Educational Effectiveness at the End of Upper Secondary School: Further Insights Into the Effects of Statewide Policy Reforms

Dissertation

zur Erlangung des Doktorgrades

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Nicolas Hübner, M.Sc.

aus Münster

Tübingen

2017

Tag der mündlichen Prüfung:	10.10.17
Dekan:	Prof. Dr. rer. soc. Josef Schmid
1. Gutachter:	Prof. Dr. Benjamin Nagengast
2. Gutachter:	Prof. Dr. Kathleen Stürmer

ACKNOWLEDGMENTS

First, I would like to thank my supervisors Prof. Dr. Benjamin Nagengast and Prof. Dr. Ulrich Trautwein for their guidance and perspective in helping me develop this dissertation project. Next, I would like to express my very great appreciation to Wolfgang Wagner for support and discussions of methodological and theoretical issues regarding this dissertation project. As a critical friend, he always reminded me of the importance to keep in mind the principles and foundations of research in general. Similarly, many thanks go to Norman Rose who provided valuable feedback regarding methodological issues and many other important aspects of research. Moreover, I would like to thank all of my colleagues at the Hector Research Institute for Education Sciences and Psychology as well as the Special Interest Group Educational Effectiveness for their excellent feedback regarding my work and numerous presentations of their own research, which kept me informed about most relevant topics and developments related to educational science. Along these same lines, many thanks go to all members of the Lead Graduate School & Research Network, who have provided valuable feedback during all stages of my research. Finally, my deepest thanks go to my family and friends for their absolute support in the past years: Renate & Horst, Vanessa, Felix, Steffen, Hendrik, Steffen, and Sven!

ABSTRACT

For several decades, educational policy reforms have been understood as major instruments of educational governance that can impact existing educational practices, for instance, in terms of changes in teaching strategies, learning materials, and students' achievements (Fullan, 1983). However, in contrast to their huge sociopolitical relevance, scientific evaluations of such reforms are scarce (e.g., OECD, 2015).

Rigorous evaluations and deeper investigations of reforms are of special societal importance for several reasons: (a) to legitimize sometimes very controversial legislative policy decisions, which are to be implemented by the educational administration, (b) to test and minimize aspects of educational policies, which are predominantly based on normative arguments and which are now implemented as trial and error policies, and (c) to increase knowledge about when educational policy reforms and curricular programs lead to intended or unintended effects for students (e.g., Black & Wiliam, 2009; McConnell, 2010; Schaffer, Nesselrodt, & Stringfield, 1997). Moreover, formative and summative evaluations of educational policy reforms against objective standards (e.g., Campbell, 1969; Konstantopoulos & Hedges, 2008) are important for decreasing the likelihood of unintended side effects right from the start of their implementation. A systematic, empirically grounded evaluation of educational policy reforms is also of special importance in the face of the high complexity of the multilayered education system, where reforms are usually focused on impacting surface structures (e.g., Elmore, 1995) but shall actually trigger students' individual educational processes, for instance, competence growth. For these reasons, the effects of policy reforms are generally very complicated to anticipate during the construction and implementation of the reforms (Fusarelli, 2002; Young & Lewis, 2015).

In the present dissertation, I investigate a variety of central psychological factors at the student level before and after the implementation of two central educational policy reforms at the end of upper secondary school. In this project, I do not merely analyze the reforms in a loose, isolated framework, but I integrate and critically reflect on them more closely in a disciplinary context. In fact, in this dissertation, I make an attempt to integrate the four studies into a larger, more general context of educational reform, which can be addressed only in an interdisciplinary way. Therefore, this dissertation also focuses on developments in educational policy and educational science in general, which define the central foundations for introducing policy reforms in the education system. Moreover, I also focus on developments related to educational governance and discussions about the increase in demands for evidence-based

policy (see Chapters 2 and 3) before outlining the need to include psychological factors and related theoretical models in reform evaluations (see Chapter 3).

The two reforms that are analyzed here are the reform of upper secondary school and the G8 reform, both of which were implemented at the beginning of the new millennium in most German states. The two reforms are still critically discussed in the society and by educational policy. In doing this, I use prominent theoretical models, for instance, a model of achievement motivation (e.g., Eccles & Wigfield, 2002) in order to generate appropriate hypotheses and integrate the results of the effects of the reforms into recent research.

The reform of upper secondary school mainly implemented mandatory course choice on an advanced course level in upper secondary school and therefore depicts a curricular intensification (CI) reform. The G8 reform reduced overall school time in high track schools (Gymnasium) from a total of 9 to 8 years by means of a compression of school time in terms of an increase in allocated time per week in lower secondary schools.

I analyzed the reform of upper secondary school using a large representative sample from Thuringia (Additional Study Thuringia of the National Educational Panel Study) and Baden-Württemberg (TOSCA study; Blossfeld, Rossbach, & Maurice, 2011; Köller, Watermann, Trautwein, & Lüdtke, 2004; Trautwein, Neumann, Nagy, Lüdtke, & Maaz, 2010). Furthermore, I conducted the analysis of the G8 reform by considering a large representative data set from Baden-Württemberg (Additional Study Baden-Württemberg of the National Educational Panel Study; Blossfeld et al., 2011).

In Study 1 (*Maximizing Gender Equality by Minimizing Course Choice Options? Effects of Obligatory Coursework in Math on Gender Differences in STEM; Journal of Educational Psychology*), differential effects of the upper secondary school reform on advanced math, math self-concept, and vocational interests were investigated. Furthermore, potential differences before and after the upper secondary school reform regarding the field of study at university in STEM (science, technique, engineering, and mathematics) subjects were focused on. Results showed that whereas gender differences in math achievement were lower after the reform, differences were larger on all other outcome variables. In spite of these results, no differences before or after the reform were found for the choice of the field of study at university.

Study 2 (*Putting All Students in One Basket Does not Produce Equality: Gender-Specific Effects of Curricular Intensification in Upper Secondary School;* Manuscript submitted for publication) expanded the results of Study 1 by considering data from another German state, namely, Thuringia. In Study 2, it was possible to analyze a broader variety of student outcome measures in English reading, mathematics, biology, and physics, as well as students' subject-specific self-concepts and interests in these subjects. The results of this study indicated no statistically significant average differences on achievement measures. However, differential effects on English reading and a higher English self-concept in favor of young men were found after the reform, whereas the reform had a negative effect on young women's math self-concept.

In Study 3 (*Comparing Apples and Oranges: Reforms can Change the Meaning of Students' Grades!;* Manuscript submitted for publication), analyses of reform effects were extended to school grades. Students' grades at the end of upper secondary school are of special importance for college and university access and later job employment. However, research has shown striking differences between teacher-assigned grades and standardized student achievement. Furthermore, grades are oftentimes assigned on a norm-referenced basis and are therefore strongly oriented toward a class' achievement composition, which changed when detracking was introduced by the CI reform. Therefore, Study 3 was focused on the research question of whether students' standardized achievement differed between before and after the CI reform, given similar grades. Results suggested considerable differences in students' standardized test achievement before and after the reform, given similar grades. Compared with basic courses, standardized achievement given a similar grade in core courses was higher. However, the opposite pattern was found when comparing achievement between advanced and core courses, given a similar grade. Furthermore, for math these effects were found to vary among high and low grade levels.

Finally, Study 4 (*The G8 reform in Baden-Württemberg: Competencies, Well-Being, and Leisure Time Before and After the Reform; Zeitschrift für Erziehungswissenschaft*) is one of the first studies to investigate effects of the G8 reform at the end of upper secondary school. In contrast to the reform of upper secondary school, the G8 reform increased the time allocated in lower secondary school in order to reduce the total time spent in school by 1 year. Study 4 therefore focused on potential changes in student achievement in mathematics, English reading, biology, and physics from before to after the reform. In addition, potential effects on variables related to students' well-being (stress and health) and leisure time use were analyzed. Differences between G8 and G9 students were found in English reading, biology, and in well-being measures in favor of the G9 students.

All studies in this dissertation investigated the research questions using advanced statistical methods such as multidimensional multiple-group IRT models or structural equation models with continuous indicators and considered survey weights, missing data, and the clustered structure of the data. The reforms that the dissertation focused on were chosen specifically in order to investigate central individual aspects but also have an exemplary, more general function in the context of investigating changes in specific surface structures of the education system on specific psychological factors related to achievement. Similarly, all reforms were implemented in the highest track school, the Gymnasium, which is currently the most frequently attended school type in lower and upper secondary school. The Gymnasium is important because the results of the upper secondary school examination strongly determine whether a student is eligible to enroll in university. In the beginning of this dissertation, I will first provide a general introduction regarding the meaning and expectations of educational policy reforms. I will subsequently integrate this material into the central findings and developments of educational effectiveness research and educational governance in Chapters 2 and 3. After presenting Studies 1 to 4 in Chapter 4, I will outline the strengths and limitations and implications of the dissertation in Chapter 5.

ZUSAMMENFASSUNG

Bildungsreformen werden seit einigen Jahrzehnten als zentraler Bestandteil der politischen Steuerung des Bildungswesens verstanden, die Einfluss auf die schulische Bildungspraxis nehmen können und beispielsweise Veränderungen der bestehenden Unterrichtstrategien, Lernmaterialien und Schülerleistungen intendieren (Fullan, 1983). Trotz ihrer hohen gesellschaftlichen Relevanz sind diese Reformen nur selten Gegenstand systematischer Untersuchungen (OECD, 2015).

Profunde Evaluationen und vertiefende Analysen von Reformen sind aus verschiedenen Gründen von besonderer gesellschaftlicher Bedeutung: (a) zur Legitimierung der teilweise sehr umstrittenen, von der Legislative getroffenen und den Instanzen der Bildungsverwaltung umzusetzenden politischen Entscheidungen, (b) zur Prüfung und Minimierung derjenigen bildungspolitischen Programmanteile und Folgen, die überwiegend unter normativen Aspekten festgelegt wurden und anschließend zur Erprobung bestimmter Reformmaßnahmen implementiert werden sollen und schließlich (c) zur Erweiterung des allgemeinen Wissens darüber, wann Bildungsreformen und curriculare Programme für Schülerinnen und Schüler eine erwünschte oder eine unerwünschte Wirkung erzielen (Black & Wiliam, 2009; McConnell, 2010; Schaffer et al., 1997).

Darüber hinaus sind begleitende wie summative Evaluationen bildungspolitischer Reformen mittels objektiver Standards bedeutsam (z.B., Campbell, 1969; Konstantopoulos & Hedges, 2008), um noch während des Umsetzungsprozesses im Sinne einer formativen Evaluation, Möglichkeiten zu identifizieren und die Auftretenswahrscheinlichkeit nicht intendierter Nebenwirkungen zu verringern. Eine systematische, empirisch fundierte Begleitforschung von Bildungsreformen ist in besonderer Weise relevant, da in Anbetracht der Komplexität und mehrdimensionalen Struktur des Bildungswesens Effekte von Bildungsreformen einerseits Oberflächenstrukturen betreffen (z.B., Elmore, 1995), jedoch andererseits bei den Schülerinnen und Schülern jeweils auch individuelle Bildungsprozesse auslösen und z.B. Kompetenzzuwächse bewirken sollen, die ohne Analysen und wissenschaftliches Wissen nur schwer im Entwicklungsprozess der Reform zu antizipieren und im Umsetzungsprozess zu erkennen sind (Conley, 1994; Fusarelli, 2002; Young & Lewis, 2015).

Die vorliegende Dissertation untersucht verschiedene zentrale Schülervariablen vor und nach zwei zentralen bildungspolitischen Reformprogrammen am Ende der Sekundarstufe II. Die Reformen sollen in vier Beiträgen nicht nur hinsichtlich ihrer jeweiligen Spezifität und

inhaltlichen und methodischen Qualität dargestellt, bzw. in einem engeren disziplinären Kontext kritisch reflektiert und verortet werden. Vielmehr wird darüber hinaus der anspruchsvolle Versuch unternommen, die zugrundeliegenden vier Beiträge in einen größeren und im Grundsatz nur interdisziplinär zu bearbeitenden Kontext einzuordnen. Daher fokussiert die Dissertation ebenfalls zentrale bildungspolitische und wissenschaftliche Entwicklungstendenzen, die den Rahmen von reformpolitischem Handeln im Bildungssektor definieren. Hierzu zählen beispielsweise Entwicklungen im Bereich der Bildungssteuerung (vgl. Kapitel 2 und 3) und Diskussionen zu einem zunehmend von wissenschaftlicher Seite geforderten evidenzbasierten bildungspolitischen Handeln (vgl. Kapitel 3).

Bei den beiden analysierten Reformen, die im Fokus der Fachbeiträge stehen, handelt es sich einerseits um die große Reform der gymnasialen Oberstufe und andererseits um die G8-Reform. Diese beiden Reformen, die Anfang der 2000er Jahre in der überwiegenden Mehrheit der Länder der Bundesrepublik Deutschland eingeführt wurden, sind auch aktuell gesellschafts- und bildungspolitisch nicht unumstritten.

Die Reform der gymnasialen Oberstufe implementierte eine curriculare Intensivierung (engl.: curricular intensification), indem sie eine Veränderung der Wahlmöglichkeiten in der Sekundarstufe II im Sinne von verpflichtenden Vorgaben zur Kurswahl auf erhöhtem Anforderungsniveau zugrunde legte. Die G8-Reform führte zu einer Reduktion der regulären Schulzeit an Gymnasien von neun auf acht Schuljahre durch eine Schulzeitkompression, im Sinne einer Verlängerung der wöchentlichen Unterrichtszeit in der Sekundarstufe I.

Die Oberstufenreform wurde im Rahmen meiner Fachbeiträge auf der Grundlage großer repräsentativer Datensätze aus Thüringen (Zusatzstudie Thüringen des Nationalen Bildungspanels; Blossfeld et al., 2011) und Baden-Württemberg (TOSCA Studie; Köller et al., 2004; Trautwein et al., 2010) untersucht. Die Analyse der G8-Reform erfolgte unter Verwendung eines repräsentativen Datensatzes aus Baden-Württemberg (Zusatzstudie Baden-Württemberg des Nationalen Bildungspanels; Blossfeld et al., 2011).

In Studie 1 (*Maximizing Gender Equality by Minimizing Course Choice Options? Effects of Obligatory Coursework in Math on Gender Differences in STEM;; Journal of Educational Psychology*) standen differenzielle Effekte der Oberstufenreform mit besonderem Blick auf voruniversitäre Mathematik, das mathematische Selbstkonzept und die beruflichen Interessen im Fokus der Analysen. Weiterhin wurden mögliche Unterschiede vor und nach der Oberstufenreform in Bezug auf die Studienfachwahl an der Universität in MINT-Fächern (Mathematik, Ingenieurwissenschaften, Naturwissenschaften und Technik) genauer betrachtet. Die Ergebnisse legen nahe, dass Geschlechterunterschiede in der voruniversitären

Mathematikleistung nach der Reform kleiner waren, während sich die Unterschiede auf den übrigen Merkmalsdimensionen vergrößerten. Trotz dieser Befunde zeigten sich vor und nach der Reform keine Unterschiede hinsichtlich des Wahlverhaltens der Fächer beim späteren Studium.

In Studie 2 (*Putting All Students in One Basket Does not Produce Equality: Gender-Specific Effects of Curricular Intensification in Upper Secondary School;* Manuskript zur Publikation eingereicht) wurden die Ergebnisse der ersten Studie unter Rückbezug auf Daten zur Oberstufenreform in Thüringen erweitert. Darüber hinaus ermöglichte die zweite Studie eine deutliche Erhöhung der Anzahl der untersuchten Variablen. So konnten hier die standardisierten Leistungen in Englisch-Lesen, Mathematik, Biologie und Physik sowie die fachspezifischen Selbstkonzepte und Interessen der Schülerinnen und Schüler in diesen Fächern näher untersucht werden. In der Studie fanden sich zwar keine statistisch signifikanten Unterschiede in den Leistungen, dennoch zeigten sich differenzielle Effekte in Englisch-Lesen und ein höheres Selbstkonzept in Englisch zu Gunsten der männlichen Schüler, während das mathematische Selbstkonzept bei Schülerinnen nach der Reform statistisch signifikant niedriger war als zuvor.

In Studie 3 (*Comparing Apples and Oranges: Curricular Intensification Reforms can Change the Meaning of Students' Grades!;* Manuskript zur Publikation eingereicht) wurden die Analysen zu Reformeffekten schließlich um eine nähere Betrachtung der Schulnoten erweitert. Die Noten von Schülerinnen und Schülern am Ende der Sekundarstufe II sind von besonderer Bedeutung für die Zulassung zu einer Universität und den späteren Beruf. Allerdings zeigen verschiedene Studien markante Differenzen zwischen der Notenvergabe von Lehrerinnen und Lehrern und den Schülerleistungen auf Basis standardisierter Tests, was häufig auch auf die soziale Bezugsnormorientierung bei der Notenvergabe zurückgeführt wird. Aus diesem Grund basiert die dritte Studie auf der erkenntnisleitenden Fragestellung, ob sich die mittleren standardisierten Leistungen von Schülerinnen und Schülern in Mathematik und Englisch bei vergleichbaren Noten vor und nach der Oberstufenreform, die eine Veränderung in der leistungsbezogenen Schülerkomposition einführte, unterscheiden. Die Ergebnisse legen nahe, dass die Schülerleistung vor und nach der Reform auch bei gleichen Schulnoten teilweise sehr deutlich differiert, insbesondere im Unterrichtsfach Mathematik. Im Vergleich zum Grundkurs vor der Reform war die auf der Basis eines standardisierten Tests gemessene Leistung im Kernfach nach der Reform, bei einer vergleichbaren Note, tendenziell höher. Im Vergleich zum Leistungskurs vor der Reform fiel dagegen die Leistung im Kernfach nach der Reform, bei einer vergleichbaren Benotung, geringer aus. Darüber hinaus zeigte sich, dass diese Effekte in Abhängigkeit der Notenstufe variierten.

Studie 4 (*Die G8-Reform in Baden-Württemberg: Leistungen, Wohlbefinden und Freizeitverhalten vor und nach der Reform; Zeitschrift für Erziehungswissenschaft*) erweitert schließlich die Befunde zur Einführung von Effekten der Oberstufenreform am Ende der Sekundarstufe II um eine Untersuchung möglicher Effekt der G8-Reform am Ende der Sekundarstufe II. Im Gegensatz zur Oberstufenreform lag der Fokus der G8-Reform auf einer Erhöhung der nominalen Lernzeit in der Sekundarstufe I, um damit die Gesamtschulzeit um ein Schuljahr zu verringern. Die vierte Studie fokussiert daher auf potenzielle Veränderungen der Schülerleistung in Mathematik, Englisch-Lesen, Biologie und Physik vor und nach der Reform. Zusätzlich wurden mögliche Effekte auf Variablen untersucht, die mit dem Wohlbefinden der Schülerinnen und Schüler (Beanspruchung und Gesundheit) und ihren Freizeitaktivitäten zusammenhängen. Die Ergebnisse der Studie deuten auf Unterschiede zwischen G8- und G9-Schülerinnen und Schülern in Englisch-Lesen, Biologie und dem Wohlbefinden zu Gunsten von G9-Schülerinnen und Schülern hin.

Alle Studien untersuchen die jeweiligen forschungsleitenden Fragestellungen mittels anspruchsvoller statistischer Verfahren, wie mehrdimensionalen Mehrgruppen-IRT Modellen oder Strukturgleichungsmodellen mit kontinuierlichen Indikatoren und unter Berücksichtigung von Surveygewichten, fehlenden Werten sowie der hierarchischen Datenstruktur. Die berücksichtigten Reformen wurden gezielt ausgesucht, um wesentliche Kernaspekte von Reformen näher zu untersuchen, erfüllten aber gleichzeitig auch eine exemplarische Funktion, Effekte von Veränderungen bestimmter Oberflächenstrukturen des Bildungswesens auf spezifische Schüleroutcomes näher zu untersuchen. Alle untersuchten Reformen fokussieren das Gymnasium und damit die aktuell am stärksten besuchte Schulform in der Sekundarstufe I. Die besondere Relevanz der Gymnasien in Deutschland resultierte traditionell aus der mit dem bestandenen Abitur verbundenen Vergabe des Zugangs zu den Universitäten.

Zu Beginn der Dissertation wird eine erste Einführung zur Bedeutung von und Erwartungen an Bildungsreformen geboten, bevor anschließend in Kapitel 2 und Kapitel 3 eine Einordnung in die zentralen Erkenntnisse und Entwicklungslinien der Effektivitätsforschung und Bildungssteuerung erfolgt. Nachdem in Kapitel 4 die Studien vorgestellt werden, werden die Ergebnisse, Limitation und Implikationen abschließend in Kapitel 5 diskutiert.

CONTENTS

List of Figures

1 Introduction

In 2013, OECD countries invested, on average, 3.7% of their gross domestic product (GDP) into primary to postsecondary education. This percentage varied from 2.5% in Hungary (5,486 million US dollars) to 4.8% in the United Kingdom (112,856 million US dollars). In Germany, investments amounted to 3.1% of GDP or approximately 104,194 million US dollars (OECD, 2016a). Besides other arguments, it is possible to identify at least three strands that can contribute to explaining such huge investments in education.

First, from a perspective of education philosophy and anthropology, education fulfills a central part of societal renewal through a transmission of knowledge. The philosopher and educator John Dewey had outlined this perspective in the beginning of the 20th century:

> With the growth of civilization, the gap between the original capacities of the immature and the standards and customs of the elders increases. Mere physical growing up, mere mastery of the bare necessities of subsistence will not suffice to reproduce the life of the group. Deliberate effort and the taking of thoughtful pains are required. Beings who are born not only unaware of, but quite indifferent to, the aims and habits of the social group have to be rendered cognizant of them and actively interested. Education, and education alone, spans the gap. (Dewey, 1916, p. 3)

As stated by Dewey, education satisfies the specific need for societal renewal, as children are not born with the specific subset of behaviors that are needed to fit into society. Furthermore, the discrepancy between a child’s abilities and the social objective of abilities increases continuously due to the growth of civilization. However, children are born with important precursor abilities and can be shaped to meet these social objectives.

Second, from a legal, ethical perspective, since 1948, global intergovernmental organizations such as the UN proclaimed that education is a human right in the Universal Declaration of Human Rights: “Everyone has the right to education. Education shall be free, at least in the elementary and fundamental stages” (United Nations General Assembly, 1948, para. 26). However, as outlined in the report of the United Nations regarding the Millennium Development Goals (MDG), this goal seems to be far from being reached. In 2015, approximately 57 million children were still not offered primary education, and in developing regions, there was a considerably smaller chance (25%) for children in poor households to participate in primary education. However, great improvements are also visible, as the rate of illiterates in between the ages of 15 to 25 years has decreased by 8%, and the number of children who are not in school has greatly decreased by about 43 million since 2000 (United Nations, 2015).

Third, from an economic perspective, research has underscored the importance of education for a variety of outcomes later in life on an individual and an aggregated, national level. Examples of such variables, which are often mentioned in the economic literature, are human capital, labor market returns, and economic growth (e.g., Hanushek & Woessmann, 2010). From this perspective, it seems reasonable for societies to identify and promote variables that have a positive effect on student learning and achievement. As outlined by Hanushek and Woessmann (2010), school quality in particular, measured by averaging mathematics and science achievement data observed in international assessments, seems to have a considerable impact on economic growth. On the basis of this finding, the authors argued that educational reforms that are able to increase student achievement (e.g., by about 0.5 SDs over 20 years) would in turn exponentially increase GDP. Although this example seems to be very theoretical as it considers neither the complex nature of public policy making (Sabatier, 2007) nor the challenges of successfully implementing education reforms in the education system (Porter, Fusarelli, & Fusarelli, 2015), it provides an interesting starting point for further consideration of the relevance of reforms in the field of education.

As is evident from above, education has a fundamental role in societal life, which can be, among others, defined with different emphases from a philosophical, anthropological, ethical, legal, or economic perspective. However, there are theoretical approaches that implicitly link these seemingly different strands.

From a perspective of German school theory (e.g., Fend, 2009), formal education fulfills four specific objectives: (a) cultural reproduction, (b) qualification, (c) allocation, and (d) integration and legitimation: peace-keeping.[1] Cultural reproduction and qualification are strongly related to the economic theories of economic growth as well as to an ethical and philosophical perspective of qualifying individuals and societal renewal. Allocation in turn focuses instead on the objective of sorting individuals into specific positions and occupations in a society by means of certificates, which are used as indicators of individuals’ abilities. The function of integration and legitimation finally addresses the transmission of values and norms, for instance, to consolidate political structures (Fend, 2009).

Especially in the last couple of decades, specific efforts have been made to raise the standards for education, for instance, in terms of educational attainment or achievement levels (e.g., The National Commission on Excellence, 1983). Policy reforms such as the No Child Left Behind (NCLB) Act, introduced under George W. Bush in 2001, or the Every Student

[1] Translated by the author.

Succeeds Act (ESSA) signed into law by Barack Obama in 2015 in the United States, can be seen as extensions of these general movements toward a stronger focus on high student competencies.

Knowledge about how to raise student education standards seems to be somewhat comparable to the search for the “Holy Grail” (e.g., Terhart, 2011). Education science and related disciplines have played a prominent role in recent decades in searching for this grail (e.g., Reynolds et al., 2014), and educational policy reforms are frequently proposed to be able to alter the education system in this regard (e.g., Hanushek & Woessmann, 2010; OECD, 2015). Lately, attempts have been made to exchange such knowledge between education science and education policy and practice, for instance, from initiatives such as the What Works Clearinghouse (e.g., Slavin, 2008). However, research and practice still seem to have a strong coexistence in many regards, and the transfer of research evidence into policy and practice is far from standard (e.g., Bromme, Prenzel, & Jäger, 2014; Cooper, Levin, & Campbell, 2009; Davies, 2000; Qi & Levin, 2013; Slavin, 2002; Slavin, 2008). In line with this, few educational policy reforms are accompanied by rigorous scientific evaluations or follow output-based funding strategies (OECD, 2015; Slavin, 2002).[2] However, as I will further outline in this dissertation, it is essential for educational interventions to be evaluated against objective standards in order to identify potential opportunities to further improve interventions or eliminate unintended side effects (e.g., Black & Wiliam, 2006; McConnell, 2010). Not evaluating educational policy reforms might be neither effective nor accountable, and this becomes especially visible when considering cases where either policy interventions have a negative impact or the status quo has an unknown negative impact on students (e.g., Torgerson & Torgerson, 2001).[3] From this perspective, rigorous evaluations of variables such as student achievement and factors related to achievement such as motivation, for instance, in terms of expectancies and value beliefs (e.g., Eccles, 1983; Eccles & Wigfield, 2002; Marsh et al., 2008; Wigfield & Eccles, 2002) should not be optional but mandatory in order to counter opinions and normative judgments of “what works” with profound knowledge (see Chapter 3).

[2] For German exceptions to this, see, for instance, evaluations of all-day schools (Ganztagsschulen) policy reforms (e.g., Fischer, Kuhn, & Tillack, 2016; Decristan & Klieme, 2016; Lossen, Tillmann, Holtappels, Rollett, & Hannemann, 2016). Trautwein, Neumann, Nagy, Lüdtke, and Maaz (2010) and Wagner, Rose, Dicke, Neumann, and Trautwein (2014) have already published extensive evaluations of the reform of upper secondary school (Oberstufenreform) with a focus on main effects. Recently, Neumann, Becker, Baumert, Maaz, and Köller (2017) published an extensive evaluation of the structural reform in Berlin. For reforms that are part of extensive evaluations in the United States, see, for instance, Borman, Hewes, Overman, and Brown (2003) for a meta-analysis on effects of comprehensive school reform.

[3] The arguments outlined by Torgerson and Torgerson (2001) did not focus explicitly on reforms but on randomized controlled trials (RCTs). However, they can be perfectly integrated into the debate on the need for rigorous educational investigations and evaluations per se.

To adequately address the aspects outlined above, in the face of the huge complexity of the education system, this dissertation is organized into four major sections:

First, I provide the theoretical foundations in order to enable the reader to embed the findings of the studies into a more general framework of the education reform movement and the German education system. To do this, I outline the *Theoretical Foundations of Educational Governance* (Chapter 2), including subchapters on the *German Education System and Current Monitoring Strategies* (Chapter 2.1), *Formal Education in Germany* (Chapter 2.2), and a chapter on *Educational Governance and Educational Change* (Chapter 2.3). As evident in Chapter 2, I outline foundations of the German education systems as these are important for a deeper understanding of the general framing conditions of the system in which the policy reforms analyzed in this dissertation are implemented.

Next, in Chapter 3, I provide deeper insights on *Educational Effectiveness and Educational Policy* by presenting a chapter on *The Intersection of Educational Effectiveness Research, Large-Scale Assessments, and Educational Policy Reforms* (Chapter 3.1), which offers an international perspective on the emergence of educational policy reforms and demonstrates relations to standards-based reforms and large-scale assessments. Next, in Chapter 3.2, which is called *Evidence of Effectiveness Research and Relations to Educational Policy*, I extend this first perspective by providing information on the more general discussion regarding research evidence and evidence-based policy making, which is centrally relevant in the context of educational policy reforms and their evaluations. Furthermore, in this chapter, I offer insights into relations between educational effectiveness research (EER) and the process of public policy making. In *A Taxonomy of Educational Policy Reforms* (Chapter 3.3), I describe several models and identify specific dimensions along which policy reforms can be distinguished and categorized more closely. In this chapter, I therefore offer a more general framework in which past, recent, and future reforms can be integrated. Finally, in Chapter 3.4 on *The Interplay between Educational Policy Reforms and Student Outcomes*, I link educational policy reforms to specific student outcomes. To do this, I use prominent effectiveness models and other related models to theoretically identify potential channels of policy reforms. This chapter underscores the importance of taking a closer look at effects on psychological factors whenever reforms are implemented.

Chapter 3 ends with the foundation of the dissertation project in terms of the *Research Questions*. Subsequently, I present four studies in Chapter 4 that all investigate different educational reforms at the end of upper secondary school with a special focus on psychological factors: In Study 1, the reform of upper secondary school in the state of Baden-Württemberg is analyzed for its effects on math achievement, vocational interests, self-concept in math, and subject choice at university (Hübner, Wille et al., 2017). The second study takes a closer look at the reform of upper secondary school in another state (Thuringia) and thereby provides an investigation of differences between students before and after the reform regarding further achievement measures as well as subject-specific interests and self-concepts in mathematics, English, biology, and physics (Hübner, Wagner, Nagengast, & Trautwein, 2017). Third, a special focus is placed on changes in grades related to standardized student achievement to obtain a more holistic perspective on potential effects of the upper secondary school reform in Baden-Württemberg and Thuringia on teacher-assigned grades (Hübner, Wagner, Hochweber, Neumann, & Nagengast, 2017). The last study analyzes effects of the G8 reform at the end of upper secondary school. The reform went along with a compression of overall school time from 9 to 8 years in the highest track schools (Gymnasium). In this study, in addition to standardized student achievement, constructs such as students' subjective health and stress as well as leisure time use are focused on (Hübner, Wagner, Kramer, Nagengast, & Trautwein, 2017).

In Chapter 5, I summarize the findings from Studies 1 to 4 and outline the *Strengths and Limitations of the Present Dissertation* before outlining *Implications for Future Research on Educational Policy Reforms*, and *Implications for Policy and Practice*. Central to this chapter is the recapitulation of the importance of rigorous evaluations, especially the consideration of psychological factors right from the beginning of the process of constructing policies in order to test the effectiveness of reforms and obtain information on aspects that can be improved.

2 Theoretical Foundations of Educational Governance

2.1 The German Education System and Current Monitoring Strategies

Traditionally, the legally binding authority of formal education in schools in Germany has resided with the 16 different states (Länder). This right, also referred to as cultural sovereignty, has been guaranteed by the constitutional law of the German Federal Republic since 1949. Depending on the size of the state, in most states, educational governance can be differentiated into different layers of government (see Figure 1). The foundation of education at the state level is built upon the Act of Education in each respective federal state. Within the constraints of the laws of each state, each state has the right to make its own decisions about educational matters such as the school curriculum, teacher education, introduction of new school types, and decisions about school tracking and educational standards (e.g., Füssel & Leschinsky, 2008; van Ackeren, Klemm, & Kühn, 2015).

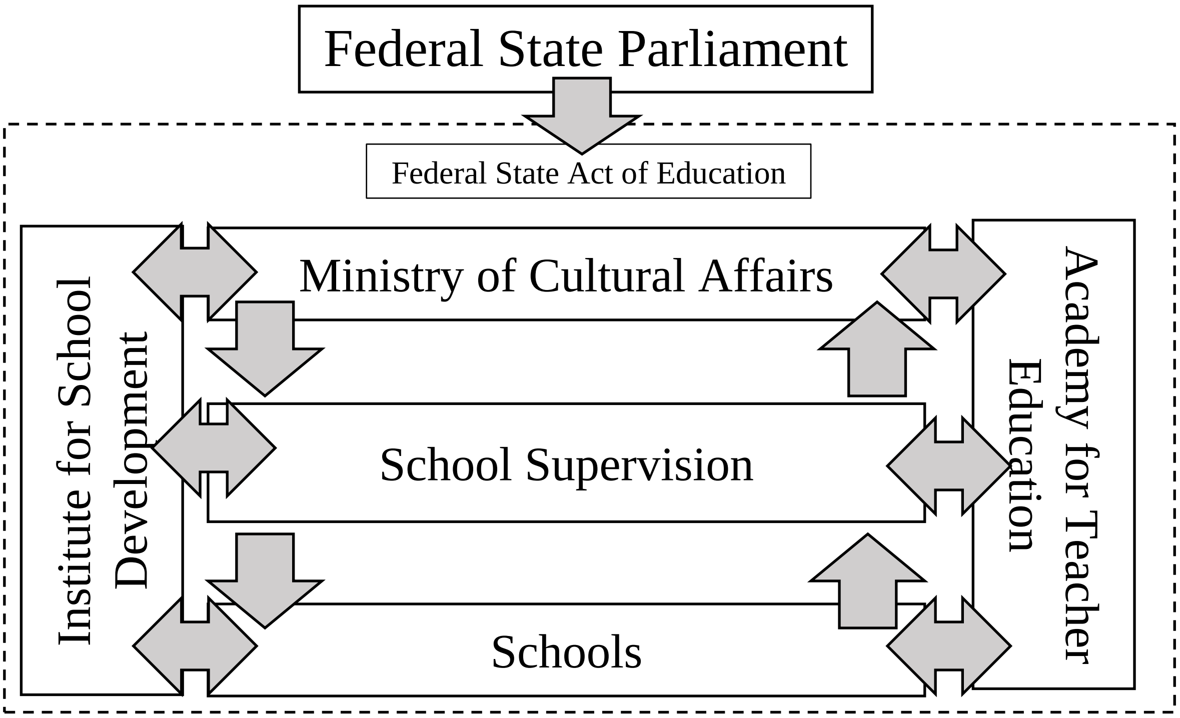
Federal State Parliament
Federal State Act of Education
Ministry of Cultural Affairs
School Supervision
Schools
Institute for School Development
Academy for Teacher Education

Figure 1. Central elements of the German educational government on the federal state level.

As there are approximately up to 6,000 schools in large German states (e.g., MSW NRW, 2016), schools are usually controlled by the school's own supervision rather than being directly controlled by the Ministry of Cultural Affairs. In larger states, school supervision is separated into upper supervision and lower supervision. This distinction is primarily oriented around different school types, which are then supervised by a different part of school supervision (e.g., van Ackeren et al., 2015), for instance, in Baden-Württemberg or North

Rhine-Westphalia. Institutes for School Development are typically strongly engaged in monitoring and developing competence standards and other issues related to school improvement and quality assurance.

Until the beginning of the new millennium, education policy in Germany was strongly oriented around inputs (e.g., regarding resource allocation and organizational guidelines). This suggests that teaching was strongly oriented toward subject-specific curricula, which provided guidance on which content areas should be taught to which kinds of students (Niemann, 2016). In 2001, the first PISA (Program for International Student Assessment) results created a “shock” in the German public and media due to the unexpected and comparably bad achievement of German students, who achieved below the OECD average in reading literacy, mathematics, and science. Because of this “shock,” a wave of structural reforms were initiated in favor of a more output-based governing strategy (Niemann, 2016). A central element of this strategy, which was related to student achievement, was the introduction of the common educational standards. Furthermore, the infrastructure for evaluating student outcomes was strongly expanded, for example, by means of rigorous monitoring strategies. Most of the enacted reforms, which are oftentimes referred to as standards-based reforms (e.g., Bellmann & Weiß, 2009; Hamilton, Stecher, & Yuan, 2009) were enacted on the state level and had their starting point at the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder (KMK). This joint conference follows specific tasks: The agenda of the Standing Conference of the Ministers of Education and Cultural Affairs is to address “educational, higher education, research and cultural policy issues of supraregional significance with the aim of forming a joint view and intention and of providing representation for common objectives” (KMK, 2017).

It is important to mention that the KMK usually passes resolutions and suggestions that are not legally binding: Only the individual states have the legal power to implement reforms in education in the states. However, it is visible that the KMK oftentimes sets the standards and foundations for initiating changes in the states for large-scale reforms (e.g., Fullan, 2000), for instance, regarding the reform of upper secondary school in Germany (Trautwein & Neumann, 2008), and the states often follow these resolutions.

As mentioned above, Germany moved from a governing strategy based on inputs to a rather output-oriented strategy. In this regard, the KMK was an important stakeholder as it adopted national standards and strategies for monitoring the educational achievement of students in the states (KMK, 2006, KMK, 2016). Educational standards can be understood as instructions on the competencies that students should possess at a specific time (e.g., at the end of lower secondary school). Furthermore, educational standards are subject-specific and describe expected achievement outcomes for students. Finally, these standards can be linked to specific competence levels in order to clarify how standards are achieved (KMK, 2005).

The core of the German monitoring strategy builds on evaluations to assess students' competencies. According to the monitoring strategy, four components are important: (a) participation in international student assessments (e.g., PISA, the Trends in International Mathematics and Science Study [TIMSS]), (b) national assessments to monitor educational standards, which are conducted by the Institute for Educational Quality Improvement (IQB; e.g., Stanat, Böhme, Schipolowski, & Haag, 2016), (c) quality assurance on the class and school levels, mainly carried out by comparative testing on the state level (VERA; e.g., Landesinstitut für Schulentwicklung, 2016), and (d) a National Educational Report, which is published every 2 years (Autorengruppe Bildungsberichterstattung, 2016). Taking a closer look at results from these four monitoring components, it is possible to get the first insights into the current status and trends of student achievement in Germany from national- and state-level perspectives.

First, regarding the participation of German students in international student assessments, the results of the last four cycles of the PISA study (OECD, 2007, OECD, 2010, OECD, 2013, OECD, 2016b) are displayed in Figure 2. As can be seen, with some exceptions in reading literacy, students have generally performed above the OECD average in all competence areas in recent years. Similar results can be found in the TIMS study (Martin, Mullis, Foy, & Olson, 2008; Mullis, Martin, Foy, & Arora, 2012; Mullis, Martin, Foy, & Hooper, 2016a, 2016b), where Germany's eighth graders have consistently performed above average in science and mathematics in studies conducted in the last decade.

In order to monitor the educational standards, the second part of the German monitoring strategy is based on national German achievement tests, which offer insights into potential state disparities in Germany. The national assessment studies are conducted in Grade 4 of elementary school and in Grade 8 in lower secondary school.

As reported in the IQB National Assessment Study 2015 (Stanat et al., 2016), there are considerable differences between German countries on most competencies. For instance, whereas students in Saxony achieved an average scale score of *M* = 528 points (*SD* = 90) in reading, amounting to 28 points above the German average (*M* = 500, *SD* = 100), students in the city state of Bremen showed an average scale score of *M* = 458 points (*SD* = 115; Böhme & Hoffmann, 2016). Results in listening and orthography were comparable in this regard. Most interesting, as the National Assessment Study follows a 3-year cycle, and similar competencies are assessed every 6 years, it is possible to identify trends in students' achievement within states and in the German average.

