Putting all students in one basket does not produce equality: gender-specific effects of curricular intensification in upper secondary school

Nicolas Hübner, Wolfgang Wagner, Benjamin Nagengast and Ulrich Trautwein

Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany

ABSTRACT

In recent decades, several countries have made an effort to increase the enrollment rates and performance of students in science and mathematics by means of mandatory, rigorous coursework, which is often referred to as curricular intensification (CI). However, there is a lack of research on intended and unintended effects of CI reforms on achievement and motivation. Using representative data from the National Educational Panel Study, we examined effects of a prototypical CI reform in 1 German state. We compared data from the last student cohort before and the first student cohort after the reform at the end of upper secondary school. There was no statistically significant effect on average achievement. However, we found first evidence for differential effects on English reading and a higher English self-concept of young men after the reform, whereas the reform had a negative effect on young women’s math self-concept.

KEYWORDS

Reform; curricular intensification; differential effects; achievement; motivation

Introduction

In recent decades, several countries have made an effort to increase the enrollment rates and performance of students in school subjects that are believed to be of specific importance to individuals and society. For instance, in A Nation at Risk, The National Commission on Excellence in Education (1983) proposed a New Basics curriculum, which emphasized compulsory lessons in English (4 years), mathematics (3 years), and science (3 years) for all high school students and called for higher standards to be achieved by all. This report can be seen as a major starting point for the ongoing debate about curricular intensification (CI). CI comprises actions that are aimed at increasing the number of students enrolled in specific courses in order to increase the average level of student achievement and harmonize performance among all students (Crosnoe & Benner, 2015).

More recently, in many countries around the world, CI reforms have focused on mathematics and the sciences as two of the so-called STEM (science, technology,
Engineering, and mathematics) subjects (Domina & Saldana, 2012; Organisation for Economic Co-operation and Development [OECD], 2015; Osborne & Dillon, 2008; Penner, Domina, Penner, & Conley, 2015; Stein, Kaufman, Sherman, & Hillen, 2011; Volante, 2016). High competencies in science and mathematics are assumed to provide a foundation that is essential for addressing issues of major individual and sociopolitical relevance and for building a prospering competitive economy (Hanushek & Woessmann, 2008; Mullis et al., 1998). However, other domains such as reading competence and foreign languages have also been the target of CI in some countries (e.g., Callahan, Wilkinson, & Muller, 2010; Wagner et al., 2011).

Research on CI effects has been mixed (e.g., Penner et al., 2015). One possible reason for this mixture is that CI reforms are often complex and might not work in the same way across different subjects, and more studies are needed to understand the effects of the various factors that are involved. Moreover, CI studies typically focus on achievement outcomes and neglect other important effects such as motivational outcomes. Finally, CI effects might differ between groups of students, and these differential effects are also understudied (e.g., Domina, McCEachin, Penner, & Penner, 2015; Domina & Saldana, 2012; Hübner et al., 2017; Nomi & Raudenbush, 2016; Penner et al., 2015). Hence, going beyond prior research and using representative data, we report effects of a statewide introduction of CI in one German state on both achievement and motivational outcomes in STEM subjects as well as English as a second language, with a special emphasis on differential effects on young women and young men.

**Curricular intensification: development and definition**

As outlined above, in the United States, CI had an importing starting point in 1983 with the publication of *A Nation at Risk* (The National Commission on Excellence in Education, 1983), which, although not a legal mandate, underscored the importance of high curricular standards and related them to compulsory enrollment in core subjects. Consequently, for the US, the development of CI is usually described as a result of this report and the many policy decisions that followed (e.g., Domina & Saldana, 2012). In this regard, the No Child Left Behind (NCLB) act can be seen as a culmination of such developments (e.g., the standards-based reform movement) as this federal law strongly fostered the introduction of high curricular standards, high-stakes testing, and the monitoring of learning with systematic accountability systems (e.g., Hess & Petrilli, 2006; Ravitch, 2011). In the US literature, CI is oftentimes quantified by the increased enrollment of students in more demanding courses and the mandatory enrollment in core subjects (e.g., Domina & Saldana, 2012).

In Germany, the term “curricular intensification” is not typically used to describe developments in education, although several recent reforms have been perfectly in line with the CI framework. From a general perspective, the publication of the results of the Programme for International Student Assessment (PISA; OECD, 2001) created a “shock” in German society as a consequence of the unexpectedly low performance of German students, and this shock can be seen as the starting point of CI-related developments on the policy level. The developments following the publication of these results resembled developments in the US that had started much earlier. Standards-based reforms introduced curricular standards that defined what students should
achieve in a specific grade, and since then, these standards have been monitored with a systematic monitoring strategy at the national level and at the level of federal states (e.g., Kultusministerkonferenz [KMK], 2005, 2006; Niemann, 2016). In line with movements toward a better comparability of students’ competencies at specific ages, most German states (re-)introduced mandatory enrollment in core subjects (math, foreign languages, and German) in upper secondary school (e.g., Trautwein & Neumann, 2008). Comparable trends have been visible in many other countries (e.g., Volante, 2016). Conceptually and on the basis of recent research, we will differentiate between four CI aspects that are especially important with respect to the reform of upper secondary school (e.g., Hübner et al., 2017; Trautwein, Neumann, Nagy, Lüdtke, & Maaz, 2010; Wagner et al., 2011; Wagner, Rose, Dicke, Neumann, & Trautwein, 2014). Previous research has mainly discussed CI in terms of detracking (e.g., Domina & Saldana, 2012; Kelly, 2007). In our case, however, CI is also critically linked to further theoretical discussions that go beyond detracking, such as (a) instructional time, (b) the quality and content of instruction (curriculum standards), and (c) the relevance of subjects for the final upper secondary school examination grade (see Table 1).

First, CI can be understood “as a form of detracking” of students (Domina & Saldana, 2012, p. 687), which can be further characterized in terms of different tracking components (inclusiveness, electivity, selectivity, scope; Sørensen, 1970). CI is based largely on the idea that students’ achievement improves when they take advanced courses at school (Domina et al., 2015; Penner et al., 2015) and that CI might therefore reduce the negative side effects of tracking on low-track students’ achievement (e.g., Hanushek & Woessmann, 2006; Lee & Bryk, 1988) and increase opportunities to learn in general (cf. Chmielewski, Dumont, & Trautwein, 2013). CI might take effect as one or more of these components is changed, for instance, through the elimination of course-level differences or the implementation of mandatory enrollment.

Second, related to mandatory enrollment, CI often involves increased instruction time in the specific subjects. Hence, CI is tied to scientific debates on instruction time, learning, and achievement (e.g., Lavy, 2015) because the mandatory enrollment of students who would not have taken a specific course otherwise typically increases their instructional time in this subject, and detracking students leads to a similar amount

| Table 1. Subject-specific dimensions of CI. |
|--------------------------------------------|
| **Dimension of CI** | **Tracking** | **Instruction time** | **Curriculum standards** | **Importance after the reform** |
|---------------------|-------------|----------------------|--------------------------|-------------------------------|
| Math                | Complete    | Reduced on average   | Increased on average     | High for all students        |
| English             | Almost complete| Stable on average  | Increased on average     | High for 94.9% of the students |
| Physics             | Partial     | Stable on average   | Increased on average     | High for 31.5% of the students |
| Biology             | Partial     | Stable on average   | Increased on average     | High for 58.5% of the students |

Note: All percentages and information regarding instructional time were taken from Tables 3 and 5. Although instruction time was stable on average for all courses beside math, note that it still may have increased for traditional minority groups (e.g., young men in English).
of instructional time for all students (e.g., Cortes, Goodman, & Nomi, 2015; Nomi & Raudenbush, 2016).

Third, CI can also mean that a more demanding curriculum is introduced (in combination with an increase in instruction time or independent of it), and both time and quality seem to impact student achievement (Hanushek & Woessmann, 2006; Lavy, 2015).

Fourth, even without changing the amount of time allocated to a subject or the contents of the curriculum, CI in a broad sense can cause specific subjects to become “more important” relative to other subjects, for instance, because they count more heavily toward important placement decisions (e.g., grade retention, final examinations, or university access). This is also in line with expectancy-value theory (EVT; e.g., Eccles & Wigfield, 2002), which would suggest that students’ utility values might be influenced by CI as course grades count more heavily toward the final examination grade after the reform, which is important for university access and employment.

