Reitsma (1998) as well as Parrila *et al.* (2005) found a cumulative development according to more basic reading skills such as word recognition but not according to comprehension. Extrapolating from these results on the individual level, the development of reading skills might also result in varying track differences for each of the two components of reading. Thus, it seemed plausible to carefully expect that tracking would particularly result in different levels of decoding speed, whereas the expected effect of tracking on reading comprehension was rather ambiguous.

Method

Sample

The initial sample comprised $N = 1{,}508$ 5th graders (49% girls; age at T1: $M = 10.88$ years, $SD = .56$) from 60 schools that were drawn so as to be representative of the federal state of Schleswig-Holstein, Germany. The majority of our sample ($n = 940$ students, 62%) attended non-academic track schools comprising 13 lower track (Hauptschule), 22 middle track (Realschule), and 4 comprehensive schools (Gesamtschule). The academic track sample consisted of $n = 568$ high-track students (Gymnasium) from 21 schools. Data were collected by trained research students and took place as group tests carried out in class during regular lessons. In our study, achievement tests and questionnaires were administered at the beginning of 5th grade (T1), at the end of 6th grade (T2), and at the beginning of 8th grade (T3). Hence, T1 took place right after students were assigned to different tracks (up to 4th grade there is no tracking in the German school system). The time intervals between each measurement point were approximately 18 months. Each phase of data collection took place within a time slot of 14 days.

Measures

Reading comprehension

Age-appropriate reading tests from the German PIRLS study (Progress in International Reading Literacy Study, Bos *et al.*, 2005) and the German large-scale study 'Aspects of Students' Initial Level and Development at Schools in Hamburg' (Lehmann, Gänsfuß, & Peek, 1999) were used. These achievement tests have been well developed in the context of large-scale studies. The students' task was to read several texts and answer questions on them. The questions focused mainly on students' skills in forming a broad and general understanding of the texts and retrieving information from them. The questions mainly comprised of multiple-choice items, but some open-format questions have also been included. Since some items were scored polytomously, the item parameters were estimated by applying the partial credit model. A common scale for the varying reading tests at the three testing occasions was obtained by means of test linking using an anchor item design (Kolen & Brennan, 2004). Using ConQuest (Wu, Adams, & Wilson, 1998), weighted likelihood estimates (WLEs) were estimated as subjects' ability scores. The WLE reliabilities of the reading tests on all occasions were sufficient ($\geq .80$).

Decoding speed

To measure decoding speed, we used a test in which the students had 2 minutes of time to read a 740-word fairytale containing a great deal of numerals (e.g., 'seven', 'twenty-two') and which has been developed in accordance with the German PISA decoding speed test (Schneider, Schlagmüller, & Ennemoser, 2007). The students' task was to underline these numerals. Thus, this test places a particular emphasis on the mere decoding ability since text comprehension is not needed in recognizing the target words. As the text was too long to finish within the 2-minute time limit, the number of read words marked by the students indicated the speed of decoding (measured as number of read words per 2 minutes). In an additional data collection, a satisfying correlation of $r = .54$ was recorded between this test and the PISA-decoding speed test, which also assesses comprehension with one and the same reading task. The decoding speed test was repeated on each occasion where data were collected. Referring to Schneider, Schlagmüller *et al.* (2007), tests measuring the speed of reading provide valuable information about basic reading skills such as decoding.

Reasoning

The subtest 'Figure Analogies' from the 'Cognitive Abilities Test for grades 4 to 12' (Heller & Perleth, 2000) was used at T1 to test children's reasoning skills as an indicator of intelligence. WLEs have been estimated as ability scores. The test's WLE reliability was satisfactory (.87).

School grades

The grades from the final report card of elementary school were collected for four school subjects: German as the first language, mathematics, science, and physical education. The German grading system includes grades from 1 (outstanding) to 6 (fail). Thus, lower grades indicate better performance.

Socio-economic status

Several aspects of the families' SES were assessed by means of a parent questionnaire. First, the 'HISEI' (Ganzeboom & Treiman, 1996) was derived from the parents' occupation. This index ranges between 16 and 90 and indicates the socio-economic background that is usually associated with a particular occupational area with higher scores indicating higher incomes. As an overall family indicator, we used the higher value obtained, irrespective of whether it belonged either to the mother or father. Second, we asked the students' parents to detail the highest level of educational achievement they had attained (graduation plus apprenticeship). As per Baumert, Watermann, and Schümer (2003), their responses were coded from 1 (lower track graduation without any apprenticeship) to 7 (university degree).

Ethnic background

We captured students' ethnic background by asking if their mother and father were born in Germany. Ethnic background was dummy coded (0 = *at least one parent was not born in Germany*, 1 = *both parents were born in Germany*).

School career recommendation and parents' educational aspirations

In Germany, teachers recommend an appropriate school track for each individual student at the end of elementary school. This recommendation, however, is not obligatory. Nevertheless, it serves as a strong indicator of the school track that parents chose since many of them follow the teachers' recommendations. Therefore, we asked the parents to report the recommendation given for their children with higher values indicating a higher school track. Parents were also asked to respond to a range of vocational aspirations regarding their children's careers. This range started with 1 (*apprenticeship*) and continued through to 5 (*university studies*).

Preschool/kindergarten time and age at school entrance

Parents were asked if and how long their children had attended preschool or kindergarten. They rated their answers from 0 ('not at all') to 5 ('more than 2 years'). Finally,

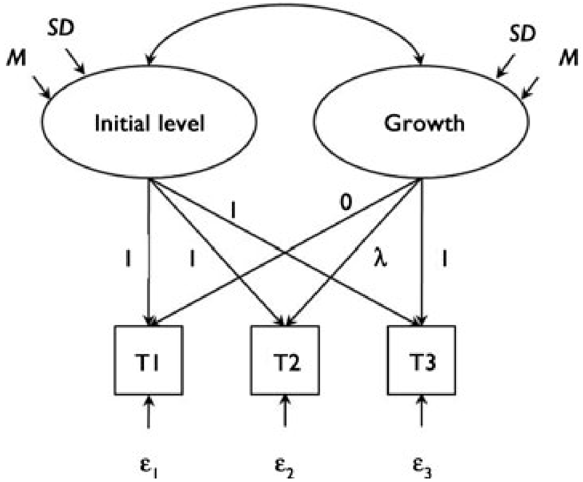

Figure 1. Basic latent growth curve model.

the parents had to indicate their children's age on school entrance on a 4-point scale anchored 1 ('5 years or younger'), 2 ('6 years'), 3 ('7 years'), and 4 ('8 years or older'). Usually, in Germany children enter school at the age of six.