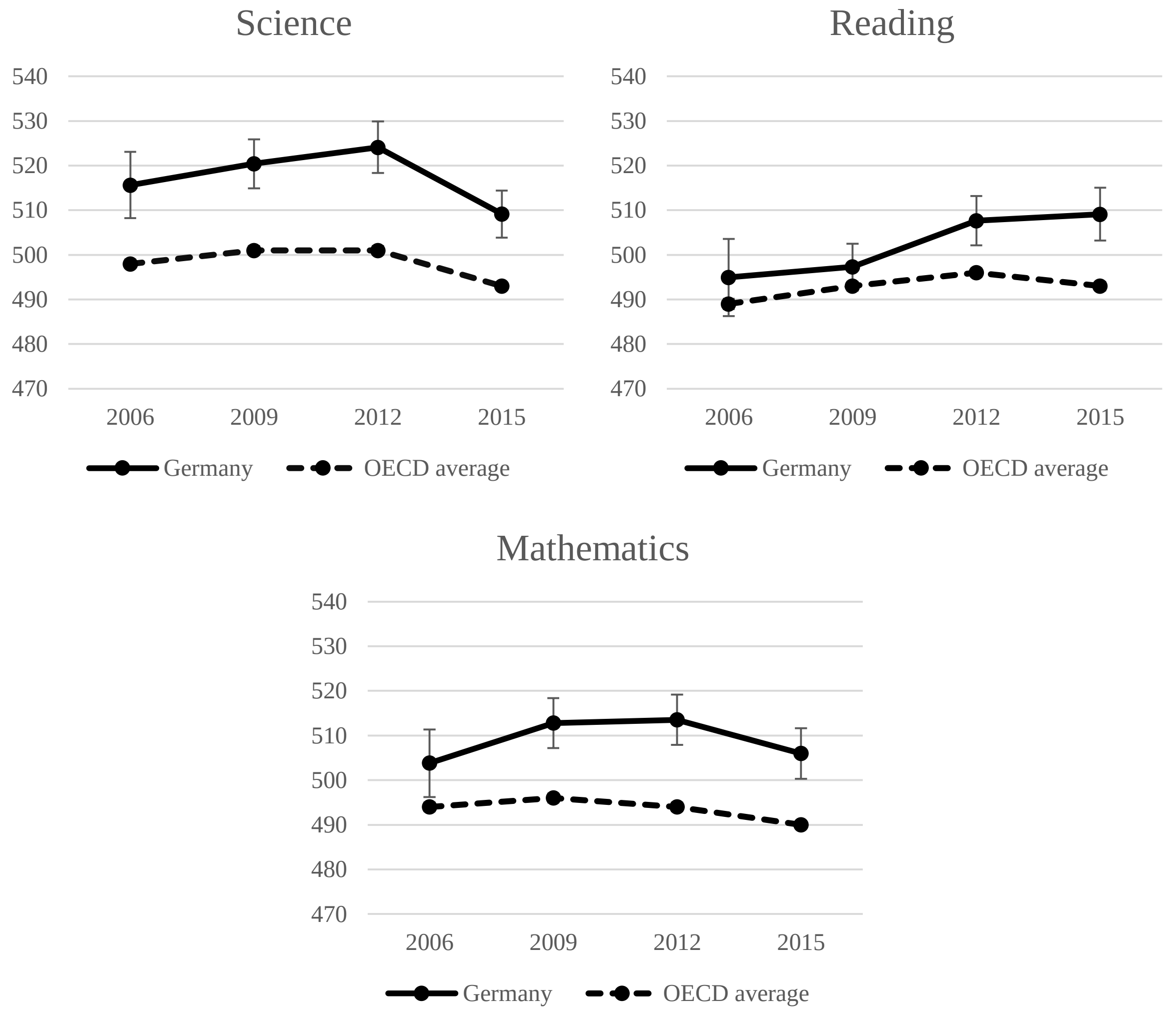

Figure 2. Achievement of German students in PISA in the last decade based on my own calculations using the PISA data, plausible values, and replicate weights. Values are identical to officially published results. OECD averages and SEs were taken from the PISA data explorer: http://pisadataexplorer.oecd.org/ide/idepisa/. The figure displays 95% confidence intervals (CIs). CIs for the OECD average are very small and fall within the grey dots. Note that recent research has suggested problems when comparing German data from 2015 with previous years due to a mode bias, which might be problematic for other countries as well (Robitzsch et al., 2017).

For reading competence, this trend shows that, on average, German students performed statistically worse in 2015 (*d* = -0.07; Cohen, 1988). Most prominent in this negative trend were students from Baden-Württemberg, who performed 23 scale scores lower in 2015, compared with 2009. Similar trends can be found for Baden-Württemberg's students' listening competence (*d* = -0.27); however, their competencies in orthography were not statistically significantly different. Baden-Württemberg is just one example of various states that showed considerable (negative) changes in their student performance. However, there are also states that showed increases in their students' achievement in reading (e.g., Brandenburg *d* = 0.19 or

Schleswig-Holstein *d* = 0.16), listening (e.g., Saxony *d* = 0.25 or Brandenburg *d* = 0.22), and orthography (e.g., Brandenburg *d* = 0.33 or Mecklenburg-Vorpommern *d* = 0.23).

In English reading, students from Bavaria performed statistically significantly above average with *M* = 515 points (*SD* = 99), whereas students in Bremen (*M* = 496, *SD* = 117), Berlin (*M* = 482, *SD* = 117), and Saxony-Anhalt (*M* = 484, *SD* = 105) performed statistically significantly below the German average (Schipolowski & Sachse, 2016). In English listening, Schleswig Holstein (*M* = 500, *SD* = 93) and Bavaria (*M* = 515, *SD* = 102) led the rankings, whereas Saxony-Anhalt performed worst (*M* = 463, *SD* = 100). It is interesting that, regarding the trend in these two areas of competence, students in all countries were able to increase their achievement, as can also be seen in the statistically significant increase in the German average performance in English reading (*d* = 0.22) and in English listening (*d* = 0.26).

Students' achievement in mathematics and the sciences were assessed in the National Assessment of 2012. Trends are not yet available for these competencies. In 2012, in mathematics, especially states from East Germany performed well (e.g., Saxony: *M* = 536, *SD* = 96), whereas students from Bremen were last in the ranking (*M* = 471, *SD* = 103). A similar pattern was found in biology, chemistry, and physics. However, as a trend analysis for languages showed considerable variation in student performance within countries, these results should be interpreted with caution.

The third component of the German monitoring strategy is related to school quality on the class and school levels and is carried out by comparative testing on the state level by means of Vergleichsarbeiten/Lernstandserhebungen (i.e., comparative assessments). These assessments take place in elementary school (VERA 3) and lower secondary school (VERA 8). According to the KMK, comparative assessments are to be used for evidence-based school improvement and quality assurance, based on individual feedback on teachers' class- and student-level achievement and information regarding school leaders' cohort-level achievement. Furthermore, so that class and school results can be compared, information on average achievement is provided on the state level (e.g., Maier, 2008; Wacker & Kramer, 2012).

Research on these comparative assessments has shown that there were considerable differences between German states in the first assessments. As outlined by Maier (2008), who assessed a total of 311 teachers from Thuringia and 825 teachers from Baden-Württemberg[4], there were considerable differences between the acceptance of comparative assessments in the two states, with Thuringia showing an advantage (*d* = -0.76). In Thuringia, teachers also

[4] No information was given on the amount of participating schools.

reported higher values on comparative assessments of diagnostic issues (*d* = -0.58) and the curricular validity of assessments (*d* = -0.66), whereas teachers from Baden-Württemberg had higher values on the evaluation of comparative assessments for grading issues (*d* = 0.20). Maier suggests that these differences might result from different reform-related implementation and feedback strategies in the two states.

In another study by Wacker and Kramer (2012), the authors assessed 914 teachers (*n* = 101 schools) at intermediate track schools before the implementation of comparative assessments in Baden-Württemberg regarding the expected effects on a variety of different outcomes. Four years later, 86 schools agreed to participate (*n* = 734 teachers) in the study again. However, now teachers were asked to rate the actual effects of the comparative assessments. In both studies, teachers were asked to rate items regarding the expected effects of the assessments in supporting lectures (e.g., oriented toward preparation or oriented toward grading). Furthermore, expected effects related to a narrowing of the curriculum (e.g., comparative assessments lead to a focus on the competence areas that are part of the assessment) and additional practicing due to the assessments (e.g., a lot of additional practice is important to prepare for the assessment). The authors found a large decrease between prospective expectations of teachers regarding the effects of the comparative assessments and teacher evaluations after the introduction of these assessments. This decrease varied from *d* = 0.66 (for narrowing the subject-related curriculum) to *d* = 1.11 (for narrowing the curriculum due to a strong orientation of the tasks toward the comparative assessment).

Overall, research on comparative assessments in Germany shows that they might indeed provide useful information for school improvement and quality assurance. However, the usefulness seems to depend greatly on the exact framing and implementation of this instrument.

The fourth component of the German monitoring strategy is the National Educational Report (e.g., Autorengruppe Bildungsberichterstattung, 2014, Autorengruppe Bildungsberichterstattung, 2016), which is published every 2 years and provides the most important information on Education in Germany. The reports always focus on a specific topic, for instance, “Education and Migration” in 2006 and 2016 or “Transitions: School – VET – University – labor market” in 2008. In detail, the report is oriented toward specific indicators of education from representative samples or official population statistics and is oriented toward three dimensions of education: (a) individual self-direction, (b) social participation, and (c) equal opportunities and human resources (Autorengruppe Bildungsberichterstattung, 2014, p. 2). According to the KMK, the report builds a foundation of policy decisions in education and increases transparency on the current status of education in Germany (KMK, 2016).

This movement toward a more output-oriented educational governance is, however, not a unique German movement but is visible worldwide. Several researchers have pointed toward problems related to the strong focus on (large-scale) assessments as the foundation for education policy decisions and quality improvement (Baird et al., 2011; Goldstein, 2004; Volante, 2016).

2.2 Formal Education in Germany

In Germany, students usually start in Grade 1 in autumn when they turn 6 until the cutoff date, which has traditionally been June 30. However, eight states introduced new regulations in the beginning of the last decade, which changed the cutoff date of the school enrollment in primary school to an earlier date. Since then, especially in these states, a lot of parents have decided to enroll their children in primary school later (e.g., in Bremen 12.7% and in Bavaria 12.4%; Autorengruppe Bildungsberichterstattung, 2016). In Germany in 2014, the enrollment of about 7% of the children was delayed, whereas only 3% were enrolled earlier in primary school (Autorengruppe Bildungsberichterstattung, 2016).

After 4 years of regular primary school (in some states, 6 years), students are differentiated into different school types. In some states such as Bavaria and Thuringia, the primary school teachers' recommendations for a specific lower secondary school are binding, but in most of the states, the recommendation are just informative in nature, and students can theoretically apply to every school type. It is interesting that there are no differences in transition rates between students in states with binding and nonbinding recommendations (Autorengruppe Bildungsberichterstattung, 2016).

Variation exists regarding the different school types between the states as is visible in Figure 3, but most students have to pick from the most demanding track (usually the Gymnasium), an intermediate track (e.g., Realschule), and the least demanding track (e.g., Hauptschule). However, there are some school types that incorporate all or some of these tracks such as the Regelschule in Thuringia, which incorporates the least demanding and intermediate tracks, or the community school in Baden-Württemberg, which incorporates all three tracks and can even contain an elementary school in its a network (e.g., KMBW, 2015). Finally, there are also some schools that specialize in educating students with specific needs (e.g., with learning disabilities or blind students).

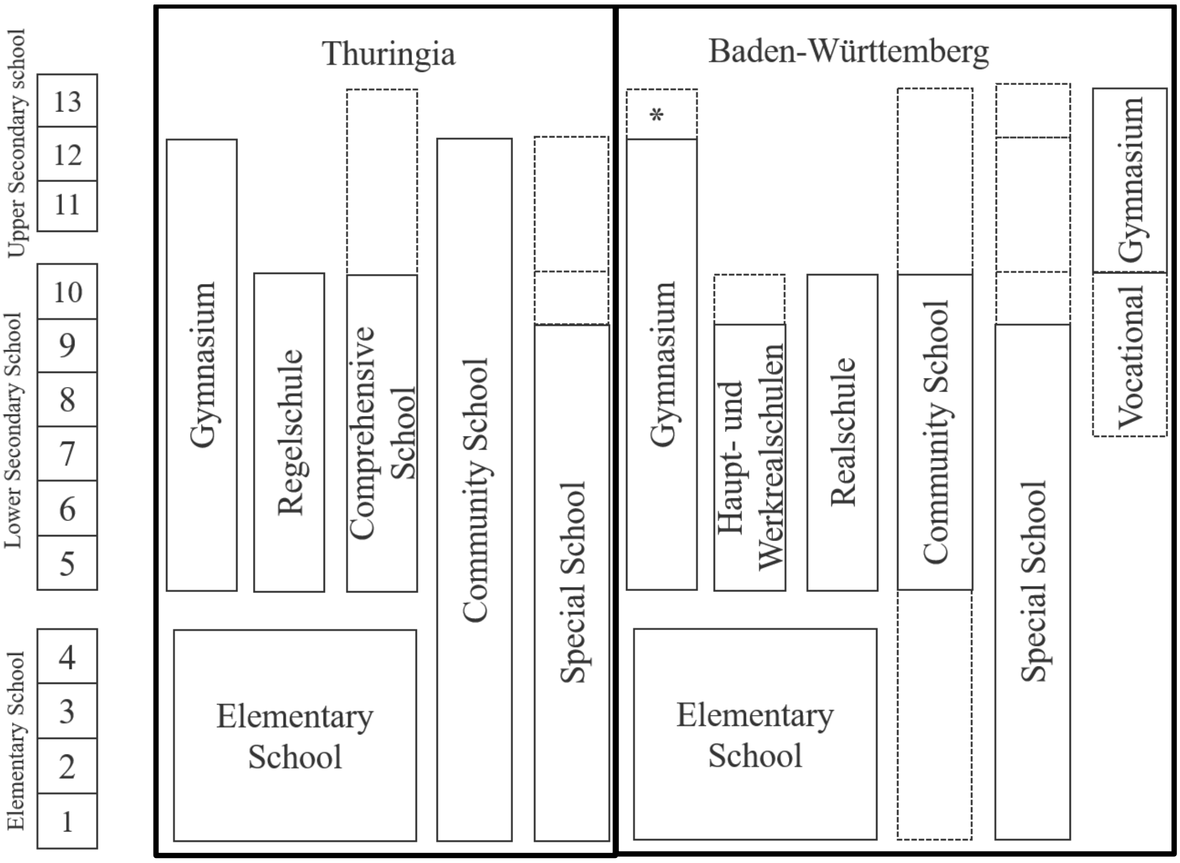

Figure 3. The central schools of the general education system in Thuringia and Baden-Württemberg. In Thuringia: Comprehensive school: Gesamtschule; Community school: Gemeinschaftsschule; Special school: Förderschule; Vocational school: Berufsbildende Schule. Please note that upper secondary school in special schools differs from upper secondary school in other school tracks (e.g., TMBJS, 2016). In Baden-Württemberg: Community School: Gemeinschaftsschule; Special school: Sonderschule. * There are 44 G9 Gymnasiums in Baden-Württemberg (e.g., KMBW, 2013). Note that other kinds of vocational schools are not displayed for the sake of parsimony. For more information on school enrollment, see the Federal Statistical Office of Germany (2017).

Currently, two major different groups of states can be identified with regard to the lower secondary school system. First, there are states that still have a more or less strong tripartite system of Hauptschule, Realschule, and Gymnasium and some form of comprehensive school (e.g., Gesamtschule or Gemeinschaftsschule), which includes more than one school track (e.g., Baden-Württemberg, Lower Saxony, or North Rhine-Westphalia). Bavaria is a special case of this group as it offers education mainly in the tripartite system. Second to this, some states focus on a dyadic system with a comprehensive school and the Gymnasium (e.g., Thuringia, Saxony, Berlin).

In 2015/2016, approximately 4.2 million students were enrolled in lower secondary school, of which 34% were enrolled in a Gymnasium, 22% in a Realschule, and 11% in a Hauptschule. The remaining students attended an integrated Gesamtschule (17%), a school with different educational tracks (11%), or some another type (4%). Around 1 million students were enrolled in upper secondary school, of which 84% attended a high track school (Gymnasium), 11% an integrated Gesamtschule, and 5% some other type (Federal Statistical Office of Germany, 2017).

2.3 Educational Governance and Educational Change

A broad variety of theoretical approaches have been concerned with questions about educational planning, guidance, and governance, especially in the German discourse on educational science (e.g., Altrichter & Maag Merki, 2016; Reinders, Ditton, Gräsel, & Gniewosz, 2011).[5]

As a starting point, it is helpful to conceptualize policy reforms in a broader framework of the process of public policy making with the aim to introduce some sort of system-wide change. This process generally consists of far more components than just the specific "reform product," which is mostly focused on empirical educational research. According to Paul Sabatier, "In the process of public policymaking, problems are conceptualized and brought to the government for solution; governmental institutions formulate alternatives and select policy solutions; and those solutions get implemented, evaluated, and revised" (Sabatier, 2007, p. 3). This view is reflected by most prominent models of public policy process (see Figure 4).[6]

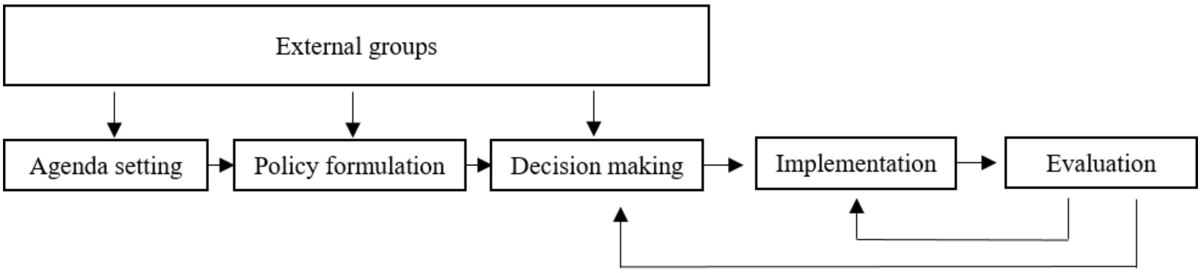

Figure 4. The policy cycle (Jann & Wegrich, 2007). For a primary version of this model, see Lasswell (1956). See also Chapter 3.4 for a more extensive version of the policy process based on Mayntz (1977).

[5] The German scientific discussion on educational policy reforms is, of course, much older and can be traced back to the end of the 1960s or early 1970s, where the educational commission of the German Advisory Council for Education published an expert opinion on this topic (e.g., Deutscher Bildungsrat, 1970) and researchers such as Saul Benjamin Robinsohn proposed a "revision of the curriculum" (Robinsohn, 1967). From that time on, there are manifold examples of scientific research on school reforms (e.g., Rolff, 1970). This time period is also related to increased research on reform implementation and school improvement, as well as research on governance and stakeholder-related accountability (e.g., Hameyer, Frey, & Haft, 1983). However, according to Terhart (1983), curriculum research was redeemed at the end of the 1970s by an increased scientific focus on teaching. Altrichter and Wiesinger (2005) again identified an increased interest in models of school reform beginning in the 1990s, and this was followed by an era of reforms, introduced after the PISA shock in Germany (e.g., Niemann, 2016). A focus on the teacher and teaching, however, seems to have remained strong over these decades (e.g., Creemers, 1994; Helmke & Weinert, 1997; Helmke, 2006; Scheerens & Bosker, 1997).

[6] Benz (2010) revived this German discussion in the general framework of governance theory, and Altrichter and Maag Merki (2016) recently published a handbook on educational governance, which transfers ideas of the governance concept to the field of education.

According to Jann and Wegrich (2007), first, numerous actors inside and outside of the government try to influence the agenda-setting according to their needs, for instance, by means of increasing attention to a specific problem or topic in the media. From this list of different topics, specific issues are selected, and the authors argue that agenda-setting is not necessarily rational. Next, specific policies that are assumed to address the problems and needs are formulated, which build the foundation of the agenda. Finally, the responsible institutions make a decision about the policy program and implement it, for instance, by means of changing a law. Finally, in the last stage of this model, the policy gets evaluated by the government itself, external scientific agents, or other actors. Over the course of the policy process, various external agents such as the unions, political opponents, the media, or other stakeholders try to shape and maybe even impede the policy. It has to be noted that the model in Figure 4 has several limitations, mostly related to a very simple representation of the far more complex policy process (Jann & Wegrich, 2007).

Based on this very global model of the policy process, one can identify different approaches related to educational planning and governance, which have been focused on in the field of education. In this regard, Berkemeyer (2010) identified major streams in the field of German educational science in recent decades, such as social-science-oriented macro-approaches, approaches involving the development of school as organizations, and approaches involving empirical educational research.

Related to this, Biehl, Hopmann, and Ohlhaver (1996; as cited in Künzli, Fries, Hürlimann, & Rosenmund, 2013), distinguished among four different models of the governmental regulation of lectures and teaching: (a) the examen-artium model, (b) the classical model, (c) the assessment model, and (d) the philanthropic model. The examen-artium model is assumed to regulate teaching and teaching contents and is based on the materials that determine whether students are admitted to higher institutions (e.g., from school to university or college). One example of this is admission tests in the United States (Scholastic Assessment Test [SAT] or American College Testing [ACT]), which strongly determine the curriculum at school. The classical model describes systems that are strongly oriented toward the curriculum as a foundation for teaching. The curriculum determines not only the content but also the time frame. This approach is comparable to models used by the Prussian school administration. However, it does not explicitly provide information to teachers about the methods that should be used for teaching. Next, the assessment model is strongly oriented toward outputs and final examinations in terms of standardized assessments. The contents of the lectures are regulated by these assessments. Compared with the examen-artium model, examinations in school determine the curriculum, rather than examinations for university excess. According to Künzli et al. (2013), the assessment model or relatives of this model are currently favored in countries that have introduced standards-based reforms. Finally, the philanthropic model focuses on inputs and is based on direct regulations of the government regarding the content and methods for teaching, rather than indirect regulations from specific assessments. Furthermore, innovations are also planned and implemented on the basis of these inputs. It is evident that these models provide only theoretical attempts to distinguish between different models of the government regulation of lectures and are therefore extreme in some regards. In practice, however, most countries have implemented characteristics of multiple models.

According to Berkemeyer (2010), approaches of empirical educational research that focus on formulating overall models of school quality can be understood as a necessary empirical correction of traditionally merely theoretical government approaches of school theory and have been exposed to much attention in recent decades. Such models typically focus on a tripartite framing of formal education in terms of inputs, processes, and outputs, and they oftentimes build the implicit or explicit foundation of educational effectiveness research (ERR; e.g., Reezigt, Guldemond, & Creemers, 1999; Scheerens, 1990). In line with Reynolds et al. (2014), in this dissertation, the objective of ERR is understood as: “It seeks to investigate all the factors within schools in particular, and the educational system in general, that might affect learning outcomes of students in both their academic and social development” (p. 197).[7]

However, before going into detail on these models and their theoretical potential for providing governance-relevant knowledge in Chapter 3, some more general aspects should be acknowledged from a larger theoretical point of view when applied to schools: First, education reform was traditionally intended to be implemented hierarchically in a loosely coupled system (Fusarelli, 2002; Porter et al., 2015).[8] The hierarchy is theoretically related to structures of the education system, where students are in the inner circle and are mainly affected by teachers, who are assumed to be directed by principals, who in turn are assumed to be instructed by the district’s education authorities (see Figure 1). These local education authorities try to implement new laws and acts, introduced by the national or federal government, the major outer circle, which includes all the other subsystems (e.g., Levin, 2000). Thinking of the educational system

[7] Note that the German term “Empirische Bildungsforschung” is referred to here as empirical educational research and is defined similarly to Gräsel (2011). The whole field of research in the area of education is referred to as educational research. EER is assumed to be one direction for educational research, which is mostly conducted on an empirical basis and focuses on aspects outlined in Reynolds et al.’s (2014) definition.

[8] For an opposing view related to the implementation of standards-based reform, see Swanson and Stevenson (2002).

from this multilevel perspective suggests that educational reforms must be able to permeate through at least some of these educational layers before they can (theoretically) impact the targeted group of students or teachers. The recognition of educational organizations as loosely coupled systems has been a central idea of researchers interested in the implementation of specific programs and effectiveness research (Swanson & Stevenson, 2002). It indicates that interactions of individuals (e.g., teachers teaching students) in the education sector, possibly in opposition to other sociopolitical systems, do not follow a very narrow scheme of instruction and are therefore greatly uneven between schools and classrooms (Fusarelli, 2002). This aspect is also related to the fact that there are no clear rules regarding a wide variety of actions within this system, unclear goals, and blurry technologies and result in a lot of pressure to truly impact instruction on the classroom level from a higher order administrative level (e.g., Swanson & Stevenson, 2002; Weick, 1976).

Next, two major aspects of policy change should be disentangled: (a) the development and characteristics of the reform itself and (b) the process of implementing the reform (see Chapter 3). Regarding the first aspect, research has indicated that reforms have a great chance to be implemented successfully if they are *flexible*, which means they can easily be modified or updated to meet the needs of the stakeholders. Furthermore, reforms also need to be *compatible*, which means they should fit in with the existing procedures and values of the system they will be implemented in (Durlak & DuPre, 2008; Rogers, 2003). Further findings by Datnow (2005), who analyzed the sustainability of the comprehensive school reform, indicated that components of reforms are explicitly useful when they actively help school leaders implement change and place few long-lasting financial demands on schools.

Regarding the second aspect, the implementation of the policy reform, Durlak and DuPre (2008) found evidence that implementation does have an effect on a variety of outcomes. Considering the results of over 500 studies, the authors identified 23 contextual factors that had a strong influence on implementation. These factors can be categorized into five larger categories: Community Level Factors, Provider Characteristics, Characteristics of the Innovation, Organizational Capacity, and Training and Technical Assistance. Many of these aspects can also be found in the extended literature review by Fixsen, Naoom, Blase, Friedman, and Wallace (2005).[9] In another study by Schaffer et al. (1997), the authors were able to identify

[9] Note that Durlak and DuPre (2008) interpreted the reform as one feature of the implementation process under “Characteristics of the Innovation,” whereas prominent policy models rather present the implementation and the reform (policy solution) as separate parts of a global policy process (e.g., Jann & Wegrich, 2007; Lasswell, 1956; Sabatier, 2007).

10 potentially handicapping issues of reform implementation, of which financial issues (e.g., reduced federal funding), commitment issues (e.g., there are no degrees of freedom for teachers to implement), and issues with the curriculum (e.g., school and state goals differ) were the three most prominent ones.

3 Educational Effectiveness and Educational Policy

3.1 The Intersection of Educational Effectiveness Research, Large-Scale Assessments, and Education Policy Reforms

As outlined in Chapter 1, the search for the holy grail to successfully increase students' achievement has a long history, with some peaks in recent decades (e.g., Hattie, 2008). However, the question is still far from having a final answer. It is interesting that different scientific disciplines have found quite different answers that might overlap only in part. As outlined in the 1966 Coleman report, which was mandated by the 1964 civil rights act, the authors summarized:

> Taking all these results together, one implication stands out above all: That schools bring little influence to bear on a child's achievement that is independent of his background and general social context; and that this very lack of an independent effect means that the inequalities imposed on children by their home, neighborhood, and peer environment are carried along to become the inequalities with which they confront adult life at the end of school. For equality of educational opportunity through the schools must imply a strong effect of schools that is independent of the child's immediate social environment, and that strong independent effect is not present in American schools. (Coleman et al., 1966, p. 325)

These findings have been updated in recent decades, and current research has shown that families indeed do matter, but, in contrast to Coleman et al. (1966), schools and especially teachers in classrooms matter as well (e.g., Campbell, Kyriakides, Muijs, & Robinson, 2003; Darling-Hammond, 2000; Hanushek & Woessmann, 2010; Heck & Hallinger, 2009; Muijs et al., 2014). Researchers such as John Hattie have provided further evidence that especially variables related to the teacher and teaching can actually explain as much variance in student achievement as individual characteristics (Hattie, 2008). Especially promising in this regard were aspects such as the teaching of metacognitive strategies (d = 0.69) or distributed learning (d = 0.71). Furthermore, formative assessments seem to have a positive effect on achievement (d = .90). It is interesting that working conditions such as within-class grouping (d = 0.28) or reducing class size (d = 0.21) seem to have less of an impact. However, these results have to be interpreted with caution (e.g., Terhart, 2011; Wecker, Vogel, & Hetmanek, 2017).

As one central starting point of EER, Reynolds et al. (2014) identified the Coleman report and related literature that has suggested that schools make little difference to student achievement over and above individual characteristics. Generally speaking, models of EER try to systemize factors related to "effective schools," mostly with a strong focus on student achievement as the central output criterion (e.g., Creemers, 1994; Reezigt et al., 1999; Scheerens, 1990; Scheerens & Bosker, 1997).

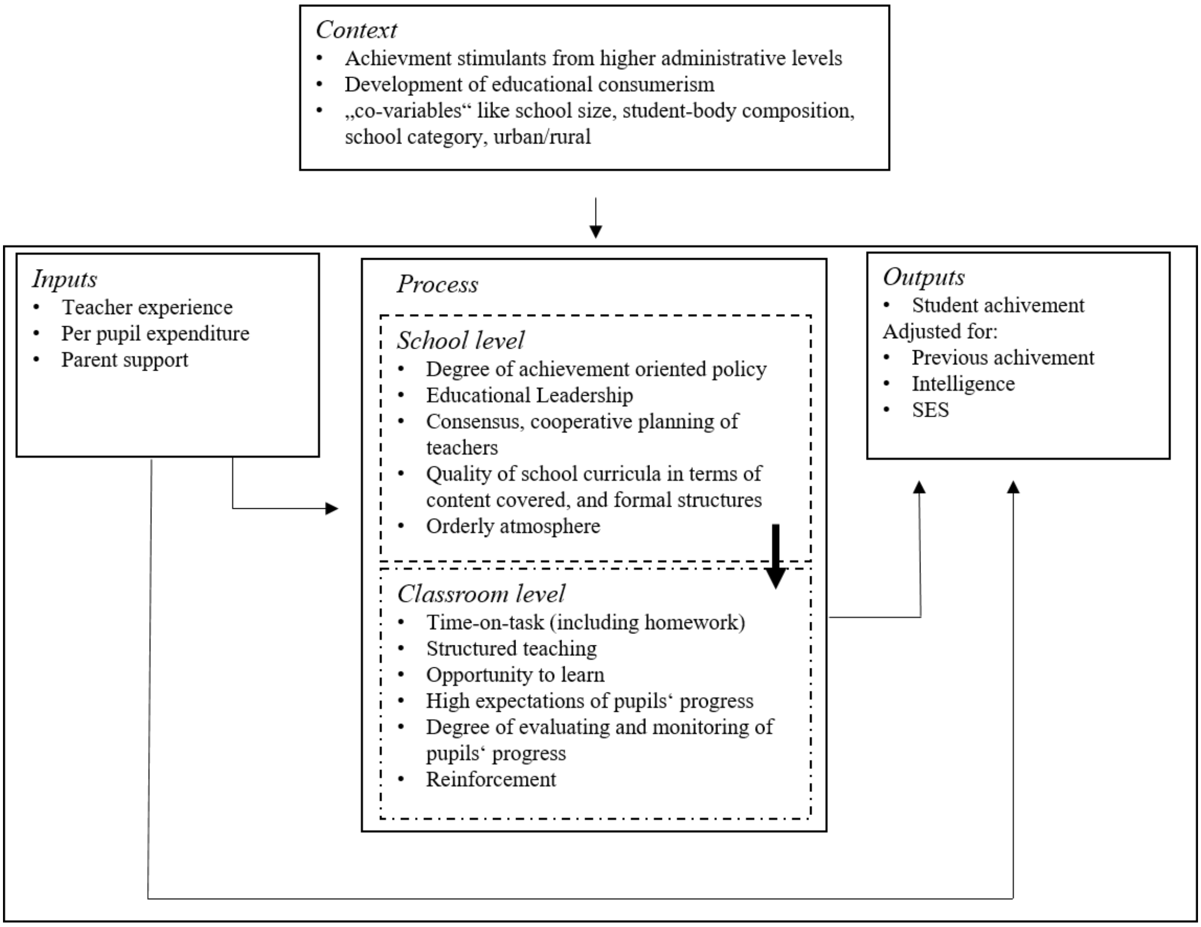
Context
• Achievment stimulants from higher administrative levels
• Development of educational consumerism
• „co-variables“ like school size, student-body composition, school category, urban/rural
Inputs
• Teacher experience
• Per pupil expenditure
• Parent support
Process
School level
• Degree of achievement oriented policy
• Educational Leadership
• Consensus, cooperative planning of teachers
• Quality of school curricula in terms of content covered, and formal structures
• Orderly atmosphere
Classroom level
• Time-on-task (including homework)
• Structured teaching
• Opportunity to learn
• High expectations of pupils‘ progress
• Degree of evaluating and monitoring of pupils‘ progress
• Reinforcement
Outputs
• Student achivement
Adjusted for:
• Previous achivement
• Intelligence
• SES

Figure 5. A model of school effectiveness (Scheerens, 1990).

As displayed in Figure 5, such models usually distinguish between three major components to explain school effectiveness and school quality, which are referred to as the input, the process, and the output (e.g., Scheerens, 1990). The core component in Figure 5 is constituted by the processes that occur in school. These processes are further distinguished into processes at the school level (e.g., educational leadership) and the classroom level (e.g., time-on-task during school lessons). The processes depend on and are influenced by specific inputs such as teacher experience or parental support as well as additional contextual variables, for instance, decisions made at higher administrative layers (e.g., Ministry of education). Finally, the processes at school lead to a specific outcome at the student level. Most important, student achievement in this model is adjusted for previous achievement, intelligence, and SES. This underscores the theoretical idea that for identifying the effect of schooling, first the impact of variables that previously affected achievement has to be controlled for. It has to be noted that this model of school effectiveness is a strongly simplified version and contains assumptions that might be more or less reasonable in the face of current research. [10]

In recent years, such models have been specifically adapted to explain determinants of student achievement and learning more accurately, and these models also provide a central foundation and framework for large-scale studies such as PISA (e.g., Baumert, Stanat, & Demmrich, 2001). As can be seen, the basic theoretical foundations of specific inputs, which influence the processes at school and in the classroom and which in turn affect student outcomes, remain similar to the models developed earlier in EER (see Figure 5). As displayed in Figure 6, these models might, however, differ in their precision regarding the variables that are considered to play an important role in the process. In this case (Figure 6), a special focus is placed on individual and family-related preconditions for learning, whereas individual characteristics are not explicitly mentioned in the model by Scheerens (1990). Grounding large-scale assessments (LSAs) on models of educational effectiveness was also important for developing standards-based reforms, as LSAs are assumed to provide important information about students' competencies and specific determinants, which can in turn be used for school and teacher accountability (e.g., Volante, 2016). Furthermore, these effectiveness models offer easy-to-read maps containing various potential variables, which, in theory, can be addressed by policy (e.g., at the school level) in order to change the school system.

According to Hamilton et al. (2009), although there is no universally accepted definition of standards-based reform, the main features can be summarized as: the setting of "academic expectations for students," "alignment of key elements of the educational system," "assessment of student achievement," "decentralization," "support and technical assistance," as well as "accountability" (p. 2). Standards-based reform has increased in importance because of A Nation at Risk (The National Commission on Excellence, 1983) with a peak following the No Child Left Behind (NCLB) act in the United States.[11]

[10] In the displayed version, which came from Scheerens (1990), the model for instance suggests that school-level variables affect classroom-level variables, thus reflecting the perspective of "top-down" processes within schools, instead of a reciprocal relationship between these two layers as suggested by the literature on distributed leadership (e.g., Heck & Hallinger, 2009; Heck & Hallinger, 2010).

[11] The standards-based reform movement (in Germany often referred to as: Outputsteuerung) is much younger in Germany and had its starting point after the PISA shock, which followed the first PISA assessment in 2000 (e.g., Niemann, 2016).

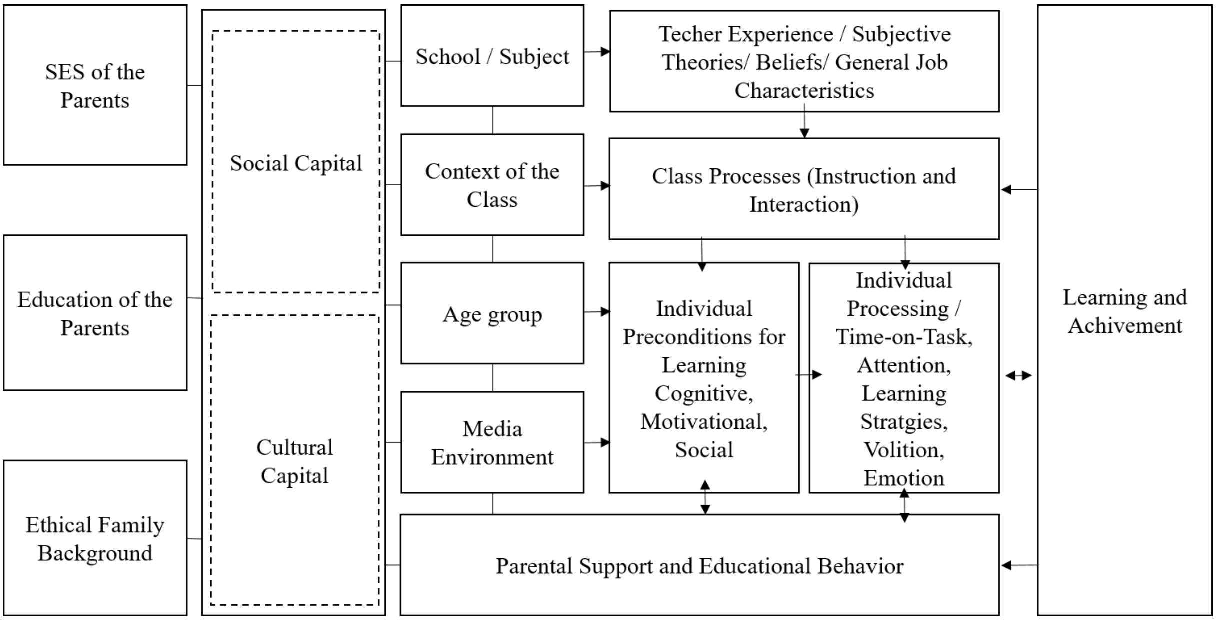
SES of the Parents
Education of the Parents
Ethical Family Background
Social Capital
Cultural Capital
School / Subject
Context of the Class
Age group
Media Environment
Techer Experience / Subjective Theories/ Beliefs/ General Job Characteristics
Class Processes (Instruction and Interaction)
Individual Preconditions for Learning Cognitive, Motivational, Social
Individual Processing / Time-on-Task, Attention, Learning Stratgies, Volition, Emotion
Parental Support and Educational Behavior
Learning and Achivement

Figure 6. Conditions for school achievement – General framework (Translated by the author; based on Baumert et al., (2001) oriented on Helmke & Weinert, 1997).

Swanson and Stevenson (2002) outlined the basic relevance of standards-based reform for educational policy: “standards-based reform possesses a process-driven conception of educational change that explicitly links schooling inputs and policy drivers to student outcomes through clearly defined mechanisms” (p. 3).

Within the framework of standards-based reform, higher order educational administration (e.g., on the state or national level) is expected to set specific goals (what students should know at a specific point in time) and monitor the status of whether these goals are reached by implementing rigorous assessment strategies (e.g., KMK, 2016).

As opposed to the United States, where many states have implemented test-based school accountability as a central part of standards-based reform (e.g., in terms of value-added models and other reward- and sanction-based mechanisms that are linked to student achievement; Ravitch, 2011), Germany has not yet followed such developments.[12] Combining the results of educational testing and accountability is oftentimes viewed as the starting point of the vast increase in standardized student assessments on national and international levels (e.g., Lee, 2015; Volante, 2016).

In their study, Swanson and Stevenson (2002) investigated (a) potential linkages between the structure of the standards-based reform movement on national and state levels, as

[12] Linking results of LSAs to accountability can influence the meaning of such assessments. If tests have severe consequences for educational administration, teachers, or students, they are oftentimes referred to as “high-stakes tests,” whereas tests without consequences are called “low-stakes tests” (e.g., Au, 2007). For features and problems linked with educational testing as a basis for education accountability, see Koretz (2008).

well as (b) associations between policies on the state level and classroom practices at schools, using a rich data set from the National Assessment of Educational Progress (NAEP) study. Overall, their findings suggest strong relations between the two levels, as they found that state activism was strongly mirrored by national movements. Furthermore, state activism had a statistically significant, independent effect on teachers' classroom practices. Their study can therefore be taken as evidence of potential positive effects of standards-based reform, and it challenges previous assumptions of a loosely coupled educational system (e.g., Fusarelli, 2002), where it was assumed that regulations are difficult (or close to impossible) to diffuse from the national or state level into the classrooms.

In line with this, the stakeholders of LSAs promoted the following: "PISA is an ongoing programme that offers insights for education policy and practice, and that helps monitor trends in students' acquisition of knowledge and skills across countries and in different demographic subgroups within each country," and in more detail, it "identif[ies] the characteristics of students, schools and education systems that perform well" (OECD, 2014, p. 24). Finally:

> The findings allow policy makers around the world to gauge the knowledge and skills of students in their own countries in comparison with those in other countries, set policy targets against measurable goals achieved by other education systems, and learn from policies and practices applied elsewhere" (OECD, 2014, p. 24).

As outlined, the framework of standards-based reform strongly relies on rigorous testing for accountability, and the OECD supports this perspective by suggesting that the results of achievement tests can be used by policy makers to shape education: Basically, from this perspective, best practice information delivered by countries that show good performance in PISA can be generalized and used as a blueprint for policy decisions in other countries.

Taking a closer look at the literature on the impact of LSAs on education policy indicates that LSAs, especially PISA, indeed impact education policy (e.g., Bieber, Martens, Niemann, & Windzio, 2014; Volante, 2016). Related to this, several authors have criticized aspects (e.g., the focus on a small range of curricular content) of the use of standardized tests and effects on policy to adapt the focus of school curricula to increase standardized achievement in LSA rankings (e.g., Koretz, 2008; Meyer & Zahedi, 2014; Volante, 2016). Moreover, as outlined by Goldstein (2014), the OECD undermines the fact that PISA results are not able to explain differences between countries in student achievement (e.g., Fend, 2004).

Volante (2016) further characterized the increased importance of the LSAs for national policy decisions:

> These contextual surveys are meant to help policymakers identify student, classroom, school, and national variables associated with student achievement. Both the OECD and

> IEA make positive statements on their respective websites on the utility of these international benchmark measures and their associated contextual surveys for informing national education policy decisions (pp. 5-6).

However, such statements stand in contrast to Baumert (2016), who argued that:

> Furthermore, empirical evidence never guarantees the practical implementation of policy decisions in a professional area of application. Basically, this is known by all actors in the policy system, even if empirical educational research is expected to make a larger contribution to policy agendas (translated; p. 223).