Effects of curricular intensification on achievement and motivation

Several studies found positive effects of intensification on achievement (e.g., Ceci, 1991; Lavy, 2015; Patall, Cooper, & Allen, 2010; Scheerens, 2014). However, there is also a great deal of literature suggesting rather mixed or zero effects (Allensworth, Nomi, Montgomery, & Lee, 2009; Domina et al., 2015; Nomi & Raudenbush, 2016; Penner et al., 2015; Stein et al., 2011). Inconsistent findings exist in particular on the effect size of the impact of CI on achievement (e.g., Penner et al., 2015). Moreover, researchers who have explored the effects of CI have usually studied changes (e.g., due to enrollment) related to subject-specific instructional time (e.g., Domina & Saldana, 2012), whereas other elements of CI have been less intensively discussed.

Domina and Saldana (2012) examined the effect of CI in mathematics, indicated by increased credits earned in math-related courses, on social stratification between the years 1982 and 2004. Their results suggested a narrowing of completion gaps by race, class, and achievement in several of these subjects (e.g., algebra II and trigonometry), whereas the gaps remained prominent in calculus courses.

Surprisingly, very few studies have explored motivational outcomes in the context of CI, even with regard to STEM reforms where the role of motivational outcomes in predicting STEM career choices is well substantiated (Jansen, Schroeders, & Lüdtke, 2014; Watt & Eccles, 2008). Further attesting to the critical role of motivational variables, achievement is reciprocally associated with students’ motivation, as academic self-concepts and interests are highly influenced by previous achievement but also predict later achievement (Marsh et al., 2015; Schurtz, Pfost, Nagengast, & Artelt, 2014).

On the basis of prior research (e.g., Marsh, 1986), one would expect to find effects of CI on motivational outcomes for at least some students as a consequence of changes in class composition. Class composition may have an effect on achievement outcomes but also on student motivation (Marsh, 1986). According to the literature on the big-fish-little-pond effect (BFLPE), the reference group has a direct impact on student motivation (Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007; Nagengast & Marsh, 2012; Seaton, Marsh, & Craven, 2009). This suggests that, on the one hand, average-achieving students who would have chosen to enroll in a basic course under prereform conditions should, on average, have a lower self-concept in core courses (after the reform) in which all
students were grouped together. On the other hand, average-achieving students who would have enrolled in an advanced course under prereform conditions should, on average, have a higher self-concept in core courses (after the reform). Changing course assignment mechanisms, as inherent in CI, can lead to a more heterogeneous composition of students regarding their achievement and should have an impact on students’ domain-specific self-concepts and interests, as both constructs are strongly related (Denissen, Zarrett, & Eccles, 2007; Trautwein, Lüdtke, Marsh, & Nagy, 2009). In this regard, one could expect increased side effects (e.g., lower self-concepts in comparably low-achieving students) due to different reference groups.

Finally, as CI is aimed at decreasing differences in student achievement, it is important to also take a look at differential effects of intensification (e.g., on gender differences). Meta-analyses in which the results of several large-scale assessment studies in lower secondary school have been combined have suggested no or small average effects to the advantage of young men in mathematics (e.g., Else-Quest, Hyde, & Linn, 2010), whereas school grades tend to favor young women over young men regardless of the subject or material (Voyer & Voyer, 2014). As can be seen from the results of the National Assessment Studies in Germany (e.g., Pant et al., 2013; Stanat, Böhme, Schipolowski, & Haag, 2016), there are still substantial differences between young women and men regarding their achievement. In math at the Gymnasium (the most demanding school track in Germany), young men outperform young women in Grade 9 in all areas by about 0.13 to 0.34 SDs. In biology, young women tend to outperform young men by about 0.10 SDs, whereas in physics, young men outperform young women (0.13–0.20 SDs; Schroeders, Penk, Jansen, & Pant, 2013). Similarly, in English, young women outperform young men (0.13–0.21 SDs; Böhme, Sebald, Weirich, & Stanat, 2016). Regarding domain-specific self-concept and interest, gender differences have consistently been reported in various countries and samples, with higher self-concept and interest in math for young men, but higher ratings in reading and foreign language for young women (Jansen et al., 2014).

For gender differences in subject-specific self-concepts and interests in Germany, Jansen, Schroeders, and Stanat (2013) and Böhme et al. (2016) showed that, besides young women’s self-concept in English and biology, which did not differ from young men’s self-concepts in these areas, motivation was generally higher for the higher performing gender (e.g., self-concept and interest for young men in math and physics). According to Nagy et al. (2010), at least for math self-concept, such differences seem to remain relatively stable over the course of students’ school careers.

Theory strongly suggests that breaking up traditional course-choice patterns and forcing students who would typically not enroll in a certain advanced course (e.g., young men in English) to enroll in this course along with students who would typically enroll in this course (e.g., young women in English) should have an impact on both motivational and related achievement constructs (e.g., Hübner et al., 2017).

The German education system and the reform of the upper secondary school system

The development of CI in the United States is the best-known example, but the trend can be observed worldwide (e.g., Hughes, 1997).
In Germany, a trend toward CI in STEM subjects has been easy to identify since the beginning of the new millennium for upper secondary, pre-university education. Although math and the sciences have played central roles in the curriculum for a long time (Hofstein, Eilks, & Bybee, 2011), the results of the Trends in International Mathematics and Science Study (TIMSS) in 1998 (Mullis et al., 1998), and the “PISA-shock” in 2001, were the starting points of an ongoing discussion on how to further increase the roles of these subjects.

In the years between 2001 and 2012, 11 of the 16 German states reformed their upper secondary school systems (Trautwein & Neumann, 2008) by reducing course choice and by introducing mandatory participation in core subjects on an advanced course level (e.g., mathematics, one subject from the field of natural sciences, and one foreign language). The implementation of these reforms was made possible by the Husumer resolutions from 1999 in which the Standing Conference of the Ministers of Education and Cultural Affairs of the States in the Federal Republic of Germany (KMK) decided to reduce the advanced course hours from 5 to 4 hr per week, if students took three or more advanced courses (Trautwein et al., 2010). The reform of upper secondary school (Grades 11 and 12) in Thuringia was implemented 10 years after the Husumer resolutions, in 2009. Therefore, the first cohort after the reform graduated in 2011, after 2 years of upper secondary school. This was the first cohort that took part in the changed course system (in Grades 11 and 12), and all cohorts that followed (2010, 2011, etc.) did the same. The cohorts from before the reform (i.e., who started upper secondary school in 2008 and graduated after 2 years in 2010) were all taught according to the old regulations before the reform. Students had been taught according to the old regulations from the late 1990s until the reform in 2009 (e.g., Thüringer Kultusministerium, 1999b). The reform had two goals: first, to increase the comparability of final examinations within and between states by focusing on specific subjects, and second, to increase students’ performance in these core subjects. In Thuringia, the reform mainly affected detracking, instructional time, course level, and the importance of core subjects.

Regarding the four dimensions of CI mentioned above, the reform clearly affected detracking (see Table 1): Whereas students were enrolled in an advanced course in either math or German before the reform and in a basic course in the other, they were all enrolled in both courses on an advanced course level afterwards. Furthermore, after the reform, students were also almost all together in one advanced course in English, whereas they were clearly tracked before the reform (see Table 2).

Regarding the second aspect, the increase in instructional time, before the reform, students self-selected into two advanced (6 hr per week) and two basic courses (4 or 3 hr per week, respectively) at the beginning of upper secondary school (Grade 11) for the rest of upper secondary school (Grades 11 and 12). Besides these four courses, students also had to participate in several other basic-level courses during their time in upper secondary school. After the reform, an upper secondary school system with reduced choice options was implemented: Since then, all students have had to participate in obligatory advanced courses in mathematics and German and have had to choose three other advanced courses: one foreign language, one science, and one social studies course (all courses 4 hr each per week; see Table 3).

Generally speaking, advanced courses followed a more demanding curriculum, were taught for more hours (about 6 hr per week) before the reform, and counted more
heavily toward the final upper secondary examination grade. Basic courses were less demanding, were taught for fewer hours per week (about 2 to 4 hr per week), and counted less toward the final examination grade, compared with advanced courses. However, the curricular content between basic and advanced courses overlapped, although the required level of engagement with the content was higher in advanced courses (e.g., Thüringer Kultusministerium, 2007). This means that in math, for instance, basic-course students had to know and apply basic concepts (e.g., basic knowledge about differential calculus), whereas advanced-course students were additionally required to be able to write proofs for specific concepts (e.g., proofs for the specific rules of differential calculus) or show further, deeper knowledge of these concepts (e.g., how to apply the nth derivative of a function; e.g., Thüringer Kultusministerium, 1999a).

Third, the curriculum in these five subjects resembled the advanced-course curriculum from before the reform (cf. Wagner et al., 2011). This means that after the reform, the requirements of these courses were similar to those of the advanced courses from before the reform (see Tables 1 and 3).