Analytical issues

Missing data

Missing data is a common problem in longitudinal studies. One recommended solution to handle this problem is multiple imputation (e.g., Graham, 2009; Schafer & Graham, 2002). In our present study, on average about 13.5% of the data were missing per variable. We used the STATA implementation (ICE, Royston, 2004) of the MICE program (Van Buuren & Oudshoorn, 1999) to create $m = 5$ complete datasets. All information available has been used to obtain a good imputation model including squares and interaction terms of core variables as these often lead to better solutions and were also used for the estimation of propensity scores (see below). The matching procedure and all subsequent analyses were then conducted five times and the results were combined in accordance with Rubin (1987).

Latent growth curve modelling

When dealing with reading development, an important issue next to the amount of growth is the shape of growth. To analyse whether students at different school tracks differ in this area, we applied LGCMs (e.g., Duncan, Duncan, & Strycker, 2006) in addition to simple mean comparisons of T2 and T3 achievement scores. The idea of LGCM is to describe trajectories over time in terms of initial level (intercept) and growth (slope). In our study, LGCMs were specified by means of structural equation modelling using the software M*plus* 5.2 (Muthén & Muthén, 2008).

In Figure 1, the intercept factor describes the initial level. Likewise, the growth factor reflects the overall change. In this study, positive values for the slope mean indicate increases in reading comprehension and decoding speed. The loadings on the growth factor to T1 and T3 were constrained to 0 and 1, respectively. The loading to T2 (λ) was freely estimated. This enabled the model to capture any shape of non-linear change. Moreover, λ reflects the proportion of change between T1 and T2 relative to the total change between T1 and T3 (e.g., Bollen & Curran, 2006). As we were interested in group differences due to school track, multi-group LGCMs were applied. This kind of analysis allows for testing of a number of invariance hypotheses. With regard to our research questions, the equalities of shape, initial level, and growth between academic and non-academic track schools were of particular interest. To test the significance of the group differences, the relevant parameters were constrained to be equal between the two groups and chi-square difference tests were conducted. Put concisely, we first tested if the initial level differed between tracks. Next, the differences in the shape of growth (i.e., the free loading λ) were investigated, and, finally, the group difference of the slope factor was tested to obtain the difference in absolute change. To account for the multiple imputed datasets, all analyses were conducted five times with the results being combined at the end. For the model comparisons, we merged the chi-square statistics using Allison's (2001) formula. The result of this procedure is a test statistic that is approximately *F*-distributed.[1]

Hierarchical data structure

For all mean comparisons with the complete (unmatched) sample, we used the M*plus* option 'type = complex' to obtain corrected standard errors for the hierarchical data structure (students in schools). Ignoring such a hierarchical data structure might lead to incorrect standard errors and, thus, to biased significance tests (Hox, 2002).

Propensity score matching

When comparing different treatment conditions (in our case different school tracks), a sufficient condition for analysing treatment effects would be that the treatment assignment is uncorrelated with the outcome (Morgan & Winship, 2007). The most common way to achieve this prerequisite is by randomization. With regard to school track differences, however, it is neither feasible nor ethically sound to randomly assign students to the particular treatment groups. Thus, we cannot decide if a widening achievement gap between students on the different school tracks is a result of the type of school or if it is quite simply the result of previously existing differences among the students.

One prevalent strategy for a statistical adjustment of this confounding issue is the use of analysis of covariance or multiple regression. These strategies, however, underlie several limitations that are difficult to overcome in research practice. First, the number of covariates that can be used is rather limited. Therefore, important pre-existing differences between two groups may be disregarded. Second, the results from these kinds of analyses depend on the pre-specified form (typically linear). Finally, this method does not ensure that two groups are comparable, that is, a lack of covariate overlap remains undetected and, thus, the persons in the two groups are actually too different to compare.

Therefore, the use of strategies such as propensity score matching to estimate causal effects with non-experimental data has been proposed (e.g., Rosenbaum & Rubin, 1983) and has recently begun to enter educational psychology research (e.g., Schneider,

[1] *The numerator degrees of freedom (df) are obtained by the difference of df between the two compared models (Δ df = 1 in all our analyses). The denominator degrees of freedom (ddf) were approximated by applying the according formula including the number of imputations (m), the numerator df, and the square roots of the chi-square statistics over the m datasets (see Allison, 2001 for more details).*

Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). The general idea of these matching methods is to model the treatment assignment process directly and to create subgroups which match in their likelihood to either belong to treatment or control group – a situation which experimental research achieves through randomization. Propensity score matching includes two fundamental steps: the estimation of the propensity score and the matching procedure.

Based on background information relevant for the group assignment, the probability of group membership (the propensity score) is estimated, typically by applying logistic regression analysis. In this research, the probability of attending academic versus non-academic track schools (i.e., the propensity score) was estimated by means of logistic regression using STATA 10. In accordance with Rosenbaum and Rubin (1985), we defined the logit of the propensity score rather than the probability itself due to better distributional properties.

Treatment and control group individuals were then matched using the obtained propensity score. For this matching procedure, we pursued two strategies. First, we used the presumably most straightforward approach by matching one child on the academic track with one child on the non-academic track via nearest neighbour matching method with caliper. This approach requires a random order of the subjects in the treated (academic) and control (non-academic) group. Then, the first treated subject is selected and the control subject within a tolerance region (caliper) is sought. We chose a caliper size of $C = .05 \times SD$ (propensity score logit). As this approach often involves quite a large sample reduction, we also applied nearest neighbour matching with caliper and replacement as a second approach. In this procedure, control subjects are allowed to be used as a match for several treatment subjects and thus, a larger proportion of the treatment group is preserved. This approach, however, results in different sample sizes for treatments and controls and thus, we used frequency weights for all subsequent analyses.

Both matching strategies were implemented using the STATA module PSMATCH2 (Leuven & Sianesi, 2003). Thereby, we used bootstrapping to obtain more reliable standard errors for all subsequent analyses (e.g., Caliendo & Kopeinig, 2005). This correction is preferable since classical standard errors underestimate the variance due to the error in the estimation of the propensity score and the additional variance induced in the matching process. As the informational status of standard errors is also challenged by the fact that propensity score matching can change sample sizes and lead to progressive significance testing only due to changes in sample size, other indicators of group differences are recommended, which are not sensitive to the sample size, for example, standardized mean differences (Imai, King, & Stuart, 2008).

When applying propensity score matching, ideally, the correlation between treatment assignment and outcome is removed and, therefore, treatment effects can be estimated (for an extensive description, see Morgan & Winship, 2007). Propensity score matching also requires the assumption that selection into treatment is based on observable characteristics, as does regression analyses. Yet, it has been found to be more appropriate regarding the aforementioned limitations and to be more robust against misspecifications (Drake, 1993; Zhao, 2008).