Related to this, Bieber et al. (2014) suggested that two aspects in particular are relevant for the strong diffusion of the "OECD agenda" to the national level, which are *transnational communication* (especially policy emulation and policy learning) as well as *competitive pressure*. Policy emulation is the process of transferring internationally accepted policy models into the national context in order to legitimize national agendas and decisions. This aspect is also underscored by recent research by Dedering (2016). By contrast, policy learning rather describes the rational process of finding policy solutions, and considering experiences from other countries and the OECD offers such information comprehensively. Competitive pressure finally describes the mechanism by which competition between countries results in mutual adaptions of policy strategies of other countries to foster success (Bieber et al., 2014). Related to PISA, such success is mainly defined in terms of achievement measures.

It has been noted that this perspective of whether LSAs are a valuable instrument for informing, substantiating, and steering policy decisions strongly relies on the assumption that differences in students' achievement between countries and educational systems can be reasonably explained and are indeed affected by educational policy and administration (e.g., Goldstein, 2014; Volante, 2016). Some authors have argued that debates oftentimes ignore the assumption that student achievement is also the result of system characteristics, which are the result of extensive, long, cultural and historical traditions and are therefore not easy to change or adopt. These ideas are in line with research that has indicated problems and limitations in transferring policies across states (e.g., Fend, 2004; Stein, Hubbard, & Mehan, 2004).

To sum up, there is a major controversy regarding the status of LSAs for educational policy. This controversy is important to consider as most LSAs are strongly oriented toward central theoretical models from EER, and both educational effectiveness models and the related results of LSAs therefore strongly impact the way people working in educational policy and administration think about education and how to reform it. Proponents of standards-based reform would argue that LSAs can provide reasonable knowledge for policy decisions (e.g., OECD, 2015), whereas opponents would strongly doubt this, for instance, because LSAs fail to clearly identify reasons for differences in student achievement between countries (e.g., Goldstein, 2014). However, what is now clear is that policy indeed integrates results from LSAs into their policy agendas, and this is why many arguments for national reforms in Germany are based on comparisons with different countries, which succeed in LSAs (e.g., Bieber et al., 2014; Dedering, 2016).

Finally, from an intermediate perspective, one could relativize both previous prospects by assuming that LSAs can provide important knowledge, which might, however, not be directly useful for public policy making (e.g., Baumert, 2016). This discussion can therefore be integrated into the larger topic of the drawbacks and opportunities of scientific evidence for policy decisions, which I will outline in the next chapter.

3.2 Evidence of Effectiveness Research and Relations to Educational Policy

After discussing the area of conflict described above, it is important to consider the contributions that the results of EER can make to educational policy and practice. The following chapter will focus in more detail on the kinds of knowledge that EER can reasonably provide and on current opportunities and limitations when using such research evidence from the perspective of research, policy, and practice. In a first step, I will outline different perspectives on evidence before linking these perspectives to specific dimensions of knowledge. Finally, I will link these different dimensions of knowledge to the process of public policy making.

As outlined by Robert E. Slavin, the first time in history that educational funding through policy was explicitly linked to the effectiveness of a program was only a little less than two decades ago. At that time, the US Congress offered $150 million p.a. to fund comprehensive reform models, the effectiveness of which had to be demonstrated in an experimental framework with standardized tests (Slavin, 2002). Two aspects were somewhat startling here: (a) Evidence-based policy seems to be surprisingly young in educational research, and (b) The methods that policy chose to judge effectiveness initially seemed to follow a traditional psychological perspective. As also suggested by Slavin (2002), “Educators and policymakers legitimately ask, ‘If we implement Program X instead of Program Y, or instead of our current program, what will be the likely outcomes for children?’ For questions posed in this way, there are few alternatives to well-designed experiments” (p. 18).[13]

Related to Slavin’s (2002) observations, various initiatives have been implemented since the beginning of the new millennium, such as the What Works Clearinghouse (WWC) or

[13] See Campbell (1969) for an older, comparable contribution to this topic.

the Best Evidence Encyclopedia (BEE). These initiatives have been implemented in an attempt to synthesize educational research and offer practitioners more profound answers to the question outlined above by relying on high-quality research (e.g., Slavin, 2008).[14] The development of these institutions appears to constitute one solution to a specific problem, already outlined earlier by Hedges and Waddington (1993): "The problem is how to convert evidence into knowledge and such knowledge into policy" (p. 345). In a response to Slavin (2008), Derek C. Briggs (2008) further disentangled two aspects that seemed to be of fundamental importance in this regard:

> The evidence necessary to answer the question, what is the magnitude of the effect of a program on student outcomes is best provided by a randomized controlled experiment, the clear gold standard (although a strong quasi-experimental design may come close). However, for the evidence necessary to answer the question, how does a program produce an effect on student outcomes? there is no clear gold standard for a methodological approach. (p. 15)

This also fits in with a critique outlined by Goldstein (2014) about LSAs (see Chapter 3.1). For Briggs (2008), initiatives such as the WWC or BEE strongly focus on the internal validity and statistical conclusion validity of research and somewhat neglect aspects of generalizability (external and construct validity).[15]

From a broader perspective, a central question related to this discussion seems to be the question of what counts as "evidence". Different opinions and definitions of what is usually referred to as "causal" exist in EER and the social sciences in general. These were outlined by Goldthorpe (2001). In this overview article, the author distinguishes between three different perspectives on causality, which he refers to as (a) *causation as robustness dependence*, (b) *causation as consequential manipulation*, and (c) *causation as a generative process*. The first approach suggests that causality can be thought of, as might be known from regression analytical modeling and as referred to by Granger (1969), as *Granger causality*. Very basically, according to Goldthorpe (2001), the idea behind this type of causality is that if a variable X is still predictive of future values of a variable Y, after controlling for everything but X, this

[14] In the face of recent developments regarding the replication of scientific evidence in disciplines that strongly rely on experimental research (e.g., Open Science Collaboration, 2015), the field of education will have to discuss the implications of these recent developments for their own field and research paradigms more strongly in the future (e.g., Deaton & Cartwright, 2016; Malouf & Taymans, 2016).

[15] For an older, quite comparable view on this issue, see Cronbach (1980; as cited in Chen & Rossi, 1987). Note that recent literature such as Hitchcock, Kratochwill, and Chezan (2015) suggests that WWC indeed provides information on generalizability. However, these seem to focus on external validity (generalizability of cause-effect relations over persons, settings, and so forth) rather than construct validity (generalizability of constructs across persons, settings, and so forth).

variable X Granger-causes Y. Although somewhat old, this idea of causality can still be found in various current publications in EER.

The second approach, causation as consequential manipulation, indicates that causality can be thought of as anything that is achieved by the application of rigorous randomized experiments. The basic idea of randomized experiments has a long tradition, especially in psychology (e.g., Rubin, 1974), and depends on the identification and manipulation of a specific factor (the independent variable), holding constant potentially confounding variables, whereas the desired outcome (the dependent variable) is traditionally measured before and after manipulation. Various different designs exist (e.g., Shadish, Cook, & Campbell, 2002). However, the basic idea follows a treatment-control comparison on the outcome variable.

Finally, as outlined by Goldthorpe (2001), several authors have criticized both concepts for the minor relevance of a theory of an underlying social process and developed a new perspective on causality in order to tie “the concept of causation to some process existing in time and space, even if not perhaps directly observable, that actually generates the causal effect of X on Y and, in so doing, produces the statistical relationship that is empirically in evidence” (p. 9). This perspective follows three steps: (1) “establishing phenomena that form the explananda; (2) hypothesizing generative processes at the level of social action; and (3) testing the hypothesis” (p. 10). The first step in this model can be purely descriptive in nature, however researchers should have evidence that the phenomena “express sufficient regularity to require and allow explanation” (p.10). Afterwards, potential causes of social regularities are considered on a more concrete level. From Goldthorpe’s (2001) perspective, the second step cannot be based merely on statistical procedures but requires “a crucial subject-matter input” (p. 11). Finally, the established models of the generative process are tested with adequate designs and statistical models.[16]

Based on the information outlined above, the question of the extent to which evidence from EER can be used for policy decisions has not yet received a final answer. Conversely, it has actually become a more sophisticated question with many different answers: What is defined as “evidence” and as “causal” strongly varies between and even within scientific disciplines (e.g., Goldthorpe, 2001), for instance, apparent in mix-ups of aspects such as correlation and causation. As outlined by Reinhart, Haring, Levin, Patall, and Robinson (2013), a large number of correlational studies in major educational research journals have made

[16] Note that Baumert (2016) ascribes LSAs a function, which is strongly related to the first step of the causation as a generative process model.

recommendations for practice, even though such a practice is not valid from the most influential, current methodological points of view. In practice, whereas economists traditionally rather make use of approaches such as instrumental variables, difference-in-differences or regression-discontinuity designs (e.g., Murnane & Willett, 2011) to estimate causal effects using nonexperimental data, psychologists are traditionally trained to conduct randomized experiments in their studies. It is evident that the two perspectives share a strong focus on internal validity, whereas external validity is often seen as secondary or not important at all. As outlined by Briggs (2008), however, external validity is, from a slightly different perspective on causality, of central importance, and this point is related to the distinction between efficacy and effectiveness (e.g., Wortman, 1983). Regarding the framework of evidence-based or evidence-informed policy (e.g., Bowen & Zwi, 2005), this means that what is claimed to be "evidence" strongly differs between different subsystems of science, and depending on these different definitions and perspectives, "universal definite evidence" does not exist.[17]

In this regard, Bromme et al. (2014) introduced a useful differentiation by distinguishing between different dimensions of knowledge provided by EER.[18] These knowledge dimensions are also related to the different types of typical research designs that are needed to generate such knowledge. The four dimensions are (a) Description, (b) Explanation, (c) Change, and (d) Evaluation and can simultaneously represent functions and knowledge dimensions of educational research.

Description and Explanation. Whereas the first dimension (*descriptive knowledge*) is generated, for instance, via rigorous educational monitoring on national and international levels (e.g., using LSAs), the second function is focused on explaining specific phenomena (*explanatory knowledge*), which might have been detected during the description process. The distinction between these two types of knowledge is not of an arbitrary theoretical nature but is also related to different types of research designs and methodologies: Explaining why things work or behave in a specific way focuses far more on processes and mechanisms that are potentially established in series of laboratory experiments or specific quasi-experimental designs using advanced methodologies to identify causal effects. On the other hand, description

[17] However, there is of course at least some sort of order between the strength between different types of evidence, whereby randomized experiments are usually seen as a gold standard (e.g., Lohr , 2004; Murnane & Willett, 2011). But as shown by Briggs (2008), even when research is committed to the highest available standards such as the WWC and the BEE, they might differ considerably in their judgment of a study's effectiveness.

[18] Note that Bromme, Prenzel, and Jäger (2014) define the functions for the German "Bildungsforschung" (educational research) and not explicitly for EER. However, in this dissertation, EER is understood as one large area of research within the larger field of educational research (see above for a definition of EER). The functions outlined by Bromme, Prenzel, and Jäger (2014) can be perfectly generalized to EER.

works perfectly without knowing about ongoing processes in detail and therefore rather depends on representative data sets.

As is obvious from this example, the two dimensions of knowledge (describing and explaining knowledge) that are potentially provided by educational science research can be strongly related to and highly relevant for practice and for public policy making. Descriptions of potentially problematic phenomena (e.g., differential achievement between girls and boys) will, however, need to be explained correctly in order to be addressed adequately, and explanations of specific processes and mechanism will have to be generalized and will need to fit into broader contexts of descriptions.

Change. The aspect of relating EER to policy and practice is especially visible in the third dimension of the model. The third dimension outlined by Bromme et al. (2014) is referred to as *change knowledge*, which can potentially result from knowledge about causal mechanisms of specific phenomena. However, Bromme et al. (2014) also pointed out that descriptive knowledge can be used as a foundation for change within a specific feedback system (see evaluation function).

Traditionally, policy makers identify problems (e.g., from descriptions) and search for appropriate explanations and solutions on the administration level as a foundation for change in terms of specific policy programs and reforms (e.g., Jann & Wegrich, 2007). As shown in a study by Dedering (2016), who investigated how the German educational administration typically uses knowledge provided by LSAs (in this case, PISA), descriptive information from PISA was used to legitimize or to preserve political power. This aspect is perfectly related to the dimension of policy emulation, described in the model of diffusion of international policies by Bieber et al. (2014).

From this perspective, it becomes more evident that changing the traditional logic of action toward a logic of action suggested by authors such as Slavin (2008), whereby politicians' decision making, related to reforms, should depend on a strict, rigorous evidence base, might be challenging for various reasons. First, movements toward strict evidence-based decision-making is likely to result in stagnancy in educational fields, where no or very limited knowledge is available.[19] This was also outlined by Slavin (2008), who argued that:

> A key requirement for evidence-based policy is the existence of scientifically valid and readily interpretable syntheses of research on practical, replicable education programs.

[19] On the homepage of the WWC (https://ies.ed.gov/ncee/wwc/), it can indeed be seen that there are many fields in education where there is too little or no strong evidence base at all.

> Education policy cannot support the adoption of proven programs if there is no agreement on what they are (p. 5).

Second, stagnancy itself stands in contrast to the behavioral logic of the policy system, where politicians are limited in the time they have available to leave their mark on the education system, and stagnation is labeled negatively (in the sense of “no progress”), especially in relation to education and economic growth (e.g., Easterly, 2001; Hanushek & Woessmann, 2010).[20] Third, sticking to an evidence base in a strict sense would potentially lead to a decrease in the power of politicians (e.g., Bennett & Howlett, 1992) as they would depend on external evidence or would be prompted to choose between only different external pieces of evidence. Furthermore, this stands in clear contrast to a long tradition regarding the logic of action of the political administration, who have traditionally had to identify the causes of problems without being able to rely on an external research base such as the WWC.[21] In such cases, if politicians are somewhat forced to choose (only) from among a specific set of scientifically justified policy options, agents who are not democratically legitimized would implicitly make decisions about policy matters, and this would stand in opposition to legal frameworks.

Based on these considerations, it seems to be more reasonable to promote *evidence-informed policy* in some situations rather than to promote strict, evidence-based policy. The idea of evidence-informed policy is in line with Hedges and Waddington’s (1993) earlier considerations: “We agree that there is a vast amount of evidence … that should be used to inform educational policy decisions” (p. 345). Furthermore, evidence-informed policy reflects the rather realistic picture of a potential broad variety of evidence that stakeholders can and have to choose from, whereby research tends to emphasize one potential source out of many (e.g., Bowen & Zwi, 2005).

However, introducing change (e.g., by means of policy reforms), based on empirical evidence, is in no sense straightforward, even if “strong evidence” is at hand. The true length of the list of potential “change killers” seems to be unknown as of yet, and the process of introducing change is demanding. This is the case not only in education (e.g., Durlak & DuPre, 2008; Schaffer et al., 1997) but also in other disciplines such as medicine (e.g., Glasgow & Emmons, 2007). However, considering research on the implementation of policy reforms could further increase the awareness of potential challenges among politicians. The third function of

[20] The resulting discrepancy might be striking, especially when considering a scientific perspective on evidence whereby effects of a reform are not necessarily expected to be positive in advance (e.g., Campbell, 1969).

[21] This is especially the case in Germany. Other education systems that have introduced more sophisticated accountability systems (e.g., the United States or England) make use of different incentive structures (e.g., Baker & O´Neil, 2016; Thomas, Gana, & Muñoz-Chereau, 2016).

the model (Bromme et al., 2014) therefore possesses the complex hybrid between the two functions of description and explanation and a new form of knowledge that is defined as the implementation or transfer knowledge (e.g., Fullan, 1983, 2016; Gräsel, 2010; Rogers, 2003).[22]

Evaluation. Finally, Bromme et al. (2014) suggested that empirical educational research also offers the evaluative function of monitoring specific changes introduced by educational policies. According to Rossi, Freeman, and Lipsey (2004), program evaluations can be described as “the use of social research methods to systematically investigate the effectiveness of social intervention programs in ways that are adapted to their political and organizational environments and are designed to inform social action in ways that improve social conditions” (p. 29). The authors further defined a social program as “an organized, planned, and usually ongoing effort designed to ameliorate a social problem or improve social conditions” (p. 29).[23]

Educational evaluations are of major importance because, as outlined, independent of the status of evidence, changes are constantly introduced in the education system by the political administration (e.g., by means of specific reforms; e.g., OECD, 2015). Furthermore, even if an innovation that has shown “strong evidence” in research or the synthesis of research is implemented, uncertainty exists about how the program will work out, given the environmental specificities of the school system. Furthermore, whether or not the specific mechanisms that have been shown to impact the desired outcomes in previous research can be addressed in a similar way in practice remains an open question to some extent (e.g., Briggs, 2008).

Related to this, Wortman (1983) distinguished between the *efficacy*, *effectiveness*, and *efficiency* of interventions. In this triad, efficacy can provide an answer to the question of whether a program *can* work (e.g., tested in randomized experiments), whereas effectiveness answers the question of whether the program indeed *does* work (e.g., in the field). Finally, efficiency focuses on the question of whether a program is cost-efficient. From this, it can be summarized that rigorous evaluations in the field can generate knowledge, for instance about the effectiveness and efficiency of a program or reform, and these two aspects are directly linked to the major functions of *accountability* and *sustainability.*

From a perspective of accountability, summative evaluations are a reasonable option to provide knowledge regarding the effectiveness of a program that can in turn be used to justify policy decisions to the taxpayer in general and parents and students more specifically (e.g., Rossi et al., 2004). Furthermore, evaluations can also be used to get a close-up on specific

[22] See also Chapter 3.3 for more detailed information on implementation.

[23] It has to be noted that this is a rather broad definition of a program, and it might differ from more specific definitions of programs in other contexts (e.g., Slavin, 2002).

changes within the education system, for instance, changes implemented by specific schools, to provide information for the justification of these decisions to educational authorities.

From a perspective of sustainability, a rigorous monitoring of reforms and specific programs is also important in order to prevent seemingly random trial-and-error policy implementation of reforms and programs and to truly learn from the interventions (e.g., Torgerson & Torgerson, 2001). This is true for both policy and science, both of which can increase knowledge about "what does work" and identify unintended side effects of specific intervention reforms (e.g., Black & Wiliam, 2006; McConnell, 2010). Moreover, a cost-efficiency analysis can provide important knowledge for future programs and reforms, which provide the foundation for a responsible use of resources needed to implement the reform.

From this perspective, evaluations can be understand as practical evidence that is based on evidence that was found previously in rather controlled, potentially artificial settings. Of course, evaluations are not only an important tool for monitoring reform effects in the education sector but are also a quite frequently chosen option for monitoring the outcomes of specific policy interventions in many different fields of policy (Rossi et al., 2004). As outlined by the European Commission, in its interinstitutional agreement on better law-making, "The three Institutions [the European Parliament, the Council and the Commission] consider that public and stakeholder consultation, ex-post evaluation of existing legislation and impact assessments of new initiatives will help achieve the objective of Better Law-Making" (Interinstitutional Agreement between the European Parliament, the Council of the European Union and the European Commission on Better Law-Making, 2016, para. 6). In more detail, impact assessments "are a tool to help the three institutions reach well-informed decisions and not a substitute for political decisions within the democratic decision-making process." However, "In the context of the legislative cycle, evaluations of existing legislation and policy, based on efficiency, effectiveness, relevance, coherence and value added, should provide the basis for impact assessments of options for further action" (Interinstitutional Agreement between the European Parliament, the Council of the European Union and the European Commission on Better Law-Making, 2016, para. 22). As shown above, the basic idea is that evaluations provide the foundation for more specific impact assessments, which are some sort of combination of various information and research on specific legislations, and, maybe even more important, both of these tools therefore provide important instruments for informed decision making. This strong commitment to rigorous assessments and evaluations is also visible in numbers because, since 2003, a total of 975 impact assessments, and since 2010, about 688 evaluations were completed (European Commission, 2016).[24]

As can be seen, the EU shows a strong commitment to the quality control of regulations using evaluations and other forms of output-oriented assessments. This is interesting to see because it underscores the idea that the standards-based reform movement (e.g., Swanson & Stevenson, 2002) seems to impact all areas of policy making and is not a unique solution for the field of education as has sometimes been suggested (e.g., Bellmann & Weiß, 2009).

For the field of education, as is evident from the Educational Policy Outlook of the OECD, however, few policy reforms are accompanied by rigorous scientific evaluations or follow output-based funding strategies (OECD, 2015; Slavin, 2002). In the report, the OECD distinguished between six major education reform types, which are (a) Equity and quality, (b) Preparing students for the future, (c) School improvement, (d), Evaluation and Assessment, (e) Governance, and (f) Funding. Most of the reforms implemented in OECD countries between 2008 and 2014 were related to the second (29%) and third types (24%), although only 10% of all reforms were accompanied by evaluations (OECD, 2015).

Caplan, Morrison, and Stambaugh (1975; as cited in Wollmann, 2014) outlined potential causes of the misfit between policy decisions and social science research. On the one hand, policy follows a (simplified) rationale to gain and keep power to accomplish the desired objectives within a given time frame, which might conflict with the objectives of multiple other stakeholders (Bennett & Howlett, 1992). On the other hand, science tries to search for a(n) (idealized) "truth" that is independent of moral and social values (e.g., Weber, 1919; Wollmann, 2014). From this perspective, evaluations of specific interventions might go along with strongly differing outcomes for the group of scientists "just evaluating it" and the politicians who are in charge of conceptualizing and implementing it. Campbell (1969) formulated the following:

> Given the inherent difficulty of making significant improvements by the means usually provided and given the discrepancy between promise and possibility, most administrators wisely prefer to limit the evaluations to those outcomes of which they can control, particularly insofar as published outcomes or press releases are concerned. (p. 410)

[24] The European Parliament, the Council, and the Commission outline three specific tools for better law-making, which are *Impact assessment, Public and stakeholder consultation and feedback,* as well as *Ex-post evaluation of existing legislation.* According to the Agreement, "*Impact assessments should cover the existence, scale and consequences of a problem and the question of whether or not Union action is needed. They should map out alternative solutions and, where possible, potential short and long-term costs and benefits, assessing the economic, environmental and social impacts in an integrated and balanced way and using both qualitative and quantitative analyses*" (Interinstitutional Agreement between the European Parliament, the Council of the European Union and the European Commission on Better Law-Making, 2016, para. 12).

He further concluded:

> Ambiguity, lack of truly comparable comparison base, and lack of concrete evidence all work to increase the administrator`s control over what gets said, or at least to reduce the bite of criticism in the case of actual failure. There is safety under the cloak of ignorance. (p. 410)

This logic of action described by Campbell (1969) nearly half a century ago seems to still hold today to some extent (e.g., Dedering, 2016; OECD, 2015). However, as outlined above, there are also visible improvements (e.g., Slavin, 2008) that indeed show a trend toward "experimental administrators" and away from "trapped administrators" (Campbell, 1969, p. 426).

In the face of the discrepancy between the large number of educational policy decisions and reforms and the small number of rigorous educational evaluations, it seems especially important to outline the links between research and policy. This was done in terms of the alliance model (Figure 7), which explicitly combined the stages of the policy cycle (e.g., Jann & Wegrich, 2007; Mayntz, 1977) and the different dimensions of knowledge (Bromme et al., 2014) that EER can provide.

As can be seen in Figure 7, there are multiple intersections (labeled a to g) where EER can reasonably provide knowledge during the process of public policy making. The core of the alliance between EER and political administration and policy is assumed to be built on the first three types of knowledge that are assumed to be important for (a) agenda setting, (b) policy formulation, (c) decision making, and (d) implementation:

Problems can be identified by applying descriptive knowledge, for instance, related to disadvantages of specific subgroups of students whose performance is below average. Such descriptive results can include, for example, the finding that boys perform considerably worse in languages compared with girls (e.g., Stanat et al., 2016) or an increased association between family background and student achievement (e.g., Gustafsson & Yang Hansen, 2017). However, not only can EER "identify new problems," but it can also help to identify and test for the potential causes of these problems. In doing this, it is thus possible to provide knowledge about the potential specific mechanisms behind descriptive findings, along with further knowledge about the effectiveness of specific interventions that might increase boys' language achievement or increase the achievement of low SES students. Identifying and testing for potential factors that might cause such undesirable developments has a long tradition in educational science and especially in EER. However, not only is it possible to identify studies in which the potential causes have been investigated and studies that have been concerned with implementing and testing specific intervention programs. Research in the field of education is also largely build upon specific theories and models that explicitly name the relevant (e.g., psychological) factors of student achievement, and such theories can be used to inform policy decisions. Furthermore, they might be especially helpful in situations in which no general knowledge exists in either policy or in research, for instance, when a new policy reform is formulated, implemented, and evaluated (see Chapter 3.4 for examples of such theory and models). This knowledge can therefore be used to suggest and create more specific policy options for changes that can be considered by the political administration and politicians when they formulate a new policy agenda and make decisions about it. It might also be helpful in the anticipation of negative side effects, as I will outline in the following chapters. Furthermore, knowledge from implementation and transfer research can be used when implementing the reform or program (e.g., Durlak & DuPre, 2008; Fullan, 1983, 2016; Gräsel, 2010; Rogers, 2003; Schaffer et al., 1997).[25]

Based on this, (e) rigorous policy evaluations can be conducted, and at this stage, educational research can either execute this process as an external agent or provide knowledge for self-evaluations. The evaluation can focus on both short- and long-term impacts. Finally, results from evaluations can again be integrated into the general knowledge framework of reforms by policy and research and be used to further adjust or replace previous decisions regarding (f) the characteristics of the reform or (g) related to the process of implementation. As is obvious, the model cannot reflect the complexity of this process, especially regarding judgments about the definition of explanatory knowledge and evidence that might be used to inform policy. Therefore, the model critically depends on the assumption that there is evidence that can be transferred, and this might be true and effective only in some areas. However, in any case, there is a vast amount of knowledge that can be used for policy decisions and that can increase the likelihood of successful educational policy making whenever reforms need to be implemented. In any case, the alliance model also suggests major challenges for research and policy, especially in terms of a convergence of the logic of the actions of these two systems as outlined by Caplan et al. (1975; as cited in Wollmann, 2014).

[25] This research will be outlined in more detail in Chapter 3.3.

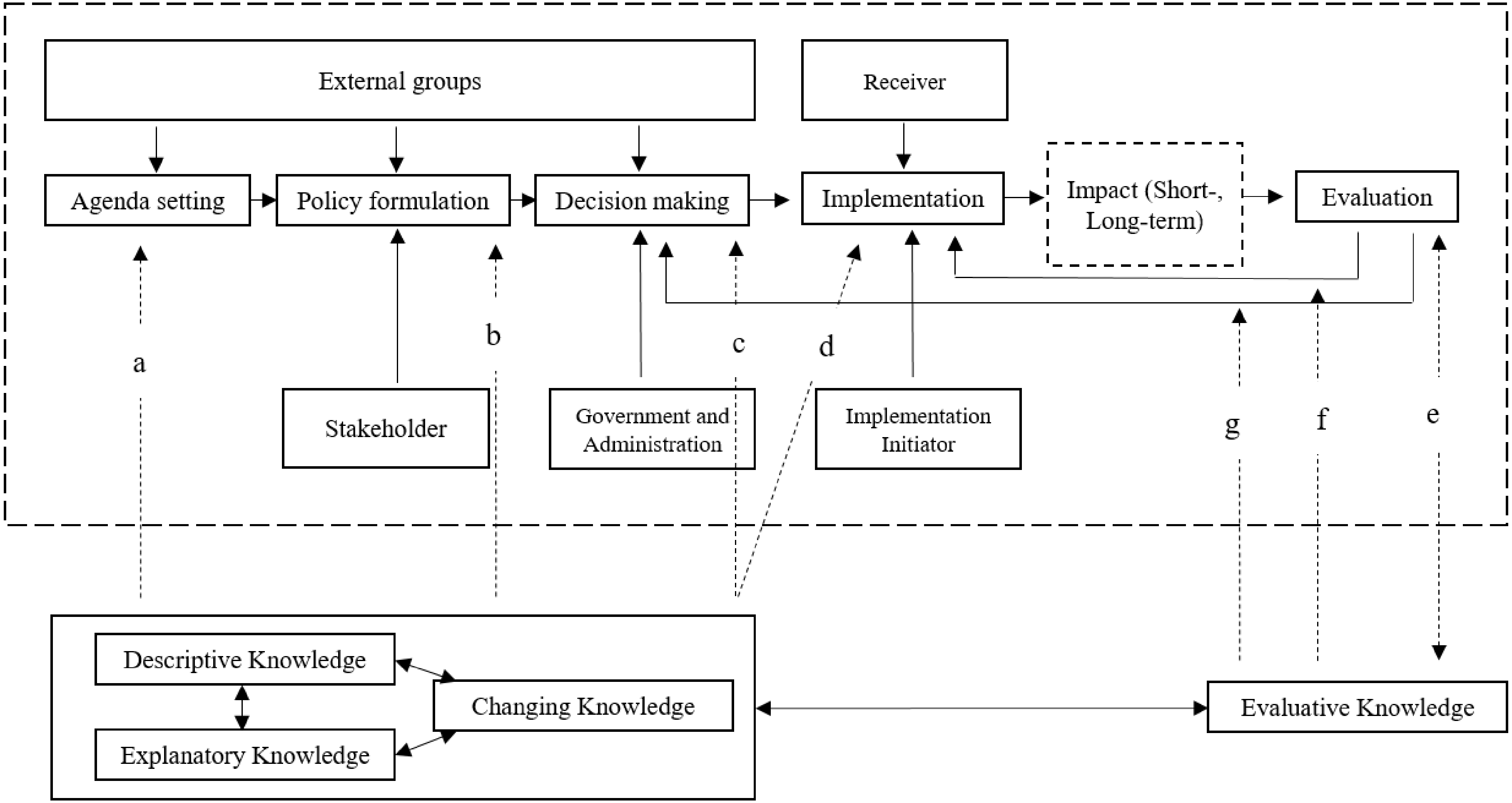

Figure 7. The alliance model of the policy process and knowledge of educational effectiveness research. For the sake of clarity, potentially mediating channels (e.g., via the stakeholder or administration) are not displayed (adapted from Mayntz, 1977, Jann & Wegrich, 2007, and Bromme et al., 2014).

3.3 Taxonomies of Educational Policy Reforms

As outlined, evaluative knowledge is one central dimension of knowledge that EER can reasonably provide and that might inform different stages of the policy process. In the model (see Figure 7), evaluative knowledge is displayed somewhat farther from the three other knowledge dimensions, which are linked together more closely. This representation was chosen because evaluative knowledge is a rather broad term that not only incorporates knowledge about how to evaluate a specific educational intervention or knowledge about how to provide the best examples of characteristics related to implementation success (e.g., Schaffer et al., 1997) but can also fall back on or even produce content that is related to the other forms of knowledge displayed in Figure 7.[26]

As is obvious from this description, evaluative knowledge has a special relevance when it comes to implementing specific reforms in the system, however, evaluative knowledge strongly depends on detailed knowledge on the specific reform. Therefore, in the following, I will provide a taxonomy on educational policy reforms in order to be able to better categorize specific reforms. Before doing this, it is important to emphasize that reforms can usually be thought of as "packages of interventions" rather than individual and strongly isolated changes (e.g., McLaughlin, 1987; Young & Lewis, 2015). This is why it is challenging to distinguish between specific types of reforms, and until now, there has been no consensus that there is one specific model that best categorizes different aspects of educational reforms. Therefore, the models presented in the following can be thought of as different, simplified models, which focus on similar and different reform features and have the potential to provide a better classification of education reforms in terms of different dimensions of reforms.

First, it is important to note that a broad variety of different terms exists, all used to describe intentions to make changes in the status quo at schools: Terms such as "school reform," "educational change," "school transformation," "school development," "school improvement," "school restructuring," are just a few among many others.[27] As defining all of these constructs

[26] Imagine a simplified example of a curricular reform in which a change in the current curriculum in mathematics is implemented. In the course of the reform evaluation, teachers are asked how much they complied with the new curriculum, and the results suggest that most teachers are not really aware of the changes introduced by the new curriculum. In this case, the evaluation would provide descriptive knowledge about teacher-related compliance in implementing the changes introduced by the reform. If reforms are evaluated with strong quasi-experimental designs and methods for estimating causal effects, these evaluations might also provide some preliminary explanatory knowledge. Nomi and Raudenbush (2016), for instance, made use of a regression-discontinuity design to identify causal effects of changes in the composition of math classes, introduced by a "Double-dose Algebra" reform. This type of knowledge is especially related to explanatory knowledge.

[27] For an extensive volume related to this topic, see Rogers (2003). Rogers offers a more general, theoretical concept for explaining the diffusion of innovations. He understands diffusion as "the process in which an

would fall outside the scope of this dissertation, I will use and define the terms *educational policy reform* and *educational change* or a combination of these terms.

In line with Haddad and Demsky (1995), a policy can be understood as: “An explicit or implicit single decision or group of decisions which may set out directives for guiding future decisions, initiate or retard action, or guide implementation of previous decisions” (p. 18).

In general, what is meant by “policy reform” can be understood best by taking a closer look at the process of public policy making (e.g., Jann & Wegrich, 2007; Mayntz, 1977). Here, in a first step, problems have to be defined and recognized before they can be further addressed by means of specific policies. It becomes clear that policy reforms, in general, address specific problems (e.g., by means of new governmental regulations). There are numerous examples of such potential “problems” in the education sector such as specific tracking structures, which are assumed to explain differences in student achievement and are related to inequality (e.g., Hanushek & Woessmann, 2006) or effects of classroom size on student achievement (e.g., Angrist & Lavy, 1999). Attempts have been made to address some of these critical problems through specific policy reforms (e.g., Nomi & Allensworth, 2009).

As outlined by Brunsson (2009), it is important to note the difference between what is often called “change” and what is referred to as “reform.” Whereas an institution or organization may be the target of numerous reforms, there might be little change following these reforms. By contrast, even if there are no explicit reforms, a specific institution might still face changes (e.g., Cerna, 2013). In line with these ideas and as outlined by Cuban (1990), change does not necessarily indicate improvement, and similar reforms on the surface can lead to similar or different effects in practice, (even) if the educational context and the implementation processes vary (e.g., Stein et al., 2004). Finally, judgments of the effects of a school reform can differ according to the framework that was used to judge the change (e.g., Konstantopoulos & Hedges, 2008).

On the basis of these points, it becomes evident that reform models that describe reforms in terms of dimensions of intended change provide a promising option for categorizing educational policy reforms. Furthermore, as can be seen above, the central aim of politicians is to change specific perceived problems, and reforms are seen as a central tool for introducing change.

innovation is communicated through certain channels over time among the members of a social system” (p. 5). In turn, “diffusion is a kind of social change, defined as the process by which alternation occurs in the structure and function of a social system” (p. 6). Fullan (2016) presented another extensive approach in the framework of educational change.

Fullan (1983) presented a first very general model related to reforms. He distinguished between four dimensions, namely, (a) the change, (b) factors affecting implementation, (c) its use in practice, and (d) outcomes. This model can be understood as a less specific version of the policy process models presented by Lasswell (1956) or Sabatier (2007). However, it comes with a more detailed differentiation of the aspects that might face change. In this regard, Fullan (1983) distinguished between changes in (a) materials, (b) structures, (c) teaching approaches, and (c) beliefs. Whereas materials refer to aspects such as textbooks or other learning materials that might foster change in class, structure is rather concerned with surface aspects of teaching and learning (e.g., ability grouping). Teaching approaches refer to aspects related to the core of the lecture, namely, introducing new strategies to teach. Finally, reforms can intend to change or revise teachers' beliefs, for instance, about general questions of student learning and teaching. In more recent literature, teachers' beliefs are also seen as a central determinant for enacting policy reforms in general (e.g., Coburn, 2005).

Next, in a more general, less specific model, Cuban (1990) distinguished between first- and second-order changes:

> First-order changes in schools would include recruiting better teachers and administrators, raising salaries, allocating resources equitably, selecting better textbooks, adding (or deleting) content and coursework, scheduling people and activities more efficiently and introducing new versions of evaluation and training. First-order changes try to make what already exists more efficient and more effective, without disturbing the basic organizational features, without substantially altering the ways in which adults and children perform their roles. Second-order changes seek to alter the fundamental ways in which organizations are put together. They reflect major dissatisfactions with present arrangements. Second-order changes introduce new goals, structures and roles that transform familiar ways of doing things into new ways of solving persistent problems. (p. 73)

As is obvious from the quote above, Cuban (1990) categorized changes according to two major dimensions that are related to the size of the changes a reform intentionally introduces. First-order changes refer to aspects that do not introduce major changes in the education system but rather try to introduce changes in surface structures to make school more effective and efficient. By contrast, if the basic foundation and structure of the school system is reformed fundamentally, this is referred to as second-order changes. Although this model was quite useful in distinguishing between what Cuban (1990) described as rather short-term surface reforms compared with large-scale reforms, it is obviously strongly limited to making a distinction in reforms in only these two dimensions. Elmore (1995) argued that there are three reasons in particular for reformers to focus on structural changes, for instance, the symbolic value of structures and the ease with which the reform can impact structures from a policy or administrative level. He also pointed out that a closer look below changes in the surface structures of the education system might be promising to obtain greater knowledge about the mechanisms that the reforms introduce for teachers, teaching, and student learning.

A more extensive model was published by Conley (1994), who distinguished between a total of 12 different dimensions of restructuring, which were further subsumed into the three large dimensions of central variables, enabling variables, and supporting variables (see Figure 8). At the heart of this model are central variables that are typically affected by specific policies. These dimensions are learner outcomes, curriculum, instruction, and assessment/evaluation. Whereas learner outcomes focus on the aspects that are related to students' actual achievements, which might be the subject of the interest of a reform, curriculum reflects a central variable that reforms might want to affect (e.g., when changing the contents and level of a specific curriculum). Next, reforms can be intended to change the way teachers teach their classes, and this intention is reflected in the dimension of instruction. When reforms are intended to change the way the results of learning are quantified in terms of student achievement, this can be captured by central variables related to assessment and evaluation. Central variables, in this model, can therefore be understand as rather broad dimensions that can help to categorize rather narrowly defined objectives of reforms into the broader perspective of the objectives of a reform.

Variables that are assumed to have an impact on central variables of learning are mentioned in the second layer and referred to as enabling variables. These proximal variables, which are assumed to "bring the change," include the learning environment, technology, time, and school-community relationship. The learning environment is related to reforms that impact central variables by changing the environment of teaching and learning, for instance, by changing the student composition in classrooms or by changing or introducing different tracks in the school system.

In a broad sense, technology focuses on the way teachers make use of specific methods to teach, and students make use of specific actions to process information. School-community relationships display features of reforms that are applied to try to change the participation of parents or other external agents. Finally, the time dimension subsumes characteristics that are related to altering the number of hours students spend in school per week, per day, or per year (Conley, 1994).

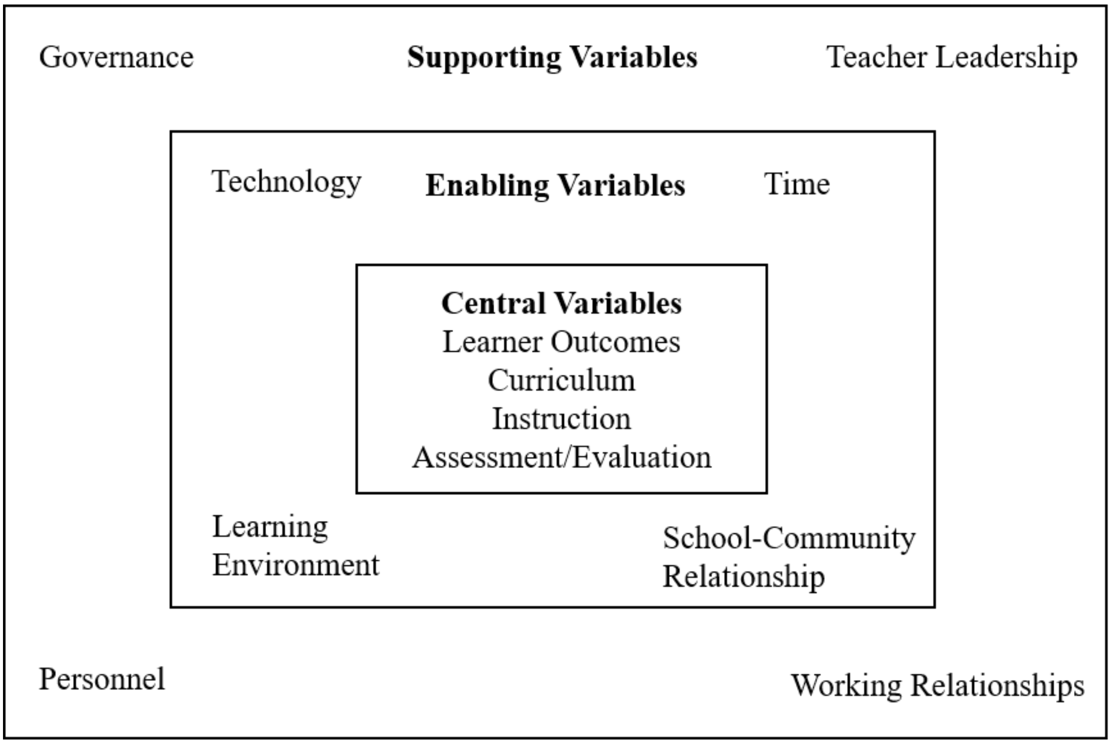

Figure 8. Dimensions of restructuring (Conley, 1994).