Finally, the changes in tracking procedures, allocated time, and course curriculum led to a change in the importance of these subjects for postsecondary education selection, which is mainly based on final examination grades. Whereas before the reform, students
were able to build a rather unique profile of advanced courses, which were given larger weights in the final examination grades, after the reform, students’ course profiles were much more similar, and thus, the weights of the final examination grades from these courses were also more similar for students’ final grades in upper secondary school.

All of the changes mentioned above were enacted by law and implemented by means of a top-down state policy reform by the ministry of education in Thuringia.

**Research questions**

This study was conducted to shed light on the differential effects of a CI reform on achievement in STEM subjects, English reading competence, and motivation. We analyzed representative data of students collected just before and right after a CI reform in one German state, making use of a cohort control design (Shadish, Cook, & Campbell, 2002). We had three major goals: First, we investigated whether there would be main effects of CI in upper secondary school. Previous research has mostly focused on effects in lower secondary school (e.g., high school). Regarding achievement, it was difficult to anticipate main effects because the reform led to multiple changes related to detracking, instructional time, the introduction of mandatory advanced courses, and the different importance of subjects for postsecondary education (Research Question 1: Are there main effects and differential effects of the CI reform in upper secondary school on students’ achievement?).

Second, not only did we include achievement measures in our evaluation, but we also analyzed potential effects on motivational variables. Motivational factors play a major role in further achievement and should be sensitive to aspects of CI such as changing classroom composition. However, only a limited amount of research has investigated the effect of mandatory enrollment on motivational variables. As suggested by prominent theories on student motivation, increasing external control (e.g., by reducing individual course choice options and introducing mandatory enrollment) might lead to decreases in intrinsic motivation (e.g., Deci, Koestner, & Ryan, 2001; Eccles & Wigfield, 2002). Furthermore, as suggested by the BFLPE literature (e.g., Marsh et al., 2007), a higher achieving reference group might result in a reduction in students’ self-concept given comparable individual student achievement. Based on this, CI might in fact increase negative reference group effects on students’ self-concept after the reform, particularly if they would have enrolled in basic courses before the reform but were grouped together with all students (including students who would have enrolled in advanced courses in the old system) afterwards. Finally, according to expectancy-value theory (e.g., Eccles & Wigfield, 2002), this might also trigger negative effects on students’ values as competence-related beliefs and values are positively related. Hence, we expected effects for at least some of the students. At the same time, we were not sure whether we would find main effects of motivation (Research Question 2: Are there main effects on motivational variables such as subject-specific self-concepts and subject-specific interests?).

Third, we evaluated differential reform effects, focusing on potential differences between young men and women, both before and after the reform. Generally, as evident from Tables 1 and 3, CI went along with mandatory course enrollment in German, mathematics, one foreign language, and one science subject on an advanced level. On this basis, we expected that advanced course achievement would generally
decrease due to increased student heterogeneity and reduced instructional time and that young men’s achievement in English would increase, due to, on average, increased instructional time for this subgroup. For motivational outcomes, we expected reference group effects and therefore, for example, that young women’s average academic self-concept would decrease in mathematics. Recent research has shown substantial differences in the achievement and motivation between young men and women (e.g., Hübner et al., 2017), and this has oftentimes been offered as a reason for different course choice patterns (e.g., Nagy et al., 2008; Nagy, Trautwein, Baumert, Köller, & Garrett, 2006; Watt, Eccles, & Durik, 2006). In line with this, it was a central goal of the reform to decrease variability in achievement between different subgroups (e.g., gender; Research Question 3: Are there differential effects of the reform on student achievement, subject-specific self-concepts, and subject-specific interests?).

**Method**

**Description of study and sample**

We used data from the Additional Study Thuringia (Blossfeld, Rossbach, & von Maurice, 2011; Wagner et al., 2011) from the National Educational Panel Study (NEPS), included in the Scientific Use File 2.0.0. This data set contains representative data from the last cohort before (2010) and the first cohort after the reform (2011), collected at the end of upper secondary school – a cohort control design (e.g., Shadish et al., 2002). Thus, the implementation of the upper secondary school reform provided a foundation for a natural experiment setting.

Overall, 32 schools were randomly drawn from a population of 105 upper secondary schools in Thuringia, and all students from the specific cohort of interest at the school were asked to participate in the study. In the end, 30 schools participated at both time points, with approximately 2,000 students; Cohort 1: \( N = 1,316 \) (participation: 70.9%, age: \( M = 18.4 \) years); Cohort 2: \( N = 886 \) (participation: 63.6%, age: \( M = 18.3 \) years). There are two reasons for the lower number of participants at the second measurement point: First, the gross sample decreased by about 25% due to lower birth rates. Second, at the second assessment point, the participation ratio decreased by about 7.6%. As described in the Results section, this did not have an impact on cohort differences in observed covariates.

**Instruments**

In this study, we analyzed effects of the reform on competencies in mathematics, English reading competence, physics, and biology, as well as on domain-specific self-concept and interest. Further details regarding the instruments and statistical analysis can be found in the supplemental online material.

**Competence in mathematics**

The mathematics test focused on mathematical literacy, which is also referred to in the assessment of education standards and PISA (e.g., OECD, 2004). Students had 30 min to work on this part of the test. Reliability was acceptable (reliability of the weighted likelihood estimator: WLE = .67).
**Competence in English reading**
The English reading test was based on items that were developed by the Institute for Educational Quality Improvement (IQB; Rupp, Vock, Harsch, & Köller, 2008). Students had 30 min to work on 21 items (in each booklet) out of 33 overall items in a multiple-matching or multiple-choice format (NEPS, 2011). The reliability of this test was good (WLE reliability = .77).

**Competence in biology**
Competence in biology was measured with items from the EVAMAR II-study (Evaluation der Maturitätsreform; Eberle et al., 2008). Students had 45 min to work on a subset of 18 items out of a total of 126 items, which were presented in a multiple-choice and open-answer format (NEPS, 2011). The reliability of this test was acceptable (WLE reliability = .61).

**Competence in physics**
Students had 45 min to work on a competence in physics test that was comprised of 55 items (17 to 18 items in each booklet). Some items were taken from the TIMSS study (Baumert, Bos, & Lehmann, 2000), and some were developed for the NEPS Additional Study Thuringia (WLE reliability = .55).

**Domain-specific self-concept**
Domain-specific self-concept was measured with a four-item test that was based on the Self-Description Questionnaire III (Marsh & O’Neill, 1984). The internal consistencies of the four scales (e.g., “I get good marks in mathematics”; “I have never done well in mathematics”) were high in our sample (math: Cronbach’s $\alpha = .94$; English: $\alpha = .94$; biology: $\alpha = .93$; physics: $\alpha = .93$). Negatively formulated items were reverse coded.

**Domain-specific interest**
Domain-specific interest was measured with a four-item test that was based on Eccles and Wigfield (2002) and adapted for mathematics, English, biology, and physics. The scales (e.g., “When I do mathematics, I sometimes get totally absorbed”; “Mathematics is simply an exciting subject”) showed high internal consistencies in our study (math: Cronbach’s $\alpha = .88$; English: $\alpha = .86$; biology: $\alpha = .91$; physics: $\alpha = .93$) and in previous studies (e.g., Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006).

**Covariates**
We controlled for further variables in the adjusted models such as gender, socioeconomic background, number of books available at home, migration background, class repetition, and cognitive ability.

**Statistical analysis**
First, we analyzed differences in central covariates between the two cohorts (i.e., before vs. after the reform) by computing separate bivariate regression models with the covariates as the dependent variables and a reform dummy as the independent variable as well as survey weights of the Additional Study Thuringia. This was done in order to
identify potential differences between the two cohorts on these covariates. Next, we investigated grade-repetition rates, school-leaving rates after lower secondary school, and transition rates using data from the Statistics Agency of Thuringia to test for possible threats to validity.

To test course choices for students before versus after the reform in English reading, biology, and physics, we additionally specified multinomial logistic regression models with course-level participation (basic, advanced, dropout) as the dependent variable and cohort membership as the independent variable. We could not test for differences in mathematics because the advanced course was mandatory after the reform (all students had to take the same math course). That is, the population parameter for the choice of an advanced course in mathematics after the reform was $\pi = 1.0$. Therefore, if the sample probability before the reform was not $p = 1.0$ (which was clearly the case as can be seen in Table 2), we could conclude that there were differences between the cohorts.

In these models, we further specified Wald tests to test the null hypothesis of no differences between cohorts in course-choice patterns. On the basis of the results of these models, we specified logistic regression models to test for differences in course-choice patterns for each subject and course level.