In terms of school track differences, propensity score matching involves perceiving school tracks as treatment conditions (in our study: treatment = academic track; control = non-academic track). The aim of matching is then to eliminate initial differences between students in the two conditions in order to investigate how comparable individuals develop in different conditions. Thus, if matching succeeds, this approach allows the tentative conclusion that divergent developmental trajectories are rather more a result of features of different school tracks than they are due to previously existing differences. The success of the matching procedure is tested by the check of balance between the two groups (whether matching homogenized the groups successfully) and the inspection of the area of common support (whether comparable individuals are available for the whole sample or only in some parts of the propensity score distributions).

Essential for the success is that the prediction of the treatment assignment is sufficient, that is, that predictors of assignment and outcome can be assumed to be equally distributed between groups and, therefore, assignment can be assumed to be random regarding these variables (Augurzky & Schmidt, 2001; Morgan & Winship, 2007). The predictors of assignment have to be selected by theoretical assumptions and/or results from previous research. With regard to school track assignment background characteristics such as previous achievement or SES have been proven to explain substantial proportions of assignment variance (see above).

Subsequent to the matching procedure, we estimated the treatment effect only for those students who stem from the population that actually attended the treatment (average treatment effect for the treated), but not for the whole population. This approach seemed to be more appropriate because it only requires that students are similar in their expected baseline but can differ in their expected treatment effects and does not require full identification of the treatment effect over the whole population (for further details, see Morgan & Winship, 2007).

With regard to LGCM for the matched samples, we did not expect a difference in the mean initial level because the T1 scores were part of the matching procedure. Therefore, we tested if the initial levels significantly differed between the two groups. If this test resulted in non-significance, as it should due to the matching procedure, the initial levels for both groups were constrained to be equal to obtain a more parsimonious model.

Results

Track differences before matching

Reading comprehension

To estimate track differences for the unmatched sample, we analysed differences in the according parameters of the LGCM: initial level, shape of growth (λ), and amount of growth. As presented in Table 1, there were large differences in the initial level of reading comprehension ($F(1,91) = 242.04$, $p < .001$, $d = 1.29$). The shape of growth did not significantly differ between school tracks ($F(1,82) = 2.07$, ns) and was $\lambda = .64$, meaning that about 65% of the overall growth in reading comprehension took place until T2 indicating that comprehension growth was slightly negatively accelerated. This model with different initial levels and an equal shape of growth fitted the data well for reading comprehension: $F(4,28) = 0.51$, ns, Comparative Fit Index (*CFI*) = 1.00, Root Mean Square Error of Approximation (*RMSEA*) = .007, Standardized Root Mean Square Residual (*SRMR*) = .012. Finally, the mean growth was constrained to be equal across tracks; no significant difference resulted ($F(1,34) = 1.22$, ns; see Table 1).

Decoding speed

The same procedure was then applied for decoding speed (see Table 1). Again, the difference in the initial level was striking ($F(1,402) = 69.82$, $p < .001$, $d = 0.83$).

Table 1. Means, standard errors, effect sizes, and invariance tests for initial level and growth factors of reading comprehension and decoding speed for unmatched sample ($n_{\text{academic}} = 568$; $n_{\text{non-academic}} = 940$)

		Initial level				Growth			
		M	*SE*[a]	*d*	$\Delta\chi^2$	*M*	*SE*[a]	*d*	$\Delta\chi^2$
Reading comprehension	Non-academic	−0.41	0.08	1.29	$F(1,91) = 242.04$	0.77	0.04	0.04	$F(1,34) = 1.22$
	Academic	0.59	0.05		$p < .001$	0.75	0.05		ns
Decoding speed	Non-academic	276.43	5.23	0.83	$F(1,402) = 69.82$	143.72	6.43	0.49	$F(1,43) = 6.94$
	Academic	332.45	6.35		$p < .001$	168.18	7.72		$p < .05$

Note. Applying the chi-square difference test ($\Delta\chi^2$), significant differences indicate a worse fit for models with equality constraints. For multiple imputation, the combined chi-square statistic is approximately *F*-distributed (Allison, 2001).[a]Corrected standard errors for nested data.

Moreover, a significant difference in the shape of growth was observed ($F(1,57) = 4.02$, $p < .05$). At non-academic track schools, growth was nearly linear ($\lambda = .54$), whereas about two-thirds of growth have taken place until T2 at academic track schools ($\lambda = .64$). The model with different initial levels and different shapes of growth fitted the data well for decoding speed: $F(3,444) = 1.10$, ns, $CFI = .998$, $RMSEA = .014$, $SRMR = .037$. Finally, the difference in the growth factor yielded significance ($F(1,43) = 6.94$, $p < .05$, $d = 0.49$) indicating a higher growth rate for academic track students than for non-academic track students (see Table 1).

In summary, we found substantial differences between academic and non-academic track school students in the initial levels of decoding speed and reading comprehension. Moreover, there was a difference between tracks in students' development in decoding speed, favouring academic track school students. To strengthen the conclusion that this difference was in fact due to school track, we then applied propensity score matching to account for the non-randomized treatment (school track) assignment. Even though there were no differences between both tracks in the growth factors of reading comprehension, we also tested the differences for reading comprehension for the matched sample to provide a more complete picture of our results.

Propensity score matching

To obtain a good set of background variables for computing the propensity score, we considered a total of 16 variables that are usually associated with the preference or choice of a particular school track (Arnold *et al.*, 2007; Schnabel *et al.*, 2002). This set of variables was comprised of demographics (sex, age), social background indicators (HISEI, parents' level of educational degree, ethnic background), student's achievement (elementary school grades and T1 achievement test scores), school track recommendation, time spent at preschool/kindergarten, and parents' educational aspirations for their children. Moreover, squares of all metric variables and interactions have been added to the logistic regression model to improve balance even on higher order moments. Prior to estimating the propensity scores by logistic regression, we tested the background variables for significant differences between the academic and non-academic track (see Table 2). As can be seen from Table 2, there was an average standardized bias[2] of 77.7%, indicating

[2] *The standardized bias is an indicator that assesses the absolute difference between treatment and control group in sample means divided by an estimate of the pooled standard deviation (Rosenbaum & Rubin, 1985).*

Table 2. Background variable differences between non-academic and academic track before matching (N = 1,508)