The top layer of the dimensions of restructuring model displays the dimension of supporting variables. It therefore captures variables related to the process of educational administration. Reforms are often based on the assumption that reforms introduced on this top layer will yield changes on the level of central variables. Governance, which is the first dimension of this model, captures characteristics related to school accountability and specific structures of decision making in school and on higher organizational levels of educational administration (e.g., district or state level accountability and governance). Teacher leadership, in turn, refers to the more or less explicit definition of the role of a teacher at a school, teacher authority, and school leadership in general. The personnel dimension refers to changes in the way personnel with different educational backgrounds (different members of the school's staff) are hired and paid for their work at the school. Finally, the aspect of working relationships captures structures of the work environment of the personnel hired at a school, as this structure might be changed due to reforms. This dimension places a special focus on the relationship, interaction, and communication of different agents, for instance, the school leader and teachers, and might be of special relevance as the school leaders are shown to have special relevance when introducing reforms in schools (e.g., Bogotch, Townsend, & Acker-Hocevar, 2010). In addition, working relationships also capture teacher collaborations and the teaching climate at a school (e.g., Conley, 1994).

In a broad sense and compared with Cuban’s (1990) model, first-order changes would be located at the layer of supporting variables, whereas second-order changes would instead be part of the enabling variables. However, if the whole process of school governance were to be restructured, for instance, as done during standards-based reform, such reforms could also be categorized as second-order changes.

Clearly, Conley’s (1994) model offers a broad variety of dimensions that offer extensions (e.g., compared with Fullan’s (1983), model) and is sophisticated enough to allow most reforms to be categorized. Furthermore, it underlines the importance of enabling variables and considers the multilevel structure of the school system in terms of the different layers.

Besides the suggestions made by these models, one can also distinguish between further reform-framing conditions. As outlined by Haddad and Demsky (1995), one can, for instance, distinguish between the scope of a specific policy in terms of complexity (low vs. high), decision environment (precise vs. imprecise), number of alternatives (low vs. high), and decision criteria (narrow vs. broad). These dimensions, in turn, are related to the overall scope of a policy in terms of an issue-specific policy, a program, a multiprogram, or a large-scale policy strategy. In this model, low values on the four dimensions indicate issue-specific, short-term interventions, whereas high values display characteristics of large-scale policy strategies.[28]

Recent models based on empirical data have linked specific “policy levers” to specific policy options. For instance, the Education Policy Outlook of the OECD (2015) distinguishes between six different types of policy levers, which are (a) equity and quality, (b) preparing students for the future, (c) school improvement, (d) evaluation and assessment, (e) governance, and (f) funding. These six types of reforms are subsumed into three higher dimensions, which are (a) students: raising outcomes, (b) institutions: enhancing quality, and (c) systems: governing effectively. As the OECD model is based on real data, it also provides the typical policies that were implemented on each dimension. For instance, in the field of equity and quality, defined as: “Policies to ensure that personal or social circumstances do not hinder achieving educational potential (fairness) and that all individuals reach at least a basic minimum level of skills (inclusion)” (OECD, 2015, p. 30), one policy option would be to support low performing and disadvantaged schools and students. Furthermore, in the field of school improvement, defined as: “Policies to strengthen delivery of education in schools that can influence student achievement” (p. 30), one option among others would be to recruit and select

[28] Related to this, Fullan (2000) further distinguished between three different types of large-scale policy reforms on the basis of their size: (a) whole district reforms, (b) whole school reforms, possibly including multiple districts, and (c) state or national reform initiatives.

high quality teachers. However, although this taxonomy of reforms was based on real data, information in terms of a general effectiveness and efficiency of specific policy levers over others remains unclear. Furthermore, only 10% of all reforms reported in the policy outlook (around 450 reforms) have been reported to be part of rigorous evaluations (OECD, 2015). To sum up, both models only partially addressed explicit links between policy options and specific outcomes, for instance, results on effects of specific reform characteristics on student outcomes. Nevertheless, they provide reasonable options to further classify educational policy reforms according to their objectives and the mechanisms that are expected to improve their effectiveness.

Another aspect that is not explicitly part of the reform itself but is strongly related to it is the implementation process. Imagine a case where students showed high competencies in advanced algebra, and there was a reform that introduced a completely new curriculum in math with a stronger focus on advanced algebra. However, imagine that, due to limited support and limited teaching material, very few teachers ended up teaching according to the new standards. An evaluation of the reform might suggest that the reform did not have a positive effect on students' achievement in advanced algebra. However, in this case, the misfit would result from issues related to the process of implementing the reform rather than to the reform itself being poorly constituted.

According to Chin and Benne (1969), there are generally three different strategies that can be applied to introduce change in human systems and that should be distinguished: (a) the empirical-rational approach, (b) the power-coercive approach, and (c) the normative-reeducative approach. According to the first approach, change will be adopted by institution members if it is rationally justified in terms of an individual benefit. In this case, change can reasonably be introduced only by informing the target who is guided by rational motives and will introduce and process the change in the institution (Quinn & Sonenshein, 2008). According to the second strategy, which is related to the hierarchy in human systems, change will be introduced when a person provides instructions to a person at a lower level in this hierarchy. The person higher in the hierarchy will use his or her power to monitor the process of change implementation and will penalize wrong behavior when needed. Finally, in the normative-reeducative approach, the focus is on the individuals who will introduce change, and their behavior is viewed as guided by social interaction and norms. Therefore, this approach is used not only to try to introduce change by informing targets about the rational benefits of the reform but also by influencing targets values, habits, and normative beliefs (Quinn & Sonenshein, 2008).

Richardson and Placier (2001) attempted to transfer the rather broad model for developing organizational theory into the context of changing schools. According to the authors, in this context, the empirical-rational approach has been shown to be especially prominent. It is based on the idea of a process model of research, which is conducted by researchers or academics and is delivered to the teacher, who will use research for practice. However, they also describe a shift toward the normative-reeducative approach, where the target individuals introduce change by reflecting on beliefs and recent practices. Related to this, Gräsel (2010) identified different strategies for the transfer of innovations in the education sector. Although not explicitly linked to research in the field of policy or policy administration and therefore based on slightly different constructs such as “innovation” and “transfer” (e.g., Rogers, 2003), it has many links to what is understood as “reform” and “implementation” in this dissertation. In her article, Gräsel (2010) identified four different strategies for transferring innovations, which are: (a) top-down strategies, (b) evidence-based strategies, (c) participative strategies of transfer-development research, and (d) transfer using design-based research. When top-down strategies are used, change can be achieved by providing input (e.g., in terms of new regulations from the educational administration) to schools, which are expected to implement the change. Furthermore, if this input that is provided to schools is based on evidence, this would display some sort of evidence-based strategy. If innovations explicitly consider the ideas of practitioners during the process of development, this is oftentimes referred to as a participative strategy, also called bottom-up theory. Finally, Gräsel (2010) distinguished another type of strategy, which is called design-based research and which is oriented more strongly toward the symbiotic, formative development and extensive exchange of both research theories and practical problems.

Most interesting, Gräsel (2010) also described the various aspects that impact the successful transfer of innovations: (a) characteristics of the innovation, (b) characteristics of the teachers, (c) characteristics of the school, and (d) characteristics of the environment and support. Characteristics of the innovation are, for instance, related to the perception of the reform among teachers, who should generally see the advantages of an innovation in order to implement it successfully. Furthermore, the innovation has to be compatible, and therefore it has to fit with existing values and structures. Next, if the complexity of an innovation is low, the likelihood that it will be successfully implemented increases (e.g., Rogers, 2003). These aspects are somewhat related to the normative-reeducative approach, which underlines not only the rational aspects but also the values of the targets in order to introduce change (Quinn & Sonenshein, 2008).

Related to the first dimension are characteristics of teachers who are often seen as the "ultimate enactors of any change effort" (Porter et al., 2015, p. 5). Based on this, if innovations are to be introduced in schools, this strongly depends on how teachers perceive the innovation and if they are willing to implement it, and this might include additional effort. This assumption is also in line with other previous research (e.g., Coburn, 2005). Furthermore, on the basis of Hall and Hord (2000), Gräsel (2010) assumed that implementing change is a process rather than a single event, and there might be variability among teachers who run through this process differently. This aspect is also important in the face of research that has suggested that the classroom level (e.g., effective teaching) greatly influences student achievement (Campbell et al., 2003; Darling-Hammond, 2000; Muijs et al., 2014; Scheerens & Bosker, 1997).

The third dimension is related to what is called "characteristics of the individual school," meaning a school's leadership and cooperation among teachers, both of which are essential aspects for a successful transfer of innovations. As also outlined by Pont, Nusche, Moorman, and Hopkins (2008), the third dimension highlights the assumption that if leaders do not identify with the purposes of the policy reform, they will most likely not be engaged in implementing it adequately. This is also important for the process of public policy making outlined above because disregarding central stakeholders (e.g., school leaders) when developing a reform can result in a considerable misfit between the reform and the context and can reduce acceptance (e.g., McDermott, Fitzgerald, & Buchanan, 2013). Finally, the characteristics of the environment and support are important for the successful implementation of innovations. In this regard, Gräsel (2010) emphasized in particular the importance of the stability of personnel and support for enacting future innovations. Furthermore, additional teacher training and the building of school networks have been shown to have a positive impact on the implementation of the innovation (e.g., Berkemeyer, Manitius, Müthing, & Bos, 2009). The results outlined by Gräsel (2010) are in line with previous research by Schaffer et al. (1997) who identified several comparable factors of reform failure. Furthermore, Gräsel's (2010) results are in line with a perspective on reforms in accordance with Rogers' (2003) theory of diffusion of innovations.

To sum up, although an extensive amount of research exists in the field of education reforms, it is rather difficult to provide one universal model that allows for a final separation of different reforms into different types. What seems to be promising, however, are characterizations of reforms based on specific dimensions and mechanisms (e.g., Conley, 1994) and in terms of the scope, the content, and the implementation process of reforms. Furthermore, it is evident that in judging the effects of reforms, not only is the framework of the judgment important but also information regarding the processes of implementation and the impact of the reform.

3.4 The Interplay between Educational Policy Reforms and Student Outcomes

In this chapter, I will outline in more detail the importance of integrating central dimensions of educational policy reforms into theoretical models of educational effectiveness. In doing this, I will describe some useful theories and models and will further integrate potential channels of policy reforms into these models.

As stated in the model of the policy process, reform-related action on the administrative level is usually initiated by the articulation of a specific problem (Jann & Wegrich, 2007; Mayntz, 1977). In the education sector, it was shown that such problems, for instance, related to disparities in student achievement (e.g., depending on gender or socioeconomic status), are oftentimes targeted by policy reforms that are expected to address student achievement (e.g., Conley, 1994; OECD, 2015). Even if reforms are located on the upper level of the education system (e.g., the federal state level), student achievement is suggested to be a major variable for judging the effects of school reforms (Konstantopoulos & Hedges, 2008).

Among others, there are two rationales in particular that support the importance of achievement measures when judging and analyzing school reform effects:

First, achievement measures have been shown to be a useful retrospective variable, as they capture various individual characteristics, determinants, preconditions, and processes at school, for instance, aspects such as students' socioeconomic status and motivation or aspects related to learning and teaching (e.g., Eccles, 1983; Helmke, 2006; Wigfield & Eccles, 1992). Investigations of student achievement (e.g., in terms of standardized student achievement) therefore contain the promising option to judge some sort of "overall effectiveness" of a policy reform (Konstantopoulos & Hedges, 2008; Wortman, 1983) by quantifying specific change on the student level. However, as already evident at this point, achievement itself provides a measure of descriptive knowledge (Bromme et al., 2014; Goldstein, 2014) and provides little information about the mechanisms that influence it.

Second, achievement measures are also useful for prospective matters, as they are useful for predicting a variety of additional individual-level outcomes later in life, for instance, post-school choices or socioeconomic success (e.g., Parker et al., 2012; Strenze, 2007). Furthermore, economic research underscores the importance of student competencies and achievement for predicting economic growth and therefore suggests that student achievement is also an important variable from an aggregated national perspective (e.g., Hanushek & Woessmann, 2008, 2010). It is interesting that whereas previous research has shown that there is a statistically significant positive association between years of schooling and economic growth, this association becomes statistically nonsignificant when quality of education, in terms of achievement test scores, is included in the model. On the basis of this, Hanushek and Woessmann (2012) conducted a variety of simulation studies in which they modeled changes in the average level of student achievement of 0.5 standard deviations, introduced by a hypothetical reform that would take 20 or 30 years to fully lead to such a change in student achievement. The results confirmed the theoretically proposed idea that after 50 years, increases in GDP of more than 10% would occur.[29]

The two outlined perspectives therefore underscore the idea that achievement has been shown to be an important comprehensive measure that results from different precursors and determinants and is also a central predictor of a broad variety of individual and aggregated outcomes later in life. However, this might not come as a surprise because, as previously outlined, models of educational effectiveness have a long tradition of describing achievement as a central outcome variable that strongly depends on several other individual determinants and processes that occur in school (e.g., Baumert et al., 2001; Creemers, 1994; Scheerens, 1990).

Another model that has been shown to be very useful in this regard and that I did not introduce earlier in the dissertation is the supply-use model (Helmke, 2006). The model (see Figure 9 for an adapted version) is based on the assumption that lectures are an offer (supply) that can be used (use) by the students (or not), and this decision, in turn, results in a specific achievement outcome. The model therefore describes lectures in terms of a potential option that results in a desired learning outcome only if the student decides to actively and adequately engage in class. Furthermore, there are a variety of mediating factors that fall between the supply of the lecture and students' learning outcomes (e.g., an individual student's motivation and perception of the lecture). Finally, the model includes specific variables that frame the supply-use process, such as the school and class climate and individual students' preconditions (e.g., learning strategies or intelligence). [30]

[29] From the more general perspective of German school theory (Fend, 2009), these aspects are also strongly related to the reproduction and quality objective of formal education (see also Chapter 1).

[30] Please note that Helmke's (2006) model is comparable to traditional models of school effectiveness by Scheerens (1990) or Creemers and Reezigt (1996) in many regards. However, it was not presented very prominently to an English-speaking audience (for exceptions, see Brühwiler and Blatchford, (2011) or Seidel, (2015). The supply-use model is generally more exhaustive in identifying specific variables and ordering the mutual processes that influence achievement, and therefore, it is more useful for displaying potential channels of educational policy reform at this point.

In order to expand this model and offer a perspective on how educational policy reforms might theoretically impact student achievement and related variables, several grey arrows have been added to the model to indicate potential channels of education reforms (see Figure 9). Although other potential channels might be reasonable, for the sake of clarity, I explicitly display three channels on which I will focus in the following. Before going into detail regarding the three specific channels, I will mention two general observations.

What first becomes evident when introducing potential channels of educational policy reforms into the supply-use model is the large complexity of the model itself, reflecting the large complexity of educational effectiveness. Even though not all potential interfaces and not all relevant variables are displayed (especially additional variables on the school level or further contextual variables), the model is already very complex in nature, and this complexity even increases when theoretical channels of reform effects are introduced. At the same time, the model still provides a simplification of the determinants and consequences of lectures and depends on many different assumptions (e.g., that the assumed order of the process holds or that major variables were not ignored).

Next, the supply-use model also underscores the large number of assumptions needed to be taken into consideration when introducing a hypothetical policy reform to increase student achievement. As suggested, what especially matters for affecting student achievement is the lecture itself and the related processes that follow (Helmke, 2006). However, whereas the lecture is very closely linked to student achievement, it is relatively far from what educational policy and educational administration actually have a direct effect on and can therefore reasonably and directly control. From this perspective, promising factors for educational policy and administration need to exhibit at least two specific characteristics. They should be (a) manipulable by educational governance (e.g., due to legal amendments) and (b) closely related to or even display important determinants of student achievement themselves. However, even if policy can identify such factors (e.g., allocated time), in most cases, the factors will be only remotely related to achievement (e.g., time on task was shown to be more important than allocated time for student achievement; (e.g., Hendriks, Luyten, Scheerens, & Sleegers, 2014), and therefore, these factors will still depend on very strong assumptions, for instance, that they will diffuse in a certain manner through the separate instances (see Figure 9) to finally lead to the desired impact on student achievement. What this culminates in is the strong dependence of educational policy reforms on intermediate factors such as how teachers adopt, judge, and implement the reform in the lecture or how individual student characteristics interact with the changes that are introduced. This underscores both the importance and the potential benefit of theoretically exploring potential channels of educational policy reforms before introducing them by means of models and theories from educational research and the need for rigorous evaluations to test these hypotheses in practice. The theories, models, and research also reported in this dissertation might contribute to this objective.

Three potential channels are outlined as examples in Figure 9: (a) effects of policy reforms on contextual components, (b) effects of policy reforms on allocated time, and (c) effects of policy reforms on the teacher. In these examples, all reforms are believed to follow the objective of influencing student achievement. Integrating reforms into specific models and theories related to educational effectiveness might be especially useful for deriving specific hypothesis in terms of competing explanatory knowledge (e.g., Bromme et al., 2014), which can be explicitly tested in research studies. Although these models are simplified, they provide a good starting point from which to reflect on potential effects of educational policy reforms.

As outlined above, many policy reforms follow the principle of influencing supporting or enabling variables in order to introduce change in central variables (e.g., Conley, 1994; OECD, 2015). This pattern is also displayed in Figure 9. As can be seen, the first potential channel (a) follows the idea of reforming contextual variables in order to change aspects of schooling beyond this surface. In many cases, contextual variables might be targeted by reforms because they are oftentimes related to changes in structures that are comparably easy to affect through policy (e.g., regulations regarding age thresholds for elementary school enrollment or grade-related admission restrictions; e.g., Elmore, 1995).

For the sake of parsimony, I will focus on effects of a reform of catchment areas here. The discussion of schools' catchment areas has a long tradition, especially in the United States, where it is also strongly related to discussions and research on free school choice (e.g., Peterson, Howell, Wolf, & Campbell, 2003). According to Ravitch (2011), this topic had its starting point early in 1950, with discussions related to school segregation and school voucher programs, and reached its peak in recent decades with the development of a variety of different school types (voucher schools, charter schools, etc.). Especially Milton Friedman's piece about "The Role of Government in Education" (Friedman, 1955) promoted school voucher programs so that students would be truly able to freely choose schools. Central to the idea of free choice is the assumption that it has a positive effect on students' performance, and there is evidence that voucher programs do lead to such effects (e.g., Shakeel, Anderson, & Wolf, 2016). To explore a potential channel, I will focus on the prominent PACES program, which was introduced by the Columbian government in the early 1990s (Angrist, Bettinger, Bloom, King, & Kremer, 2002).

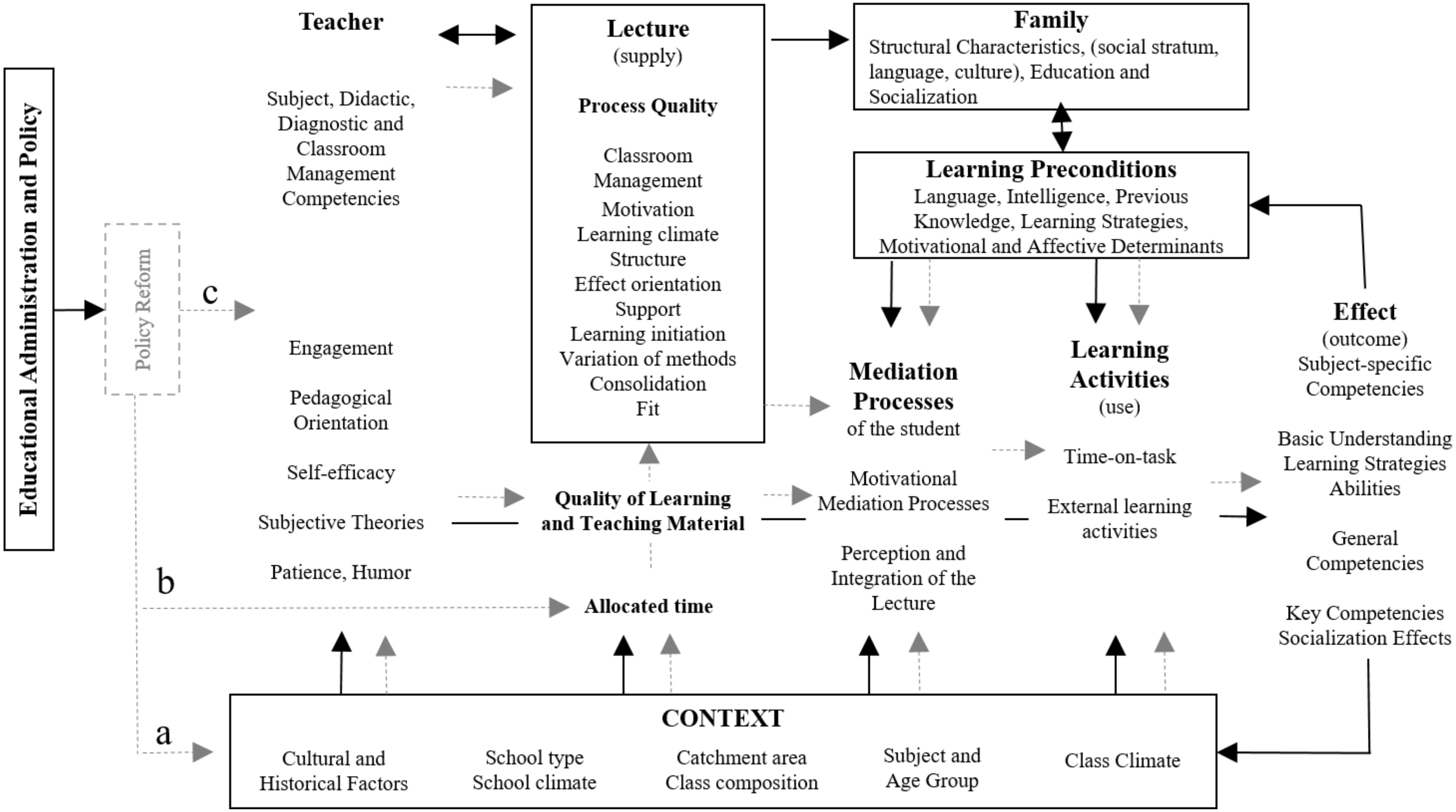
Educational Administration and Policy
Policy Reform
a
b
c
Teacher
Subject, Didactic, Diagnostic and Classroom Management Competencies
Engagement
Pedagogical Orientation
Self-efficacy
Subjective Theories
Patience, Humor
Lecture
(supply)
Process Quality
Classroom Management
Motivation
Learning climate
Structure
Effect orientation
Support
Learning initiation
Variation of methods
Consolidation
Fit
Quality of Learning and Teaching Material
Allocated time
Family
Structural Characteristics, (social stratum, language, culture), Education and Socialization
Learning Preconditions
Language, Intelligence, Previous Knowledge, Learning Strategies, Motivational and Affective Determinants
Mediation Processes
of the student
Motivational Mediation Processes
Perception and Integration of the Lecture
Learning Activities
(use)
Time-on-task
External learning activities
Effect
(outcome)
Subject-specific Competencies
Basic Understanding Learning Strategies Abilities
General Competencies
Key Competencies Socialization Effects
CONTEXT
Cultural and Historical Factors
School type School climate
Catchment area Class composition
Subject and Age Group
Class Climate

Figure 9. Adapted supply-use model (Helmke, 2006; translated), including three potential channels of educational policy reforms. Note that for the sake of clarity, all possible channels are not displayed in this figure. For the same reason, no recursive relationships are displayed for the potential reform channels.

The voucher program introduced here was expected to increase secondary school enrollment rates and therefore targeted low-income families living in low socioeconomic status neighborhoods in a voucher lottery. Over the course of the program, the voucher's value of $190, which was first determined by averaging the tuition of low to average cost private schools, was reduced due to inflation. Therefore, private school tuition had to increasingly be supplemented by private funds. In an in-depth analysis of the reform effect, Angrist et al. (2002) found an increase in achievement of 0.20 standard deviations for students who had received a school voucher. The authors attributed the results to three causal channels, which were (a) increased participation in (assumingly better) private schools for lottery winners, (b) a trend toward attending even more expensive private schools if a previous intention to attend a private school existed, and (c) an increase in lottery winners' effort and motivation to perform well in school because failing a grade would exclude students from the program.

Integrating these findings into Figure 9 would address at least two aspects. First, private schools in Colombia might have provided a better school environment in the early 1990s, and this might have be partly related to better teaching and learning. As outlined in the supply-use model, teaching quality seems to be especially related to learning outcomes, and private schools might more rigorously select teachers in this regard. However, additional aspects might also contribute to private schools as better learning environments such as the class composition of students with a comparably higher socioeconomic status or school leadership. As the PACES program did not select entire low SES student groups but rather randomly selected the students, who were then placed into classes with higher SES students, this might have led to positive effects in terms of a more fruitful learning environment. Furthermore, students' external learning activities might have changed as peers had a higher SES on average. Finally, and this might be especially related to individual student characteristics, the reform might have changed individual students' motivation in some regard. As students who did not maintain a satisfactory performance in school did not receive a voucher for the next school period, this might have especially triggered students' motivational mediation process during the lecture as a precursor of later student outcomes (e.g., utility value). Furthermore, as the value of attending private school was underscored by receiving money for funding, and oftentimes, additional private funding was used to finance private school, the students might have had additional motivational incentives to perform appropriately (Angrist et al., 2002).[31] This first example of a voucher

[31] It is important to note, however, that results from other studies have suggested that introducing voucher programs can also increase segregation among students (e.g., Brunner, Imazeki, & Ross, 2010).

reform therefore showed which potential channels a reform of contextual variables might take in order to affect student achievement.

Another important variable that is amenable to policy interventions is the time allocated in school or class (e.g., Scheerens, 2014b), which is displayed as (b) the potential channel in Figure 9. The relevance of time has already been emphasized in previous publications, for instance, in the Carroll model (Carroll, 1963, 1989). In his model, Carroll distinguished between five classes of variables that could explain variation in students' learning. These variables are (a) aptitude, (b) opportunity to learn, (c) perseverance, (d) quality of instruction, and (e) ability to understand instruction. In this case, opportunity to learn is especially related to time. Usually, within-school time is further distinguished into *allocated time*, defined as time allocated to a specific subject by the schedule; *instructional time*, defined as the net measure of being exposed to teaching, excluding time for organizational issues; and *time-on-task*, defined as the time a student is engaged in learning tasks (Berliner, 1990; Scheerens & Hendriks, 2014).[32] What policy can reasonably affect and control is allocated time, whereas instructional time and especially time-on-task would have to be assumed to also profit from changes in allocated time. For the two reforms in the focus of the dissertation, however, it seems very reasonable to assume that increases in allocated time strongly resemble increases in instructional time and therefore result in increased time-on-task. This relates back to the fact that the reforms were legally binding, and per-week increases were rather small and should therefore not have dramatically changed the quality of teaching or learning in this additional time. According to a meta-analysis by Hendriks et al. (2014), effects of increasing time at school tend to be statistically significant but rather weak. Furthermore, in line with previous findings by Lavy (2015), recent research by Cattaneo, Oggenfuss, and Wolter (2017) using data from LSAs suggested that instructional time has a positive effect on student achievement.[33] The integration of this line of research into the supply-use model would build on the assumption that if allocated time is increased, then time-on-task increases, which is in turn very closely related to students' achievement outcomes (e.g., Carroll, 1989). However, it has been noted that this relation between time and learning is rather nonlinear (e.g., Levin, 1986; Scheerens & Hendriks, 2014), which indicates that constantly increasing the length of a school day might result in adverse effects compared with extending the number of years spent in school (as cited in Carroll, 1989).

[32] Berliner (1990) distinguished between other more specific components of time (e.g., transition time or waiting time); however, these are not mentioned here for the sake of parsimony. Please see Berliner (1990) for further information.

[33] Please note that different definitions of time are often mixed up. Definitions of the different dimensions of time in this dissertation are based on the recently published extensive work by Scheerens and Hendriks (2014).

The third potential channel displayed in Figure 9 is related to effects of educational policy reforms on teachers. As can be seen in the model, the teacher and especially the lecture led by him or her has a central relevance for student outcomes.[34] Reforms related to teachers can either target teacher education in general or improve the education of teachers who are already employed. Among others, two arguments are currently especially prominent in the discussion on teacher education in the United States. As outlined by Wang, Odell, Klecka, Spalding, and Lin (2010), these are (a) quality of teaching is the most important factor that has an impact on student learning, and (b) teacher education can have an impact on teaching quality.

In general, according to Aebli (1961; Klieme, 2006), lectures can be described in dimensions of surface structures and deep structures. Whereas surface structures are related to aspects of the organization of the lecture or teaching methods, deep structures of the lecture have a closer link to learning and effective teaching and reflect aspects such as classroom management, cognitive activation, or constructive support (e.g., Good, Wiley, & Florez, 2009; Klieme, 2006). Based on this distinction, a variety of reforms seem to be reasonable for improving teaching quality, all of which could be directly included in teacher education and teacher training or in terms of further education on the job. As outlined by Kunter and Trautwein (2013), additional training to increase classroom management abilities could be provided, for instance, in terms of trainings to introduce rules and routines or trainings to set adequate sanctions for misbehavior. Other aspects of a reform of teacher education might introduce opportunities for training cognitive activation (e.g., gaining knowledge about how to cognitively activate students) or constructive support (e.g., giving adequate feedback). According to the supply-use model, introducing reforms to increase teaching quality could therefore improve students' mediation processes and engagement in learning activities. Therefore, student competencies would increase via this channel if teaching quality were to increase. However, compared with other structural reforms (e.g., Elmore, 1995), increasing teaching quality might be a very demanding and time consuming reform, but various research results have suggested that such reforms might be very promising because they could produce large, sustainable effects (e.g., Darling-Hammond, 2000; Hattie, 2008; OECD, 2015).

Related to this, Swanson and Stevenson (2002) analyzed effects of a standards-based reform movement on state-level policy activism and related the activism on the state level to teachers' instructional practices using NAEP data. Most interesting, they found an increase in

[34] In the German context, there has also been a strong focus on teacher cooperation as an important foundation of school development in the last decade (e.g., Fussangel & Gräsel, 2012; Gräsel, Fussangel, & Pröbstel, 2006; Steinert et al., 2006). However, the German focus on the relevance of the teacher and teaching is much older (e.g., Terhart, 1983).

instructional practices that were promoted by the standards-based reform movement on national and state levels. These findings suggest that national reform movements, at least in the United States, can impact state policies and in turn even impact instructional processes in class.

Compared with other potential reforms, it is evident that improving teacher training and teacher education provides a much deeper, more fundamental approach to school improvement because, as outlined above, their behavior is linked more closely to students' performance compared with other structures of the education system. Therefore, most other reforms that are implemented to increase student achievement will have to coercively anticipate how the intended reform program will affect teachers and teaching (e.g., Coburn, 2005) directly or indirectly in order to succeed as teachers are the "ultimate enactors of any change effort" (Porter et al., 2015, p. 5).

As can be seen from the three examples outlined above, different policy options can be reasonably integrated into the supply-use model (Helmke, 2006), and this integration can be helpful for formulating a hypothesis regarding the specific channels that might lead to increased student achievement or not. However, the integration, identification, and anticipation of potential advantages and challenges strongly depends on the accuracy of the theories and model. The model displayed above has a specific focus, and all plausible interactions of students' characteristics and perceptions are not explicitly considered in it. For this reason, it seems reasonable to consider other theories and more specific models according to the characteristics of specific reforms.

One of these promising theories for reforms of CI, also part of Studies 1, 2, and 3, is the expectancy-value theory (EVT; e.g., Eccles, 1983; Wigfield & Eccles, 2002). Although this theory does not explicitly consider the link between characteristics of the school, lecture, teaching, and achievement, it puts a specific focus on relations between students' motivation and students' performance. Both the supply-use model and the expectancy-value model are generally in line with what is suggested in Conley's (1994) model of restructuring; however, the two models have a different focus.

In the following, I will integrate typical dimensions of a CI reform into the expectancy-value model in order to derive hypothesis about potential channels of the CI reform on student outcomes (see Figure 10). As suggested in the EVT (e.g., Eccles, 1983; Eccles & Wigfield, 2002; Wigfield & Eccles, 2002), students' performance and related choices are influenced by both expectation of success and subjective task values. Expectation of success refers to what was already termed self-efficacy by Bandura (1997), and in EVT, it is assumed to be theoretically influenced by students' goals and general self-schemata such as the self-concept of one’s abilities. However, according to Eccles and Wigfield (2002), ability beliefs (e.g., academic self-concepts) and expectations for success (e.g., self-efficacy) are not empirically distinguishable. In line with this, Guo et al. (2016) summarizes that the use of self-concept in studies on EVT has become standard. The other EVT components that directly influence student performance are four different value beliefs, which reflect the student’s desire to engage in the task: intrinsic value, attainment value, utility value, and cost. Intrinsic value refers to the enjoyment derived from engaging in an activity and is closely related to what Ryan and Deci (2000) defined as intrinsic motivation, whereas attainment value defines the degree to which it is important for a person to perform the activity well. Utility value finally describes the perceived degree of usefulness of a given task, and cost defines perceived negative outcomes of engaging in the activity (Eccles & Wigfield, 2002).

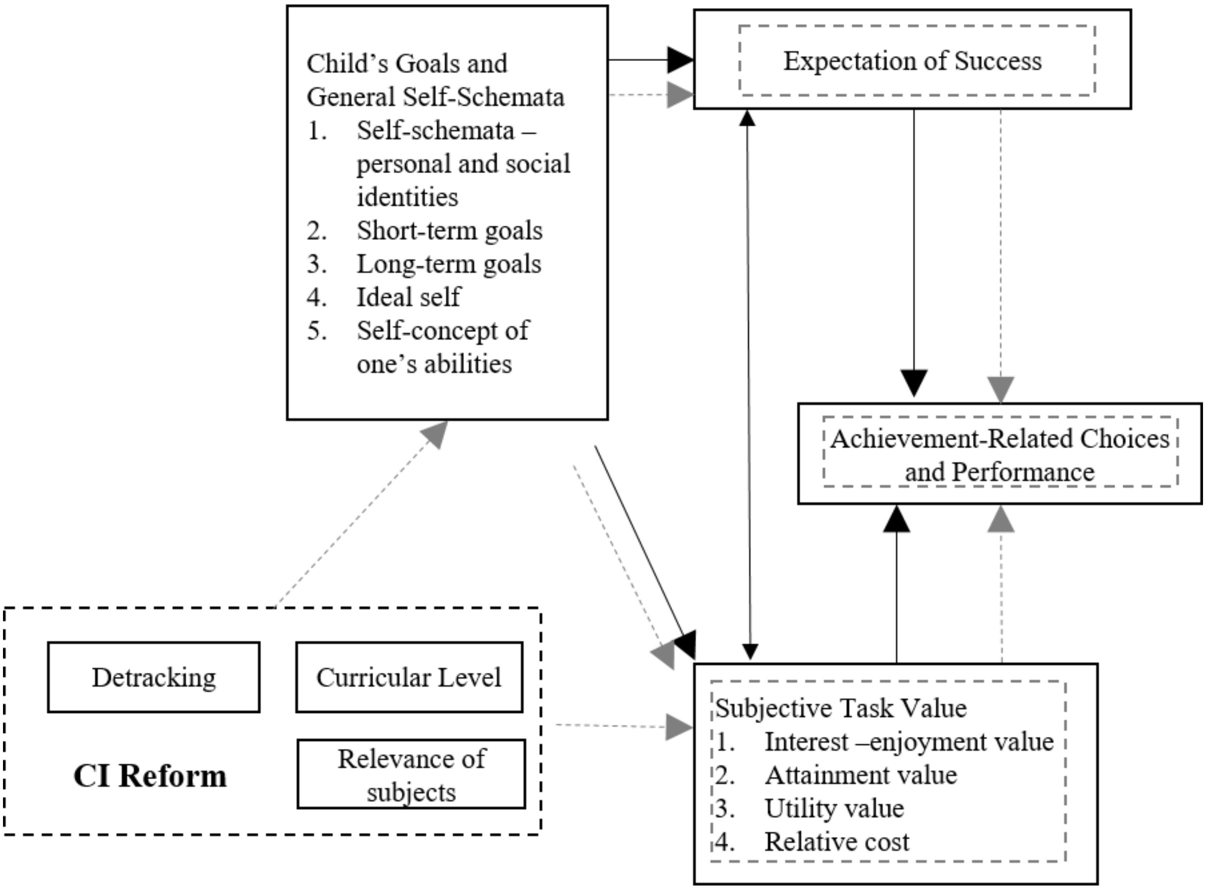

Figure 10. Adapted expectancy-value model (Eccles & Wigfield, 2002) including potential channels of reforms that directly and indirectly affect student achievement. Note that for the sake of clarity, several factors in the expectancy-value model are not displayed. See Eccles and Wigfield (2002) for the complete version of the model.

The components focused on in EVT were treated rather broadly in Helmke’s (2006) model in terms of the learning potential component. As can be seen in Figure 10, three components of the CI reform have been integrated into the model (see also Study 2). These components are (a) detracking, (b) curricular level, and (c) relevance of subjects. The first component is related to the change in reference groups introduced by the CI reform. When introducing mandatory enrollment in specific courses, all students are typically tracked together (e.g., previous nonenroller and previous enroller or previous basic and previous advanced course students). This aspect is especially related to the class composition of students. The second component (b) refers to an, on average, increase in curricular level, when, after the reform, for instance, nonenrollers and enrollers are enrolled in a specific course, or all students have to participate in one course on an advanced level. The third component (c) is related to the relevance of a subject that was made a mandatory part of students' time table. Therefore, grades count more heavily for the group of students who would traditionally not enroll in this subject or would enroll in a basic course (see Study 2).

CI reforms that introduce detracking can lead to a different performance distribution in class, with high and potentially more lower achieving students. Changing the performance distribution can result in differences in the grading of students (see Study 3), as grades are oftentimes assigned on a norm-referenced basis (e.g., Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006). Research has suggested that grades are a major source of feedback for students' academic self-concepts (e.g., Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007), and if grades change, related measures such as students' self-concepts are also likely to be influenced. This effect was often investigated in studies on the big-fish-little-pond effect (BFLPE; e.g., Marsh et al., 2008), which indicates that given similar achievement, a stronger reference group decreases the subject-specific self-concept, whereas a weaker reference group increases it (see Studies 1 and 2). If now the self-concept of one's ability in a specific subject is lower (e.g., due to a stronger reference group), EVT would suggest that this could reduce achievement-related performance. Simultaneously, prior research and theory suggest that subject-specific self-concept in math affects subsequent interest in math, whereas effects of interest on self-concept are rather small (e.g., Eccles & Wigfield, 2002; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005). Therefore, if detracking changes self-concepts, this could also result in different subjective task values (e.g., lower intrinsic values). Eccles and Wigfield (2002) argue that an increase in value in activities that students perform well should result from (a) classical conditioning (the positive effect of doing well in a specific subject) and (b) reducing values in tasks where students do not perform well for reasons linked to self-preservation (high self-esteem and efficacy).

Next, if the curricular level (b) increases, for a majority of potential nonenrollers (e.g., due to the introduction of mandatory [advanced] courses for all students), this might further contribute to decreasing the self-concepts of this potentially lower achieving group. This effect might therefore remain pronounced even if these potentially lower performing students can increase their standardized achievement. However, as can be seen, the upper secondary school reform implemented detracking and increases in the curricular level for potential basic course students simultaneously, and this is why all changes that were introduced were perfectly confounded and presumably interacted with each other. Based on this, it is possible that increasing the curricular level for basic course students alone would introduce no or slight changes, but combining detracking with an increased level might lead to stronger effects.

Finally, especially in upper secondary schools, CI reforms can also (c) introduce changes in the relevance of subjects, for instance, if the relevance of a subject for a student's final GPA changes because the course becomes mandatory and therefore counts toward the student's GPA. This might especially have an effect on specific aspects of task values, for instance, for utility values, which might increase if performance in a specific course has a higher weight for GPA and is therefore more strongly related to grade-based admission processes at college or university. Along these same lines, if performance-related self-perception is lower due to a perceived high-performing reference group, this might increase the perceived costs of engaging in learning-relevant tasks to achieve a good result and decrease intrinsic value. Furthermore, being forced to participate in a subject in which enrollment was not mandatory before and which would have been deselected before or which would have been selected on a lower level before might also result in effects on intrinsic values (e.g., enjoyment). As self-determination theory would suggest, experiencing the abolishment of free course choice (increased external control) might reduce perceived competence and autonomy and therefore decrease students' intrinsic motivation (e.g., Deci, Koestner, & Ryan, 1999; Ryan & Deci, 2000).

Of course, the links between the specific components of reforms and educational research outlined above provide only some potential channels that have to be tested in further studies and were tested to some extent in this dissertation. However, as can be seen, linking educational policy reforms to models and theories that describe determinants of educational effectiveness and in terms of student learning can be quite promising for two reasons in particular:

First, related to the traditional framework and potential channels of educational policy reforms outlined above, a direct effect of educational policy reforms on student achievement does not seem to be very likely (e.g., Conley, 1994; Elmore, 1995). Due to this, most policy reforms rely on the strong assumption that changes introduced in distal variables (e.g., related to the structure of the education system) follow specific channels and diffuse through the different layers of the system in a specific manner (e.g., Conley, 1994). Increasing the awareness of these assumptions and discussing the plausibility of them from a policy and research perspective appears mandatory. Furthermore, what this implies for research is that there is a need to focus on a broad variety of psychological factors that are proximal to achievement (e.g., self-concepts or interests) when analyzing and explaining effects of educational policy reforms. According to the suggested models, it becomes evident that extensive investigations of student achievement as a central outcome variable seem especially reasonable whenever some sort of "overall reform effect" is of interest (e.g., Konstantopoulos & Hedges, 2008). However, in terms of the different dimensions of knowledge outlined above (Bromme et al., 2014), this would rather result in descriptive knowledge. For a more coherent picture in terms of explanatory knowledge, considering variables that are precursors of achievement (e.g., self-concepts or interests) seems to be both necessary and promising and will likely lead to deeper insights into mechanisms of specific reforms, especially if based on strong research designs. Theories and models of or related to educational effectiveness can be used to help identify the most important variables and to anticipate and formulate specific hypotheses on potential effects of policy reforms.