Achievement outcomes were analyzed with unidimensional and multidimensional two- (2PL) and one-parameter (1PL) logistic item response theory (IRT) models. We estimated 1PL and 2PL multiple IRT (MIRT) models, respectively, each in a single model with cohort-specific structural models (multiple group) and measurement models held constant across groups using a latent class mixture modeling framework, implemented in Mplus 7.4 (Muthén & Muthén, 1998–2012), to adequately address the unreliability of the achievement measures. The quality of the test was evaluated beforehand with regard to reliability, item fit, as well as uniform and nonuniform differential item functioning (DIF) for sex, cohort, migration background, and socioeconomic status. Results for domain-specific self-concept and domain-specific interest were based on simple structural equation models in which the indicators were assumed to be metric.

As recommended by McNeish, Stapleton, and Silverman (2017), we used survey weights and cluster sampling by robust standard errors to consider the selection probability in all models. We used the Benjamini-Hochberg procedure to correct for multiple testing (Benjamini & Hochberg, 1995). All analyses of adjusted and unadjusted (M)IRT models were conducted with full information maximum likelihood (FIML) as there is a growing consensus that multiple imputation (MI) or FIML estimation is superior to traditional methods (e.g., Enders, 2010; Graham, 2009).

In order to partly overcome the limitations and to strengthen the assumption to which our study resembled a natural experiment, we included many additional analyses and robustness checks: First, the cohort control design, whereby we investigated two consecutive, representative cohorts (one cohort before and one cohort after the reform), supported our expectation that differences (e.g., due to selection) would be small (e.g., Shadish et al., 2002). This expectation was further supported by a check for differences on observed covariates (e.g., gender, migration background, highest international socioeconomic index of occupational status, number of books in the home, and general cognitive ability; see supplemental online material), of which none were statistically significantly different between the cohorts, minimizing the threat of potential selection. Furthermore, as the reform was announced when students were already enrolled in
lower secondary school (when they were in Grade 9; e.g., Thüringer Kultusministerium, 2007), differential student selectivity from elementary to lower secondary school was expected to be very small. We extended tests of student selectivity to upper secondary school using additional data from the Federal Statistical Office related to grade repetition behavior and school leaving after lower secondary school, all of which were, in line with our expectations, very small. Furthermore, we specified adjusted models (including further covariates) and unadjusted models (without further covariates), and the two models revealed comparable results in most cases. Finally, for the achievement tests, we also tested for differential item functioning (DIF), which was rare and would have indicated item bias or potential differences between subgroups on specific items.

**Results**

**Preliminary analysis**

We first investigated possible differences between students who participated before versus after the reform on the assessed covariates. None of the differences between the two groups were statistically significant (see Table A1 in the supplemental material).

Next, we took a closer look at the process of transitioning to upper secondary school and analyzed possible differences with regard to grade repetition behavior and school leaving after lower secondary school, using population data from the Statistics Agency of Thuringia. Comparing data from the last 5 years before the reform with data collected since 2010, we found minor differences in school transition rates. Before the reform, according to the population data, on average, 94.4% of students in Grade 10 moved to Grade 11, whereas around 91.9% of the students moved to Grade 11 after the reform. Regarding grade-repetition rates, an average share of 2.3% of students repeated Grade 10 before the reform, whereas 1.6% of students repeated Grade 10 after the reform. Before the reform, an average of 3.7% of students left school after Grade 10, whereas afterwards, this share came to 4.2%. We also checked for possible differences in transition and grade repetition shares during upper secondary school but did not find substantial differences between students measured before versus after the reform.

**Course choice and allocated time**

After taking a closer look at potential student selection during the process of transitioning to upper secondary school (e.g., in terms of grade repetition and school leaving), we investigated differences in allocated time and course choice behavior before and after the reform. Following the selection analysis, we tested for differences in course choice before versus after the reform, using multinomial logistic regression models and Wald tests (Table 2). As expected, we found statistically significant differences in course-choice rates for all subjects before versus after the reform; English: $\chi^2(2) = 42.82, p < .001$, physics: $\chi^2(2) = 49.86, p < .001$, biology: $\chi^2(2) = 86.30, p < .001$. We did not test for differences in mathematics because advanced math was mandatory after the reform. Inspecting these cohort differences in more detail, we found statistically significant differences for advanced and basic courses in all subjects (see Table 2).
Controlling the false discovery rate (FDR) by applying the Benjamini-Hochberg procedure separately for each course level did not change these results.

Examining course-choice patterns in advanced courses by gender (see Table 4) revealed two things. First, we found increases in participation rates in advanced courses for young men and young women in all subjects ($p < .001$). Second, gender differences were not statistically significant only for English and mathematics after the reform.

As expected, although participation in advanced courses increased on average, we found a decrease in the average time allocated to mathematics of 41.4 min (see Table 5). For all other subjects, we did not find statistically significant changes when comparing time allocated before versus after the reform.

**Achievement before and after the reform**

Next, we addressed the first research question, regarding potential main effects of the CI reform on overall (before vs. after the reform) and course-specific (basic vs. advanced vs. core course) achievement. Differences in achievement between the two cohorts ranged from $d = 0.04$ to $d = 0.12$ in the unadjusted model and from $d = 0.00$ to $d = 0.08$ in the adjusted model across the achievement tests (see Table A2 in the supplemental material). However, none of these differences were statistically significant after we controlled the FDR.

In addition, we tested for potential differences in achievement variability before versus after the reform. Here, no statistically significant differences were found for any of the subjects. We also specified 2PL MIRT models and models without items with

### Table 4. Choice of advanced courses by gender.

| Subject   | Cohort 1: AC in % | Cohort 2: AC in % |
|-----------|-------------------|-------------------|
|           | Young men        | Young women       | Young men | Young women |
| Mathematics | 58.1             | 34.7              | 100       | 100        |
| English    | 21.4             | 41.3              | 93.1      | 96.4       |
| Physics    | 18.1             | 3.9               | 46.0      | 18.4       |
| Biology    | 16.3             | 23.8              | 41.6      | 73.2       |

Note: All differences within genders were statistically significant ($p < .001$). We did not find significant gender differences ($p < .05$) between young men and young women for English in Cohort 2 only. We did not test for differences in mathematics because advanced math was mandatory after the reform. If differences were not statistically significant after the BH correction, they are labeled with BH. Cohort 1 = Cohort before the reform; Cohort 2 = Cohort after the reform. The results are from analyses in which the weights and cluster structure of the data were taken into consideration.

### Table 5. Average time allocated per week.

| Subject   | Cohort 1 | Cohort 2 | Difference | $p$  | $p_{adj}$ |
|-----------|----------|----------|------------|------|-----------|
| Mathematics | 4.91     | 4        | -0.91      | 4    | -         |
| English    | 3.80     | 3.89     | 0.09       | 0.287| 0.430     |
| Physics    | 1.55     | 1.73     | 0.18       | 0.192| 0.430     |
| Biology    | 2.68     | 2.66     | 0.02       | 0.867| 0.867     |

Note: Average hours were calculated in accordance with official information on obligatory course hours. We did not test for significant differences in math because advanced math was mandatory after the reform (i.e., the population parameter for the choice of advanced courses in mathematics after the reform was $n = 1.0$). Therefore, if the sample probability before the reform was not $p = 1.0$ (which was clearly the case as can be seen in Tables 4 and 5), we could conclude that there were differences between the cohorts. Cohort 1 = Cohort before the reform; Cohort 2 = Cohort after the reform; $p_{adj} =$ Benjamini-Hochberg-corrected $p$ values. The results are from analyses in which the weights and cluster structure of the data were taken into consideration.
severe DIF to check the robustness of our results, but results remained stable. Note that items exhibiting severe DIF were found only for physics and biology.

Taking a closer look at course-specific student achievement before versus after the reform (see Table 6) indicated a statistically significant decline in all advanced courses. We expected this effect due to the increased heterogeneity and reduction of 2 hr per week in advanced courses. Differences between advanced courses before versus after the reform were very prominent in physics ($d = -0.77$, $p = .011$) but also clearly visible in mathematics ($d = -0.50$, $p < .001$), biology ($d = -0.48$, $p = .001$), and English reading ($d = -0.39$, $p < .001$). Comparing course-specific achievement by cohort, we found a statistically significant Course Level × Cohort interaction in English reading ($d = 1.05$, $p = .001$), indicating an increase in average achievement in basic courses and a decrease in average achievement in advanced courses after the reform. In addition, we found a statistically significant Course Level × Cohort interaction in biology ($d = 0.62$, $p = .001$). Here, achievement in advanced courses decreased, whereas achievement in basic courses remained constant (see also Figure A1 in the supplemental online material).