	Non-academic track (n = 568)		Academic track (n = 940)				
Background variable	M	SE	M	SE	t	p	% bias
School career recommendation	1.75	0.02	2.81	0.02	−30.77	.000	−157.2
Sex (0 = boys)	0.49	0.02	0.50	0.02	−0.46	.648	−2.0
Age at T1	10.99	0.02	10.71	0.02	9.63	.000	44.9
HISEI	44.87	0.49	56.73	0.68	−14.37	.000	−60.9
Parents' highest educational degree	3.71	0.05	5.27	0.07	−18.29	.000	−79.7
Ethnic background	0.81	0.01	0.88	0.01	−3.13	.002	−14.3
Preschool/kindergarten time	4.14	0.05	4.51	0.04	−5.56	.000	−26.5
Age at school entrance	2.24	0.02	2.19	0.02	1.85	.065	8.1
Parents' educational aspirations	2.63	0.05	4.47	0.04	−25.80	.000	−124.2
German grade[a]	3.86	0.03	4.89	0.02	−27.34	.000	−130.3
Mathematics grade[a]	3.93	0.03	4.91	0.03	−22.99	.000	−109.1
Science grade[a]	4.24	0.03	5.13	0.02	−23.37	.000	−110.9
Physical education grade[a]	4.94	0.02	5.19	0.03	−6.74	.000	−29.9
Reasoning	−0.48	0.05	0.61	0.06	−13.42	.000	−58.6
Decoding speed T1	276.20	2.89	332.63	3.68	−12.03	.000	−52.3
Reading comprehension T1	−0.40	0.03	0.58	0.03	−19.79	.000	−93.3
Propensity score (logit)	−4.28	0.10	2.46	0.09	−44.87	.000	−218.1
Total bias (Mean [|% bias|])					77.7		

Note. HISEI, Highest International Socio-Economic Index of Occupational Status.
[a]Grades are recoded (larger scores = better grade).

a large total bias between the two groups. Although this is not the primary criterion in evaluating the success of the propensity model, the logistic regression equation led to a quite good prediction of group membership (Nagelkerke pseudo-R^2 index of .79).

The area of common support is the range where treatment and comparison group overlap. In the case of propensity score matching, this region is defined by the propensity score and indicates for which sub-sample the estimated treatment effects apply. As pictured in Figure 2, the distributions of the propensity score were quite different between academic and non-academic track students. The proportion of academic track students in the area of common support (i.e., within the two vertical lines), however, was relatively high (478 out of 568 or 84%), whereas only about 62% (582 out of 940) from the non-academic track was in this area.

Our matching procedure without replacement resulted in a sample of 139 matched pairs. Table 3 shows that the standardized bias after matching – that should be below 3–5% (e.g., Caliendo & Kopeinig, 2005) – was 2.2%. Applying matching with replacement demanded the additional consideration of Mahalanobis distances for the interactions between reasoning × decoding speed, and school career recommendation × reading comprehension to reach a satisfying balance. This matching approach resulted in a

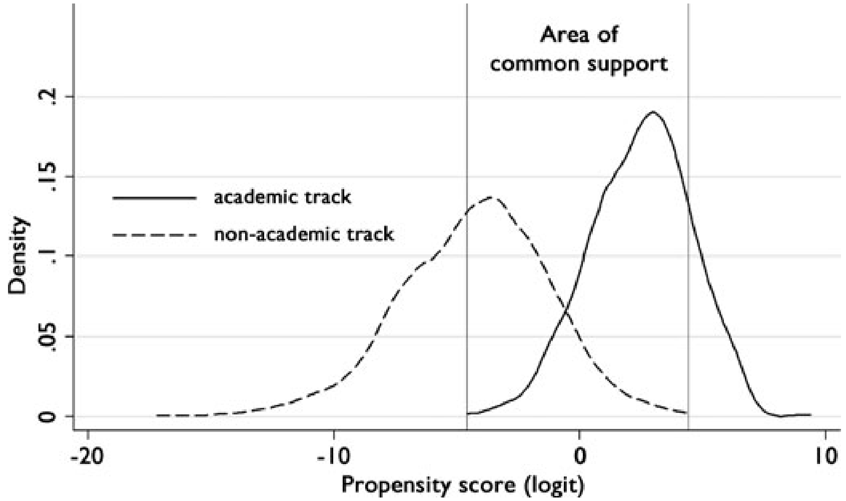

Figure 2. Area of common support.

sample of $n = 297$ academic track students matched to $n = 111$ non-academic track students. For all subsequent analyses, we used frequency weights to account for the fact that the latter ones were repeatedly used as matches. Despite there still being some considerable bias for single covariates (see Table 3), the overall standardized bias was sufficiently reduced (4.4%). When comparing students within the area of common support that were used for matching with those that were not used for matching, the bias between both groups also was low (<5% for matching with and without replacement). Thus, the results of the subsequent analyses may be taken for representative for students within the area of common support.

Track differences after matching[3]

Decoding speed

To test the effect of ability grouping, we first compared the means of decoding speed at T2 and T3. Due to the propensity score procedure we did not control baseline differences for these analyses. By and large, findings did not differ between the two matching procedures. As presented in Table 4, the mean differences for decoding speed at T2 and T3 were significant for both matching procedures resulting in higher scores for academic track students.

We then applied multi-group LGCM. In the first step, we tested differences between the matched samples of both tracks in the initial levels of decoding speed. These differences were not significant for both matching approaches ($p \geq .525$) again indicating the good balance between both groups. As a second step, we tested the differences in the shape of growth (λ in Figure 1) between the two groups. Therefore, we tested an LGCM, in which the loading was constrained to be equal between the treatment group (academic track students) and the control group (non-academic track students), against a model,

[3] *We also tested track differences in reading comprehension for the matched samples. As was true for the full (unmatched) sample, no significant differences in the shape and the amount of growth were observed. Full results are available from the first author on request.*

Table 3. Background variable differences between non-academic and academic track after matching

	Matching without replacement (n = 139 pairs)						Matching with replacement (n = 297 pairs)[c]					
	Non-academic		Academic				Non-academic		Academic			
Background variable	*M*	*SE*[b]	*M*	*SE*[b]	*t*	%	*M*	*SE*[b]	*M*	*SE*[b]	*t*	%
School career recommendation	2.44	0.06	2.45	0.04	−0.07	−0.7	2.67	0.04	2.70	0.03	−0.64	−4.5
Sex (0 = boys)	0.54	0.04	0.53	0.04	0.25	2.4	0.54	0.03	0.55	0.03	−0.19	−1.2
Age at T1	10.74	0.04	10.75	0.04	−0.18	−1.8	10.73	0.02	10.69	0.02	1.08	7.0
HISEI	52.77	1.29	53.12	1.25	−0.20	−1.9	55.04	0.86	55.19	0.92	−0.12	−0.8
Parents' highest educational degree	4.90	0.14	4.83	0.14	0.31	3.1	5.06	0.10	4.92	0.09	1.03	6.9
Ethnic background	0.85	0.03	0.83	0.03	0.46	4.5	0.87	0.02	0.89	0.02	−0.66	−4.4
Preschool/kindergarten time	4.38	0.09	4.38	0.10	−0.02	−0.2	4.51	0.07	4.41	0.06	1.08	7.4
Age at school entrance	2.18	0.04	2.17	0.04	0.08	0.8	2.18	0.03	2.17	0.03	0.25	1.7
Parents' educational aspirations	3.92	0.10	3.87	0.12	0.30	2.8	4.26	0.06	4.17	0.07	1.07	7.0
German grade[a]	4.52	0.05	4.52	0.05	0.03	0.3	2.26	0.03	2.30	0.03	−0.84	−5.6
Mathematics grade[a]	4.52	0.06	4.49	0.06	0.26	2.5	2.28	0.04	2.26	0.04	0.33	2.2
Science grade[a]	4.89	0.05	4.85	0.05	0.54	5.2	1.96	0.04	1.87	0.03	1.86	12.7
Physical education grade[a]	5.12	0.06	5.12	0.06	−0.01	−0.1	1.82	0.04	1.81	0.04	0.15	1.0
Reasoning	0.22	0.13	0.19	0.13	0.16	1.6	0.66	0.08	0.66	0.09	−0.01	−0.0
Decoding speed T1	300.17	7.17	297.67	6.97	0.25	2.5	306.01	4.01	304.26	4.16	0.30	2.0
Reading comprehension T1	0.20	0.07	0.17	0.07	0.23	2.3	0.48	0.04	0.46	0.04	0.48	3.2
Propensity score (logit)	0.02	0.14	−0.05	0.14	0.40	4.0	1.38	0.10	1.22	0.10	1.07	7.2
Total bias (Mean [|%bias|])						2.2						4.4