Second, it becomes evident that policy reforms most likely introduce a large number of intended and unintended effects on a broad variety of outcome variables. From this perspective, educational policy reforms are much more of a gamble regarding their effects than a rigorous, accountable decision per se, and this holds even if policy reforms are planned and implemented with great caution in the education system (e.g., Gross, Booker, & Goldhaber, 2009). This points to the idea that reforms must be rigorously evaluated for their effects in order to shine at least some light into the black box and on the mechanisms resulting from reforms (e.g., Campbell, 1969), thus mediating the effects on the outcomes. This is important not only from a perspective of general scientific interest but also from a perspective of evidence-informed, accountable policy making.

To sum up, as shown above, the integration of central dimensions of reforms into models of or related to educational effectiveness is fruitful for anticipating potential intended and unintended effects of educational policy reforms. Nevertheless, research rigorously integrating such psychological factors for generating and testing specific hypotheses is currently very limited. Therefore, a central part of this dissertation project was dedicated to increasing such knowledge.

3.5 Research Questions

The present dissertation project investigated effects of major educational policy reforms in Germany that were enacted in the first decade of the new millennium. To do so, first, I outlined the theoretical foundations of educational governance (Chapter 2) before I presented the relation between EER and educational policy (Chapter 3). These first two chapters therefore provide the central foundations for the following studies (Chapter 4) and the knowledge that is needed to better integrate the increased emergence of reform-related political action.

The first policy reform that was the focus of the empirical studies in this dissertation was the reform of upper secondary school, which introduced CI in terms of mandatory course choice in mathematics, German, and a foreign language on an advanced course level. Before the reform, students were able to choose between advanced and basic courses in these subjects, and advanced courses were offered for 5 (Baden-Württemberg) or 6 hr (Thuringia) and basic courses for less than 4 hr (except mathematics in Thuringia) per week. After the reform, all of the core courses were taught for 4 hr per week.

Next, the G8 reform of lower secondary school was analyzed. The G8 reform basically reduced the total number of years spent in school from 9 years at high academic track schools (Gymnasium) before the reform to a total of 8 years after the reform. However, the total time spent in school was kept equal, which is why students' weekly hours in lower secondary school were increased.

As outlined above (see Chapter 3.4), a rigorous consideration of the links between the surface characteristics of the education system (e.g., tracking or allocated time) and individual psychological factors appears promising for a better understanding of effects of educational policy reforms. Based on this, a specific focus of this dissertation was placed not only on analyzing student achievement as a central outcome variable but also on investigating the effects of reforms on additional variables that are assumed to be strongly related to student achievement. Related to this, models of educational effectiveness suggest that different reforms can affect achievement via different potential channels (see Chapter 3.4). This is why theory implies that there should be effects of some reforms on variables that would be less likely to be affected by other reforms. At the same time, the amount of research on different reforms might differ, and knowledge about specific variables might be available for some reforms but not for others.

Therefore, it seemed especially promising to investigate differential effects on variables such as subject-specific achievement, self-concepts, and interests in the presence of CI reforms, whereas subject-specific achievement, stress, and subjective health seem to be of special relevance in the context of the G8 reform (e.g., Creemers, 1994; Eccles & Wigfield, 2002; Helmke, 2006; Huebener, Kuger, & Marcus, 2017; Kühn, van Ackeren, Bellenberg, Reintjes, & Im Brahm, 2013). Analyzing reforms by integrating them into established psychological theories and models that explain student achievement and student learning and investigating a broad variety of related variables will not only generate descriptive knowledge (e.g., Bromme et al., 2014) on reform effects but will also provide further insight into potential channels of the policy reforms.

According to EVT (e.g., Eccles & Wigfield, 2002), there are two components that are particularly important for students' achievement from a motivational perspective. These two aspects are the expectation of success, which was defined by Eccles and Wigfield (2002) in terms of Badura's theory of self-efficacy (e.g., Bandura, 1997) and value beliefs. Expectation of success is in turn related to students' goals and general self-schemata, for instance, students' self-concepts (ability beliefs). Self-concepts are believed to be strongly influenced by students' previously perceived performance and are built into an external frame of reference (e.g., Marsh, 1986; Marsh et al., 2007). As outlined by Eccles and Wigfield (2002), in practice, ability beliefs and expectations of success are strongly related and not empirically distinguishable, and this is why we focused on measures of self-concept as a competence-related belief, in line with recent research in this field (e.g., Guo et al., 2016).

One central question that was targeted in this dissertation was whether reforms of CI can lead to changes in students' academic self-concepts, interests, and their related achievement. Furthermore, the studies in this dissertation also took a closer look at differential long-term effects, based on potential changes in career choices in one study, as several studies have found gender differences in self-concepts (e.g., Marsh & Yeung, 1998; Watt & Eccles, 2008), course selection, and related career choices (e.g., Ma & Johnson, 2008; Nagy et al., 2008), which might reasonably be impacted by CI, as further outlined above and in the subsequent studies. In addition, I also shed light on one specific foundation of self-concepts, namely, school grades. Teacher-assigned grades are oftentimes assigned "on a curve" and therefore strongly depend on the composition of the reference group. As the composition of the reference group was changed over the course of the CI reforms, the studies in this dissertation took a closer look at whether this also had an impact on grades, related to standardized test achievement (see Chapter 3.4).

Second, the studies in this dissertation also took a closer look at another major German policy (G8 reform, i.e., the reduction in secondary schooling in high academic track schools of

1 year) and its related effects on student outcomes, which introduced changes in allocated time (e.g., Kühn et al., 2013). As outlined above, the relevance of time has already been described in previous research, for instance, in the Carroll Model (Carroll, 1963, 1989). Time in school can be distinguished into *allocated time*, defined as time allocated to a specific subject by the schedule; *instructional time*, defined as the net time students are exposed to instructions, excluding time for organizational issues; and *time-on-task*, defined as the time a student is engaged in learning tasks (Berliner, 1990; Scheerens & Hendriks, 2014). As outlined, policies oftentimes affect surface structures such as allocated time, whereas instructional time or time-on-task are usually assumed to be impacted by changes (increases or decreases) in allocated time in general. Both Hendriks et al. (2014) and Lavy (2015) found that increases in instructional time increased student achievement. Furthermore, Scheerens and Hendriks (2014) suggested, based on results of different meta-analyses, that time-on-task has a positive effect on student achievement. However, uncertainty still exists about how to exactly influence time-on-task, for instance, in terms of longer school days or years or possibly in terms of summer school, as the relation between time and performance is not linear (e.g., Scheerens, 2014a). In line with this, up to now, very few studies have investigated whether or not the caution that Levin (1986) suggested regarding longer school days might be appropriate (as cited in Carroll, 1989).

Both reforms were analyzed by considering data from the end of upper secondary school. Compared with other periods in the education system, the end of upper secondary school traditionally plays a special role for the subsequent transition process to employment or university access (Trautwein & Neumann, 2008). All four studies conducted here made use of rich, representative data sets in order to analyze effects of the upper secondary school reform in Baden-Württemberg and Thuringia as well as to investigate reform effects of the G8-reform in Baden-Württemberg. The four studies of this dissertation were perfectly suited to answer the outlined questions for three reasons:

First, all studies investigated effects of major German educational policy from the most recent decade, effects that are still discussed controversially in public. The results of this dissertation can therefore be used to inform policy and the public and enrich ongoing discussions with recent results from educational research.

Second, all reforms were investigated with a specific focus on student achievement and relevant, related psychological factors. Up to now, studies investigating effects of policy reforms on psychological factors have been rather scarce, and therefore, knowledge about effects of reforms on student outcomes is oftentimes limited to loosely described changes in achievement measures.

Finally, all studies analyzed the reforms according to theories and models of or related to educational effectiveness and therefore provide examples of how to generally link surface changes in educational policy reforms to potential mechanisms related to the class or to the individual student. This might be especially useful for anticipating and explaining specific intended and unintended effects of policy reforms by means of profound previous research. Along the same lines, this provides an important first step toward a more holistic perspective of what educational policy reforms actually change in school.

Study 1 (*Maximizing Gender Equality by Minimizing Course Choice Options? Effects of Obligatory Coursework in Math on Gender Differences in STEM*) investigated effects of the reform of upper secondary school on achievement in advanced mathematics, math self-concept, realistic and investigative vocational interests, and field of study at university. A special focus in all analyses was placed on potential differences between young women and young men on all these variables before and after the reform. The study is especially useful for increasing knowledge about potential differential effects of policy reforms on achievement and related subject-specific self-concepts and vocational interests. These potential changes were integrated into larger theoretical concepts (e.g., Eccles & Wigfield, 2002). The study was conducted on a rich representative data set from the TOSCA study (Köller et al., 2004; Trautwein et al., 2010).

Study 2 (*Putting All Students in One Basket Does not Produce Equality: Gender-Specific Effects of Curricular Intensification in Upper Secondary School*) estimated effects of the reform of upper secondary school in another German country, namely, Thuringia. Although the reform of upper secondary school was introduced somewhat later (2010/2011), the principles of the reform were very similar to the reform introduced earlier in Baden-Württemberg. Compared with the outcomes analyzed in Study 1, the second study took a closer look at a broader variety of measures such as achievement in English reading, mathematics, biology, and physics as well as students' subject-specific self-concepts and interests. Using data from the Additional Study Thuringia of the National Educational Panel Study (NEPS; Blossfeld et al., 2011), Study 2 further investigated both main effects and potential gender disparities before and after the reform.

Study 3 (*Comparing Apples and Oranges: Reforms can Change the Meaning of Students' Grades!*) took a closer look at the meaning of student grades at the end of upper secondary school before and after the reform of upper secondary school. Student grades are an important variable for college or university access and employment. However, research has shown that teacher-assigned grades and standardized student achievement are less than perfectly related to each other. As grades are oftentimes assigned by making use of norm-references, and the CI reform introduced changes in students' reference groups, Study 3 focused on the question of whether students' standardized achievement differed before and after the reform, given similar grades. Compared with Studies 1 and 2, Study 3 focused on a central precursor variable of students' self-concept in mathematics and English and therefore further increased knowledge about the potential mechanisms found in Studies 1 and 2.

Finally, Study 4 (*The G8 Reform in Baden-Württemberg: Competencies, Well-Being and Leisure Time Before and After the Reform*) is one of the very first studies to investigate effects of the G8 reform at the end of upper secondary school. In contrast to the reform of upper secondary school, the G8 reform did not change the class composition of students in highly demanding upper secondary schools but rather led to increases in allocated time in lower secondary schools in order to reduce the total number of years spent in school by 1 year. The last study therefore focused on potential changes in student achievement in mathematics, English reading, biology, and physics before and after the reform, but it also took a closer look at changes in variables related to students' well-being (stress and health) and leisure time use.

In the General Discussion, I integrate the results of this dissertation into the broader framework of educational policy reform and policy evaluation. Research that satisfies both claims of scientific standards and claims of practical relevance for the policy process is, although strongly needed, still not common in the field of educational science (Thiel, 2014).

4 The Empirical Studies

4.1 Study 1

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

Abstract

Math achievement, math self-concept, and vocational interests are critical predictors of STEM careers and are closely linked to high school coursework. Young women are less likely to choose advanced math courses in high school, and encouraging young women to enroll in advanced math courses may therefore bring more women into STEM careers. We looked at a German statewide educational reform that required all students to take advanced math courses and examined differential effects of the reform on young men and women's math achievement, math self-concept, vocational interests, and field of study at university. We compared data from 4,730 students before the reform and 4,715 students after the reform. We specified multiple regression models and tested main effects of gender and cohort as well as the effect of the Cohort × Gender interaction on all outcomes. All outcomes showed clear gender differences favoring young men before the reform. However, the reform was associated with different effects for young men and women: Whereas gender differences in math achievement were smaller after the reform, differences between young men and women in math self-concept and realistic and investigative vocational interests were larger after the reform than before. Gender differences in the field of study at university did not differ between before and after the reform. Results suggest that reducing course choice options in high school does not automatically increase gender equality in STEM fields.

Keywords: gender differences, school reform, math achievement, math self-concept, vocational interests

Maximizing Gender Equality by Minimizing Course Choice Options? Effects of Obligatory Coursework in Math on Gender Differences in STEM

Women are underrepresented in mathematically intensive STEM (science, technology, engineering, and mathematics) domains (Ceci, Williams, & Barnett, 2009; Schoon & Eccles, 2014). Gender disparities in STEM fields are crucial for the larger economy because the presence of more women would diversify the workforce and might add to a more competitive work environment with an increased number of qualified employees in this area (e.g., NSF, 2013; OECD, 2010). In addition, women's underrepresentation also matters to gender inequity in income because STEM fields provide high-status career options (e.g., Sells, 1980; Watt, Eccles, & Durik, 2006). Advanced high school coursework in math is a key predictor of STEM career choices (Ma & Johnson, 2008), and young women are less likely to choose advanced math courses than young men (Nagy et al., 2008; Updegraff, Eccles, Barber, & O'Brien, 1996). Thus, it is important to ask whether the challenge of recruiting more women into STEM careers may be addressed by mandatory enrollment in advanced math courses in high school (e.g., by changing course assignment procedures; Ma & Johnson, 2008; Sells, 1980). However, there is limited real-world data on the effectiveness of such reforms.

In the present study, we re-analyzed representative data from a large school achievement study on the effects of a major reform of upper secondary education in a large state in Germany. More specifically, the reform required all students to take an advanced math course, which successfully eliminated a prior imbalance between young men and women in these advanced courses. We studied the effects of this school reform on gender differences in math achievement, math self-concept, and interests in realistic and investigative areas because such outcomes are critical in terms of later educational choices. Furthermore, we investigated effects on students' actual field of study at university 2 years after they completed high school.

Predictors of Gendered Career Choices in STEM

Academic Achievement and STEM Career Choices

In explaining STEM career choices for young men and women, research on educational choices has traditionally focused on the role of math achievement on career interests (e.g., Parker et al., 2012; Sells, 1980). Such work has consistently shown that math achievement is a key predictor of both high school subject choices and later career choices, particularly with respect to mathematically intensive STEM careers (Parker et al., 2012; Sells, 1980). For instance, there is evidence that high school math achievement predicts career aspirations in

STEM during high school (e.g., Ma & Johnson, 2008), field of study at university (e.g., Parker et al., 2012), and university retention (Alarcon & Edwards, 2013).

The relation between academic achievement and career choice is often explained by employing rational choice models (Gottfredson, 1986; Lubinski & Benbow, 2006). First, individuals prefer careers that provide activities they expect to be good at. Second, individuals who have the required competencies gain access to the professional field, for instance, due to admission restrictions for college majors. Third, individuals tend to leave professions if their competencies are insufficient for the specific profession. Thus, young people with high math achievement have a tendency to pursue mathematically intensive STEM careers such as physics, engineering, or informatics (Humphreys & Yao, 2002; Parker et al., 2012).

Self-Concept and STEM Career Choices

Above and beyond the effects of achievement, young people's career choices are also critically linked to their academic self-concept in high school (Schoon & Eccles, 2014; Watt & Eccles, 2008). Academic self-concept is defined as a person's self-evaluation of his or her own general ability in a specific domain, such as doing well in math (Bong & Skaalvik, 2003; Marsh, 1986). In developing a domain-specific self-concept, students refer to their own achievement in a domain but also compare their own ability with their interpretation of peers' achievements in the same domain (e.g., Marsh, 1986; Marsh et al., 2015).

In fact, self-concept has been shown to be related to future-oriented motivation and aspirations such as career choices (e.g., Schoon & Eccles, 2014; Watt & Eccles, 2008); math self-concept has been identified as positively related to various educational outcomes in the STEM area, such as high school students' educational aspirations within the STEM fields (Jansen, Scherer, & Schroeders, 2015; Schoon & Eccles, 2014) and choice and retention of mathematically intensive STEM university subjects (Perez, Cromley, & Kaplan, 2014; Schoon & Eccles, 2014) for both men and women.

It is important to mention that self-concept does not measure the same thing as self-efficacy, although they are closely related (e.g., Bong & Skaalvik, 2003). Furthermore, self-concept predicts educational biographies and trajectories, whereas self-efficacy is used for predicting success in a specific task (Jansen, Scherer, & Schroeders, 2015).

Vocational Interests and STEM Career Choices

Next to math achievement and self-concept, vocational interests are very important in predicting STEM career choices. The role of interest for achievement-related outcomes is well-established (Schoon & Eccles, 2014; Su, Rounds, & Armstrong, 2009). Whereas educational psychology has traditionally focused on children's and adolescents' interest in learning and achievement in the school context (Krapp, 1999; Wigfield & Cambria, 2010), research and theories in vocational psychology, such as Holland's theory of vocational interests (Holland, 1959, 1997), have been highly effective at addressing young people's postschool career choices with interests describing activities in fields of professions or university majors (Rounds & Su, 2014; Su & Rounds, 2015). Vocational interests are central predictors of vocational choices such as the selection of a college major or profession (Humphreys & Yao, 2002; Pässler, Beinicke, & Hell, 2014) and are also crucial for job performance and turnover (Nye, Su, Rounds, & Drasgow, 2012) as well as income (Huang & Pearce, 2013).

Holland (1966) defined vocational interests as "the expression of personality in work, hobbies, recreational activities, and preferences" (p. 3) and expected that they would directly influence goal-oriented behaviors. He posited that individuals should strive for educational and occupational environments that are in line with their interests, and there is a large body of research that supports this proposition (e.g., Humphreys & Yao, 2002; Strong, 1943). Vocational interests are therefore defined as trait-like preferences for activities, and these preferences are captured on a very general level (Holland, 1997; Rounds & Su, 2014). In this regard, vocational interests differ from the term *interest* in educational psychology. Interest in educational psychology is usually defined as a motivational variable that "refers to the psychological state of engaging or the predisposition to reengage" (Hidi & Renninger, 2006, p. 112). Contrary to conceptualizations of interest in educational psychology, which usually focus on domain-specific interest in single (school) subjects (e.g., Hidi & Ainley, 2002), vocational interests emphasize broad sets of activities and experiences that go with different kinds of professions. Thereby, Holland's model represents six interest domains, which describe activities that are related to different careers: *realistic*, *investigative*, *artistic*, *social*, *enterprising*, and *conventional*. In our study, we focused on the realistic and investigative dimensions because they have been shown to be related to mathematically intensive STEM fields (Ackerman & Heggestad, 1997;Su et al., 2009). People with high realistic interests tend to like working with things and prefer activities that involve the manipulation of objects, tools, and machines. People with high investigative interests are likely to be interested in understanding how physical and biological phenomena function and tend to prefer activities that include analyzing and problem solving on a more abstract level (Holland, 1997). Consequently, young people with realistic and investigative interests are likely to choose mathematically intensive STEM careers such as physics, engineering, or informatics (Su & Rounds, 2015; Su, Rounds, & Armstrong, 2009).

Gender Differences in Math Achievement, Math Self-Concept, and Realistic and Investigative Interests

Gender differences in math achievement have often been used to explain gendered career choices in the STEM domains (e.g., Hyde, Fennema, Ryan, Frost, & Hopp, 1990; Reilly, Neumann, & Andrews, 2015). Historically, there has been a pattern of young men outperforming young women in math achievement (e.g., Hyde, Fennema, & Lamon, 1990). However, more recent research has provided mixed evidence: Some studies have suggested no or only slight differences in math achievement between young women and men in high school (e.g., Hyde, Lindberg, Linn, Ellis, & Williams, 2008; Voyer & Voyer, 2014), whereas others have indicated that such differences still exist and that the magnitude of the differences between young men and women varies between countries and according to the educational requirements of the system (e.g., Else-Quest, Hyde, & Linn, 2010; Reilly et al., 2015). For German samples, previous research has consistently indicated that young men still perform better in math in high school than young women (e.g., Else-Quest et al., 2010; Nagy et al., 2008).

Regarding math self-concept, previous research has shown that—after achievement is controlled for—boys tend to report higher math self-concept than girls even in primary school, and such gender differences remain constant across high school (e.g., Marsh & Yeung, 1998; Nagy et al., 2008) .

With respect to realistic and investigative interests, previous research has consistently shown that men score higher on both interest dimensions than women (e.g., Lippa, 1998; Su et al., 2009).

Relations between Achievement, Self-Concept, and Vocational Interests

Academic achievement, the self-evaluation of academic achievement (i.e., self-concept), and interests have been found to be interrelated, which means that, in general, people are interested in and feel competent in domains they are good at. The relations between these constructs have been described in different theoretical frameworks, such as Eccles et al. (1983) expectancy-value theory and Lent, Brown, and Hackett’s (1994) social cognitive career theory. According to these theories, prior achievement influences an individual’s evaluation of his or her achievement (e.g., self-concept), as well as his or her interests in the same domain. A person’s interests are furthermore influenced by his or her perception of competence, and both self-concept and interests are believed to predict later achievement. The rationale behind these relations is that individuals who have positive previous achievement-related experiences in one domain will feel more competent and will develop interests in the same domain. Furthermore, if they feel competent and are interested, they will engage more frequently and intensely in tasks and activities related to that domain, and thereby, they will show high levels of persistence and effort. In the end, this leads to better performance in the same domain (Wigfield, Tonks, & Klauda, 2009).

There is a lot of empirical support for such relations between achievement, self-concept, and interests. With respect to the relation between achievement and self-concept, several studies have indicated that achievement and self-concept are positively correlated (e.g., Chen, Yeh, Hwang, & Lin, 2013), and bidirectional relations have been found, indicating that students' prior achievement influences their self-concept and that their self-concept influence their later achievement (for a review, see Marsh, 2007). Furthermore, there is evidence that self-concept predicts changes in interests (Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005; Wigfield et al., 1997) and that interests and achievement are also interrelated. Thereby, correlation-based research has shown positive relations between achievement and interests for various conceptualizations of interest, such as individual interest (see Schiefele, Krapp, & Wintler, 1992) or task values (see e.g., Updegraff, Eccles, Barber, & O'Brien, 1996), but also for vocational interests, where positive correlations between math achievement and realistic as well as investigative interests have been found (Ackerman et al., 1997; Rolfhus & Ackerman, 1996). Furthermore, self-concept has been found to predict later interests (e.g. Denissen, Zarrett, & Eccles, 2007; Marsh et al., 2005), and a reciprocal relation has been found between interests and achievement (e.g., Denissen et al., 2007; Jansen, Lüdtke, & Schroeders, 2016). However, prior studies have so far focused on subject-specific conceptualizations of interest, and less is known about directional relations between these constructs and realistic and investigative interests.

Effects of Course Level on Achievement, Self-Concept, and Vocational Interests

Students' achievement, self-concept, and vocational interests have been linked to their enrollment in advanced and basic courses in high school (e.g., Köller, Baumert, & Schnabel, 2001; Marsh, 2005). The effects of high school coursework on achievement, self-concept, and interests have been explained by variability in the benefits for and constraints on students taking basic and advanced courses (e.g., Köller, 2001; Marsh, 2005). In Germany, as in most school systems in developed countries, students in upper secondary school self-select into basic and advanced math courses, which differ in terms of curricular content and level as well as in class composition (Schnabel, Alfeld, Eccles, Köller, & Baumert, 2002). These differences between advanced and basic coursework have been found to lead to differential effects on students' achievement, self-concept, and interests, after students' previous performance was controlled for (e.g., Köller, et al. 2001; Trautwein, Köller, Lüdtke, & Baumert, 2005). Regarding students' academic achievement, course level and achievement have been found to be positively associated; students in advanced courses have typically shown higher achievement at the end of high school than those in basic courses, even after students' prior achievement was taken into account (e.g., Gamoran & Mare, 1989; Köller et al., 2001).

Effects of course level on self-concept and vocational interests are less clear. Regarding self-concept, positive associations have been found between a student's own achievement and his or her self-concept in the same domain, as described in the previous section (Marsh, 1986; Marsh et al., 2014). Thus, students showed higher self-concept in advanced courses than in basic courses in general (Chmielewski, Dumont, & Trautwein, 2013). However, students tend to compare their own achievement with the perceived achievement of their classmates and consequently judge their own achievement as relatively lower when they are surrounded by students with higher achievement. Therefore, students in advanced courses have shown a lower self-concept than students with comparable achievement in basic courses (Chmielewski et al., 2013; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006).

With respect to vocational interests, research has shown that students in advanced and basic courses differ in their vocational interests because their course choices are based on their vocational interests (Nagy & Husemann, 2010; Patrick, Care, & Ainley, 2011). However, it is less clear if or how course level might also predict vocational interests. First, a positive association has been identified between achievement and vocational interests as described above, on which basis one might speculate that course level in math might positively influence realistic and investigative interests (Ackerman & Heggestad, 1997; Anthoney & Armstrong, 2010). Second, initial findings have indicated effects of the average level of class achievement on students' vocational interests. Cambria, Brandt, Nagengast, and Trautwein (2016) investigated 10th graders' achievement in several domains and their vocational interests. They found that achievement in math was positively associated with realistic and investigative interests and that students with the same individual math achievement level had higher realistic and investigative interests when they were in a class with a higher mean level of achievement. To sum up, math achievement, math self-concept, and vocational interests are central predictors of mathematically intensive STEM career choices, and these predictors explain gendered career choices in these fields. The findings regarding gender differences in math achievement have been inconsistent, but a considerable amount of research has shown that young men demonstrate higher math self-concept and STEM-related vocational interests than young women. Furthermore, the existing literature indicates that students' achievement and self-concept in math as well as their STEM-related interests are closely related to high school coursework.

The Present Study

In the present study, we examined the effects of a reform in upper secondary high school on gender differences in central predictors of STEM career choices and students' choice of STEM university subjects by reanalyzing representative data from 9,545 German students. Math high school coursework has been found to be closely linked to achievement, self-concept, and interests in the STEM fields (Nagy et al., 2008; Updegraff et al., 1996), all of which are central predictors of STEM career choices (Ma & Johnson, 2008; Nagy et al., 2008). A lower percentage of young women than men had chosen advanced math courses before the reform took place, but this difference was completely eliminated by the reform because the reform required all students to take advanced math courses. Thus, we expected effects of the reform on gender differences in STEM-related outcomes.

There is ample evidence of such effects of high school coursework on achievement, self-concept, and interests, but previous research has not addressed how gender differences in math achievement, self-concept, and interests as key predictors of STEM career choices may be influenced by requiring all students to enroll in advanced courses in math. The present study takes a major step toward filling this gap by investigating such an educational policy and its effects on women's participation in the STEM fields. We examined how changes in high school coursework are related to gender differences in predictors of STEM career choices and students' subjects of study at university after school. To do so, we evaluated effects of a school reform that was introduced in 2002 in one of the largest German states. The reform included the abolition of different math courses. Before the reform, students had been allowed to take math as either an advanced or a basic course. After the reform, all students had to take an obligatory advanced-level math course (Ministry of Education and Cultural Affairs, Youth and Sport Baden-Württemberg, 2002).

Because high school course level tends to predict students' achievement and self-concept, and because young women were less likely than young men to choose advanced courses in math before the reform, we expected that the effects of the reform on these outcomes would differ between the young women and men in the current study. As positive effects of course level on students' achievement have been documented, we hypothesized that gender differences in math achievement would be smaller after the reform (when all young men and women took advanced math courses) compared with before the reform (when more young men than young women had taken advanced math courses). Here, we assume that the smaller gender differences in achievement expected after the reform would be based on the higher achievement of young women after the reform compared with before. Regarding gender differences in math self-concept, we hypothesized that gender differences would be larger after the reform than before. This proposition was based on the finding that course level tends to have negative effects on a student's self-concept, and there was a higher percentage of young men than young women in advanced courses before the reform, whereas all students took advanced courses after the reform. We therefore expected that young women's self-concept would be lower after the reform than before on average, which would lead to greater gender differences in math self-concept. So far, there is less work on effects of high school coursework on vocational interests, and it is therefore not clear whether and how the reform might be related to gender differences in realistic and investigative interests. However, if we were to find similar effects of course level on STEM-related vocational interests as on self-concept and subject-specific interest, we would tentatively expect larger gender differences in realistic and investigative interests after the reform than before.

Because we expected differential reform effects on central predictors of STEM career choices (math achievement, math self-concept, realistic and investigative interests), we did not specify what the effects on the actual choice of STEM university subjects would be.

Method

The Reform of Upper Secondary School in the German School System

Before the reform of upper secondary school education, students in most German states self-selected their courses and were given the choice between math as an advanced course (about five hours per week) or a basic course (about three hours per week). In total, each student was required to select two advanced courses and typically six basic courses in different subjects. The individual combination of advanced and basic courses represented an individual profile for each student for all of their upper secondary school trajectories, and students were not able to choose different courses each semester. Beginning in 2002, most German states enacted reforms of their higher secondary education systems and implemented a course program. This program can be characterized by a reduction in the number of options in favor of a higher subject-related average amount of time allocated across all students to specific compulsory core subjects (e.g., German, mathematics, and foreign language). In most states, students were no longer able to self-select into different courses from that point in time on but were instead required to take a total of five courses from specific fields (e.g., math, foreign language, science) for a similar amount of time (4 hr per week). Besides these compulsory courses, students had to participate in other courses for a reduced number of hours (2 hr per week; e.g., arts, science, or social studies; Köller, Watermann, Trautwein, & Lüdtke, 2004; Trautwein, Neumann, Nagy, Lüdtke, & Maaz, 2010). To sum up, the two major changes of the reform were (a) an increase in the number of courses that had to be chosen for final examinations in upper secondary school on an advanced course level and (b) written exams in the first four of these courses (instead of the first three).

Description of Study and Sample

Data were drawn from the study “Transformation of the Secondary School System and Academic Careers” (TOSCA; Köller et al., 2004; Trautwein et al., 2010). The TOSCA study was designed to assess a representative sample of students in the last 4 months of their final year of upper secondary school in one German state (Baden-Württemberg). The data from the first waves of TOSCA 2002 and TOSCA 2006 are representative for all students in the final year of upper secondary school in the state of Baden-Württemberg. We considered data from N = 149 schools in the first wave of the first cohort (TOSCA 2002; N = 4,730; 54.5% female) as well as data from N = 146 schools in the first wave of the second cohort (TOSCA 2006; N = 4,715; 54.1% female). Over the course of the reform, another school type (biotechnological Gymnasium) was introduced. Robustness checks revealed no differences in results when students from this type of school were included versus not included. In our sample, roughly 60% of the students were enrolled in a general higher secondary school, and 40% were in a vocational upper secondary school. The time between the start of the course and our measurement was approximately 1.5 years. The measurement took place right at the end of the course. Data collection was executed by trained research assistants who visited every class and lasted for approximately one day per school. The first cohort contains data from students who chose basic and advanced courses in upper secondary high school, whereas the second cohort consists of data from students who all took the obligatory advanced math courses. The data from the two cohorts were drawn from the same schools. In both cohorts, a second assessment took place 2 years after the first measurement point via questionnaires that were sent to the participants. Overall, 80% of all students agreed to participate in the first wave of TOSCA 2002, and 82% of all students agreed to participate in the first wave of TOSCA 2006. At the second assessment, which followed 2 years after the first assessments for TOSCA 2002 and TOSCA 2006, respectively, information was obtained about students’ field of study at university from N = 1,741 students from TOSCA 2002 and N = 2,157 from TOSCA 2006 (see Figure 1).

Instruments

Math achievement. The *Advanced mathematics test* was based on items from the Third International Mathematics and Science Study (TIMSS; Mullis et al., 1998). According to Mullis et al. (1998), the advanced mathematics test takes into account “current thinking and priorities in the field of mathematics” (p. 284). The advanced mathematics test contained a total of 68 items from the areas of (a) Numbers, Equations, and Functions, (b) Analysis, (c) Geometry, (d) Propositional Logic and Proofs, as well as (e) Probability and Statistics. Most of the items were related to the first area and directly tested competencies from upper secondary school. Approximately two thirds of all of the items were multiple-choice questions, whereas the other items were administered in an open-ended format. A multimatrix design was used to administer the items; therefore, the students did not work on all 68 items but on a subset of items in one of four booklets that contained six different item clusters that were rotated systematically. In order to be able to compare the two different cohorts, items were scaled by applying item response theory (IRT; Rasch model) to account for the multimatrix design and to test for differential item functioning. As reported by Nagy, Neumann, Trautwein, & Lüdtke (2010), we used five completed data sets with *plausible values* (*PVs*), which were estimated in Mplus 5.2. These PVs were based on multiply imputed data, which was imputed previously with NORM (Schafer, 1997). As reported by Nagy et al. (2010), the psychometric properties of the test are good (PV reliability TOSCA 2002: .88; PV reliability TOSCA 2006: .90).

Mathematics self-concept. Mathematics self-concept was measured with four items from the Self Description Questionnaire III (SDQ III; Marsh & O’Neill, 1984; Marsh & Shavelson, 1985; Marsh, 1992), using the German translation by Schwanzer, Trautwein, Lüdtke, and Sydow (2005). The translated items focused on the evaluation of cognitive aspects (e.g., “I was always good in mathematics,” e.g., Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007). The scale showed high internal consistency in both samples, TOSCA 2002 (Cronbach’s α = .89) and TOSCA 2006 (Cronbach’s α = .90).

Vocational interests. Vocational interests were assessed with the Revised General Interest Structure Test (AIST-R; Bergmann & Eder, 2005), which is based on Holland’s (1997) RIASEC model. This instrument categorizes students with regard to six different dimensions of interest, namely, realistic (R), investigative (I), artistic (A), social (S), enterprising (E), and conventional (C) interests by using a total of 60 items (six 10-item scales). Students were asked to rate how interested they were in the described activities on a 5-point Likert scale ranging from 1 (*not at all*) to 5 (*very much*). An example item of realistic interests is “Working with machines or technical devices” and “Doing physically challenging work,” whereas investigative interests were assessed with items such as “Dealing with unexplored things” and

"Working in an experimental laboratory." The realistic and investigative facets, which were of specific interest in the present context, showed high internal consistencies (realistic interests—TOSCA 2002: Cronbach's α = .86; TOSCA 2006: Cronbach's α = .87; investigative interests—TOSCA 2002 and 2006: Cronbach's αs = .85).

Field of study at university. The field of study at university was assessed for each cohort 2 years after they graduated from high school. Students were able to report their subject of study or a combination of study subjects. Students' data were coded according to the official classification system of the Federal Statistical Office, the Fachserie 11 (Federal statistical office, 2008). In the current study, we used information about the field of study and computed one variable for which mathematics, engineering, computer science, and physics were coded as STEM subjects only if they were indicated as the first subject of study. In addition, we also specified various alternative codings where only the first, the first two, or all three subject indications were used to calculate the dependent variable and included biology, chemistry, or both as STEM subjects. The general pattern of results was identical across all these different analyses. Furthermore, we did not find any significant differences in STEM-related course change or student withdrawal patterns when comparing the first and second assessments between TOSCA 2002 and TOSCA 2006. The results were based on analyses in which mathematics, engineering, computer science, and physics were coded as STEM subjects.

Covariates. We controlled for the influence of several variables described below.

School types. Because students from different school types (e.g., vocational higher secondary schools and general higher secondary schools) usually differ in cognitive and noncognitive aspects (Trautwein et al., 2010), we included a dummy variable to be able to distinguish between vocational and general higher secondary schools.[1]

Socioeconomic background. Socioeconomic background was measured with information about the highest level of occupation in the family (of either the father or mother) and coded in accordance with the International Standard Classification of Occupations (ISCO-88). The ISCO scores were in turn converted into International Socio-Economic Index of Occupational Status (ISEI) 88 scores (Ganzeboom, De Graaf, & Treiman, 1992; Ganzeboom & Treiman, 1996). The highest ISEI value between the two parents was used to characterize the socioeconomic background of the students.

Number of books available in the home. The number of books available in the home was measured on a 7-point scale ranging from zero books available to more than 500 books

[1] Due to the different vocational school types that were considered in the TOSCA studies, we also specified models with dummy-coded variables for every type of vocational school as additional robustness checks. The results did not differ meaningfully.

available. This variable has been shown to be a good indicator of a family's cultural capital (e.g., Evans, Kelley, Sikora, & Treiman, 2010).

Age. The age of the students at the time of the assessment was calculated on the basis of information about students' year and month of birth.

Immigration background. Students with at least one parent born outside of Germany were coded as students with an immigration background.

Statistical Analyses

In order to test for reform effects, we specified multiple regression models involving the TOSCA study survey weights and tested gender as a moderator of the effect of the reform on the different STEM-related outcomes. The models contained the variables gender and cohort as well as socioeconomic background (HISEI), cultural capital (number of books), immigration background, type of school (general Gymnasium vs. type of vocational Gymnasium), and age as covariates. We controlled for these covariates to eliminate the influence of these potential confounders and to increase the precision of our estimation. In addition, we added the Cohort × Gender interaction in order to examine whether the reform had differential effects on young women and men. Because students from different types of schools usually differ in their cognitive and behavioral outcomes (Trautwein et al., 2010), we also controlled for this differential impact by including the three-way interaction between Cohort × School Type × Gender as well as the interaction between School Type × Cohort.

We also specified a multivariate model with a Wald test for the interaction effects and controlled for the false discovery rate of all parameter estimates in each multiple regression afterwards by applying the Benjamini-Hochberg adjustment (Benjamini & Hochberg, 1995).

We additionally investigated students' actual field of study at university 2 years after they completed high school. Of special interest in the current analysis were potential differences with regard to whether or not students chose a STEM-related field of study. We therefore specified models to predict field of study in STEM versus other fields of study in multiple logistic regressions.

We used the statistical software R (R Development Core Team, 2014) and the survey package (Lumley, 2014) to inspect the data. The final models were specified in Mplus 7.4 (Muthén & Muthén, 2012). All models took into consideration survey weights to obtain representative results for students in upper secondary schools in Baden-Württemberg.

In order to report meaningfully interpretable coefficients, we present fully standardized coefficients, meaning that both the dependent and continuous independent variables were standardized. We also present partially standardized coefficients, meaning that only the dependent variable was standardized (also referred to as Cohen’s *d*; Cohen, 1988). Continuous variables were centered. The partially standardized coefficients might be especially useful for interpreting effects of dichotomous variables. With regard to the fully standardized solution, the interaction terms were standardized before we included them in the regression models. In order to explore and interpret possible interaction effects, we additionally estimated simple main effects between the two cohorts for young women and men and school types for statistically significant three-way interactions by using the model constraint option in M*plus* 7.4. Estimating simple main effects to interpret interactions is also recommended by Jaccard, Wan, and Turrisi (1990). Furthermore, we also calculated structure coefficients (e.g., Courville & Thompson, 2001) to gain further insights into the dynamics of our data. Structure coefficients indicate the proportion of the multiple correlation that can be accounted for by the first-order correlation. When multicollinearity is high, the beta weights might be relatively small. However, structure coefficients are able to indicate this more precisely.

Effect sizes. Regarding the interpretation of effect sizes and on the basis of a literature review, as suggested by Henson (2006), we argue that effect sizes of d > 0.05 should be considered practically relevant. As can be seen in the literature, this seems to be the average amount of growth that can be expected from a half to 1 year of schooling (e.g., Hill, Bloom, Black, & Lipsey, 2008; Low, Yoon, Roberts, & Rounds, 2005; Low, 2009; Nagy et al., 2010; Wagner, Rose, Dicke, Neumann, & Trautwein, 2014). However, as stated in Henson (2006), benchmarks should be used cautiously.

Cluster structure. Students from the same class or school cannot be treated as independent observations because they are more similar to each other than they are to students from other classes or schools. Not considering this cluster structure leads to overestimated standard errors (Snijders & Bosker, 2012). To address the clustered data structure (students were nested within classes), standard errors were adjusted by applying a design-based correction as implemented in Mplus (Muthén & Muthén, 1998-2012), which automatically takes the multilevel structure into account and makes use of a sandwich estimator (see e.g., Asparouhov, 2005; Muthén & Satorra, 1995). Here, we followed McNeish, Stapleton, and Silverman’s (2016) recommendations as they pointed out that alternative design-based methods (or population-averaged methods) can be more intuitive and do not rely on assumptions that are inherent in the specification of random effects in hierarchical linear modeling. Design-based methods allow the researcher to adjust the standard errors of estimates and fit statistics for the nested structure of the data and have been shown to perform well in various different nested data settings (e.g., Stapleton, Yang, & Hancock, 2016).

Missing values. Missing values are a common problem in the social sciences, and several approaches have been implemented to account for missing values in a meaningful way (e.g., Enders, 2010; Graham, 2009). There is a growing consensus that approaches such as multiple imputation (MI) or full information maximum likelihood (FIML) estimation are superior to traditional methods (e.g., complete case analysis or pairwise deletion). For all outcomes except math achievement and all independent variables, missing values were addressed with full information maximum likelihood in Mplus 7.4 (Muthén & Muthén, 2012). There were no missing values on the math achievement tests as we used plausible values that were generated for every student and the primary analysis of the TOSCA study (Nagy et al., 2010).