In the adjusted model, the interaction effect in English reading was statistically significant but changed its direction ($d = 0.14$, $p < .001$), indicating that students in basic courses performed higher on average after the reform than students in advanced courses.

### Table 6. Mean levels of student achievement by course and by gender.

#### Unadjusted results of unidimensional 1PL IRT model

| Course       | Cohort 1 | Cohort 2 |
|--------------|----------|----------|
|              | AC_a     | BC_b     | CS/AC_c | BC_d     |
| Mathematics  | 55.1bc   | 44.7(0.9)ac | 50.1 (0.9)ab | –        |
| English reading | 52.8bc   | 45.8(0.9)acd | 48.9 (0.9)ab | 52.4 (3.6)b |
| Physics      | 60.5bcd  | 45.3(3.0)acd | 52.8 (2.4)abcd | 41.4 (2.4)abc |
| Biology      | 54.0bcd  | 47.7(1.4)a    | 49.2 (1.5)a  | 49.1 (1.6)a |

#### Adjusted results of unidimensional 1PL IRT model

| Course       | Cohort 1 | Cohort 2 |
|--------------|----------|----------|
| Mathematics  | 52.7bc   | 47.2 (0.6)ac | 50.1 (0.7)ab | –        |
| English reading | 53.5bcd  | 47.2 (0.8)bcd  | 49.9 (0.7)ab  | 45.0 (0.9)ab |
| Physics      | 57.2bcd  | 47.1 (2.7)bcd  | 52.1 (2.2)bd  | 43.6 (2.1)bd |
| Biology      | 54.0bcd  | 47.9 (1.0)bc   | 49.9 (1.1)ab  | 48.1 (1.7)a |

#### Mean levels of student achievement by gender

| Course       | Cohort 1 | Cohort 2 |
|--------------|----------|----------|
|              | Young men_a | Young women_b | Young men_c | Young women_d |
| Mathematics  | 52.7bd   | 46.6 (0.1)ac  | 53.9 (1.2)bd  | 46.8 (1.0)ac  |
| English reading | 48.1bcd  | 50.6 (0.9)ac  | 50.5 (0.8)a  | 50.7 (0.8)a  |
| Physics      | 54.0bd   | 45.4 (0.9)ac  | 53.9 (1.5)bd  | 46.7 (1.3)ac  |
| Biology      | 49.4     | 50.2 (0.8)ac  | 49.8 (1.0)ac  | 50.6 (0.9)ac  |

#### Adjusted results of unidimensional 1PL IRT model

| Course       | Cohort 1 | Cohort 2 |
|--------------|----------|----------|
| Mathematics  | 52.1bd   | 47.6 (0.6)ac  | 53.2 (1.0)bd  | 47.2 (0.7)ac  |
| English reading | 47.6bcd  | 51.4 (0.6)ac  | 48.4 (0.6)bd  | 51.2 (0.6)bd  |
| Physics      | 53.6bcd  | 46.1 (0.9)ac  | 53.3 (1.3)bd  | 47.0 (1.1)bd  |
| Biology      | 49.0     | 51.0 (0.7)ac  | 49.1 (0.8)ac  | 50.9 (0.8)ac  |

Note: Cohort 1 = Cohort before the reform; Cohort 2 = Cohort after the reform. AC = Advanced Course; BC = Basic Course; CS = Core Subject. Results of 1PL models are displayed with and without controlling for differences on further covariates. Due to small sample sizes in the basic English course, variances and covariances were not estimated in this group for gender, socioeconomic background, migration, and grade repeaters to avoid singularity in the information matrix. Intercepts for advanced courses and the basic course before the reform were identical in models that did and did not consider the students in the basic courses after the reform. The metric of the latent variable was transformed to $M = 50$ and $SD = 10$ on the basis of pooled means and standard deviations. Indices indicate two-sided statistically significant group differences ($p < .05$). If differences were not statistically significant after the BH correction, they are labeled with BH. The results are from analyses in which the weights and cluster structure of the data were considered.
after the reform. This most likely resulted from a small group of students who had a
special focus on foreign languages (a different first foreign language in addition to English
as a basic course). However, the interaction effect in biology remained stable ($d = 0.43,
p = .017$). Results from 2PL IRT models and models without items exhibiting severe DIF did
not differ meaningfully. Controlling the FDR did not change any of these results.

**Gender-specific achievement before and after the reform**

In the next step, we investigated differential effects of the CI reform on achievement
measures. Regarding gender-specific achievement (Table 6), we expected that gender
differences would be very prominent for subgroups in which a potentially huge share of
students would be affected by the reform, namely, young men in English.

Our analysis revealed that in English reading, young women outperformed young
men before the reform ($d = -0.25, p = .005$), but this did not hold afterwards ($d = -0.02,
p = .804$). Here, we found a statistically significant Cohort × Sex interaction in the
adjusted model ($d = -0.10, p = .009$), indicating a decrease in the gender disparity
after the reform: Whereas young women outperformed young men before the reform,
the achievement levels of the two groups did not differ afterwards. After controlling the
FDR, this effect was still statistically significant in the adjusted model ($p = .019$). In the
unadjusted model, this effect was not statistically significant ($d = -0.23, p = .066$).

Regarding math, young men performed better than young women before ($d = 0.61,
p < .001$) and after the reform ($d = 0.71, p < .001$). However, the Cohort × Sex interaction was
not statistically significant for mathematics ($d = -0.10, p = .154$), indicating no statistically
significant change in the gender gap from before to after the reform in mathematics (see
Figure 1). Considering achievement in physics, we again found gender differences before
($d = 0.86, p < .001$) and after the reform ($d = 0.72, p < .001$), but the change in achievement
differences between young men and women in physics before versus after the reform,
displayed by the Cohort × Sex interaction effect, was not statistically significant ($d = 0.14,
p = .386$). These interaction effects were not different in the 2PL MIRT models.

**Domain-specific self-concept and interest before and after the reform**

We completed our evaluation by considering two social cognitive constructs, namely,
domain-specific self-concept and domain-specific interest. We first took a closer look at
average differences before and after the reform, and then we analyzed gender-specific
differences more closely. First, we did not find any differences in average domain-specific
self-concept before or after the reform for any of the subjects. Second, we did find gender-
related statistically significant differences in domain-specific self-concept: Whereas young
men had higher self-concepts in mathematics and physics, young women had higher self-
concepts in English and biology. This pattern was robust for all comparisons except for
English after the reform, where we did not find a statistically significant difference
between young men and young women ($d = -0.08, p = .320$). Our most interesting finding
was a statistically significant Cohort × Sex interaction for mathematics self-concept
($d = -0.35, p = .012$), driven by a lower self-concept of young women after the reform.
By contrast, the same interaction for English self-concept was not statistically significant
($d = -0.20, p = .078$), although young men’s achievement was statistically significantly
higher after the reform than before the reform ($d = -0.22, p = .017$). These effects remained stable in the adjusted models and when we controlled the FDR.

Concerning domain-specific interest, similar to the results for self-concept, we did not find any statistically significant average differences between young men and young women before versus after the reform. However, in all subjects except mathematics, all gender differences within a cohort were statistically significant (see Table A3 and Figures A2 and A3 in the supplemental material).

**Discussion**

This study sheds light on differential effects of a CI reform on main and differential effects on achievement and motivation in STEM subjects and English in upper secondary school. We investigated differences in student achievement before versus right after the policy reform was implemented for all upper secondary schools in the state of Thuringia, showing that, overall, the reform had no statistically significant impact on average student achievement.

For the dimensions of CI, we found strong evidence for changes in tracking patterns, which resulted from increased enrollment in advanced courses. This finding was
prominent for subgroups in which potentially huge shares of students were affected by CI (e.g., young men in English).

Furthermore, we did find evidence for increased achievement in English for young men. Results indicate that, besides subject-specific differences, changing course level alone did not lead to changes in achievement. This held for both groups that were traditionally the majority (young men) and groups that were traditionally the minority (young women) in advanced courses in mathematics. In English, however, all aspects of CI were affected, including instructional time. This seemingly had an impact on young men who have traditionally been the minority in advanced English courses.

**Practical implications**

Besides finding poor support for the positive effects of this reform on achievement measures, we did find subgroup effects that might be cause for some concern. In line with previous research (e.g., Hübner et al., 2017), our results appear to suggest that the reform had somewhat of an adverse effect on math self-concept and that this effect seemed to be triggered in students who tended to enroll in advanced courses less often before the reform (e.g., young women) and were faced with a generally stronger reference group after the reform. As outlined in the theory, motivational constructs, especially math self-concept, plays a major role in future STEM career choices (Eccles, 1983; Jansen et al., 2014; Parker et al., 2012); however, in this regard, the results of our study instead indicate a potential widening of the STEM career gap. These findings are also in line with Hübner et al.’s (2017) results, which pointed to negative effects of a similar reform in a different state on young women’s math self-concept.