Note. HISEI, Highest International Socio-Economic Index of Occupational Status.

[a] Grades are recoded (larger scores = better grade).

[b] Bootstrapped standard errors.

[c] Frequency weights accounted for subjects serving as repeated matches.

Table 4. Means, standard errors, and effect sizes for the treatment effects for decoding speed for the matched samples

		Non-academic track		Academic track			
Matching		*M*	*SE*[b]	*M*	*SE*[b]	*t*	*d*
Without replacement	T2	376.70	8.21	407.50	7.52	−2.77**	0.33
(n = 139 pairs)	T3	441.05	7.98	463.63	8.15	−1.98*	0.24
With replacement	T2	397.23	5.43	415.10	5.41	−2.33*	0.19
(n = 297 pairs)[a]	T3	461.27	6.64	479.08	5.61	−2.05*	0.17

Note. [a]Frequency weights accounted for subjects serving as repeated matches.
[b]Bootstrapped standard errors.
*$p < .05$; **$p < .01$ (two-tailed).

Table 5. Means, standard errors, effect sizes, and invariance tests for the growth factors of decoding speed for the matched samples

Matching		*M*	*SE*[b]	*d*	$\Delta\chi^2$
Without replacement	Non-academic	144.58	7.55	0.40	$F(1,538) = 3.90$
(n = 139 pairs)	Academic	161.61	7.23		$p < .05$
With replacement	Non-academic	156.92	5.54	0.22	$F(1,51) = 4.09$
(n = 297 pairs)[a]	Academic	170.59	4.96		$p < .05$

Note. Applying the chi-square difference test ($\Delta\chi^2$), significant differences indicate a worse fit for models with equality constraints. For multiple imputation, the combined chi-square statistic is approximately *F*-distributed (Allison, 2001).
[a]Frequency weights accounted for subjects serving as repeated matches.
[b]Bootstrapped standard errors.

in which this loading was allowed to vary between both groups. For both matching approaches, this test yielded significance (matching without replacement: $F(1,194) = 6.16, p < .05$; matching with replacement: $F(1,27) = 4.31, p < .05$) indicating that the shape of growth was different between both groups. Students at academic track schools showed a higher growth rate until T2 ($\lambda = .62$) compared to students at non-academic tracks ($\lambda = .54$). The same was true for matching without replacement ($\lambda_{\text{academic track}} = .68$; $\lambda_{\text{non-academic track}} = .53$). Thus, for the comparison of the slope means, we started with a model with equal initial levels for academic and non-academic track students but with different shapes of growth. The overall fit for the model was good (matching without replacement: $F(4,59) = 0.45$, ns, $CFI = .999$, $RMSEA = .009$, $SRMR = .054$; matching with replacement: $F(4,16) = 1.80$, ns, $CFI = .987$, $RMSEA = .077$, $SRMR = .065$). We then constrained the slope means to be equal across the two groups resulting in a significant decrease of model fit (see Table 5). Thus, larger growth rates for students at the academic track than for students at the non-academic track were recorded.

Discussion

A core feature of Germany's secondary school system is explicit between-school tracking. The idea of tracking is that it should help to provide an educational setting which is appropriate for fostering students with regard to their individual skill level. However, there is a growing body of literature that questions such positive effects of tracking. As reading skills are an important prerequisite not only for success in an academic context but also in daily life, tracking effects on reading development are of particular interest to politicians, practitioners, and scientists. In this study, we used multi-group LGCMs to compare the developmental trajectories of reading comprehension and decoding speed between students on academic and non-academic tracks during the first 3 years of secondary school. To strengthen the conclusion that different developments are actually a result of school track, we applied propensity score matching. Drawing on differences between students on different tracks in mathematics achievement (e.g., Argys *et al.*, 1996; Becker *et al.*, 2006), we expected that students at academic track schools attain larger reading growth rates than students at non-academic track schools. However, according to the ambiguous results for studies on a cumulative versus compensatory individual development of reading, it was not obvious if tracking effects would have similar results for different reading skills. With regard to findings from earlier studies (Bast & Reitsma, 1998; Parrila *et al.*, 2005), it seemed plausible to expect larger track differences for decoding speed than for reading comprehension.

The results for the analyses before and after matching were quite similar as regards to the growth factors: no significant track differences between the slope means for reading comprehension and a significantly higher slope mean for academic track students' decoding speed than for their non-academic counterparts. Regarding the initial level, however, the findings for the complete sample were naturally quite different from the results for the matched sample. At the beginning of secondary school, academic track students achieve largely higher levels of reading comprehension as well as decoding speed than students in non-academic track schools. These initial level differences disappeared when propensity score matching was applied, indicating a good balance between both tracks. An additional finding of our study indicated a slowing down in reading growth. Students in our sample gained more than two-thirds of overall reading comprehension growth during the first 18 months of our study (until T2). With regard to decoding speed, this was only true for students on academic tracks while non-academic track students' development was nearly linear. It seems as if a student's full learning potential for reading comprehension – and similarly so for decoding speed in academic track schools – might be released at some point during their secondary school years. The slowing down in development indicates that students in our study may approach this point. Non-academic track students, however, still seem to have some more room to improve their decoding skills.