Results

In the first step, the two cohorts were compared with respect to possible differences in the covariates (Table 1). Overall, these pre-existing differences between the two cohorts seemed to be of small practical relevance. Differences were found for age (d = -0.22, p < .001), largely due to the TOSCA 2002 assessment taking place a little bit later in the school year because of an organizational issue. However, because this difference applied equally to young women and men, it should not have had any effect on the results. Furthermore, we controlled for age in all analyses. In addition, a difference in the number of books available in the home (d = -0.06, p = .021) was significant, whereas differences on all other variables (including gender) were not significant.

Next, we compared the lengths of time (in hours per week) allocated to mathematics by gender between the two cohorts before and after the reform. Table 2 shows a difference in the average amount of time allocated to math for both young men (3.5 min per week) and young women (19.7 min per week) and an average increase in the total sample after the reform (12.2 min). As expected, the average amount of time allocated to mathematics increased more for young women than for young men as shown by a significant Gender × Cohort interaction (B = 16.20, p < .001).

Test of Advanced Mathematics Achievement

We hypothesized that the gender difference in math achievement in favor of young men would be smaller after the reform that introduced the obligatory advanced mathematics course for both young men and women. To test our prediction, we used multiple regression analyses to explore a possible difference between the two cohorts in advanced mathematics achievement (Table 3).

The Cohort × Gender interaction was statistically significant (*d* = 0.14, *p* = .025, 95% CI [0.01, 0.26]). In line with our hypothesis, the interaction indicated a smaller difference between young women and men after the reform than before (see Figure 2). This was mainly due to a higher average level of young women's achievement after the reform (*d* = 0.14, *p* = .002, 95% CI [0.05,0.22]), whereas young men's achievement did not differ before and after the reform (*d* = 0.00, *p* = .988, 95% CI [-0.11, 0.11]). The Cohort × School Type interaction (*d* = 0.08, *p* = .255, 95% CI [-0.08,0.22]) was not statistically significant, but the Cohort × Gender × School Type interaction had a significant regression weight (*d* = -0.19, *p* = .029, 95% CI [-0.35,0.02]), indicating that the effects of the reform differed between the different school types. Our results indicate a three-way interaction between Cohort × Gender × School Type. Exploring this interaction revealed statistically significant differences for young women, but not for young men, before versus after the reform for general gymnasiums but not for vocational gymnasiums, in favor of the cohort that was measured after the reform. However, for young men, the effect of the reform was not statistically significantly different between vocational gymnasiums and general gymnasiums.

Math Self-Concept

With regard to math self-concept, we expected a larger gender difference after the reform. In line with our expectations, and as shown in Table 4, the moderating effect of gender on the relation between cohort and self-concept was statistically significant (*d* = -0.16, *p* < .001, 95% CI [-0.27, -.04]). The larger gender difference after the reform was the result of a statistically significantly lower average math self-concept for young women after the reform (*d* = -0.19, *p* < .001, 95% CI [-0.27, -0.11]) compared with before the reform. For young men, math self-concept did not differ significantly before versus after the reform (*d* = 0.04, *p* = .433, 95% CI [-0.18, 0.08]). The other two interaction effects, Cohort × School Type (*d* = -0.03, *p* = .619, 95% CI [-0.16, 0.09]) and Cohort × Gender × School Type (*d* = 0.11, *p* = .157, 95% CI [-0.04,0.27]), were both not statistically significant.

Realistic and Investigative Vocational Interests

According to our hypotheses, we expected larger gender differences in realistic and investigative interests after the reform. As reported in Table 5, we found a significant and negative interaction between cohort and gender in predicting realistic vocational interests (*d* = -0.15, *p* = .007, 95% CI [-0.26, -0.04]), thus indicating a larger gender difference after the reform than before. This larger gender difference resulted from a significantly higher mean score for young men (d = 0.27, p < .001, 95% [0.19, 0.35]) and a smaller, albeit also significantly higher mean score for young women (d = 0.12, p < .001, 95% [0.05, 0.19]) after the reform (see Figure 2).

In addition to realistic vocational interests, we tested for a gender difference in investigative interests (Table 6). Taking a closer look at our results, we found a significant interaction effect (see Figure 2), indicating a larger gender difference in investigative vocational interests after the reform (d = -0.12, p = .019, 95% CI [-0.23, -0.02]). No significant difference between before and after the reform was found for young women in investigative interests (d = -0.01, p = .773, 95% CI [-0.09, 0.07]), but young men showed, on average, a higher level of interest after the reform (d = 0.11, p = .01, 95% CI [0.03, 0.21]. For both outcomes, the Cohort × School Type interaction and the Cohort × Gender × School Type interaction were not statistically significant (see Table 6).

The results for the multivariate approach were similar to the results for the univariate approach: The Wald test for the interaction effect was statistically significant, $\chi^2(12) = 55.06$, p < .001. Furthermore, even after the Benjamini-Hochberg corrections, all interaction effects remained statistically significant in the multivariate and univariate approaches. Overall, we found that the structure coefficients supported our results regarding multiple linear regression models and the interpretation of the relevance of the Cohort × Gender interaction for all outcome variables (see Table 8).

Field of Study at University

Whether or not the upper secondary school reform had an effect on university subject choices was handled as an open research question. Therefore, we did not formulate an explicit hypothesis with regard to this construct. The results presented here are based on an analysis that considered only students who did not intend to become teachers.[2] As reported in Table 7, none of the additional interaction effects were statistically significant. Thus, a potential shift, which would go along with an increase in women enrolling in STEM subjects at university was not found in our data set (Cohort × Gender: OR = 1.01, p = .838, 95% CI [0.86, 1.21]). We

[2] The pattern of gender differences in the literature varies with respect to different professions within the STEM fields. Whereas a larger percentage of young men than women tend to choose mathematically intensive STEM subjects, gender differences are much less pronounced with regard to STEM teaching professions (Watt, Richardson, & Devos, 2013). To meet this objective, we excluded teaching students from our analysis. However, robustness checks did not reveal any substantial difference between the results of these two groups of students. Furthermore, although men tended to start their studies a bit later (e.g., due to mandatory community or military services), we did not find significant gender differences before and after the reform regarding students who attended university and those who did not.

further tested for potential differences between students who provided information about their university subject and those who did not. Results revealed that women (*OR* = 0.73, *p* < .001) and students from vocational schools (*OR* = 0.54, *p* < .001) as well as older students (*B* = -.20, *p* < .001) were less likely to report their subject, whereas students with a higher HISEI (*B* = .28, *p* < .001), more books at home (*B* = .33, *p* < .001), and higher cognitive abilities (*B* = .28, *p* < .001) reported their subject more often. We controlled for these variables in all analyses. It is important to note that these differences did not differ significantly between the two cohorts, as shown by the Wald test, $\chi^2(7) = 7.75$, *p* = .356.

Discussion

In the current study, we examined effects of a higher secondary school education reform on STEM-related outcomes in a large and representative sample. The reform is of high theoretical and practical interest because it abolished a prior imbalance between young men and women in taking advanced math courses. High school coursework in math has been shown to be related to STEM career choices as well as to math achievement, math self-concept, and vocational interests, all of which are important predictors of STEM career choices. Therefore, we expected that the effects of the reform on these outcomes would differ by gender. Overall, the results supported most of our predictions. First, there were significant gender differences in all outcomes before the reform, with higher scores for young men than for young women. Second, we found differential effects of the reform for young women and men in all outcomes except field of study at university. However, the direction of the effects differed: The gender difference in math achievement was smaller after the reform, but gender differences in math self-concept and STEM-related vocational interests were even larger after the reform than before. However, the larger gender difference after the reform in math self-concept was based on young women’s lower scores, whereas young men’s scores did not differ. Also, the greater differences in vocational interests were due to young men’s higher interests after the reform, whereas young women’s interests were only slightly higher (realistic) or did not differ (investigative). Third, we found no overall effect of the reform on gender differences in the choice of STEM subjects at university.

Differential Reform Effects for Young Men and Women

The effects of the reform on math achievement are in accordance with previous research that reported positive effects of course level on achievement, which can be attributed to more demanding curricula, more teaching time, and larger weights from grades in advanced courses with respect to their contribution to final GPA (e.g., Brunello & Checchi, 2007; Gamoran &

Mare, 1989; Hanushek & Wössmann, 2006; Kelly, 2004; Lucas, 2001). Presumably because a larger proportion of young men than women had chosen advanced courses in math before the reform, but all students took the same math course after the reform, young women were able to come closer to young men's math achievement, although there was still a significant gender difference after the reform. In addition, there was a difference in teaching time between the courses before versus after the reform, with more lessons taught per week in the advanced course (five lessons) than in the basic course (three lessons). Although meta-analyses do not suggest a clear pattern with regard to the effects of extended learning time on achievement, most studies have shown zero to small positive effects (e.g., Patall, Cooper, & Allen, 2010; Scheerens & Hendriks, 2014). Thus, the difference in teaching time might provide a possible explanation for the differential effects of the reform on young women's and men's math achievement. However, we cannot explicitly test for or disentangle the effects of instructional time or course level on our results at this point.

Against this background, the larger gender differences after the reform with respect to math self-concept and STEM-related interests might come as a surprise at first glance.

A change in reference group provides a good explanation for the larger gender difference in math self-concept after the reform: It is a common finding that social comparisons are central for the development of students' self-concept. In evaluating their own abilities, individuals refer not only to their own prior achievement in a domain, but also to the level of achievement they perceive in their surroundings (e.g., Marsh, 2005; Niepel, Brunner, & Preckel, 2014; Trautwein et al., 2006). As discussed above, students' achievement differs between advanced and basic courses; thus, both courses provide different frames of reference for social comparisons. Higher course levels are usually associated with negative effects on students' evaluations of their own abilities after individual ability is controlled for (Marsh, 2005; Trautwein et al., 2006). Before the reform, young women tended to choose basic courses in math where they were surrounded by an (on average) a weaker reference group, compared with students in advanced courses. Therefore, they perceived their own math ability in comparison with other, on average, lower achieving classmates. After the reform, all students were instructed at the same course level. Consequently, after the reform, young women could compare their own achievement with the achievement of all other students in their class, which included students with relatively lower achievement but also those with relatively higher achievement. It is therefore likely that the reason why young women's evaluation of their own math abilities was somewhat lower was due to the, on average, higher achieving reference group. There was no significant difference in young men's self-concept after the reform, which can be explained by the proportions of young men in advanced and basic courses before the reform, as they participated in advanced and basic courses in almost equal parts before the reform. According to the literature described above, it is therefore likely that possible effects of course level on young men's self-concept cancelled each other out. These explanations are further supported by the fact that young women's math self-concept in basic courses before the reform was statistically lower, compared with young women's self-concept after the reform (*d* = -0.12, *p* < .001), whereas the reverse was true for young women in advanced courses before the reform (*d* = 0.83, *p* < .001). Furthermore, the difference between young men and women in basic courses was not statistically significant (*d* = - 0.07, *p* = .086), whereas the gender gap for advanced course students was statistically significant, favoring young men (*d* = -0.13, *p* = .001).

In our study, we found larger gender difference after the reform in realistic and investigative interests as well, but in contrast to math self-concept, the greater differences were based on young men's higher levels of interests after the reform, whereas young women showed only slightly higher interests (realistic) or even similar scores (investigative) after the reform. There is a gap in research on how vocational interests might be related to course level. However, as reported in the Introduction, previous research has indicated positive relations between individual levels of math achievement and realistic and investigative interests (Ackerman & Heggestad, 1997; Anthoney & Armstrong, 2010). Furthermore, previous research has shown negative effects of the mean level of achievement on domain-specific levels of interest (Köller, Trautwein, Lüdtke, & Baumert, 2006; Schurtz, Pfost, Nagengast, Artelt, 2014; Trautwein et al., 2006) and initial findings with respect to vocational interests. These findings indicate that there might be positive effects of the mean level of math achievement on realistic and investigative vocational interests (Cambria et al., 2016). However, as these findings provide only initial indications on how vocational interests might be related to class level, they enable us to discuss our findings only on a speculative basis. Thereby, one could argue that there might be a positive association between class level and students' realistic and investigative interests, but this association differs by gender, with larger associations for young men than for young women. Previous research on vocational interests has indeed indicated differential associations between ability and vocational interests, although such findings have so far been limited to general cognitive ability and have not been applied to math (e.g., Reeve & Heggestad, 2004). However, more research is needed to explore the relation between class level and vocational interests for young women as well as for young men.

Although we found differential effects of the reform on central predictors of STEM career choices, we found no difference in gender ratios in the numbers of students who chose to study STEM university subjects. There are two aspects to consider when interpreting the absence of effects of the reform on gender differences in STEM university subject choices. First, we found opposite effects of the reform on gender differences in four important predictors of STEM career choices: Whereas differences in math achievement were eliminated, differences in math self-concept and both interest facets were larger. Consequently, it is possible that the effects of the reform on the predictors cancelled each other out, with the consequence that no effect on the choice of STEM subjects remained. Second, choosing a university subject is a complex process that involves numerous factors (see Schoon & Eccles, 2014). The reform influenced students' upper secondary high school coursework, but it did not directly affect other structural factors or the wider context they grew up in, such as their family structure, the role models they perceived, or their stereotypical views of STEM professions.

Practical Implications

Our study adds to the increasing number of studies that have found intended as well as unintended effects of educational reforms. In fact, educational policy reforms do not necessarily improve educational outcomes but can instead result in numerous unintended consequences. In addition, the aspects of the reforms most likely interact differently with different student characteristics, even if such aspects are well-structured and carefully planned (Gross, Booker, & Goldhaber, 2009). For instance, studies by Gross et al. (2009), Domina, McEachin, Penner, and Penner (2015), and Lee and Reeves (2012) showed that school reforms could have differential effects for minority students (e.g., African American and Hispanic students) or could vary for specific school districts. The results show that school reforms can have differential effects on several outcomes, and such outcomes can even differ for particular subgroups such as young women and men; not every well-intentioned reform will reach all goals, and some might even backfire.

Unintended consequences of reforms can be attributed to, amongst other factors, the complex nature of establishing and especially of implementing reforms (e.g., McLaughlin, 1987; Young & Lewis, 2015) in the education sector as a "loosely coupled system" (Porter, Fusarelli, & Fusarelli, 2015, p. 114). Conversely, with regard to the current study, one might argue that the higher achievement and realistic interests that came with this reform came at a price—a lower math self-concept for young women—which had to be expected given the change in reference group.

Although high school coursework is central to young people's career choices, and although we found differential effects of the reform on central predictors of STEM career choices for young men and women, we did not find effects of the reform on gender differences in the choice of STEM university subjects, which indicates that one single reform might not significantly influence students' career choices. In the complex context that young people grow up in, there is a cumulative process of multiple experiences that shape young people's academic attitudes and behavior, such as career choice (cf. Schoon & Eccles, 2014). Influencing gender differences in high school course selection by restricting choice options might be one way to balance some gender differences in the STEM context, namely, gender differences in math achievement. However, reforming course choice options does not necessarily impact any of the reasons for why young women are less likely to choose advanced math courses than young men (e.g., gender stereotypes, different expectancies of parents, teachers, peers; cf. Schoon & Eccles, 2014; Wigfield & Eccles, 2000). Such high school reforms might therefore be "too little too late" to increase gender equity in the STEM fields in a meaningful and sustainable way. Furthermore, although course-taking gaps in other countries have narrowed in recent decades (e.g., Domina & Saldana, 2012; Osborne & Dillon, 2008), subsequent changes in STEM career plans do not seem to be of considerable size (Jerrim & Schoon, 2014).

Limitations and Further Research

The current study demonstrates that intensifying school curricula and providing equal access to advanced courses "does not necessarily level the [educational] playing field" with regard to all important outcomes (Domina & Saldana, 2012, p. 688). Although our investigation was based on a strong data set, some limitations should be kept in mind when interpreting the results. First, our results were limited to the domain of math. Math is a key domain within the STEM fields (Ma & Johnson, 2008; Sells, 1980), and math achievement, self-concept, and interests are very important for math-intensive STEM career choices (e.g., Parker et al., 2012; Schoon & Eccles, 2014). Nevertheless, other STEM domains such as physics or chemistry are also meaningful for later math-intensive STEM career choices (e.g., Hazari, Sonnert, Sadler, & Shanahan, 2010), and gender differences in such high school courses are often even larger than in math (e.g., NSF, 2015). Evaluating the effects of a reform on central STEM outcomes in these domains might therefore provide additional information about effects on important predictors of math-intensive STEM career choices.

Second, the current study was based on cross-sectional data. According to Shadish, Cook, and Campbell (2002), quasi-experiments lack "random assignment of units to conditions" (p. 104), which may lead to selection bias. We attempted to address these challenges by using a lagged cohort control design that should have led to relatively small selection differences between cohorts (drawn from the same schools). We additionally checked for potential differences between cohorts and used covariates to control for these.

Third, besides these methodological issues, there are other possible reasons for the results that we found. Our results may be explained by the multidimensional structure of the reform. As stated by Malen and Knapp (1997), “policy takes many forms, performs many functions, and begets many effects,” which is why “it is difficult to get a fix on the boundaries, let alone the ‘workings’ of a policy or a set of policies” (p. 419). In our case, as mentioned, not only did time vary between the groups before and after the reform, but the reference groups and course levels also varied. Therefore, the effects of the reform cannot be directly attributed to one specific aspect or mechanism of the reform in a causal manner but must be interpreted from within the multilayered framework of the entire policy reform.

However, as society is constantly changing, it would be reasonable to expect main and interaction effects that indicate the increased participation of young women in STEM classes because they are now as able to do so as young men. However, the results of our study instead indicate the opposite pattern. Regarding this point, it is also important to mention that society’s growing interest and all resulting efforts had already increased in the beginning of this century and not just between these two cohorts in particular (National-State-Commision for Educational Planning and Scientific Promotion, 2002; NSF, 2000). In addition, we checked closely whether any other educational reforms had been implemented between the two cohorts, but this was not the case.

Further research should address the question of whether effects, such as the drop in self-concept, can be found in different subsamples. This refers to questions such as whether such effects can be found for all young women or only the subsample of those who would have chosen basic courses if they had been allowed to, and whether similar effects can be found for young men who would have chosen basic courses if they had been allowed to.

Fourth, our results are limited to the issue of gender differences in STEM career choices at the end of secondary education, and more research is needed to explore the complex pattern of gender differences in the STEM fields throughout students’ educational careers. In our study, we focused on important predictors of STEM career choices as well as students’ choice of university major in the STEM fields. Therefore, our results provide insights into various effects of the reform on central STEM outcomes. However, regarding the issue of gender differences in the STEM area, not only do women tend to choose such majors less frequently than their male counterparts, but women also drop out of university at higher rates (Ackerman, Kanfer, & Beier, 2013; Perez-Felkner, McDonald, & Schneider, 2014). Considering social comparison processes, one could possibly argue that women entering the STEM fields are likely to experience such comparison processes during their studies, where they need to deal with other high-achieving students. Experiencing such comparison processes at an earlier point in high school might therefore make women less likely to pursue such careers and—consequently—less vulnerable to dropping out of STEM fields during college. Furthermore, prior work on the development of interest suggests that interest takes time to develop (see Hidi & Renninger, 2006) and that such a change in upper secondary high school coursework as investigated in the present study might be less related to students' vocational interests than to their achievement and self-concept or that such effects might take longer. In this study, we investigated effects of changes in coursework requirements on students' interests 1.5 years after they started taking these high school courses. It might be the case that such a time period is insufficient to fully study effects on interest developments and that effects would be different or more pronounced if more time could have elapsed between when the students began taking these high school courses and the measurement point. Further research spanning a longer time frame is needed to test such propositions as well as to develop more potent remedies for the gender differences that still exist.

Conclusion

The present study was aimed at taking a closer look at effects of high school coursework on gender differences in math-intensive STEM fields. To this end, we investigated effects of a statewide educational reform in Germany with a large representative sample. The reform required all students to take advanced courses in math and eliminated the prior imbalance between young men and women in choosing such courses. Our results showed that it is crucial to take multiple aspects into consideration in order to obtain insights into possible differential effects of changes in coursework requirements. Although requiring all students to take advanced math courses appears to be adequate for eliminating gender differences in math achievement, it seems that young women were not aware of this: Young men and women's achievement differed less after the reform, but young women showed an even lower self-concept compared with young men than had been there before the reform. With respect to realistic and investigative interests, although young women showed no or only slightly higher interests after the reform, the interests of young men were substantially higher after the reform. Mechanisms that ensure that all students will benefit in comparable ways from such school reforms and impede negative side effects, such as those found for young women's self-concept, should be a primary focus of future research.

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*(2), 219–245. doi:10.1037/0033-2909.121.2.219

Ackerman, P. L., Kanfer, R., & Beier, M. E. (2013). Trait complex, cognitive ability, and domain knowledge predictors of baccalaureate success, STEM persistence, and gender differences. *Journal of Educational Psychology, 105*(3), No–. doi:10.1037/a0032338

Alarcon, G. M., & Edwards, J. M. (2013). Ability and motivation: Assessing individual factors that contribute to university retention. *Journal of Educational Psychology, 105*(1), 129–137. doi:10.1037/a0028496

Anthoney, S. F., & Armstrong, P. I. (2010). Individuals and environments: Linking ability and skill ratings with interests. *Journal of Counseling Psychology, 57*(1), 36–51. doi:10.1037/a0018067

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12*, 411–434.

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

Bergmann, C., & Eder, F. (2005). *AIST-R: Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-R) - Revision. [General Interest Structure Test and Environmental Structure Test - Revision]*. Göttingen: Beltz.

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review, 15*(1), 1–40. doi:10.1023/A:1021302408382

Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy, 22*(52), 782–861. doi:10.1111/j.1468-0327.2007.00189.x

Cambria, J., Brandt, H., Nagengast, B., & Trautwein (2016). Vocational interests. The impact of class achievement and gender. *Manuscript in preparation*.

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women’s underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin, 135*(2), 218–261. doi:10.1037/a0014412

Chen, S.-K., Yeh, Y.-C., Hwang, F.-M., & Lin, S. S. J. (2013). The relationship between academic self-concept and achievement: A multicohort–multioccasion study. *Learning* *and Individual Differences*, *23*, 172–178. https://doi.org/10.1016/j.lindif.2012.07.021

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. doi:10.3102/0002831213489843

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc. doi:10.1234/12345678

Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is not enough. *Educational and Psychological Measurement*, *61*(2), 229–248.

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.

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. doi: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. doi:10.3102/0002831211426347

Eccles, J. S., Adler, T., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. C. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–121). San Francisco, CA: W. H. Freeman & Co.

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. doi:10.1037/a0018053

Enders, C. K. (2010). *Applied missing data analysis. Methodology in the social sciences*. New York, NY: Guilford Press.

Evans, M. D. R., Kelley, J., Sikora, J., & Treiman, D. J. (2010). Family scholarly culture and educational success: Books and schooling in 27 nations. *Research in Social Stratification and Mobility*, *28*(2), 171–197. doi:10.1016/j.rssm.2010.01.002

Federal statistical office (Ed.). (2008). *Bildung und Kultur - Studierende an Hochschulen Wintersemester 2007/2008. Fachserie 11, Reihe 4.1*. Wiesbaden: Statistisches Bundesamt.

Gamoran, A., & Mare, R. D. (1989). Secondary school tracking and educational inequality: Compensation, reinforcement, or neutrality? *American Journal of Sociology*, *94*(5), 1146.

doi:10.1086/229114

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21*(1), 1–56. doi:10.1016/0049-089X(92)90017-B

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*(3), 201–239. doi:10.1006/ssre.1996.0010

Gottfredson, L. S. (1986). Occupational aptitude patterns map: Development and implications for a theory of job aptitude requirements. *Journal of Vocational Behavior*, *29*(2), 254–291. doi:http://dx.doi.org/10.1016/0001-8791(86)90008-4

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

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. doi:10.3102/0162373709333886

Hanushek, E. A., & Wössmann, L. (2006). Does educational tracking affect performance and inequality? Differences- in-differences evidence across countries. *The Economic Journal*, *116*(510), C63–C76. doi:10.1111/j.1468-0297.2006.01076.x

Hazari, Z., Sonnert, G., Sadler, P. M., & Shanahan, M.-C. (2010). Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. *Journal of Research in Science Teaching*, *47*(8), 978–1003. doi:10.1002/tea.20363

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist*, *34*(5), 601–629. doi:10.1177/0011000005283558

Hidi, S., & Ainley, M. (2002). Interest and Adolescence. In F. Pajares & T. C. Urdan (Eds.), *Academic Motivation of Adolescence* (pp. 247–275). Greenwich, Conn: Information Age Pub.

Hidi, S., & Renninger, K. A. (2006). The Four-Phase Model of Interest Development. *Educational Psychologist*, *41*(2), 111–127.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. doi:10.1111/j.1750-8606.2008.00061.x

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology*, *6*(1), 35–45. doi:10.1037/h0040767

Holland, J. L. (1966). *The psychology of vocational choice: A theory of personality types and model environments*. Waltham, MA: Blaisdell.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, Fla.: Psychological Assessment Resources.

Huang, J. L., & Pearce, M. (2013). The other side of the coin: Vocational interests, interest differentiation and annual income at the occupation level of analysis. *Journal of Vocational Behavior*, *83*(3), 315–326. doi:10.1016/j.jvb.2013.06.003

Humphreys, L. G., & Yao, G. (2002). Prediction of graduate major from cognitive and self-report test scores obtained during the high school years. *Psychological Reports*, *90*(1), 3–30. doi:10.2466/PR0.90.1.3-30

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*(2), 139–155. doi:10.1037/0033-2909.107.2.139

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect. *Psychology of Women Quarterly*, *14*, 299–324. doi:10.1111/j.1471-6402.1990.tb00022.x

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Diversity: Gender similarities characterize math performance. *Science*, *321*(5888), 494–495. doi:10.1126/science.1160364

Jaccard, J., Wan, C. K., & Turrisi, R. (1990). The detection and interpretation of interaction effects between continuous variables in multiple regression. *Multivariate Behavioral Research*, *25*(4), 467–478.

Jansen, M., Lüdtke, O., & Schroeders, U. (2016). Evidence for a positive relation between interest and achievement: Examining between-person and within-person variation in five domains. *Contemporary Educational Psychology*. https://doi.org/10.1016/j.cedpsych.2016.05.004

Jansen, M., Scherer, R., & Schroeders, U. (2015). Students’ self-concept and self-efficacy in the sciences: Differential relations to antecedents and educational outcomes. *Contemporary Educational Psychology*, *41*, 13–24. doi:10.1016/j.cedpsych.2014.11.002

Jerrim, J., & Schoon, I. (2014). Do teenagers want to become scientists? A comparison of gender differences in attitudes toward science, career expectations, and academic skills across 29 countries. In I. Schoon & J. S. Eccles (Eds.), *Gender Differences in Aspirations and Attainment* (pp. 203–223). Cambridge, U.K.: Cambridge University Press.

Kelly, S. (2004). Are teachers tracked? On what basis and with what consequences. *Social Psychology of Education*, *7*(1984), 55–72. doi:10.1023/B:SPOE.0000010673.78910.f1

Köller, O., Baumert, J., & Schnabel, K. U. (2001). Does interest matter? The relationship between academic interest and achievement in mathematics. *Journal for Research in Mathematics Education*, *32*(5), 448–470. doi:10.2307/749801

Köller, O., Trautwein, U., Lüdtke, O., & Baumert, J. (2006). Zum Zusammenspiel von schulischer Leistung, Selbstkonzept und Interesse in der gymnasialen Oberstufe. *Zeitschrift Für Pädagogische Psychologie*, *20*(1), 27–39. doi:10.1024/1010-0652.20.12.27

Köller, O., Watermann, R., Trautwein, U., & Lüdtke, O. (2004). *Wege zur Hochschulreife in Baden-Württemberg: TOSCA - eine Untersuchung an allgemein bildenden und beruflichen Gymnasien*. Opladen: Leske + Budrich.

Krapp, A. (1999). Interest, motivation and learning: An educational-psychological perspective. *European Journal of Psychology of Education*, *14*, 23–40. doi:10.1007/BF03173109

Lee, J., & Reeves, T. (2012). Revisiting the Impact of NCLB High-Stakes School Accountability, Capacity, and Resources: State NAEP 1990-2009 Reading and Math Achievement Gaps and Trends. *Educational Evaluation and Policy Analysis*, *34*(2), 209–231. doi:10.3102/0162373711431604

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a Unifying Social Cognitive Theory of Career and Academic Interest, Choice, and Performance. *Journal of Vocational Behavior*, *45*, 79–122.

Lippa, R. (1998). Gender-related individual differences and the structure of vocational interests: the importance of the people-things dimension. *Journal of Personality and Social Psychology*, *74*(4), 996–1009.

Low, K. S. (2009). *Patterns of Mean-Level changes in vocational interests: A quantitative review of longitudinal studies*. Unpublished doctoral dissertation of the University of Illinois at Urbana-Champaign.

Low, K. S., Yoon, M., Roberts, B. W., & Rounds, J. (2005). The stability of vocational interests from early adolescence to middle adulthood: a quantitative review of longitudinal studies. *Psychol Bull*, *131*(5), 713–737. doi:10.1037/0033-2909.131.5.713

Lubinski, D., & Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years. *Special Section: Doing Psychological Science*, *1*(4), 316–345. doi:10.1111/j.1745-6916.2006.00019.x

Lucas, S. R. (2001). Effectively maintained inequality: Education transitions, track mobility, and social background effects. *American Journal of Sociology, 106*(6), 1642–1690. doi:10.1086/321300

Lumley, T. (2014). *"Survey: Analysis of complex survey samples". R package version 3.30*.

Ma, X., & Johnson, W. (2008). Mathematics as the critical filter: Curricular effects on gendered career choices. In H. M. G. Watt & J. S. Eccles (Eds.), *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences* (pp. 55–83). Washington, DC: American Psychological Association.

Malen, B., & Knapp, M. (1997). Rethinking the multiple perspectives approach to education policy analysis: Implications for policy-practice connections. *Journal of Education Policy, 12*(5), 419–445. doi:10.1080/0268093970120509

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23*(1), 129–149. doi:10.3102/00028312023001129

Marsh, H. W. (1992). *Self Description Questionnaire (SDQ) III: A theoretical and empirical basis for the measurement of multiple dimensions of late adolescent self-concept: A test manual and a research monograph.* Macarthur.

Marsh, H. W. (2005). Big-fish-little-pond effect on academic self-concept. *Zeitschrift Für Pädagogische Psychologie, 19*(3), 119–129. doi:10.1024/1010-0652.19.3.119

Marsh, H. W. (2007). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology.* Leicester, UK: British Psychological Society.

Marsh, H. W., Abduljabbar, A. S., Morin, A. J. S., Parker, P. D., Abdelfattah, F., Nagengast, B., & Abu-Hilal, M. M. (2015). The big-fish-little-pond effect: Generalizability of social comparison processes over two age cohorts from Western, Asian, and Middle Eastern Islamic countries. *Journal of Educational Psychology, 107*(1), 258–271. doi:10.1037/a0037485

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J. S., Abdelfattah, F., Nagengast, B., … Abu-Hilal, M. M. (2014). 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. doi: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. doi:10.1111/j.1745-3984.1984.tb00227.x

Marsh, H. W., & Shavelson, R. J. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist, 20*(3), 107–123. doi:10.1207/s15326985ep2003_1

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. doi:10.3102/0002831207306728

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, *76*(2), 397–416.

Marsh, H. W., & Yeung, A. S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and English constructs. *American Educational Research Journal*, *35*(4), 705–738. doi:10.3102/00028312035004705

McLaughlin, M. W. (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis*, *9*(2), 171–178. doi:10.3102/01623737009002171

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2016). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*. doi:10.1037/met0000078

Ministry of Education and Cultural Affairs, Youth and Sport Baden-Württemberg. (2002). *Die neue gymnasiale Oberstufe in Baden-Württemberg. Infodienst Schule spezial*. Stuttgart: Ministerium für Kultus Jugend und Sport Baden-Württemberg.

Mullis, I. V. S., Martin, M. O., Beaton, A., Gonzalez, E., Kelly, D., & Smith, T. (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.

Muthén, B. O., & Muthén, L. K. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267–316.

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 J. S. Eccles & H. M. G. Watt (Eds.), *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences* (pp. 115–143). Washington, DC, US: American Psychological Association. doi:10.1037/11706-004

Nagy, G., & Husemann, N. (2010). Berufliche Interessen vor und nach dem Übergang in die gymnasiale Oberstufe. In W. Bos, E. Klieme, & O. Köller (Eds.), *Schulische Lerngelegenheiten und Kompetenzentwicklung. Festschrift für Jürgen Baumert* (pp. 59–84). Münster, New York, München, Berlin: Waxmann.

Nagy, G., Neumann, M., Trautwein, U., & Lüdtke, O. (2010). Voruniversitäre Mathematikleistungen vor und nach der Neuordnung der gymnasialen Oberstufe in Baden-Württemberg. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke, & K. Maaz (Eds.), *Schulleistungen von Abiturienten. Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand.* (pp. 147–180). Wiesbaden: VS Verl. für Sozialwissenschaften. doi:10.1007/978-3-531-92037-5_6

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. doi:10.1111/j.1532-7795.2010.00644.x

National-State-Commision for Educational Planning and Scientific Promotion. (2002). *Frauen in den ingenieur- und naturwissenschaftlichen Studiengängen. Bericht der BLK vom 2 . Mai 2002*. Retrieved from http://www.blk-bonn.de/papers/heft100.pdf

Niepel, C., Brunner, M., & Preckel, F. (2014). The longitudinal interplay of students' academic self-concepts and achievements within and across domains: Replicating and extending the reciprocal internal/external frame of reference model. *Journal of Educational Psychology*, *106*(4), 1170–1191. doi:10.1037/a0036307

NSF, (National Science Foundation). (2000). *Summary Report on the Impact Study of the National Science Foundation's Program for Women and Girls*.

NSF, (National Science Foundation). (2013). *Women, minorities, and persons with disabilities in science and engineering*. Retrieved from http://www.nsf.gov/statistics/wmpd/2013/pdf/nsf13304_full.pdf

NSF, (National Science Foundation). (2015). *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2015 Digest*. *Special Report NSF*. https://doi.org/Special Report NSF 11-309

Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2012). Vocational interests and performance: A quantitative summary of over 60 years of research. *Perspectives on Psychological Science*, *7*(4), 384–403. doi:10.1177/1745691612449021

OECD. (2010). *OECD information technology outlook 2010*. OECD Publishing. doi:10.1787/it_outlook-2010-en

Osborne, J., & Dillon, J. (2008). *Science education in europe: Critical reflections. A report to* *the Nuffield Foundation*. Retrieved from http://efepereth.wdfiles.com/local--files/science-education/Sci_Ed_in_Europe_Report_Final.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 multicontext study. *Developmental Psychology*, *48*(6), 1629–1642. doi:10.1037/a0029167

Pässler, K., Beinicke, A., & Hell, B. (2014). Gender-Related Differential Validity and Differential Prediction in Interest Inventories. *Journal of Career Assessment*, *22*(1), 138–152. doi:10.1177/1069072713492934

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. doi:10.3102/0034654310377086

Patrick, L., Care, E., & Ainley, M. (2011). The relationship between vocational interests, self-efficacy, and achievement in the prediction of educational pathways. *Journal of Career Assessment*, *19*, 61–74. doi:10.1177/1069072710382615

Perez, T., Cromley, J. G., & Kaplan, A. (2014). The role of identity development, values, and costs in college STEM retention. *Journal of Educational Psychology*, *106*(1), 315–329. doi:10.1037/a0034027

Perez-Felkner, L., McDonald, S.-K., & Schneider, B. (2014). What happens to high-achieving females after high school? Gender and persistence on the postsecondary STEM pipeline. In I. Schoon & J. S. Eccles (Eds.), *Gender Differences in Aspirations and Attainment* (pp. 285–320). Cambridge, U.K.: Cambridge University Press.

Porter, R. E., Fusarelli, L. D., & Fusarelli, B. C. (2015). Implementing the common core: How educators interpret curriculum reform. *Educational Policy*, *29*(1), 111–139. doi:10.1177/0895904814559248

R Development Core Team. (2014). *R*. Vienna, Austria: R Foundation for Statistical Computing.

Reeve, C. L., & Heggestad, E. D. (2004). Differential relations between general cognitive ability and interest-vocation fit. *Journal of Occupational and Organizational Psychology*, *77*, 385–402. https://doi.org/Doi 10.1348/0963179041752673

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology*, *107*(3), 645–662. doi:10.1037/edu0000012

Rolfhus, E. L., & Ackerman, P. L. (1996). Self-report knowledge: At the crossroads of ability, interest, and personality. *Journal of Educational Psychology*, *88*(1), 174–188. https://doi.org/10.1037/0022-0663.88.1.174

Rounds, J., & Su, R. (2014). The nature and power of interests. *Current Directions in Psychological Science*, *23*(2), 98–103. doi:10.1177/0963721414522812

Schäfer, J. L. (1997). *Analysis of incomplete multivariate data: Monographs on statistics and applied probability: Vol. 72*. New York, NY: Chapman & Hall.

Scheerens, J., & Hendriks, M. (2014). State of the art of time effectiveness. In *SpringerBriefs in Education. Effectiveness of Time Investments in Education* (pp. 7–29). Cham: Springer International Publishing.

Schiefele, U., Krapp, A., & Winteler, A. (1992). Interest as a predictor of academic achievement: A meta-analysis of research. In S. Hidi (Ed.), *The role of interest in learning and development* (pp. 183–212). Book Section, Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Schnabel, K. U., Alfeld, C., 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*(2), 178–198. doi:10.1006/jvbe.2001.1863

Schoon, I., & Eccles, J. S. (Eds.). (2014). *Gender differences in aspiration and attainment: A life course perspective*. Cambridge, U.K.: Cambridge University Press.

Schurtz, I. M., Pfost, 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. doi:10.1016/j.learninstruc.2014.08.001

Schwanzer, A. D., Trautwein, U., Lüdtke, O., & Sydow, H. (2005). Entwicklung eines Instruments zur Erfassung des Selbstkonzepts junger Erwachsener. *Diagnostica*, *51*(4), 183–194. doi:10.1026/0012-1924.51.4.183

Sells, L. W. (1980). Mathematics: The invisible filter. *Engineering Education*, *70*, 340–341.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental designs for generalized causal inferences*. Berkeley: Houghton Mifflin.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Los Angeles, CA: Sage.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*.

doi:10.3102/1076998616646200

Strong, E. K. J. (1943). *Vocational interests of men and women*. Stanford: Stanford University Press.

Su, R., & Rounds, J. (2015). All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields. *Frontiers in Psychology, 6*(FEB), 1–20. doi:10.3389/fpsyg.2015.00189

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, *135*(6), 859–884. doi:10.1037/a0017364

Trautwein, U., Köller, O., Lüdtke, O., & Baumert, J. (2005). Student tracking and the powerful effects of opt-in courses on self-concept: Reflected-glory effects do exist after all. In H. W. Marsh, R. Craven, & D. M. McInerney (Eds.), *New frontiers for self research* (pp. 307–327). Greenwich, CT: Information Age Press.

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. doi:10.1037/0022-0663.98.4.788

Trautwein, U., Neumann, M., Nagy, G., Lüdtke, O., & Maaz, K. (Eds.). (2010). *Schulleistungen von Abiturienten: Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand.* Wiesbaden: VS Verl. für Sozialwissenschaften.

Updegraff, K. A., Eccles, J. S., Barber, B. L., & O’Brien, K. M. (1996). Course enrollment as self-regulatory behavior: Who takes optional high school math courses? *Learning and Individual Differences*, *8*, 239–259. doi:10.1016/S1041-6080(96)90016-3

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204. doi:10.1037/a0036620

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. *Zeitschrift Für Erziehungswissenschaft*, *17*(2), 345–369. doi: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. doi:10.1037/11706-000

Watt, H. M. G., Eccles, J. S., & Durik, A. M. (2006). The leaky mathematics pipeline for girls. *Equal Opportunities International*, *25*(8), 642–659. doi:10.1108/02610150610719119

Watt, H. M. G., Richardson, R. W., & Devos, C. (2013). (How) Does Gender Matter in the Choice of a STEM Teaching Career and Later Teaching Behaviours? *International Journal of Gender, Science and Technology*, *5*(3), 187–206. Retrieved from http://genderandset.open.ac.uk/index.php/genderandset/article/viewArticle/331

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, *30*(1), 1–35. doi:10.1016/j.dr.2009.12.001

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. doi:10.1006/ceps.1999.1015

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Journal of Educational Psychology*, *89*(3), 451–469. doi:10.1037/0022-0663.89.3.451

Wigfield, A., Tonks, S., & Klauda, S. T. (2009). Expectancy-value theory. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 55–57). New York, NY: Routledge.

Young, T., & Lewis, W. D. (2015). Educational policy implementation revisited. *Educational Policy*, *29*(1), 3–17. doi:10.1177/0895904815568936

Table 1

Descriptive Statistics for the Two Cohorts

Variable	TOSCA 2002	TOSCA 2006	Effect size	*p*
Gender (% female)	54.0%	53.1%	0.98	.679
Immigration background (%)	20.0%	20.8%	1.08	.237
HISEI	59.16 (15.57)	58.49 (15.73)	-0.04	.120
Books	5.64 (1.22)	5.57 (1.23)	-0.06	.021
Age	19.56 (0.79)	19.40 (0.65)	-0.22	< .001
Math achievement	50.10 (9.82)	51.07 (9.42)	0.10	.002
Math self-concept	2.76 (0.81)	2.70 (0.85)	-0.08	.003
Realistic vocational interests	2.08 (0.74)	2.24 (0.80)	0.20	< .001
Investigative vocational interests	2.60 (0.83)	2.64 (0.81)	0.04	.138

Note. Weighted results. For dichotomous dependent variables, logistic regression was used to test the differences. For continuous dependent variables, linear regression was used. HISEI = highest international socioeconomic index. Effect sizes: for dichotomous dependent variables, odds ratios (ORs) are displayed; for continuous dependent variables, Cohen’s *d* (Cohen, 1988) is displayed.