Furthermore, results of our study can be integrated into the discussion in the literature on how to shape sustainable educational change and foster educational improvement. As the OECD pointed out in their *Education Policy Outlook 2015*, there is a “need for effective education policy reforms” (OECD, 2015, p. 22) so that the current and upcoming economic and sociopolitical challenges can adequately be faced. Evaluations of educational reforms should be a natural part of a sustainable, evidenced-based accountability policy. Failing to do so might be highly problematic not only for the question of “what works” but even more so for the question of “what does not work” (e.g., Reynolds et al., 2014).

This aspect is of special importance when promoting educational policy reforms as a major instrument for change. In fact, not only do educational policy reforms generally improve educational outcomes and lead to the desired effects, but they can also introduce or foster unintended side effects as shown in this and various other studies (e.g., Domina et al., 2015; Gross, Booker, & Goldhaber, 2009; Hübner et al., 2017). In addition, the results of this study support the claim of other studies that similar reforms inherently lead to similar effects in different educational environments and for all participating students (e.g., Mehan, Hubbard, & Stein, 2005).

**Limitations and future prospects**

The study we used to analyze the impact of the CI policy reform contained cross-sectional data in a cohort control design (Shadish et al., 2002), where students were
assessed before and right after the implementation of the reform. However, lower birth rates in the population after the reform resulted in a considerably lower gross sample size compared with the sample after the reform. We tried to address this issue by introducing adjusted models, where we statistically controlled for the impact of further covariates (e.g., socioeconomic status, cognitive ability) on our outcomes, and various robustness checks regarding the selectivity and sensitivity of our results to model specification issues. Although the students did not differ on these measures, we could not formally test whether the populations differed on unobserved covariates.

Although we implemented multiple robustness checks and considered additional official data from the Statistics Agency of Thuringia, all of which strengthened the assumption that the quasi-experimental cohort control design (Shadish et al., 2002) we used in this study can be considered a natural experiment, this still remained an assumption. Furthermore, we were not able to apply further advanced methods (e.g., difference in differences approaches) due to a lack of available data. However, it has to be noted that such methods also come with a variety of different assumptions (e.g., common trend assumption; see, e.g., Angrist & Pischke, 2009), which are, depending on the federal state, not necessarily justifiable in our context. Future research should shed light on the longitudinal effects of policy reforms that reduced course-choice options in upper secondary school. Considering longitudinal data could provide important answers about the practical significance of reductions in young women’s math self-concept for future STEM career choices (e.g., Hübner et al., 2017). Furthermore, such designs could allow researchers to apply other advanced methodological approaches to further test the robustness of reform effects. Another important question that we addressed only in part involves the different CI effects of course level and allocated time on achievement. In our analyses, we found evidence that both time and course level affect achievement. However, we could not clearly disentangle the two effects from each other because the effects were confounded with other variables (e.g., change in student composition).

It is important to note that the term CI does not have a clearly defined general overlap with other terms, such as curricular reform (e.g., Fenwick, 2011; Robinsohn, 1967), as the goal of CI is primarily to achieve intensification through changes in structures. In theory, CI usually goes along with the increased enrollment of students in higher level coursework. Therefore, what oftentimes precedes processes of CI are substantial (e.g., structural) education system reforms that foster CI, for instance, through mandatory enrollment in specific advanced courses, which, however, might follow a curriculum that already existed previously (e.g., Nomi & Raudenbush, 2016; Trautwein et al., 2010). Therefore, the curricula of such courses can remain untouched.

With regard to German educational policy reforms, unfortunately, previous research has failed to link these reforms to more general debates about large-scale reform movements (e.g., the standards-based reform movement and CI reforms) or general theories of time and achievement (e.g., Carroll, 1963, 1989) or motivation and achievement (e.g., Eccles & Wigfield, 2002). Furthermore, the four elements of CI have been only loosely discussed in previous research on upper secondary school reforms in Germany (e.g., Hübner et al., 2017; Trautwein et al., 2010; Wagner et al., 2014). This can be partly traced back to the complex nature of policy reforms (e.g., Jann & Wegrich, 2007; Rogers, 2003), which oftentimes prevent findings from being generalized to a broader theoretical context. Future research should therefore address this issue in order to more strongly
integrate policy reforms and theoretical concepts. Not only could this help to overcome recent limitations in anticipating potentially intended and unintended reform effects, but it could also help in testing and developing theory in applied contexts. Adequate judgments about the effectiveness of a reform strongly rely on the modus of the implementation of the reform by the teacher in the classroom, and we have numerous reasons to believe that teachers truly implemented the reform at the specified time. First, the reform was implemented in a top-down fashion by a change in the school law. Beginning in 2009, all schools were forced (by law) to implement the new upper secondary school structure with the specific course choice restrictions (with a largely unchanged course-level-specific curriculum). All students from one cohort were officially part of the old or the new system (there were no students in the same grade who were in different systems). Second, students from before the reform were already in Grade 12 when the reform was implemented for the next student cohort in Grade 11. The regulations of the new school structure (including core courses) were published after the (prereform) students had already been enrolled in Grade 11 for half a year (Thüringer Kultusministerium, 2008). This made it highly unlikely that teachers would have already been able to implement these new regulations before the reform. However, although it seems unlikely that teachers would have decided to teach the old curriculum, also because there were new standardized examinations at the end of upper secondary school for all schools after the reform, we should mention that we have no information on the extent to which teachers were truly able to adhere to the “officially defined” advanced level of the new core courses. Such information (e.g., lesson protocols and timetables) should be gathered in future research to be able to increase knowledge on how teachers truly implemented the legally predefined curricular changes in class.

Conclusion

The results of this study showed that the CI reform in upper secondary school, whereby all students were literally “put in the same baskets (classes),” did not automatically produce the intended effects of increased achievement and less heterogeneity in achievement. In sum, we found first evidence for differential effects on English reading and a higher English self-concept of young men after the reform, whereas the reform had a negative effect on young women’s math self-concept. The study underscores the importance of carefully planning systemic reforms and strengthens the importance of conducting systematic evaluations during processes of educational change.

Acknowledgements

This paper uses data from the National Educational Panel Study (NEPS): Additional Study Thuringia, doi:10.5157/NEPS:TH:2.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

We thank Eike Wille and Thurston Domina for their useful feedback on a previous version of this manuscript.
Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by the German Research Foundation [TR 553/9–1] as part of the Priority Program: “Education as a Lifelong Process” [SPP 1646].

Notes on contributors

Nicolas Hübner is a postdoctoral researcher at the Hector Research Institute of Education Sciences and Psychology, collaborator of the National Educational Panel Study (NEPS), and associated with the LEAD Graduate School & Research Network. His research interests include educational effectiveness and evaluation, especially the interplay between educational policy reforms, student achievement, and motivation and the application of quantitative methods.

Wolfgang Wagner is a postdoctoral researcher at the Hector Research Institute of Education Sciences and Psychology. His main research interests include educational effectiveness and the assessment of characteristics of learning environments and their effects on the development of academic achievement as well as methodological issues in the field of multilevel structural equation models.

Benjamin Nagengast is professor at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen. His research interests include quantitative methods (causal inference, latent variable models, multilevel modeling), motivation and academic self-concept, educational effectiveness, and the evaluation of educational interventions.

Ulrich Trautwein is professor at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen. His research interests include the development of student motivation, personality, academic effort, and achievement.

ORCID

Nicolas Hübner http://orcid.org/0000-0003-3528-8086
Wolfgang Wagner http://orcid.org/0000-0001-9781-4630
Benjamin Nagengast http://orcid.org/0000-0001-9868-8322
Ulrich Trautwein http://orcid.org/0000-0003-0647-0057

References

Allensworth, E., Nomi, T., Montgomery, N., & Lee, V. E. (2009). College preparatory curriculum for all: Academic consequences of requiring algebra and English I for ninth graders in Chicago. Educational Evaluation and Policy Analysis, 31(4), 367–391. https://doi.org/10.3102/01623709343471
Angrist, J. D., & Pischke, J.-S. (2009). Mostly harmless econometrics: An empiricist’s companion. Princeton, NJ: Princeton University Press.
Baumert, J., Bos, W., & Lehmann, R. (Eds.). (2000). TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie: Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe [Third International Mathematics and Science Study: Mathematics and science
education at the end of school. Vol. 2: Competencies in mathematics and physics at the end of upper secondary school]. Opladen: Leske + Budrich.

Benjamini, Y., & Hochberg, J. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.