Our hypothesis proposing benefits for academic track school students was supported with regard to decoding speed. Academic track students' larger growth rate might depend on various facts. First, teachers in different school tracks vary in their use of instructional practices. In academic track schools, for example, critical thinking, cognitive activating tasks and problem solving are emphasized, whereas in non-academic track schools, exercises and repetition are stressed (Kunter & Baumert, 2006; Raudenbush *et al.*, 1993; Retelsdorf *et al.*, 2010; Van Houtte, 2004). Moreover, at the expense of academic goals controlling students' behaviour is a considerable task for teachers at non-academic track schools. Altogether, teachers in higher track schools seem to promote learning to a larger degree than teachers in lower track schools do. Basic reading skills such as decoding abilities, however, are not the focus of secondary school teachers – within neither the academic track nor the non-academic track. Thus, other explanations such as classroom or school composition might be more compelling. This idea involves the assumption that students' achievement development depends on aggregated student characteristics per school (Hattie, 2002; Thrupp, Lauder, & Robinson, 2002). Students at schools with, for example, high mean achievement levels, high mean SES, and low ethnic heterogeneity ought to benefit compared to students at schools with an adverse mixture of characteristics. This might be due to achievement-related norms and values that develop subject to the student composition. A third explanation concerns routine. Decoding speed involves high automaticity, which in turn requires great amounts of practice. Academic track students' decoding speed might grow faster because the overall preoccupation with written texts is, in general, more pronounced in academic track schools (across all school subjects) and thus, decoding might automatize to a higher extent at the academic track. As students were tested using the same test several times, we cannot rule out training effects. The time span between the three measurement occasions, however, was quite long (18 months). Moreover, even though it might be a training effect, this effect was more pronounced at academic track schools than at non-academic track schools again indicating a beneficial learning environment.

With regard to the development of reading comprehension, we did not find significant differences between academic and non-academic track students. Thus, the assumption Retelsdorf and Möller (2008) made was not supported. These authors speculated that the achievement gap between different school tracks might widen over time, since they found some diverging trend that, however, failed to show statistical significance. The absence of a tracking effect for reading comprehension might be due to the levels of comprehension our test required. As the students' task was to form a broad and general understanding of the texts and to retrieve particular information from the texts, we used tasks that rather required surface or propositional representation for comprehension (cf. Kintsch, 1998). Maybe, applying tests that demand higher level processes from students such as inferential skills or integration into background knowledge, increased benefits for academic track students might also emerge for reading comprehension.

Our tentative assumption that tracking effects might be larger for decoding speed than for reading comprehension was supported. This result fits to the equivocal findings of research investigating the development of individual differences in reading. For example, drawing on a Finnish sample, Parrila *et al*. (2005) found that comprehension developed rather compensatory, whereas a Matthew effect was observed for decoding speed. The ambiguous findings in our study might be due to the particular tests we applied. As previously stated, the reading comprehension test required relatively low levels of comprehension. These functional reading tasks are practiced by everyone in daily life so that students at different school tracks might develop very similarly. By contrast, decoding speed needs high automaticity – in particular due to the speed test character – this is only reached by extensive practice (Lundberg, 2002). As mentioned above, basic reading skills such as decoding speed, however, are neither the focus of academic nor non-academic track teachers as this skill is rather an educational objective of elementary school. However, in academic track schools, students are generally exposed to larger amounts of text throughout the spectrum of school subjects than their non-academic track counterparts. Thus, it seems plausible that basic reading skills might improve faster at academic track schools than at non-academic track schools due to the increased amount of practice.

Regardless of the explanation of our results, there are certain concerns that must be taken into account when extrapolating our findings. First, we only estimated the average treatment effect for the treated, and the results were only representative for the proportion of students within the area of common support. Within this area, however, the differences between students included in the matched samples and those not included were small. The amount of academic track students in this area was quite satisfying, though only slightly more than half of the non-academic track students lay in this area. However, this could even strengthen the conclusion that school track does matter since even positively selected non-academic students did not achieve equal growth in decoding speed when viewed against comparable students at higher track schools. Second, when extrapolating our results one should keep in mind that, in general, students' achievement level at non-academic tracks is well below that of the level at academic tracks as it becomes apparent when one views the results for the unmatched (complete) sample. As aforementioned, according to the growth factors, these analyses yielded similar results as for the analyses for the matched sample did.

Compared with the widening achievement gap in the mathematics domain (e.g., Becker *et al.*, 2006), the differences in reading development in our study overall were somewhat smaller. This might be due to the very different role these competencies play in the curriculum of secondary schools. While there is no explicit curriculum for reading in Germany's secondary schools, the acquisition of mathematical competencies depends much more on the curriculum and, thus, school's learning environment. Even though tracking effects for reading seem to be smaller than in other domains, and even though our results are somewhat equivocal, this study contributes to research that questions positive effects of tracking. Indeed, the developmental trajectories for reading comprehension are comparable within academic and non-academic track schools. However, one should keep in mind that there is still a considerable difference between students at both tracks at the entrance of secondary school, which remains stable over time (see results for the unmatched samples). In fact, students in non-academic track schools hardly reach the level of reading comprehension at grade eight that their counterparts in academic track schools have already reached at the beginning of grade 5. Moreover, for decoding speed, it turned out that even for students with a very similar composition of characteristics at the entry to secondary school – in terms of prior achievement and socio-demographic background – academic track school students gain more in decoding speed. This result is in line with previous studies that confirmed that students achieve better in higher ability groups than in lower ability groups (e.g., Fuligni, Eccles, & Barber, 1995; Hoffer, 1992; Pallas *et al.*, 1994; van Houtte, 2004; Wiliam & Bartholomew, 2004). Our results, however, allow only very cautious interpretations of tracking effects, since we only investigated the domain of reading, whereas previous research has shown that effects of ability grouping are not necessarily universal (cf. Baumert, Becker, Neumann, & Nikolova, 2010). Thus, achievement data from different domains and additional non-achievement-related outcome measures should be taken into account in future studies investigating tracking effects. Moreover, propensity score matching provides a promising approach for the comparison of similar students in different treatments. However, our analyses do not allow comparing extremely high- or low-performing students. Accordingly, the question if these students particularly benefit from tracking remains unanswered. Finally, with our current study we cannot decide whether these differences go back to instructional, compositional, or institutional causes. All in all, we may, nevertheless, conclude that – within particular domains – tracking does have undesired side effects, which may cumulate during adolescents' further development and affect their educational outcomes, and thus will increase the spread of achievement within one age cohort. Such track-specific developmental trajectories become particularly problematic, when students at a particular track will not reach minimum standards of achievement anymore.

Limitations and strengths

One benefit of using propensity score matching is the enhancement of causal inferences (e.g., West & Thoemmes, 2008). This analytical approach offers vast opportunities for the investigation of schooling effects. In practice, related research questions are always limited to non-experimental data since we cannot randomly assign students to particular schools or school tracks. Propensity score matching allows researchers to take into account the assignment procedure and offers a way to directly address the issue of selection bias for the estimation of treatment (i.e., school) effects with observational data. In this study, our matching procedure has reduced the total selection bias comprehensively ($\text{bias}_{\text{before matching}} = 78\%$ vs. $\text{bias}_{\text{after matching}} = 2\%$ and accordingly 4%). Additionally, gaining similar results for track differences in the growth of both components of reading without matching as well as applying two different propensity score matching approaches, gives our findings a certain robustness. It is still conceivable, however, that the differences in decoding speed growth between academic and non-academic track students depend on selection bias of unobservable background variables (cf. Winship & Morgan, 1999). This is a common problem of methods relying on the control of observed characteristics as regression or propensity score matching, which we cannot overcome with our data. In order to rule this problem out, studies investigating the effects of tracking should draw on methods controlling also for non-observables, for example, on longitudinal samples including several phases of data collection before track assignment and several thereafter (cf. Becker, 2009).