Table 2

Time Allocated to Mathematics Before and After the Reform

	Enrollees in advanced math courses (TOSCA 2002)	Average allocated time (TOSCA 2002)	Average allocated time (TOSCA 2006)	Increase (TOSCA 2006)	*p*
Young man	44.7%	3.92 hr (177 min)	4 hr (180 min)	3.49 min	.007
Young women	27.9%	3.56 hr (160 min)	4 hr (180 min)	19.68 min	< .001
Total	35.5%	3.73 hr (167 min)	4 hr (180 min)	12.17 min	< .001

Note. Results for TOSCA 2002 are based on self-reported course choice. The analyses took into consideration the survey weights and clustered structure of the data. One lesson lasted for 45 min. In TOSCA 2006, the average time allocated by young men and women was equal because of the mandatory advanced course.

Table 3

Predicting Advanced Mathematics Achievement: Results from Multiple Regressions Models

Predictor	B	*p*	SE	*d*[a]
Cohort (T2 = 1)	.00	.988	0.06	0.00
Gender (f = 1)	-.58	< .001	0.04	-0.58
HISEI	.00	.912	0.01	0.00
Books	.07	< .001	0.01	0.06
Immigration background (=1)	-.16	< .001	0.03	-0.16
Age	-.18	< .001	0.02	-0.24
School type (VS =1)	-.61	< .001	0.06	-0.61
Cohort × Gender	.14	.025	0.06	0.14
Cohort × School Type	.09	.255	0.07	0.08
Cohort × Gender × School Type	-.19	.029	0.09	-0.19
R^2		.23		

Note. All coefficients are fully standardized. Continuous variables are centered. T2 = TOSCA 2006; HISEI = highest international socioeconomic index; VS = vocational school.

[a] The dependent variable is standardized (Cohen, 1988).

Table 4

Predicting Advanced Math Self-Concept: Results from Multiple Regressions Models

Predictor	B	*p*	SE	*d*[a]
Cohort (T2 = 1)	-.04	.433	0.04	-0.04
Gender (f = 1)	-.29	< .001	0.03	-0.29
HISEI	.01	.370	0.01	0.00
Books	.05	< .001	0.01	0.04
Immigration background (=1)	.00	.925	0.03	0.00
Age	-.13	< .001	0.02	-0.18
School type (VS =1)	.06	.131	0.04	0.06
Cohort × Gender	-.16	< .001	0.06	-0.16
Cohort × School Type	-.03	.619	0.06	-0.03
Cohort × Gender × School Type	.11	.157	0.08	0.11
R^2		.05		

Note. All coefficients are fully standardized. Continuous variables are centered. T2 = TOSCA 2006; HISEI = highest international socioeconomic index; VS = vocational upper secondary school.

[a] The dependent variable is standardized (Cohen, 1988).

Table 5

Predicting Realistic Vocational Interests: Results from Multiple Regressions Models

Predictor	B	*p*	SE	*d*[a]
Cohort (T2 = 1)	.27	< .001	0.04	0.27
Gender (f = 1)	-.84	< .001	0.03	-0.84
HISEI	-.04	.004	0.01	0.00
Books	.04	.001	0.01	0.03
Immigration background (= 1)	-.08	.002	0.03	-0.08
Age	-.03	.013	0.01	-0.05
School type (VS = 1)	.09	.099	0.06	0.09
Cohort × Gender	-.15	.007	0.06	-0.15
Cohort × School Type	.00	.932	0.07	0.01
Cohort × Gender × School Type	.00	.948	0.09	0.01
R^2		.22		

Note. All coefficients are fully standardized. Continuous variables are centered. T2 = TOSCA 2006; HISEI = highest international socioeconomic index; VS = vocational upper secondary school.

[a] The dependent variable is standardized (Cohen, 1988).

Table 6

Predicting Investigative Vocational Interests: Results from Multiple Regressions Models

Predictor	B	*p*	SE	*d*[a]
Cohort (T2 = 1)	.11	.01	0.04	0.11
Gender (f = 1)	-.62	< .001	0.03	-0.62
HISEI	.00	.745	0.01	0.00
Books	.11	< .001	0.01	0.09
Immigration background (= 1)	.01	.668	0.03	0.01
Age	-.03	.045	0.01	-0.04
School type (VS = 1)	.07	.142	0.05	0.07
Cohort × Gender	-.12	.019	0.05	-0.12
Cohort × School Type	-.05	.371	0.06	-0.05
Cohort × Gender × School Type	.02	.770	0.08	0.02
R²		.12		

Note. All coefficients are fully standardized. Continuous variables are centered. T2 = TOSCA 2006; HISEI = highest international socioeconomic index; VS = vocational upper secondary school.

[a] The dependent variable is standardized (Cohen, 1988).

Table 7

Predicting Field of Study at University: Results from Multiple Logistic Regressions Models

Predictor	OR	CI		*p*
Cohort (T2 = 1)	0.97	0.85	1.11	.702
Gender (f = 1)	0.37	0.32	0.42	< .001
HISEI	0.90	0.82	0.98	.022
Books	0.90	0.82	0.99	.037
Immigration background (= 1)	0.97	0.88	1.07	.538
Age	0.83	0.75	0.92	< .001
School type (VS = 1)	1.08	0.90	1.31	.411
Cohort × Gender	1.02	0.86	1.21	.838
Cohort × School Type	1.01	0.86	1.20	.871
Cohort × Gender × School Type	1.01	0.87	1.16	.948
Pseudo-R^2		.24		

Note. The table displays standardized results where mathematics, engineering, computer science, and physics were coded as STEM subjects. Odds ratios significantly larger than 1 indicate a higher likelihood of studying STEM subjects. T2 = TOSCA 2006; HISEI = highest international socioeconomic index; VS = vocational upper secondary school.

Table 8

Structure Coefficients for Multiple Linear Regression Models

Predictor	Advanced mathematics	Math self-concept	Realistic interests	Investigative interests
Cohort (T2 = 1)	0.11	-0.18	0.21	0.05
Gender (f = 1)	-0.54	-0.74	-0.96	-0.95
HISEI	0.28	0.22	-0.00	0.15
Books	0.32	0.25	0.03	0.28
Immigration (= 1)	-0.25	-0.14	-0.09	-0.07
Age	-0.49	-0.50	-0.02	-0.03
School type (VS = 1)	-0.70	-0.02	0.08	0.00
Cohort × Gender	-0.23	-0.61	-0.47	-0.55
Cohort × School Type	-0.42	-0.07	0.14	0.00
Cohort × Gender × School Type	-0.47	-0.25	-0.20	-0.28

Note. The table displays structure coefficients (e.g., Courville & Thompson, 2001) for each predictor of all four multiple linear regression models. T2 = TOSCA 2006; HISEI = highest international socioeconomic index; VS = vocational upper secondary school.

Figure 1. Schematic illustration of the study's timeline. All data in Wave 1 were collected at the end of upper secondary school.

Measures	2002	2004	2006	2008
Math achievement Math self-concept Vocational interests	TOSCA 2002 Wave 1 *N* = 4,730		TOSCA 2006 Wave 1 *N* = 4,715	
Field of study at university	↳	TOSCA 2002 Wave 2 *N* = 2,318	↳	TOSCA 2006 Wave 2 *N* = 2,852

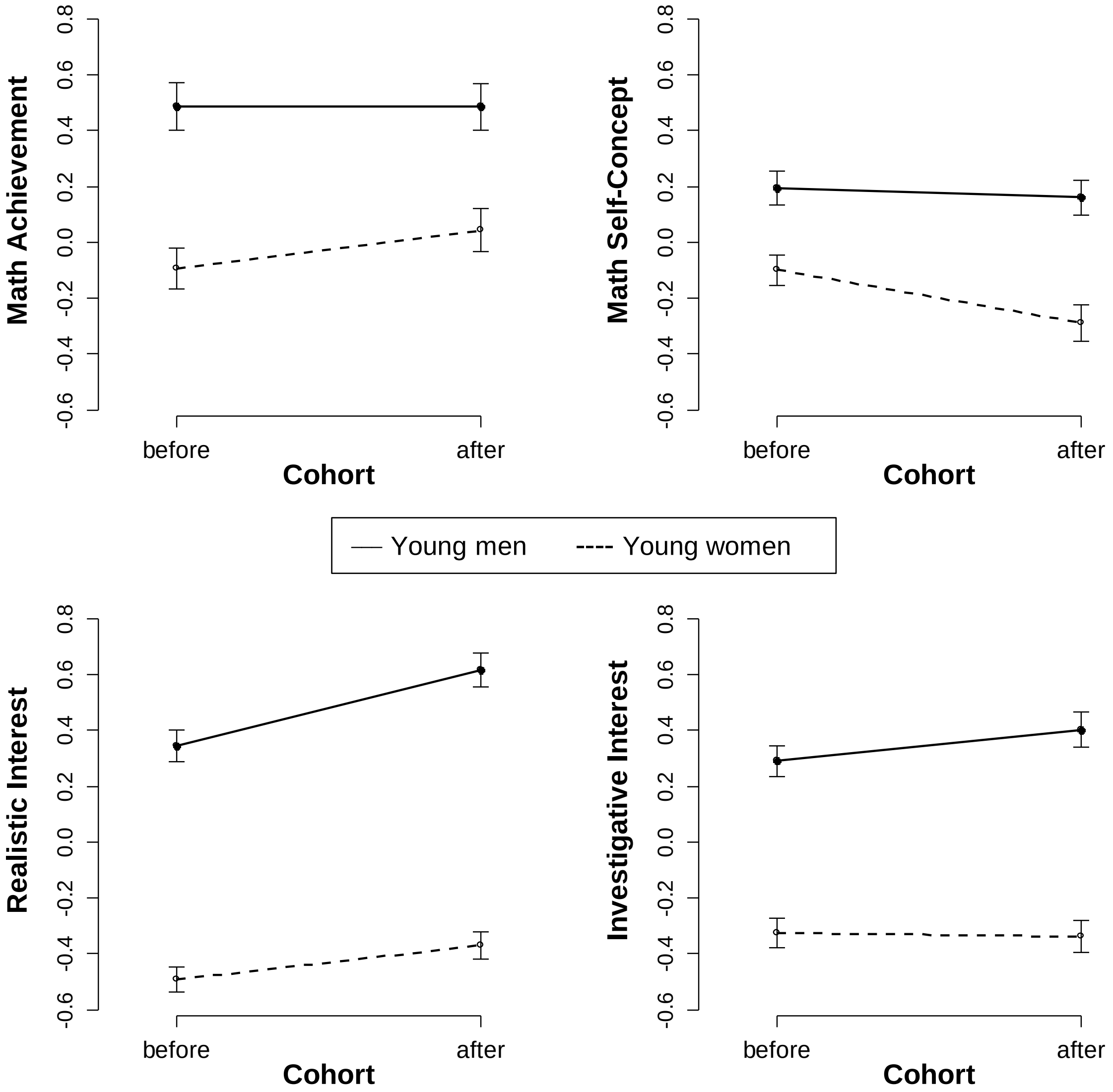

Figure 2. Plots of the moderating effect of gender on the relation between reform and math achievement, math self-concept, realistic interests, and investigative interests with 95% confidence intervals. The dependent variables are presented in standard deviation units.

4.2 Study 2

Hübner, N., Wagner, W., Nagengast, B., & Trautwein, U. (2017). Putting all students in one basket does not produce equality: Gender-specific effects of curricular intensification in upper secondary school on achievement and motivation. Manuscript submitted for publication.

The following manuscript has not been accepted and published yet. This version of the manuscript has been submitted to *School Effectiveness and School Improvement* on January 18th, 2017. The Taylor & Francis Group will have the copyright of the final version of the article, if it will be accepted. The version displayed here may not exactly replicate the final version published in the journal. It is not the copy of record.

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 course work, 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 one 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 differential effects on English reading and a higher English self-concept in favor 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; Osborne & Dillon, 2008; Stein, Kaufman, Sherman, & Hillen, 2011). 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, Domina, Penner, & Conley, 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. Hence, going beyond prior research and using representative data, we report effects of a state-wide 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: A Definition

CI can involve different elements. Conceptually, we differentiate between four aspects. 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, McEachin, Penner, & Penner, 2015; Penner et al., 2015) and that CI might therefore help students overcome the negative side-effects of tracking on low-track students' achievement (e.g. Hanushek & Woessmann, 2006; Lee & Bryk, 1988) and opportunities to learn in general (c.f. 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 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).

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, studies on the effects of CI have usually examined 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., 2014; 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). 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). 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).

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, preuniversity 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 TIMSS study in 1998 (Mullis et al., 1998) were the starting point 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 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.

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 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 3).

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 2).

Third, the curriculum in these five subjects resembled the advanced-course curriculum from before the reform (c.f. 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 2).

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 designed 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.

Second, not only did we include achievement measures in our evaluation, but we also analyzed potential effects on motivational variables. Motivation plays a major role in further achievement and should be sensitive to aspects of CI such as changing classroom composition. 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.

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 2, CI went along with mandatory course enrollment in German, mathematics, one foreign language, and one science subject on an advanced level. On the basis of this, 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.

Method

Description of Study and Sample

We used data from the Additional Study Thuringia (Blossfeld, Rossbach, & 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 = .68).

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 (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, 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 α = .94; English: α = .94; biology: α = .93; physics: α = .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 showed sufficient internal consistencies in previous studies (e.g., Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006). The scales (e.g., "When I do mathematics, I sometimes get totally absorbed"; "Mathematics is simply an exciting subject") showed comparable internal consistencies in our study (math: Cronbach's α = .88; English: α = .86; biology: α = .91; physics: α = .93).

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 3), 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- and one-parameter 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.

As recommended by McNeish, Stapleton, and Silverman (2016), 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)

Results

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 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

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 3). 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 3). 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. 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

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. 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 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.

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 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 meaningful. Controlling the FDR did not change any of these results.

Gender-Specific Achievement before and after the Reform

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 (*d* = -0.10, *p* = .009) but not in the unadjusted models (*d* = -0.23, *p* = .066), 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).

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 x 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 noncognitive constructs: domain-specific self-concept and domain-specific interest. First, we did not find any gender 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 or young women before versus after the reform. However, in all subjects except mathematics, all gender differences within a cohort were statistically significant (see supplemental material for further information).

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 a potentially huge share 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. Our results suggest that the reform seems to have somewhat of an adverse effect on self-concept: As the reference group of the students who would have chosen the basic courses if given the choice (e.g., young women) improved, math self-concept for this group was lower 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.

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). 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).

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. To sum up, the findings indicate that young men's achievement and self-concept in English reading was higher after the reform, whereas young women mostly showed a lower self-concept in math after the reform. The study underscores the importance of carefully planning systemic reforms and strengthens the importance of conducting systematic evaluations during processes of educational change.

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/10.3102/0162373709343471

Baumert, J. (Ed.). (2000). *TIMSS-III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. 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. 2. Competencies in Mathematics and Physics at the End of Upper Secondary School]*. Opladen: Leske u. 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*, *57*(1), 289–300.

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

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/10.3102/0162373709359805

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Thomson Learning.

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/10.3102/0002831213489843

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Cortes, K. E., Goodman, J. S., & Nomi, T. (2015). Intensive math instruction and educational attainment. *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, New Jersey: Wiley. https://doi.org/10.1002/9781118963418.childpsy407

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/10.3102/0002831211426347

Eberle, F., Gehrer, K., Jaggi, B., Kottonau, J., Oepke, M., & Pflüger, M. (2008). *Evaluation der Maturitätsreform 1995. Schlussbericht zur Phase II [Evaluation of the Upper Secondary School Reform of 1995. Final report for the 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: 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

Enders, C. K. (2010). *Applied missing data analysis*. *Methodology in the social sciences*. New York: Guilford Press.

Evans, M., Kelley, J., Sikora, J., & Treiman, D. J. (2010). Family scholarly culture and educational success: Books and schooling in 27 nations. *Research in Social Stratification and Mobility*, *28*(2), 171–197. https://doi.org/10.1016/j.rssm.2010.01.002

Ganzeboom, H. B. G., & Treiman, D. J. (2003). Three internationally standardised measures for comparative research on occupational status. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in Cross-National Comparison* (pp. 159–193). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-9186-7

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research.* Washington, DC, US: American Psychological Association.

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. (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

Hanushek, E. A., & Woessmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal*, *116*(510), 63-76. https://doi.org/10.1111/j.1468-0297.2006.01076.x

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

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 in STEM by minimizing course choice options? Effects of obligatory coursework in math on gender differences in STEM. *Journal of Educational Psychology.* Advance online publication. 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

Jansen, M., Schroeders, 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

Kiefer, T., Robitzsch, A., & Wu, M. (2017). *TAM: Test analysis modules. R package version 1.99999-31*. Retrieved from http://CRAN.R-project.org/package=TAM

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal*, *125*(588), 397-424. 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. https://doi.org/10.2307/2112266

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale: Lawrence Erlbaum Associates.

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/10.3102/00028312023001129

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J. S., Abdelfattah, F., Nagengast, B.,. . . Abu-Hilal, M. M. (2014). 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

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2016). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods.* Advance online publication. 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., Martin, M. O., Beaton, A. E., 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.

Muthén, B., & Muthén, L. K. (1998-2012). *Mplus User's Guide. Seventh Edition*. Los Angeles and CA.: Muthén & Muthén.

NEPS. (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

Nomi, 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

OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: OECD.

OECD. (2015). *Education policy outlook 2015*. Paris: OECD Publishing.

Osborne, J., & Dillon, J. (2008). *Science education in europe: Critical reflections: A report to the Nuffield Foundation*. Retrieved from http://efepereth.wdfiles.com/local-files/science-education/Sci_Ed_in_Europe_Report_Final.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 multicontext 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.ssresearch.2015.03.008

R Development Core Team. (2016). R. Vienna, Austria: R Foundation for Statistical Computing.

Reynolds, D., Sammons, P., de Fraine, 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

Robitzsch, A. (2015). sirt: Supplementary Item Response Theory Models. R package version 1.8-9. Retrieved from http://CRAN.R-project.org/package=sirt

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

Standards-Based Assessment Tasks for English as a First Language: Vol. 1. Münster: Waxmann.

Scheerens, J. (Ed.). (2014). *Effectiveness of time investments in education*. Cham: Springer International Publishing.

Schurtz, I. M., Pfost, 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

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

Sørensen, A. B. (1970). Organizational differentiation of students and educational opportunity. *Sociology of Education, 43*(4), 355. https://doi.org/10.2307/2111838

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, 4*(81), 453–492. https://doi.org/10.3102/0034654311423025

The National Commission on Excellence. (1983). *A nation at risk: The imperative for educational reform*. Washington: Government Printing Office.

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* (pp. 467–501). Reinbek bei Hamburg: Rowohlt.

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*(S2), 233–249. https://doi.org/10.1007/s11618-011-0196-1

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

Table 1

Subject-Specific Dimensions of CI

Dimension of CI	Tracking	Instruction time	Curriculum standards	Importance after the reform
Math	Complete detracking	Reduced on average	Increased on average	High for all students
English	Almost complete detracking	Stable on average	Increased on average	High for 94.9% of the students
Physics	Partial detracking	Stable on average	Increased on average	High for 31.5% of the students
Biology	Partial detracking	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 2 and 5. Although instructional 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).

Table 2

Typical Timetable for Students Before and After the Upper Secondary School Reform

	Before the reform (2010)			After the reform (2011)		
Final examination subject no.	Subject	Time	Level	Subject	Time	Level
1	G/M	6	Advanced	G	4	Advanced
2	FL/S/SS	6	Advanced	M	4	Advanced
3	M/G	4	Basic	FL	4	Advanced
4	FL(CS)	3	Basic	S	4	Advanced
5		-		SS	4	Advanced

Note. Typical timetable for students in upper secondary school before and after the reform. Time = Instructional time in hours per week; Level = Level of instruction based on advanced or basic curriculum. G = German, M = Mathematics, FL = Foreign Language, S = Sciences (biology, chemistry, and physics), SS = Social Sciences, CS = Computer Sciences. Choice options are indicated by multiple subjects in a cell.

Table 3

Sample Sizes and Course Choices for Students Before and After the Reform

	N	Participation of the drawn sample in %	Course choice in %		
			AC	BC	Dropout
Cohort 1	1,316	70.9			
Mathematics			45.5	54.5	-
English			32.1	62.6	5.4
Physics			10.5	30.5	59.0
Biology			20.3	48.7	30.9
Cohort 2	886	63.6			
Mathematics			100.0	-	-
English			94.9	3.3	1.8
Physics			31.5	23.2	45.3
Biology			58.5	15.8	25.7

Note. AC = Advanced course; BC = Basic course; Dropout = No selection of course. All differences in AC proportions before and after the reform were statistically significant (AC: $p < .001$; BC: $p < .01$). Only dropout rates for biology and English did not differ significantly. Differences for mathematics were not tested because the advanced course was mandatory after the reform. If differences were not statistically significant after the BH correction, they were labeled with ^{BH}. The results are from analyses in which the weights and cluster structure of the data were taken into consideration.

Table 4

Choice of Advanced Courses by Gender

	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

	Cohort 1	Cohort 2	Difference	*p*	p_{adj}
Mathematics	4.91	4	-0.91		-
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. 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.

Table 6

Mean Levels of Student Achievement by Course and by Gender

Mean levels of student achievement by course				
	Unadjusted results of unidimensional 1PL IRT model			
	Cohort 1		Cohort 2	
	AC$_{a}$	BC$_{b}$	CS/AC$_{c}$	BC$_{d}$
Mathematics	55.1$_{bc}$	44.7(0.9)$_{ac}$	50.1 (0.9)$_{ab}$	-
English reading	52.8$_{bc}$	45.8(0.9)$_{acd}$	48.9 (0.9)$_{ab}$	52.4 (3.6)$_{b}$
Physics	60.5$_{bcd}$	45.3(3.0)$_{acd}$	52.8 (2.4)$_{abd}$	41.4 (2.4)$_{abc}$
Biology	54.0$_{bcd}$	47.7(1.4)$_{a}$	49.2 (1.5)$_{a}$	49.1 (1.6)$_{a}$
Adjusted results of unidimensional 1PL IRT model				
Mathematics	52.7$_{bc}$	47.2 (0.6)$_{ac}$	50.1 (0.7)$_{ab}$	-
English reading	53.5$_{bcd}$	47.2 (0.8)$_{acd}$	49.9 (0.7)$_{ab}$	45.0 (0.9) $_{ab}$
Physics	57.2$_{bd}$	47.1 (2.7)$_{acd}$	52.1 (2.2)$_{bd}$	43.6 (2.1)$_{abc}$
Biology	54.0$_{bcd}$	47.9 (1.0)$_{ac}$	49.9 (1.1)$_{ab}$	48.1 (1.7)$_{a}$
Mean levels of student achievement by gender				
	Unadjusted results of unidimensional 1PL IRT model			
	Cohort 1		Cohort 2	
	Young men$_{a}$	Young women$_{b}$	Young men$_{c}$	Young women$_{d}$
Mathematics	52.7$_{bd}$	46.6 (0.1)$_{ac}$	53.9 (1.2)$_{bd}$	46.8 (1.0)$_{ac}$
English reading	48.1b$_{bcd}$	50.6 (0.9)$_{a}$	50.5 (0.8)$_{a}$	50.7 (0.8)$_{a}$
Physics	54.0$_{bd}$	45.4 (0.9)$_{ac}$	53.9 (1.5)$_{bd}$	46.7 (1.3)$_{ac}$
Biology	49.4	50.2 (0.8)	49.8 (1.0)	50.6 (0.9)
Adjusted results of unidimensional 1PL IRT model				
Mathematics	52.1$_{bd}$	47.6 (0.6) $_{ac}$	53.2 (1.0)$_{bd}$	47.2 (0.7) $_{ac}$
English reading	47.6$_{bcd}$	51.4 (0.6) $_{ac}$	48.4 (0.6)$_{ab}$	51.2 (0.6) $_{a}$
Physics	53.6$_{bd}$	46.1 (0.9) $_{ac}$	53.3 (1.3)$_{bd}^{BH}$	47.0 (1.1) $_{ac}^{BH}$
Biology	49.0	51.0 (0.7)	49.1 (0.8)	50.9 (0.8)

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. 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.

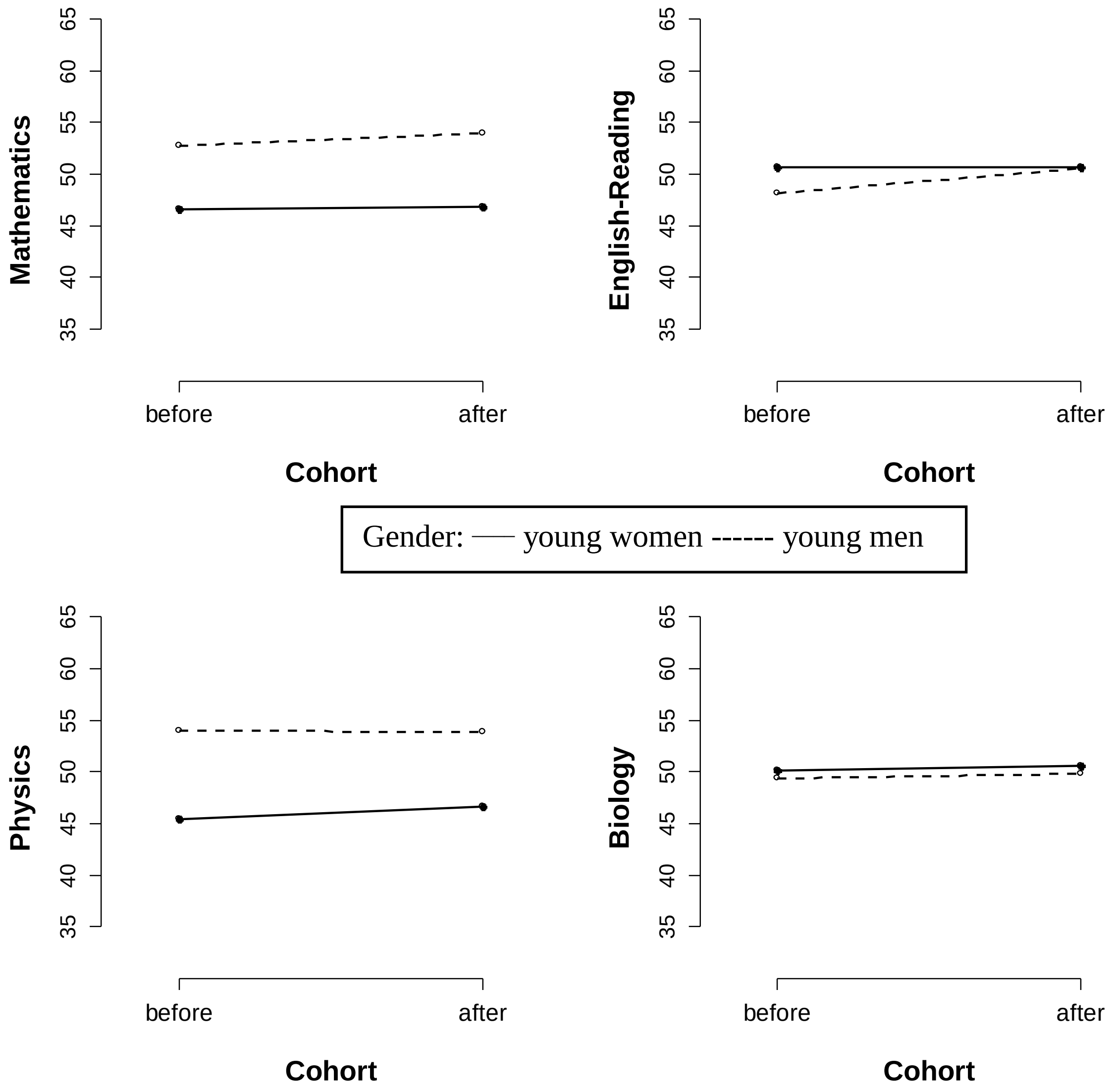

Figure 1. Student achievement in mathematics, English reading, physics, and biology by gender and cohort.

Supplemental online material

Additional information about the instruments

Competence in mathematics. The test for measuring mathematics competence focused on *mathematical literacy*, which is also referred to in the assessment of education standards and PISA (e.g., OECD, 2004). The test differentiates between four content areas: Quantity, Space and Shape, Change and Relationships, and Date and Chance. These areas are tested with regard to six different cognitive components: applying technical skills, modeling, arguing, communicating, representing, and problem solving. Competence in mathematics was assessed with 40 items in a multiple-choice or open format. Students had 30 min to work on this part of the test. Reliability was acceptable (reliability of the weighted likelihood estimator: WLE = .68).

Competence in English reading. The English reading test was based on items that were developed by the Institute for Educational Quality Improvement (IQB; Rupp et al., 2008). These items are aligned with the Common European Framework of Reference for Languages (CEF) as well as with the national education standards for English. Item difficulty ranged from level B1 to level C1 of the CFE. 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). Reliability of this test was good (WLE reliability = .77).

Competence in biology. Competence in biology was measured with items from the EVAMAR II-study (Eberle et al., 2008) in six content areas: cytology/anatomy/metabolism, information processing/behavior, immunbiology, genetics/developmental biology, ecology, and systematics/evaluation. Competencies in these areas were measured with regard to three cognitive requirements: reproducing and practice application, restructuring and transferring, and evaluating and reforming problems. Students had 45 min to work on a subset of 18 items out of 126 overall items, which were presented in multiple-choice and open-answer formats (NEPS, 2011). Reliability of this test was acceptable (WLE reliability = .61).

Competence in physics. Students had 45 min to answer 55 items (17-18 items in each booklet) that measured their competence in physics. Items were presented mostly in a multiple-choice format, whereas some were in a forced-choice or open format. Some items were taken from the TIMSS study (Baumert, 2000), and some were developed for the NEPS Additional Study Thuringia. Construction of items was aligned with the Requirements of Final Examinations (Abitur) for physics. Reliability was moderate (WLE reliability = .55). We discovered in a later IRT analysis that we had to exclude several items due to negative item discriminations. We removed seven items from the biology and physics competence test and one item from the English reading test. Negative item discriminations imply that students with lower abilities, on average, have a higher probability of correctly responding to this item than students with higher abilities and can thereby be an indicator of poor item quality. The results did not differ statistically significantly when these items were removed.

Domain-specific self-concept. Domain-specific self-concept was measured with a four-item test. They were adapted for mathematics, English, biology, and physics, respectively. All items were translated and partly modified on the basis of the Self-Description Questionnaire III (Marsh & O'Neill, 1984). The internal consistency 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 α = .94; English: α = .94; biology: α = .93; physics: α = .93). Negatively formulated items were reverse coded.

Domain-specific interest. Domain-specific interest was measured with and Eccles and Wigfield's (2002) four-item test, which was adapted for mathematics, English, biology, and physics. The scales showed sufficient internal consistencies in previous studies (e.g., Trautwein et al., 2006). Scales (e.g., "When I do mathematics, I sometimes get totally absorbed"; "Mathematics is simply an exciting subject") showed comparable internal reliabilities in our study (Math: Cronbach's α = .88; English: α = .86; biology: α = .91; physics: α = .93).

We controlled for additional variables in the adjusted models.

Socioeconomic background. The social status of the students' family was assessed with the International Socio-Economic Index of Occupational Status 2008 (ISEI-08; Ganzeboom & Treiman, 2003). The highest value of the ISEI in the family was used to characterize the socioeconomic background of the students.

Number of books available at home. The number of books available at home was measured on a 7-point scale ranging from zero books available to more than 500 books available. This variable has been shown to be a good indicator of the cultural capital of the family (e.g., Evans, Kelley, Sikora, & Treiman, 2010).

In addition, *migration background* was controlled for. Students with at least one parent born outside of Germany were coded as students with a migration background.

Cognitive ability. Cognitive ability was measured with the revised version of the test of cognitive skills for Grades 4 to 12 (KFT 4-12 + R; Heller & Perleth, 2000). This test is based on the idea of overall cognitive performance and on the Lorge-Thorndike-Intelligence-Test (c.f. NEPS, 2011). The KFT 4-12 + R measures three different cognitive dimensions: verbal, quantitative, and figural-spatial dimensions. The verbal and quantitative subscales both consist of 20 items, whereas the figural-spatial subscale consists of 25 items. All items were presented in a multiple-choice format, and students were allowed to work on them for 24 min. The reliability of overall cognitive ability was good (WLE reliability = .80).

Additional information about the statistical analysis

In order to assess the sensitivity of our results to potential (even though unexpected due to the natural experimental design of the study) differences between cohorts on relevant background characteristics and the robustness of results, we specified 1PL and 2PL models and conducted analyses in which we did and did not control for the covariates. In addition, we specified models without items with severe DIF (C) to check the robustness of our results. This was done because DIF does not necessarily imply that the respective items are “unfair” (item bias) but may also reflect valid differences between subgroups (item impact; Zumbo, 1999). We considered, sex, migration background, socioeconomic background, number of books available at home, cognitive skills, and information on grade repetition as covariates. Adjusted models allow the user to compare possible differences after controlling for group differences on these covariates. We centered all covariates in the adjusted models at their grand mean. Differences between group-specific means on the covariates therefore may be interpreted in terms of “deviations” from the average student composition. Potential differences between cohorts on a covariate were “adjusted” by regressing the dependent variables on the covariates with the regression weights freely estimated in each group. Intercepts in these models represented expected (or adjusted) group means for the average student composition under the assumption of potential cohort-specific (linear) relationships between the covariates and the dependent variables. We estimated means and standard errors on the dependent variables in unadjusted models and intercept differences and standard errors in adjusted models. Another advantage of these adjusted models over the unadjusted model was increased power due to a reduction in residual variance. We tested for interaction effects (e.g., Cohort × Course Level and Cohort × Sex) by applying the delta method (e.g., Casella & Berger, 2002), implemented in the model constraint option in Mplus. We usually reported two-sided *p*-values and used the Benjamini-Hochberg (BH) method to correct for multiple testing (Benjamini & Hochberg, 1995). For multidimensional models, we calculated adjusted *p*-values that were based on all group comparisons of all dependent variables, whereas for the unidimensional models, the adjusted *p*-values resulted from calculations for each respective dependent variable. Different specifications did not statistically significantly change the results presented here. We indexed the *p*-values that were not statistically significant after the BH correction in complex models.

For ease of interpretation, we linearly transformed the resulting parameters to a metric with M = 50 and SD = 10. Differences between students or student groups are given in standard deviation units, also referred to as Cohen’s *d* (Cohen, 1988).Results for domain-specific self-concept and domain-specific interest are based on simple structural equation models in which the indicators are assumed to be metric. Additional information about the statistical analysis

Achievement outcomes were analyzed with unidimensional and multidimensional two- and one-parameter logistic item response theory (IRT) models. In the unidimensional two-parameter logistic (2PL) IRT model, the probability that a person *s* will solve item *i* is given as follows:

$$\mathrm{P}(\mathrm{X}_{\mathrm{is}} = 1|\theta_{\mathrm{s}}, \alpha_{\mathrm{i}}, \beta_{\mathrm{i}}) = \frac{\exp(\alpha_{\mathrm{i}}(\theta_{\mathrm{s}} - \beta_{\mathrm{i}}))}{1 + \exp(\alpha_{\mathrm{i}}(\theta_{\mathrm{s}} - \beta_{\mathrm{i}}))} \qquad (1)$$

with θ_{s} representing a person's trait level, α_{i} representing the discrimination parameter for item *i*, and β_{i} representing the difficulty of item *i*. If all α_{i} are equal, the 2PL model reduces to a 1PL model. The extension from the 2PL logistic model to a case with multiple elements in the θ vector is given as:

$$\mathrm{P}(\mathrm{X}_{\mathrm{is}} = 1|\boldsymbol{\theta}_{\mathbf{s}}, \boldsymbol{\alpha}_{\mathbf{i}}, \beta_{\mathrm{i}}) = \frac{\exp(\boldsymbol{\alpha}_{\mathbf{i}}\boldsymbol{\theta}_{\mathbf{s}}' + \gamma_{\mathrm{i}})}{1 + \exp(\boldsymbol{\alpha}_{\mathbf{i}}\boldsymbol{\theta}_{\mathbf{s}}' + \gamma_{\mathrm{i}})} \qquad (2)$$

where $\boldsymbol{\alpha}$ is a 1 × m vector of item discrimination parameters, and $\boldsymbol{\theta}$ represents an m-dimensional m × 1 vector of person coordinates. The intercept γ_{i} is a scalar of the item's location.

DIF was classified according to the ETS classification system into negligible (A), slight to moderate (B), or moderate to severe (C; Longford, Holland, & Thayer, 1993). In the subsequent analyses, items with negative discriminations were excluded.

The IRT analyses were conducted in R (R Development Core Team, 2016) using the tam (Kiefer, Robitzsch, & Wu, 2017) and sirt (Robitzsch, 2015) packages, whereas the final models were specified in Mplus 7.4 (Muthén & Muthén, 1998-2012). 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 addition, we specified unadjusted models that included the covariates of the adjusted models as auxiliary variables (e.g., Collins, Schafer, & Kam, 2001; Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997).

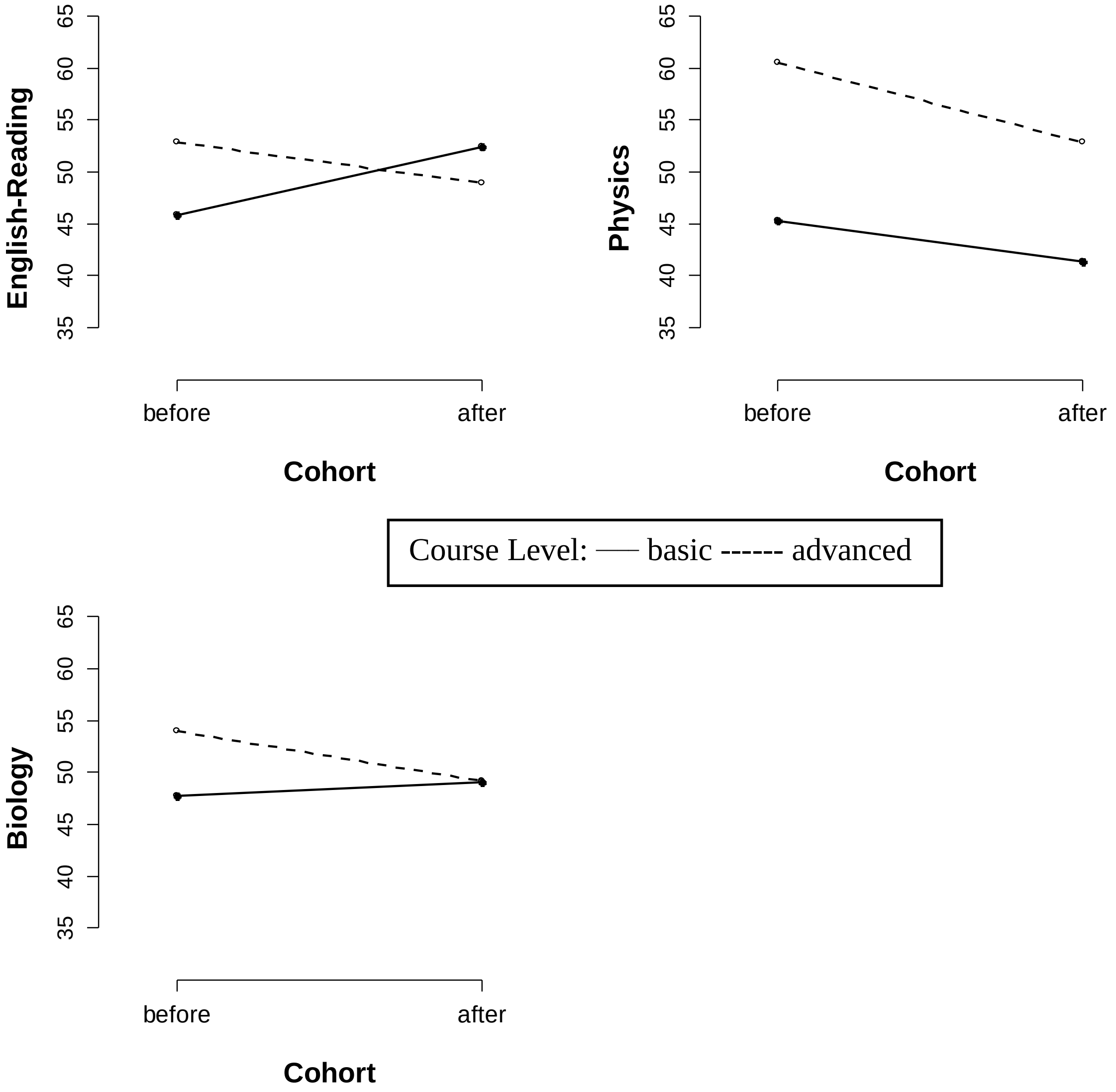

Figure A1. Student achievement in English reading, physics, and biology by course level and cohort.