Blossfeld, H.-P., Rossbach, H. G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft*, 14(Suppl 2).

Böhme, K., Sebald, S., Weirich, S., & Stanat, P. (2016). Geschlechtsbezogene Disparitäten [Gender-related disparities]. In P. Stanat, K. Böhme, S. Schipolowski, & N. Haag (Eds.), *IQB-Bildungstrend 2015: Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich* (pp. 377–408). Münster: Waxmann.

Callahan, R., Wilkinson, L., & Muller, C. (2010). Academic achievement and course taking among language minority youth in U.S. schools: Effects of ESL placement. *Educational Evaluation and Policy Analysis*, 32(1), 84–117. https://doi.org/10.3102/0162373709359980

Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, 64, 723–733.

Carroll, J. B. (1989). The Carroll model: A 25-Year retrospective and prospective view. *Educational Researcher*, 18(1), 26–31. https://doi.org/10.3102/0013189X018001026

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology*, 27(5), 703–722. https://doi.org/10.1037/0012-1649.27.5.703

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type: An international comparison of students’ mathematics self-concept. *American Educational Research Journal*, 50(5), 925–957. https://doi.org/10.3102/002831213489843

Cortes, K. E., Goodman, J. S., & Nomi, T. (2015). Intensive math instruction and educational attainment: Long-run impacts of double-dose algebra. *Journal of Human Resources*, 50(1), 108–158. https://doi.org/10.3368/jhr.50.1.108

Crosnoe, R., & Benner, A. D. (2015). Children at school. In R. M. Lerner (Ed.), *Handbook of child psychology and developmental science* (pp. 1–37). Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118963418.childpsy407

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, 71(1), 1–27. https://doi.org/10.3102/00346543071001001

Denissen, J. J. A., Zarrett, N. R., & Eccles, J. S. (2007). I like to do it, I’m able, and I know I am: Longitudinal couplings between domain-specific achievement, self-concept, and interest. *Child Development*, 78(2), 430–447. https://doi.org/10.1111/j.1467-8624.2007.01007.x

Domina, T., McEachin, A., Penner, A., & Penner, E. (2015). Aiming high and falling short: California’s eighth-grade algebra-for-all effort. *Educational Evaluation and Policy Analysis*, 37(3), 275–295. https://doi.org/10.3102/0162373714543685

Domina, T., & Saldana, J. (2012). Does raising the bar level the playing field? Mathematics curricular intensification and inequality in American high schools, 1982–2004. *American Educational Research Journal*, 49(4), 685–708. https://doi.org/10.3102/002831211426347

Eberle, F., Gehrer, K., Jaggi, B., Kottonau, J., Oepke, M., & Pflüger, M. (2008). *Evaluation der Maturitätsreform 1995 (EVAMAR): Schlussbericht zur Phase II* [Evaluation of the upper secondary school reform of 1995 (EVAMAR): Final report for Stage II]. Bern: Staatssekretariat für Bildung und Forschung SBF.

Eccles, J. S. (1983). Expectancies, values, and academic choice: Origins and changes. In J. Spence (Ed.), *Achievement and achievement motivation* (pp. 87–134). San Francisco, CA: W. H. Freeman.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, 136(1), 103–127. https://doi.org/10.1037/a0018053
Enders, C. K. (2010). *Applied missing data analysis (Methodology in the social sciences series)*. New York, NY: Guilford Press.

Fenwick, L. (2011). Curriculum reform and reproducing inequality in upper-secondary education. *Journal of Curriculum Studies*, 43(6), 697–716. https://doi.org/10.1080/00220272.2011.576773

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

Gross, B., Booker, T. K., & Goldhaber, D. (2009). Boosting student achievement: The effect of comprehensive school reform on student achievement. *Educational Evaluation and Policy Analysis*, 31(2), 111–126. https://doi.org/10.3102/0162373709333886

Hanushek, E. A., & Woessmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal*, 116(S10), C63–C76. https://doi.org/10.1111/j.1468-0297.2006.01076.x

Hanushek, E. A., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, 46(3), 607–668. https://doi.org/10.1257/jel.46.3.607

Hess, F. M., & Petrilli, M. J. (2006). *No Child Left Behind primer*. New York, NY: Peter Lang.

Hofstein, A., Eilks, I., & Bybee, R. (2011). Societal issues and their importance for contemporary science education – A pedagogical justification and the state-of-the-art in Israel, Germany, and the USA. *International Journal of Science and Mathematics Education*, 9(6), 1459–1483. https://doi.org/10.1007/s10763-010-9273-9

Hübner, N., Wille, E., Cambria, J., Oschatz, K., Nagengast, B., & Trautwein, U. (2017). Maximizing gender equality by minimizing course choice options? Effects of obligatory coursework in math on gender differences in STEM. *Journal of Educational Psychology*, 109(7), 993–1009. https://doi.org/10.1037/edu0000183

Hughes, M. (1997). The National Curriculum in England and Wales: A lesson in externally imposed reform? *Educational Administration Quarterly*, 33(2), 183–197. https://doi.org/10.1177/0013161X97033002006

Jann, W., & Wegrich, K. (2007). Theories of the policy cycle. In F. Fischer, G. J. Miller, & M. S. Sidney (Eds.), *Handbook of public policy analysis: Theory, politics, and methods* (pp. 43–62). Boca Raton, FL: CRC Press/Taylor & Francis.

Jansen, M., Schroders, U., & Lüdtke, O. (2014). Academic self-concept in science: Multidimensionality, relations to achievement measures, and gender differences. *Learning and Individual Differences*, 30, 11–21. https://doi.org/10.1016/j.lindif.2013.12.003

Jansen, M., Schroders, U., & Stanat, P. (2013). Motivationale Schülermerkmale in Mathematik und den Naturwissenschaften [Motivational student characteristics in math and sciences]. In H. A. Pant, P. Stanat, U. Schroders, A. Roppelt, T. Siegle, & C. Pöhlmann (Eds.), *IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (pp. 347–366). Münster: Waxmann.

Kelly, S. (2007). The contours of tracking in North Carolina. *The High School Journal*, 90(4), 15–31. https://doi.org/10.1353/hsj.2007.0016

Kultusministerkonferenz. (2005). *Bildungsstandards der Kultusministerkonferenz: Erläuterungen zur Konzeption und Entwicklung* [Educational standards of the KMK: Explications on conception and development]. München: Luchterhand.

Kultusministerkonferenz (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring* [Overall strategy of the KMK on educational monitoring]. Neuwied: Luchterhand.

Lavy, V. (2015). Do differences in schools’ instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal*, 125(S88), F397–F424. https://doi.org/10.1111/ecoj.12233

Lee, V. E., & Bryk, A. S. (1988). Curriculum tracking as mediating the social distribution of high school achievement. *Sociology of Education*, 61(2), 78–94. https://doi.org/10.2307/2112266

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal*, 23(1), 129–149. https://doi.org/10.3102/0028312023001129
Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J. S., Abdelfattah, F., Nagengast, B., ... Abu-Hilal, M. M. (2015). The internal/external frame of reference model of self-concept and achievement relations: Age-cohort and cross-cultural differences. *American Educational Research Journal*, 52(1), 168–202. https://doi.org/10.3102/0002831214549453

Marsh, H. W., & O'Neill, R. (1984). Self description questionnaire III: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement*, 21(2), 153–174. https://doi.org/10.1111/j.1745-3984.1984.tb00227.x

Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). The big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal*, 44(3), 631–669. https://doi.org/10.3102/0002831207306728

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, 22(1), 114–140. https://doi.org/10.1037/met0000078

Mehan, H., Hubbard, L., & Stein, M. K. (2005). When reforms travel: The sequel. *Journal of Educational Change*, 6(4), 329–362. https://doi.org/10.1007/s10833-005-2750-1

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1998). *Mathematics and science achievement in the final year of secondary school: IEA’s Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Nagengast, B., & Marsh, H. W. (2012). Big fish in little ponds aspire more: Mediation and cross-cultural generalizability of school-average ability effects on self-concept and career aspirations in science. *Journal of Educational Psychology*, 104(4), 1033–1053. https://doi.org/10.1037/a0027697

Nagy, G., Garrett, J., Trautwein, U., Cortina, K. S., Baumert, J., & Eccles, J. S. (2008). Gendered high school course selection as a precursor of gendered careers: The mediating role of self-concept and intrinsic value. In H. M. G. Watt & J. S. Eccles (Eds.), *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences* (pp. 115–143). Washington, DC: American Psychological Association. https://doi.org/10.1037/11706-004

Nagy, G., Trautwein, U., Baumert, J., Köller, O., & Garrett, J. (2006). Gender and course selection in upper secondary education: Effects of academic self-concept and intrinsic value. *Educational Research and Evaluation*, 12(4), 323–345. https://doi.org/10.1080/13803610600765687

Nagy, G., Watt, H. M. G., Eccles, J. S., Trautwein, U., Lüdtke, O., & Baumert, J. (2010). The development of students’ mathematics self-concept in relation to gender: Different countries, different trajectories? *Journal of Research on Adolescence*, 20(2), 482–506. https://doi.org/10.1111/j.1532-7795.2010.00644.x

National Educational Panel Study. (2011). *Curricular reform study in Thuringia – Main study 2009/10 (A70) – Students, 12th grade: Information on the competence test*. Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/TH/1-0-0/C_A70_EN.pdf

Niemann, D. (2016). Germany: The intersection of international achievement testing and educational policy development. In L. Volante (Ed.), *The intersection of international achievement testing and educational policy: Global perspectives on large-scale reform* (pp. 19–36). New York, NY: Routledge.