Moreover, our study did not investigate the particular characteristics of tracking that are responsible for the divergent trajectories of decoding speed. In addition to institutional aspects such as curriculum differences or instructional styles another explanation for tracking effects is classroom or school composition. Despite quite a long history of research on such compositional effects of about 30 years, there is still no consensus on the size of unique and shared variance of compositional and institutional differences (see above). Thus, future studies should account for these desiderata.

A final concern involves the number of phases of data collection. With only three occasions of measurement, the number of alternative growth shapes that can be investigated is quite limited. As the trajectories for decoding speed bend down at T2 to a higher extent for the academic track students than for the non-academic students, whereas the difference at T3 continues to remain significant, we cannot conclude if students at non-academic track schools might catch up later. A related limitation is the time span between the particular waves of data collection. The fact that there were intervals of approximately 18 months between each phase of data collection also meant that we were not able to analyse short-time effects.

In summation, we found a clear effect of school track on the development of decoding speed, whereas reading comprehension growth seems not to be directly related to school track. Thus, our results suggest that school track does make a difference, but not for every particular reading skill. Despite some limitations of our study, our study suggests that propensity score matching might be a viable way for studying school track differences when self-selection into a treatment is present.

Acknowledgements

The research reported in this paper is part of the project 'Self-concept, Motivation, and Literacy: Development of Student Reading Behavior' directed by Jens Möller (Christian-Albrechts-University of Kiel). The project was funded by the German Research Foundation (DFG; Mo 648/15-1/15-3). Furthermore, we would like to thank Declan Donaghey for language editing.

References

Aarnoutse, C., Leeuwe, J. V., 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

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage Publications, Inc.

Argys, L. M., Rees, D. I., & Brewer, D. J. (1996). Detracking America's schools: Equity at zero cost? *Journal of Policy Analysis and Management*, *15*, 623-645. doi:10.1002/(SICI)1520-6688(199623)15:4<623::AID-PAM7>3.0.CO;2-J

Arnold, K.-H., Bos, W., Richert, P., & Stubbe, T. C. (2007). Schullaufbahnpräfenzen am Ende der vierten Klassenstufe [School type preferences at the end of the 4th grade]. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, ... R. Valtin (Eds.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 271-297). Münster, Germany: Waxmann.

Augurzky, B., & Schmidt, C. M. (2001). The propensity score: A means to an end. *IZA Discussion Paper*, *271*. Retrieved from: http://ftp.iza.org/dp271.pdf

Aunola, K., Leskinen, E., Onatsu-Arvilommi, T., & Nurmi, J.-E. (2002). Three methods for studying developmental change: A case of reading skills and self-concept. *British Journal of Educational Psychology*, *72*, 343-364. doi:10.1348/000709902320634447

Barth, J. M., Dunlap, S. T., Dane, H., Lochman, J. E., & Wells, K. C. (2004). Classroom environment influences on aggression, peer relations, and academic focus. *Journal of School Psychology*, *42*, 115-133. doi:10.1016/j.jsp.2003.11.004

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., Becker, M., Neumann, M., & Nikolova, R. (2010). Besondere Förderung von Kernkompetenzen an Spezialgymnasien? Der Frühübergang in grundständige Gymnasien in Berlin [Do academic tracks with specific curricular profiles accelerate the development of achievement in reading, mathematics, and English literacy? Early transition to the academic track of secondary schooling in Berlin]. *Zeitschrift für Pädagogische Psychologie*, *24*, 5-22. doi:10.1024/1010-0652/a000001

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y.-M. (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., Nagy, G., & Lehmann, R. H. (in press). Cumulative advantages and the emergence of social and ethnic inequality: Matthew effects in reading and mathematics development within elementary schools? *Child Development*.

Baumert, J., Watermann, R., & Schümer, G. (2003). Disparitäten der Bildungsbeteiligung und des Kompetenzerwerbs. Ein institutionelles und individuelles Mediationsmodell [Disparities in educational participation and attainment: An institutional and individual mediation model]. *Zeitschrift für Erziehungswissenschaft*, *6*, 46-72. doi:10.1007/s11618-003-0004-7

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

Becker, M., Lüdtke, O., Trautwein, U., & Baumert, J. (2006). Leistungszuwachs in Mathematik. Evidenz für einen Schereneffekt im mehrgliedrigen Schulsystem? [Achievement gains in mathematics: Evidence for differential achievement trajectories in a tracked school system?]. *Zeitschrift für Pädagogische Psychologie*, *20*, 233–242. doi:10.1024/1010-0652.20.4.233

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models. A structural equation perspective*. Hoboken, NJ: John Wiley & Sons, Inc. doi:10.1002/0471746096

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R., Voss, A., & Walther, G. (Eds.). (2005). *IGLU. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente* [*Scale handbook of the German PIRLS study*]. Münster, Germany: Waxmann.

Caliendo, M., & Kopeinig, S. (2005). Some practical guidance for the implementation of propensity score matching. *IZA Discussion Paper*, *1588*. Retrieved from http://ftp.iza.org/dp1588.pdf

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*, 934–945. doi:10.1037/0012-1649.33.6.934

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, *49*, 1231–1236.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fuligni, A. J., Eccles, J. S., & Barber, B. L. (1995). The long-term effects of seventh-grade ability grouping in mathematics. *Journal of Early Adolescence*, *15*, 58–89. doi:10.1177/0272431695015001005

Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. *Social Science Research*, *25*, 201–239. doi:10.1006/ssre.1996.0010

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. doi:10.1146/annurev.psych.58.110405.085530

Hallam, S., & Ireson, J. (2003). Secondary school teachers' attitudes towards and beliefs about ability grouping. *British Journal of Educational Psychology*, *73*, 343–356. doi:10.1348/000709903322275876

Hattie, J. A. C. (2002). Classroom composition and peer effects. *International Journal of Educational Research*, *37*, 449–481. doi:10.1016/S0883-0355(03)00015-6

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision* [*Cognitive Abilities Test (CogAT; Thorndike, L. & Hagen, E., 1954–1986) – German adapted version/author*]. Göttingen: Beltz.

Hoffer, T. B. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis*, *14*, 205–227. doi:10.3102/01623737014003205

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*, 481–502. doi:10.1111/j.1467-985X.2007.00527.x

Ireson, J., & Hallam, S. (2001). *Ability grouping in education*. London: Chapman.

Ireson, J., Hallam, S., & Plewis, I. (2001). Ability grouping in secondary schools: Effects on pupils' self-concepts. *British Journal of Educational Psychology*, *71*, 315–326. doi:10.1348/000709901158541

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kirsch, I. S., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). *Reading for change: Performance and engagement across countries. Results from PISA 2000*. Paris: OECD Publishing.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Berlin, Germany: Springer.

Köller, O., & Baumert, J. (2001). Leistungsgruppierungen in der Sekundarstufe I – Ihre Konsequenzen für die Mathematikleistung und das mathematische Selbstkonzept der Begabung [Ability grouping at secondary level 1. Consequences for mathematics achievement and the self-concept of mathematical ability]. *Zeitschrift für Pädagogische Psychologie*, *15*, 99–110. doi:10.1024//1010-0652.15.2.99

Kulik, J. A., & Kulik, C.-l. C. (1992). Meta-analytic findings on grouping programs. *Gifted Child Quarterly*, *36*, 73–77. doi:10.1177/001698629203600204

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*, 231-251. doi:10.1007/s10984-006-9015-7

Lehmann, R. H., Gänsfuß, R., & Peek, R. (1999). *Aspekte der Lernausgangslage und der Lernentwicklung von Schülerinnen und Schülern an Hamburger Schulen – Klassenstufe 7. Bericht über die Untersuchung im September 1998* [*Aspects of students' initial level and development at schools in Hamburg – grade 7. Report on the study in September 1998*]. Hamburg, Germany: Landesinstitut für Lehrerbildung und Schulentwicklung.

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 3.0.0). Retrieved from http://ideas.repec.org/c/boc/bocode/s432001.html

Liu, W. C., Wang, C. K. J., & Parkins, E. J. (2005). A longitudinal study of students' academic self-concept in a streamed setting: The singapore context. *British Journal of Educational Psychology*, *75*, 567–586. doi:10.1348/000709905×42239

Lucas, S. R. (1999). *Tracking inequality. Stratification and mobility in American high schools*. New York: Teachers College Press.

Lundberg, I. (2002). The child's route into reading and what can go wrong. *Dyslexia*, *8*, 1–13. doi:10.1002/dys.204

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives*, *2*, 99–106. doi:10.1111/j.1750-8606.2008.00048.x

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*, 280–295. doi:10.1037/0022-0663.79.3.280

Maughan, B., & Rutter, M. (1987). Pupils' progress in selective and nonselective schools. *School Leadership & Management*, 7, 50–68. doi:10.1080/0260136870070110

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

Mulkey, L. M., Catsambis, S., Steelman, L. C., & Crain, R. L. (2005). The long-term effects of ability grouping in mathematics: A national investigation. *Social Psychology of Education*, *8*, 137–177. doi:10.1007/s11218-005-4014-6

Muthén, L. K., & Muthén, B. O. (2008). *Mplus* (Version 5.2) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.

Oakes, J. (1987). Tracking in secondary schools: A contextual perspective. *Educational Psychologist*, *22*, 129–153. doi:10.1207/s15326985ep2202_3

Pallas, A. M., Entwisle, D. R., Alexander, K. L., & Stluka, M. F. (1994). Ability-group effects: Instructional, social, or institutional? *Sociology of Education*, *67*, 27–46. doi:10.2307/2112748

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J.-E., & Kirby, J. R. (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

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1993). Higher order instructional goals in secondary schools: Class, teacher, and school influences. *American Educational Research Journal*, *21*, 523–553. doi:10.3102/00028312030003523

Retelsdorf, J., Butler, R., Streblow, L., & Schiefele, U. (2010). Teachers' goal orientations for teaching: Associations with instructional practices, interest in teaching, and burnout. *Learning and Instruction*, *20*, 30–46. doi:10.1016/j.learninstruc.2009.01.001

Retelsdorf, J., & Möller, J. (2008). Entwicklungen von Lesekompetenz und Lesemotivation: Schereneffekte in der Sekundarstufe? [Developments of reading literacy and reading motivation: Achievement gaps in secondary school?]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *40*, 179–188. doi:10.1026/0049-8637.40.4.179

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

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

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal*, *4*, 227–241.

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

Salmela-Aro, K., Kiuru, N., & Nurmi, J.-E. (2008). The role of educational track in adolescents' school burnout: A longitudinal study. *British Journal of Educational Psychology*, *78*, 663–689. doi:10.1348/000709908×281628

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. doi:10.1037//1082-989X.7.2.147

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313. doi:10.1037/a0014268

Schnabel, K. U., Alfeld, C., Patterson, F. D., Eccles, J. S., Köller, O., & Baumert, J. (2002). Parental influence on students' educational choices in the United States and Germany: Different ramifications – same effect? *Journal of Vocational Behavior*, *60*, 178–198. doi:10.1006/jvbe.2001.1863

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs: A think tank white paper*. Washington, DC: American Educational Research Association.

Schneider, W., Schlagmüller, M., & Ennemoser, M. (2007). *LGVT 6–12. Lesegeschwindigkeits- und -verständnistest für die Klassen 6–12* [*Reading speed and comprehension test for grades 6 to 12*]. Göttingen: Hofgrefe.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*, 360–406. doi:10.1598/RRQ.21.4.1

Thrupp, M., Lauder, H., & Robinson, T. (2002). School composition and peer effects. *International Journal of Educational Research*, *37*, 483–504. doi:10.1016/S0883-0355(03)00016-8

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, *98*, 788–806. doi:10.1037/0022-0663.98.4.788

Van Buuren, S., & Oudshoorn, K. (1999). *Flexible imputation by MICE. Report TNO-PG 99.054*. Retrieved from http://www.multiple-imputation.com

Van de Gaer, E., Pustjens, H., Van Damme, J., & De Munter, A. (2006). Tracking and the effects of school-related attitudes on the language achievement of boys and girls. *British Journal of Sociology of Education*, *27*, 293–309. doi:10.1080/01425690600750478

Van Houtte, M. (2004). Tracking effects on school achievement: A quantitative explanation in terms of the academic culture of school staff. *American Journal of Education*, *110*, 354–388. doi:10.1086/422790

West, S. G., & Thoemmes, F. (2008). Equating groups. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *Handbook of social research methods* (pp. 414–430). London: Sage.

Wiliam, D., & Bartholomew, H. (2004). It's not which school but which set you're in that matters: The influence of ability grouping practices on student progress in mathematics. *British Educational Research Journal*, *30*, 279–293. doi:10.1080/0141192042000195245

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, *25*, 659–706. doi:10.1146/annurev.soc.25.1.659

Wu, M. L., Adams, R. J., & Wilson, M. (1998). *ConQuest: Generalized item response modeling software* [Computer Software]. Melbourne: Australian Council for Educational Research.

Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications. *Economics Letters*, *98*, 309–319. doi:10.1016/j.econlet.2007.05.010

Received 11 August 2010; revised version received 9 August 2011