Nom, T., & Raudenbush, S. W. (2016). Making a success of “Algebra for All”: The impact of extended instructional time and classroom peer skill in Chicago. *Educational Evaluation and Policy Analysis*, 38(2), 431–451. https://doi.org/10.3102/0162373716643756

Organisation for Economic Co-operation and Development. (2001). *Knowledge and skills for life: First results from PISA 2000. Education and skills*. Paris: Author.

Organisation for Economic Co-operation and Development. (2004). *Learning for tomorrow’s world: First results from PISA 2003*. Paris: Author.

Organisation for Economic Co-operation and Development. (2015). *Education policy outlook 2015: Making reforms happen*. Paris: Author.
Osborne, J., & Dillon, J. (2008). Science education in Europe: Critical reflections: A report to the Nuffield Foundation. Retrieved from http://www.nuffieldfoundation.org/sites/default/files/Sci_Ed_in_Europe_Report_Final.pdf

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.). (2013). The IQB National Assessment Study 2012: Competencies in mathematics and the sciences at the end of secondary level I. Münster: Waxmann. Retrieved from https://www.iqb.hu-berlin.de/bt/lv2012/Bericht/bt/lv2012/Bericht/IQB_NationalAssesse.pdf

Parker, P. D., Schoon, I., Tsai, Y.-M., Nagy, G., Trautwein, U., & Eccles, J. S. (2012). Achievement, agency, gender, and socioeconomic background as predictors of postschool choices: A multi-context study. Developmental Psychology, 48(6), 1629–1642. https://doi.org/10.1037/a0029167

Patall, E. A., Cooper, H., & Allen, A. B. (2010). Extending the school day or school year: A systematic review of research (1985–2009). Review of Educational Research, 80(3), 401–436. https://doi.org/10.3102/0034654310377086

Penner, A. M., Domina, T., Penner, E. K., & Conley, A. (2015). Curricular policy as a collective effects problem: A distributional approach. Social Science Research, 52, 627–641. https://doi.org/10.1016/j.ssrresearch.2015.03.008

Ravitch, D. (2011). The death and life of the great American school system: How testing and choice are undermining education. New York, NY: Basic Books.

Reynolds, D., Sammons, P., De Freine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. School Effectiveness and School Improvement, 25(2), 197–230. https://doi.org/10.1080/09243453.2014.885450

Robertson, S. B. (1967). Bildungsreform als Revision des Curriculum [Educational reform through curriculum revision]. Neuwied: Leuchterhand.

Rogers, E. M. (2003). Diffusion of innovations (5th ed.). New York, NY: Free Press.

Rupp, A. A., Vock, M., Harsch, C., & Köller, O. (2008). Developing standards-based assessment tasks for English as a first foreign language: Context, processes, and outcomes in Germany. Münster: Waxmann.

Schürenst, J. (Ed.). (2014). Effectiveness of time investments in education: Insights from a review and meta-analysis. Cham: Springer.

Schroeders, U., Penk, C., Jansen, M., & Pant, H. A. (2013). Geschlechtsbezogene Disparitäten [Gender-related disparities]. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Eds.), IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I (pp. 249–274). Münster: Waxmann.

Schurtz, I. M., Pfoert, M., Nagengast, B., & Artelt, C. (2014). Impact of social and dimensional comparisons on student’s mathematical and English subject-interest at the beginning of secondary school. Learning and Instruction, 34, 32–41. https://doi.org/10.1016/j.learninstruc.2014.08.001

Seaton, M., Marsh, H. W., & Craven, R. G. (2009). Earning its place as a pan-human theory: Universality of the big-fish-little-pond effect across 41 culturally and economically diverse countries. Journal of Educational Psychology, 101(2), 403–419. https://doi.org/10.1037/a0013838

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Sørensen, A. B. (1970). Organizational differentiation of students and educational opportunity. Sociology of Education, 43(4), 355–376.

Stanat, P., Böhme, K., Schipolowski, S., & Haag, N. (Eds.). (2016). IQB-Bildungstrend 2015: Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich [National Assessment Study 2015: Language competencies at the end of Grade 9 in the second national assessment study]. Münster: Waxmann.

Stein, M. K., Kaufman, J. H., Sherman, M., & Hillen, A. F. (2011). Algebra: A challenge at the crossroads of policy and practice. Review of Educational Research, 81(4), 453–492. https://doi.org/10.3102/0034654311423025

The National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington, DC: U.S. Government Printing Office.
Thüringer Kultusministerium (Ed.). (1999a). Lehrplan für das Gymnasium: Mathematik [Curriculum for the Gymnasium: Mathematics]. Erfurt: Author. Retrieved from https://www.schulportal-thueringen.de/web/guest/media/detail?tspi=5491

Thüringer Kultusministerium. (1999b). Thüringer Lehrpläne für das Gymnasium [Curriculum for the Gymnasium in Thuringia]. Erfurt: Author. Retrieved from https://www.schulportal-thueringen.de/web/guest/media/detail?tspi=5491

Thüringer Kultusministerium. (2007). Gymnasiale Oberstufe [Upper secondary school at the Gymnasium]. Weimar: Author.

Thüringer Kultusministerium. (2008). Informationen zur neuen gymnasialen Oberstufe [Information on the new upper secondary school at the Gymnasium]. Weimar: Author.

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(4), 788–806. https://doi.org/10.1037/0022-0663.98.4.788

Trautwein, U., Lüdtke, O., Marsh, H. W., & Nagy, G. (2009). Within-school social comparison: How students perceive the standing of their class predicts academic self-concept. Journal of Educational Psychology, 101(4), 853–866. https://doi.org/10.1037/a0016306

Trautwein, U., & Neumann, M. (2008). Das Gymnasium [The Gymnasium]. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer, & L. Trommer (Eds.), Das Bildungswesen in der Bundesrepublik Deutschland: Strukturen und Entwicklungen im Überblick (pp. 467–501). Reinbek bei Hamburg: Rowohlt.

Trautwein, U., Neumann, M., Nagy, G., Lüdtke, O., & Maaz, K. (Eds.). (2010). Schulleistungen von Abiturienten: Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand [Achievement of upper secondary school graduates: The newly ordered upper secondary school on the trial]. Wiesbaden: VS Verlag für Sozialwissenschaften.

Volante, L. (Ed.). (2016). The intersection of international achievement testing and educational policy: Global perspectives on large-scale reform. New York, NY: Routledge.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. Psychological Bulletin, 140(4), 1174–1204.

Wagner, W., Kramer, J., Trautwein, U., Lüdtke, O., Nagy, G., Jonkmann, K., … Schilling, J. (2011). 15 Upper secondary education in academic school tracks and the transition from school to postsecondary education and the job market. Zeitschrift Für Erziehungswissenschaft, 14(Suppl 2), 233–249. https://doi.org/10.1007/s11618-011-0196-1

Wagner, W., Rose, N., Dicke, A.-L., Neumann, M., & Trautwein, U. (2014). Alle alles lehren – Schulleistungen in Englisch, Mathematik und den Naturwissenschaften vor und nach der Neuordnung der gymnasialen Oberstufe in Sachsen [Teaching everyone everything – Student achievement in English, mathematics and the natural sciences before and after Saxony’s upper secondary school reform]. Zeitschrift Für Erziehungswissenschaft, 17(2), 345–369. https://doi.org/10.1007/s11618-014-0492-7

Watt, H. M. G., & Eccles, J. S. (Eds.). (2008). Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences. Washington, DC: American Psychological Association.

Watt, H. M. G., Eccles, J. S., & Durik, A. M. (2006). The leaky mathematics pipeline for girls: A motivational analysis of high school enrolments in Australia and the USA. Equal Opportunities International, 25(8), 642–659. https://doi.org/10.1108/02610150610719119