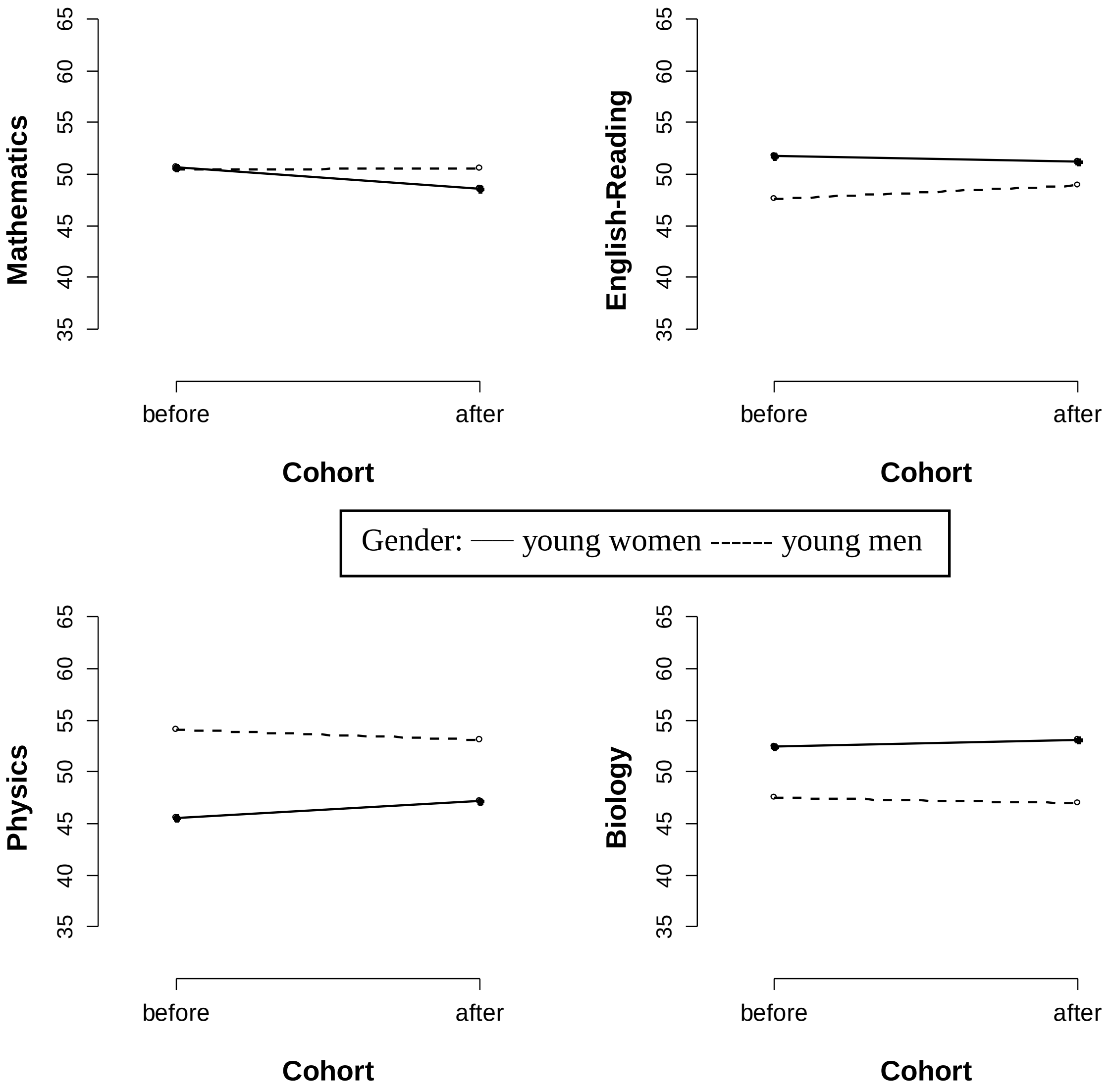

Figure A2. Domain-specific interest in mathematics, English, physics, and biology by gender and cohort.

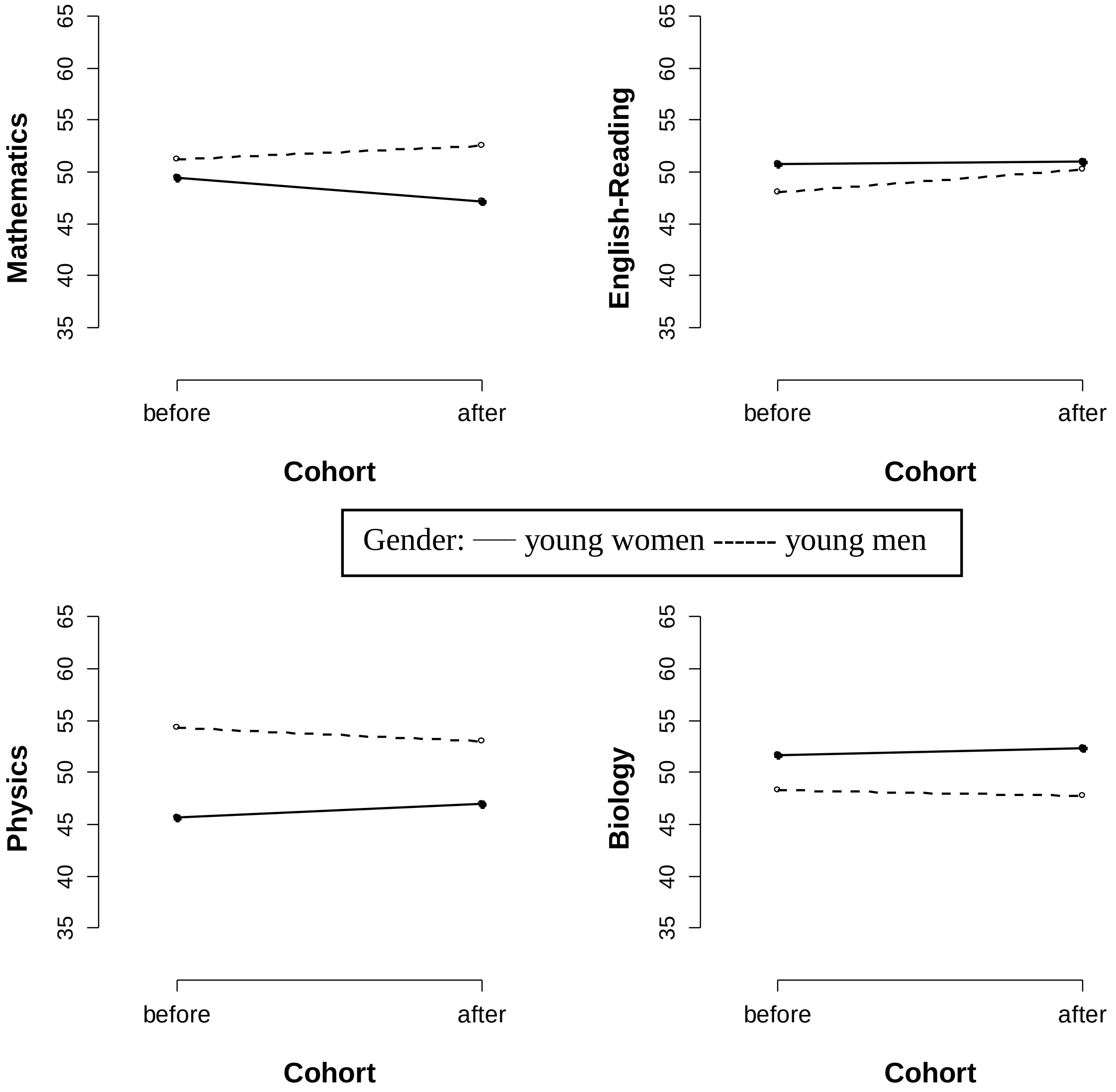

Figure A3. Domain-specific self-concept in mathematics, English, physics, and biology by gender and cohort.

Table A1

Mean Level of Student Achievement

Unadjusted results of multidimensional 1PL IRT model					
	Cohort 1	Cohort 2	Diff in SD	*p*	*p_{adj}*
Mathematics	49.7	50.3(0.8)	0.06	0.399	0.521
English reading	49.4	50.6(0.5)	0.12	0.014	0.056
Physics	49.7	50.3(0.9)	0.04	0.500	0.521
Biology	49.8	50.2(0.6)	0.06	0.521	0.521
Adjusted results of multidimensional 1PL IRT model					
Mathematics	50.0	50.0(0.6)	0.00	0.933	0.956
English reading	49.6	50.4(0.5)	0.08	0.078	0.312
Physics	49.9	50.1(0.8)	0.00	0.745	0.956
Biology	50.0	50.0(0.5)	0.00	0.956	0.956

Note. Cohort 1 = Cohort before the reform; Cohort 2 = Cohort after the reform; p_{adj} = Benjamini-Hochberg-corrected *p*-values. The displayed results are 1PL models in which we did and did not control for differences in further covariates. The metric of the latent variable was transformed to *M* = 50 and *SD* = 10 on the basis of the pooled means and standard deviations. The displayed *p*-values were based on two-tailed tests. The results are from analyses in which the weights and cluster structure of the data were taken into consideration.

Table A2

Descriptive Statistics

	Cohort 1		Cohort 2		Dif.	*p*	p_{adj}
	N	*M*	*N*	*M*			
Gender	1316	0.53	885	0.56	-0.03	0.935	0.935
Migration	1043	0.07	852	0.05	0.02	0.053	0.360
HISEI	796	55.00	723	55.74	-0.74	0.356	0.509
Books 0 – 25	51	1.71	32	1.83	-0.12	0.220	0.440
Books 26 – 200	289	3.54	241	3.58	-0.04	0.493	0.560
Books > 200	638	5.52	588	5.57	-0.05	0.072	0.360
Grade repetition	1314	0.14	885	0.10	0.04	0.303	0.505
Cognitive Ability_V	1291	49.78	863	50.00	-0.22	0.504	0.560
Cognitive Ability_Q	1291	49.46	864	49.84	-0.38	0.137	0.418
Cognitive Ability_N	1290	49.23	864	50.09	-0.86	0.167	0.418

Note. Cohort 1 = Cohort before the reform; Cohort 2 = Cohort after the reform. Dif. = Unstandardized difference between Cohort 1 and Cohort 2. p_{adj} = Benjamini-Hochberg-adjusted *p*-values. Means were estimated with survey weights. Differences for gender, migration, and grade repetition were tested with logistic regression models. Differences for cognitive abilities were tested with a unidimensional IRT model for each ability area. Significance tests were based on analyses that took into consideration the weights and cluster structure of the data. No difference was statistically significant ($p < .05$).

Table A3

Mean Level of Domain-Specific Self-Concept and Interest by Gender

Mean level of domain-specific self-concept by gender				
Unadjusted results of multidimensional structural equation models				
	Cohort 1		Cohort 2	
	Young men$_{a}$	Young women$_{b}$	Young men$_{c}$	Young women$_{d}$
Mathematics	51.2$_{bd}$	49.4 (0.8)$_{acd}$	52.5 (1.1)$_{bd}$	47.2 (1.0)$_{abc}$
English	48.0$_{bcd}$	50.8 (0.6)$_{a}$	50.2 (0.9)$_{a}$	51.0 (0.7)$_{a}$
Physics	54.3$_{bd}$	45.7 (0.8)$_{ac}$	53.0 (1.2)$_{bd}$	47.0 (0.8)$_{ac}$
Biology	48.3$_{bd}$	51.7 (0.9)$_{ac}$	47.7 (1.2)$_{bd}$	52.3 (0.8)$_{ac}$
Adjusted results of multidimensional structural equation models				
Mathematics	51.1 $_{d}$	49.9 (0.3) $_{cd}$	51.8 (1.0) $_{bd}$	47.2 (1.0) $_{abc}$
English	47.8 $_{bcd}$	51.1 (0.4) $_{a}$	49.7 (0.7) $_{a}$	51.4 (0.8) $_{a}$
Physics	54.4 $_{bd}$	46.0 (0.5) $_{ac}$	52.4 (0.9) $_{bd}$	47.1 (0.9) $_{ac}$
Biology	48.2 $_{bd}$	51.7 (0.5) $_{ac}$	47.9 (0.6) $_{bd}$	52.1 (0.9) $_{ac}$
Mean level of domain-specific interest by gender				
Unadjusted results of multidimensional structural equation models				
	Cohort 1		Cohort 2	
	Young men$_{a}$	Young women$_{b}$	Young men$_{c}$	Young women$_{d}$
Mathematics	50.4	50.6 (0.8) $_{d}$	50.5 (1.3)	48.6 (1.1)$_{b}$
English	47.6$_{bd}$	51.7 (0.7)$_{ac}$	48.9 (0.8)$_{bd}$	51.8 (0.8)$_{ac}$
Physics	54.1$_{bd}$	45.6 (0.8)$_{acd}^{BH}$	53.1 (1.3)$_{bd}$	47.2 (0.8)$_{ab}{}^{BH}{}_{c}$
Biology	47.5$_{bd}$	52.5 (0.8)$_{ac}$	47.0 (1.4)$_{bd}$	53.1 (0.6)$_{ac}$
Adjusted results of multidimensional structural equation models				
Mathematics	50.3	51.0 (0.4)$_{d}$	50.1 (1.1)	48.6 (1.1)$_{b}$
English	47.5$_{bd}$	51.8 (0.4)$_{ac}$	48.4 (0.7)$_{bd}$	52.3 (0.9)$_{ac}$
Physics	54.0$_{bd}$	46.1 (0.5)$_{ac}$	52.4 (0.6)$_{bd}$	47.5 (0.9)$_{ac}$
Biology	47.8$_{bd}$	52.3 (0.4)$_{ac}$	46.9 (0.6)$_{bd}$	53.1 (0.7)$_{ac}$

Note. Cohort 1 = Cohort before the reform; Cohort 2 = Cohort after the reform. The displayed results are 1PL models in which we did or did not control for differences on further covariates. The metric of the latent variable was transformed to M = 50 and SD = 10 on the basis of the 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 were labeled with ^{BH}. The results are from analyses in which the weights and cluster structure of the data were taken into consideration. The covariance of class repeaters was not estimated in the adjusted model for domain-specific interest to avoid singularity in the information matrix.

Additional information on Table 5

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 π = 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.

Additional information on Table 6

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.

References of the Supplemental Online Material

Baumert, J. (Ed.). (2000). *TIMSS-III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. 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. 2. Competencies in Mathematics and Physics at the End of Upper Secondary School]*. Opladen: Leske u. Budrich.

Benjamini, Y., & Hochberg, J. (1995). Controling the false discovery rate: A practical and powerful approach to multiple testing, *57*(1), 289–300.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). *Duxbury advanced series*. Australia and Pacific Grove, CA: Thomson Learning.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, *6*(4), 330–351. doi:10.1037/1082-989X.6.4.330

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

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

Enders, C. K. (2010). *Applied missing data analysis*. *Methodology in the social sciences*. New York: Guilford Press.

Evans, M., Kelley, J., Sikora, J., & Treiman, D. J. (2010). Family scholarly culture and educational success: Books and schooling in 27 nations. *Research in Social Stratification and Mobility*, *28*(2), 171–197. doi:10.1016/j.rssm.2010.01.002

Ganzeboom, H. B. G., & Treiman, D. J. (2003). Three internationally standardised measures for comparative research on occupational status. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in Cross-National Comparison* (pp. 159–193). Boston, MA: Springer US. doi:10.1007/978-1-4419-9186-7\textunderscore

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research.* Washington, DC, US: American Psychological Association.

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

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

Kiefer, T., Robitzsch, A., & Wu, M. (2015). *TAM: Test analysis modules. R package version 1.16-0*. Retrieved from http://CRAN.R-project.org/package=TAM

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale: Lawrence Erlbaum Associates.

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. doi:10.1111/j.1745-3984.1984.tb00227.x

Muthén, B., & Muthén, L. K. (1998-2012). *Mplus User's Guide. Seventh Edition*. Los Angeles and CA.: Muthén & Muthén.

NEPS. (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

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

R Development Core Team. (2016). R. Vienna, Austria: R Foundation for Statistical Computing.

Robitzsch, A. (2015). sirt: Supplementary Item Response Theory Models. R package version 1.8-9. Retrieved from http://CRAN.R-project.org/package=sirt

Rupp, A. A., Vock, M., Harsch, C., & Köller, O. (2008). *Developing standards-based assessment tasks for english as a first language: Context, processes and outcomes in Germany*. *Standards-Based Assessment Tasks for English as a First Language: Vol. 1*. Münster: Waxmann.

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. doi:10.1037/0022-0663.98.4.788

4.3 Study 3

Hübner, N., Wagner, W., Hochweber, J., Neumann, M., & Nagengast, B. (2017). Comparing apples and oranges: Curricular intensification reforms can change the meaning of students' grades! Manuscript submitted for publication.

The following manuscript has not been accepted and published yet. This version of the manuscript has been submitted to the *Journal of Educational Psychology* on July 14th, 2017. The American Psychological Association will have the copyright of the final version of the article, if it will be accepted. The version displayed here may not exactly replicate the final version published in the journal. It is not the copy of record.

Abstract

Teacher-assigned grades provide a central piece of information in the admission processes of universities and colleges and are relevant for employment decisions. Beside grades, the results of standardized achievement tests are frequently used for student selection and allocation. However, studies have shown that correlations between the two achievement measures are far from perfect, and this has been argued to result at least in part from norm-referenced grading, which is based on the performance composition of a class. In this study, we investigated whether a curricular intensification reform, which introduced changes in the performance composition of students by introducing mandatory course enrollment, resulted in changes in the relation between results of standardized student achievement tests and teacher-assigned grades. We analyzed cohort control design data from two large representative samples of students from two German states (Baden-Württemberg: *N* = 5,574; Thuringia: *N* = 2,202) before and after upper secondary school reforms, which were quite similar in the two states. Results indicate considerable differences in students' standardized test achievement scores before and after the reform, given similar grades. Furthermore, in math, course-level-specific reform effects of the association of grades and achievement were found to vary between groups of student receiving good and poor grades. Implications for educational policy and school reforms and suggestions for grading are discussed.

Keywords: curricular intensification, reform, standardized achievement, grades

Comparing Apples and Oranges: Curricular Intensification Reforms can Change the Meaning of Students' Grades!

Recently, many countries have put specific effort into increasing students' attainment and achievement levels, especially in subjects such as mathematics and languages. This movement began in 1983, with the publication of the A Nation at Risk report (The National Commission on Excellence, 1983) and has peaked in recent decades, where policy reforms such as the No Child Left Behind (NCLB) Act have claimed universal proficiency for all students in core subjects such as reading and math (e.g., Hess & Petrilli, 2006).

Although states were allowed to individually define proficiency in the United States, what followed these policies was the introduction of standards-based reforms, which consist of core components such as the rigorous standardized testing of students and the test-based accountability of schools (e.g., Ravitch, 2011; Swanson & Stevenson, 2002). In order to implement the new demands, which also appeared elsewhere (e.g., Germany or England; e.g., Volante, 2016), many countries introduced curricular intensification (CI) reforms. These reforms typically set rigorous mandatory enrollment standards regarding specific core courses (e.g., Domina & Saldana, 2012; Hübner, Wille et al., 2017).

Although an increasing amount of literature has investigated effects of such reforms on achievement measures and motivation (e.g., Domina, McEachin, Penner, & Penner, 2015; Hübner, Wille et al., 2017; Nomi & Raudenbush, 2016), less attention has been paid to the question of how CI reforms, which oftentimes lead to changes in the achievement-related composition of students within classes, might affect teacher-assigned grades and their meaning. School grades and standardized test achievement are central predictors of important life outcomes such as socioeconomic success (Strenze, 2007), college and university students' GPA and institutional retention (Koretz et al., 2016; Richardson, Abraham, & Bond, 2012; Robbins et al., 2004), and postschool choices (Parker et al., 2012). Furthermore, they comprise a central part of the admission criteria for colleges, universities, and employers (Clinedinst, Koranteng, & Nicola, 2015; Koretz et al., 2016; Robinson & Monks, 2005; Rojstaczer & Healy, 2012; Thorsen & Cliffordson, 2012) and provide an important foundation for students' academic self-concept (e.g., Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007). It has been noted that grades and test scores tend to differ when it comes to individual student achievement, indicated by a far from perfect correlation between the two achievement measures (e.g., Borghans, Golsteyn, Heckman, & Humphries, 2016; Dickinson & Adelson, 2015; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005; Neumann, Trautwein, & Nagy, 2011; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006; Willingham, Pollack, & Lewis, 2002), and this has been attributed at least in part to norm-referenced grading (e.g., Marsh et al., 2007; Trautwein et al., 2006).

Thus, in this study, we took a closer look at the CI reforms in two German states, which led to the detracking of students into core courses and thereby introduced changes in the student composition of classes. We explored whether these CI reforms might have subsequently affected students' grades and the relations between grades and standardized test achievement in cohorts before and after the CI reforms.

Grades, Test Scores, and the Frame of Reference

Willingham et al. (2002) suggested that discrepancies between grades and test scores might result from different sources, for instance, situational differences (e.g., changes in motivation over time and across contexts), or systematic (e.g., variation in grading standards) and unsystematic errors (e.g., unreliability in grades and test scores).

Further research, especially related to grading standards, suggests that when assigning grades, teachers usually make use of different frames of reference (e.g., Neumann et al., 2011). Most important in this regard is the distinction between self-referenced grading, criterion-referenced grading, and norm-referenced grading. Self-referenced grading indicates that teachers compare a student's achievement with his or her previous achievement in order to judge performance. In this case, the achievement of other students in the class or learning group are not relevant for the judgment. Criterion-referenced grading involves a comparison between a student's achievement and a specific learning goal. This type of grading is often used when students must meet specific objectives in order to get credit.

However, few teachers seem to consistently make use of absolute criteria when grading students, and in contrast to standardized achievement tests, multiple measures are available and used for grading (e.g., Cross & Frary, 1999). Research has shown that teachers make use of a variety of nonachievement-related criteria when assigning grades, such as students' interest or effort, socioeconomic status, or inappropriate behavior (Guskey, 2006; Hochweber, Hosenfeld, & Klieme, 2014; Westphal et al., 2016; Zimmermann, Schütte, Taskinen, & Köller, 2013), and teachers' use of norm-references is very common when they assign grades (e.g., Cizek, Fitzgerald, & Rachor, 1995; Marsh et al., 2007).

Norm-referenced grading (also referred to as "grading on a curve") suggests that students are first sorted (explicitly or implicitly) by their achievement. Comparably good students are then assigned As or Bs, comparably bad students are assigned Ds or Fs, and moderately performing students are graded somewhere in between (e.g., Trautwein et al., 2006). Studies have shown that students with equal levels of standardized achievement are assigned very different grades in classes with high- versus low-performing reference groups (e.g., Marsh, 1987; Marsh et al., 2007), and this reference group effect is mentioned a great deal in the literature on the "Big-Fish-Little-Pond Effect" (BFLPE; e.g., Marsh et al., 2008; Trautwein et al., 2006). In line with these findings, as shown by Neumann et al. (2011), the reference group effect can, for instance, have an impact on students' math grades: In their study, aggregated standardized school achievement had a negative effect on students' coursework grades, independent of the course level (advanced or basic course), after standardized individual student achievement was controlled for.

According to this research, changing the class composition (e.g., by means of tracking or detracking) should typically go along with a shift in the achievement-related sorting of students in class. Thus, this affects the process of grading if grades are assigned to students by comparing and rank-ordering individual achievement against the achievement of other students in the class (e.g., Brookhart, 2008, 2015; Schinske & Tanner, 2014; Trautwein et al., 2006). However, in spite of the relevance of school grades, previous research has failed to show whether CI reforms at the end of upper secondary school can foster changes in teacher-assigned grades.

Grading, Test Scores, and Curricular Intensification

In the face of recent school reform efforts dedicated to increasing student achievement (e.g., Hess & Petrilli, 2006; OECD, 2015), many countries have tried to increase the enrollment rates and achievement of students in school by implementing mandatory course enrollment policies. Such reforms are typically referred to as CI reforms and are meant to increase achievement and reduce differences between students by setting similar standards for all students (e.g., Crosnoe & Benner, 2015).

According to Domina and Saldana (2012) and in line with Sørensen (1970), CI is typically related to the detracking of students (e.g., Domina & Saldana, 2012). Detracking, which basically means that students are no longer sorted into different educational tracks or courses, can for instance go along with the mandatory enrollment of all students in core courses. In such a case, not only does CI change the academic requirements of a course, but it also affects variation in the achievement-related class composition (e.g., Hübner, Wille et al., 2017; Nagy, Neumann, Trautwein, & Lüdtke, 2010).

The mandatory enrollment of students in courses in which they would not have enrolled if they had been given a choice clearly points to the relation between CI and the scientific debate about students' achievement and instructional time (e.g., Carroll, 1989; Cortes, Goodman, & Nomi, 2015; Lavy, 2015; Nomi & Allensworth, 2009).

Currently, the results of research on effects of intensification are mixed. Whereas some studies have found positive effects of intensification, others have found no or mixed effects (e.g., Ceci, 1991; Domina et al., 2015; Lavy, 2015; Nomi & Raudenbush, 2016).

As outlined above, CI reforms can change the achievement composition of students within a class (e.g., Domina et al., 2015; Hübner, Wille et al., 2017; Nomi & Allensworth, 2009). As the achievement of the reference group is essential for grading on a curve (Brookhart, 2015; Marsh, 1987; Trautwein et al., 2006), this can lead to changes in teachers' grading. Such effects might appear, for instance, if students who were tracked into advanced and basic courses before a reform are grouped together in core courses afterwards, as done in the reform of upper secondary school in most German states (e.g., Hübner, Wille et al., 2017) or if students are grouped into classes on the basis of previous standardized achievement, as done in the "Double-Dose Algebra" reform in Chicago (e.g., Nomi & Raudenbush, 2016).

Recent evidence on effects of reform-introduced changes in class compositions on grades was published by Nomi and Allensworth (2009). The authors investigated the "Double-Dose Algebra" reform, which was implemented in 2003 in Chicago Public Schools and introduced algebra courses that offered additional support for students performing below the national median. In the course of introducing this double dose, some schools decided to group the low-performing students together in regular algebra classes as well, whereas other schools did not. The results suggested that although algebra achievement increased for students who took the additional algebra course, it had only modest effects on grades.

In another study, Nomi and Raudenbush (2016) further investigated reform effects that were related to student tracking. In doing this, they were able to show that placing students who performed at the median in homogeneous low-skilled classrooms had no or only a small negative effect on their standardized achievement, whereas placing them in heterogeneous classrooms substantially increased their achievement. Simultaneously, although the standardized achievement of students placed in homogeneous low-skilled classrooms did not change considerably, they were assigned higher grades in comparison with students placed in heterogeneous classrooms. These results underscore the importance of the reference group for achievement measures (e.g., Marsh et al., 2007; Marsh et al., 2008).

As outlined, few studies have investigated how CI reforms might change the meaning of a grade in terms of predicted standardized achievement, given a specific grade before versus after a reform. A comparable meaning of grades before and after reforms is especially important at the end of upper secondary school in order to guarantee the comparability of grades when used for employee selection and college/university admission. Furthermore, if students have similar standardized achievement scores but get different grades before versus after a reform, this might lead to general disadvantages for students from specific cohorts regarding their career prospects.

Present Study and Research Questions

Grading procedures often follow norm-references (e.g., Brookhart, 2008), and teachers might assign similar grades to students with different standardized test achievement scores or the other way round (Nomi & Raudenbush, 2016). This tendency might be especially apparent when students from basic and advanced courses in upper secondary school are assigned to the same core course, as done in the upper secondary school reform in Germany.

The change in the grading of students, given a specific achievement level, is especially important from a perspective of postschool student selection at college or university or for later employment. Furthermore, as grades are of central importance for academic self-concepts (e.g., Marsh et al., 2008), CI reforms might also impact other outcomes such as aspirations or career choices in STEM (science, technology, engineering, and mathematics) areas (e.g., Jansen, Scherer, & Schroeders, 2015; Schoon & Eccles, 2014). Therefore, if CI reforms foster changes in the grading of students, such grades will have a different meaning for different student cohorts within a state, a school, or even within different classes taught by the same teacher over time, and this could lead to an unfair and less reliable selection of students after school when grades are used as achievement indicators. On the other hand, this could also suggest that students with similar levels of achievement will get different grades, and this might result in individual disadvantages for future career prospects. Grades build a central foundation for students’ academic self-concepts (e.g., Marsh et al., 2007), and therefore, differences that have already been found in self-concepts before versus after CI reforms (e.g., Hübner, Wille et al., 2017) might have originally been triggered by changes in teacher-assigned grades.

Therefore, in the present study, we reanalyzed representative data from two German states (Thuringia and Baden-Württemberg) that enacted an upper secondary school reform, which introduced mandatory course enrollment in German, mathematics, and one foreign language.

In a first step, we tested the association between grades and standardized test achievement in math and English for the different course groups (before the reform: basic and advanced courses; after the reform: core courses). In line with previous research outlined above, we expected the detracking of students to lead to differences in the relation between teacher-assigned grades and standardized achievement scores before versus after the reform. Due to the grouping of potentially high- and low-achieving students in one class, students with identical grades were expected to perform lower in core courses compared with students in advanced courses and higher in core courses compared with students in basic courses.

Second, we extended the first analysis by taking a closer look at specific grade groups that had grades ranging from low to high grades (Grade Groups D to A). We expected that, in general, the results found in the first step would be replicated here.

Third, interaction effects of Grade Group × Course Level, for high and low grade groups (As and Ds) were tested in order to further investigate potential differences in standardized achievement depending on the grade and course level. We expected that standardized achievement for high grades would more closely resemble the relation between grades and standardized achievement in advanced courses, whereas standardized achievement for low grades would be more strongly oriented toward the basic course, due to ceiling and floor effects of student achievement.

Finally, we explored whether reducing the course choice options from two courses (basic vs. advanced) to one course (core) would generally decrease the ability to differentiate across all students, given information about their grades before versus after the reform, indicated by a reduction in the amount of variance that could be explained in standardized achievement. Differences would suggest that the reform might have changed the boundaries of the grading distributions.

Method

Descriptions of the Study and Sample

We used data from two German studies in two different states: The Transformation of the Secondary School System and Academic Careers (TOSCA) study (Köller, Watermann, Trautwein, & Lüdtke, 2004; Trautwein, Neumann, Nagy, Lüdtke, & Maaz, 2010) and the Additional Study Thuringia (Blossfeld, Rossbach, & Maurice, 2011; Wagner et al., 2011) from the National Educational Panel Study (NEPS), included in the Scientific Use File 2.0.0. Both data sets contain representative data from one cohort before the reform and one cohort after the reform, which were collected at the end of upper secondary school. This design is referred to as a cohort control design (e.g., Shadish, Cook, & Campbell, 2002) and is assumed to provide the foundation for a natural experiment. In Baden-Württemberg, a total of 88 general upper secondary schools (Gymnasium) participated at both time points, Cohort 1 (Time 1): *N* = 2,772 (age: *M* = 19.5 years); Cohort 2 (Time 2): *N* = 2,802 (age: *M* = 19.3 years). In Thuringia, 30 schools participated at both time points, Cohort 1 (Time 1): *N* = 1,316 (age: *M* = 18.4 years); Cohort 2 (Time 2): *N* = 886 (age: *M* = 18.3 years). In Thuringia students graduated after 8 years, whereas students in Baden-Württemberg graduated after 9 years (until 2012). However, both groups of students were required to spend a similar minimum number of hours in school during their years of schooling (at least 265 hr in 8 or 9 years).

The Reform of Upper Secondary School

Between 2001 and 2012, 11 of the 16 states in Germany reformed their upper secondary school systems (Trautwein & Neumann, 2008). The states introduced reduced course choice options and increased mandatory participation in specific core subjects (e.g., math, a foreign language, and natural science) that were taught at a level that was equivalent to what was the advanced course level before the reform.

Although starting points have varied slightly, depending on state regulations (e.g., the total number of years taken to graduate or average weekly hours spent in advanced and basic courses), most principles of the reforms were quite similar across states. As articulated by the ministers of education, the reform was dedicated to two specific goals. First, it was expected to increase the comparability of final examinations and resulting grade point averages between schools and states by increasing the focus on specific subjects. Second, it was expected to increase average student achievement due to the implementation of core subjects on an advanced level (Trautwein & Neumann, 2008).

Basically, before the reform, students were able to choose two advanced courses at the beginning of upper secondary school, each taught for 5 hr (Baden-Württemberg) or 6 hr (Thuringia) per week. The two advanced courses were chosen for the entire span of upper secondary school and were part of the final examinations at the end of upper secondary school. Besides participating in the advanced courses, all students had to participate in a variety of additional courses for a reduced amount of time on a basic course level. Two of these basic-level courses had to be chosen at the beginning of upper secondary school to be part of the final examinations.

After the reform, the number of choices were reduced: All students had to participate in mandatory advanced courses in the three subjects of mathematics, German, and one foreign language for 4 hr per week each and had to choose two additional courses, which were also taught for 4 hr per week on an advanced course level (e.g., Hübner, Wille et al., 2017). Similar to before the reform, students still had to participate in several other subjects in addition to these five subjects on a basic course level during upper secondary school. The reform was implemented by law in terms of a top-down state-wide policy reform. Before the reform, in contrast to students from Thuringia, who had to enroll in math as a mandatory basic course (if it was not chosen as an advanced course) for 4 hr per week, students from Baden-Württemberg had to enroll for 3 hr in math as a basic course (if it was not chosen as an advanced course).

Instruments

Math achievement. Reanalyzing data from the TOSCA study, we made use of an *Advanced mathematics test*, which was based on test items from the Third International Mathematics and Science Study (TIMSS; Mullis et al., 1998). About two thirds of all items were administered in a multiple-choice format, whereas the other items were open-ended questions. The items were administered in a multimatrix design. Therefore, students worked on a subset of items (one of four booklets) rather than on all 68 items. Item response theory (IRT) was used to scale all of the items so that student achievement before and after the reform could be meaningfully compared and so that the multimatrix design could be adequately accounted for.

The mathematics test in the Additional Study Thuringia focused on *mathematical literacy*, which is also a focus of the Program for International Student Assessment (PISA; e.g., OECD, 2004). Overall, students had to work on 35 items for 30 min (each student worked on a subset of 19 to 21 items per booklet). We applied a similar scaling strategy as for the previous tests. The reliability of both tests was acceptable (WLE reliability = .68).

English achievement. In the TOSCA study, a short version of the Test of English as a Foreign Language (TOEFL) was used to assess students' English competencies. The test consists of a total of 79 items from three subscales: Listening and Comprehension (28 items; Cronbach's α = .79), Structure and Writing Expression (23 items, Cronbach's α = .75), and Vocabulary and Reading Comprehension (28 items; Cronbach's α = .77). The latent intercorrelations between the three factors were high ($r \geq .85$). On the basis of this, we considered a unidimensional IRT model for the TOEFL test. The reliability of this score was high (WLE reliability = .87).

In the Additional Study Thuringia, an English reading test was administered with 33 items that were developed by the Institute for Educational Quality Improvement (IQB; Rupp, Vock, Harsch, & Köller, 2008). Students had to work on 21 items for 30 min, which were administered in a multiple-choice or multiple-matching format (NEPS, 2011). We applied a similar scaling strategy as for the previous tests. The reliability of this test was good (WLE reliability = .77). All standardized tests were administered in the last semester of upper secondary school.

Grades. In the TOSCA study, data were provided from the first semester of the final school year (i.e., 13.1; this notation means Grade 13, Semester 1) in mathematics and English. These grades were based on written exams and oral participation in class and ranged from 0 (*worst achievement*) to 15 (*best achievement*) points. In the Additional Study Thuringia, data from all four classes (11.1, Grade 11, Semester 1 to 12.2, Grade 12, Semester 2) in math (all *r*s ≥ .75) and English (all *r*s ≥ .79) in upper secondary school were available and were strongly correlated. Therefore, we decided to average these grades (see Table 2). Robustness checks using models considering only information from Grade 12.1 (i.e., Grade 12, Semester 1) in Thuringia revealed comparable results. We additionally created specific grade groups (D to A), in order to be able to better picture potential nonlinear trends in the data. The grade groups were built as follows: Group D: < 6.5 points, Group C: ≥ 6.5 and < 9.5 points, Group B: ≥ 9.5 and < 12.5 points, Group A: ≥ 12.5 points. This taxonomy is comparable to the traditional grading metric in the German upper secondary school system. We decided to place all students with grades lower than 6.5 points in one grade group because there were only a few students in the lowest parts of the grade distribution.

Statistical Analysis

We first present estimated correlations between standardized tests and grades for English and mathematics across all students and descriptive statistics of these measures for each course and tested them for statistically significant differences. Next, we specified multiple indicators and multiple causes (MIMIC; e.g., Jöreskog & Goldberger, 1975) item response theory (IRT) models, separately for the two states. In these models, standardized achievement was modeled in terms of a unidimensional one-parameter logistic (1PL) IRT model (measurement model) and predicted by the subject-specific grade (structure model) to avoid biased estimates due to unreliability, which would be the case if, for example, point estimates of achievement (e.g., WLEs) were used instead of latent variables. We estimated these models with cohort-specific structural models (multiple group) and constant measurement models across groups using an IRT multiple-group model, implemented in Mplus 7.4 (Muthén & Muthén, 1998-2012). Multiple groups were built on the basis of information about the course (basic, advanced, and core) as well as on the grade group (D to A). The analyses proceeded in three steps.

First, we estimated multiple-group models for three different groups for each state and subject (course-level groups: basic and advanced courses before the reform and core courses afterwards). We did this separately for each subject and state using grand-mean-centered grades. Then we compared the predicted, average standardized achievement (intercepts) between the three resulting different groups using the model constraint option implemented in Mplus.

Second, we specified multiple-group models for 12 different groups (all combinations of Course Level × Grade Group) and compared the average standardized achievement scores predicted for the four different grade groups across the three different courses. We again estimated these models separately for each subject and each state and also tested Course Level × Grade Group interactions for very good (As) and very bad (Ds) grades to check for potential differences in predicted standardized achievement, depending on the specific grade group and the course level due to ceiling and floor effects in students' achievement.

Third, to compare the variance explained in standardized achievement between the course systems before and after the reform, we specified a multiple-group model that was comparable to the one used in the first step of analysis. In this model, we made use of the model constraint option to test the differences in explained variance between all students before and after the reform for statistical significance.

The coefficient of determination for the group after the reform was calculated as a new parameter by dividing the explained variance (i.e., the squared regression weight multiplied by the variance of the grades) by the total variance (i.e., the sum of explained and residual variance). For the groups before the reform, a combined coefficient of determination (i.e., variance explained across both groups) was calculated, reflecting the R^2 from a multiple regression with grade and a dummy-coded variable for course level (0 = basic course, 1 = advanced course) and their interaction as predictors. It was based on the explained variance by grade and by the mean difference between both (sub)groups (i.e., basic and advanced level), each weighted by the relative probability of group membership (reflecting the different group sizes), divided by the total variance (i.e., the sum of explained variance and weighted residual variance). The difference between the two coefficients of determination (before the reform, after the reform) was tested against the null hypothesis of a difference of zero.

Again, we estimated all models separately for each subject and state.

In order to be able to meaningfully interpret the coefficients, the metric of the latent achievement variable was transformed so that M = 50 and SD = 10. p-values were controlled for the false discovery rate within a subject and a state (Benjamini & Hochberg, 1995).

Item analyses and selection. The quality of the standardized achievement tests was assessed beforehand with regard to reliability and differential item functioning (DIF) between cohorts. Furthermore, we specified two-parameter logistic (2PL) models in R (R Core Team, 2017) using the *TAM* package (Kiefer, Robitzsch, & Wu, 2017) to check for negative item discriminations, which would suggest that students with lower average competence had a higher likelihood of correctly solving this item. Therefore, negative item discriminations might be an indicator of incorrect coding or poor item quality. DIF was analyzed using the Mantel-Haenszel DIF method (Mantel & Haenszel, 1959) implemented in the *difR* package in the statistics package R (Magis, Beland, Tuerlinckx, & De Boeck, 2010). Besides checking for significance and the log-odds ratio statistic (e.g., Penfield & Camilli, 2006), we classified DIF according to the ETS classification system into negligible (A), moderate (B), or large (C; Holland & Thayer, 1985). Results from models using all items did not differ meaningfully from results including only items with negligible DIF (Category A). For the sake of clarity, we decided to report results of models that included all test items, but we excluded one item with a negative item discrimination beforehand.

Cluster structure and survey weights. Observations of students from similar schools cannot be treated as independent because they are more similar to each other than they are to students from different schools. Ignoring the cluster structure usually leads to an underestimation of standard errors (Snijders & Bosker, 2012). To address the clustered data structure (students were nested within classes), we adjusted the standard errors by applying a design-based correction as implemented in Mplus 7.4 (Muthén & Muthén, 1998-2012), which takes the multilevel structure into account by the use of a “sandwich” estimator (see e.g., Asparouhov, 2005; Muthen & Satorra, 1995). In doing this, we followed the recommendations of McNeish, Stapleton, and Silverman (2016), who suggested that population-averaged methods do not rely on assumptions that are inherent in the specification of random effects in hierarchical linear modeling. We also considered survey weights in order to establish the representativeness of our results for the population of students in each state at upper secondary grammar schools.

Missing values. Different approaches are available in social science research to handle missing data (e.g., Enders, 2010). All analyses were conducted with full information maximum likelihood (FIML) estimation (e.g., Graham, 2009).

Results

Preliminary Analyses

In the first step, we took a closer look at the descriptive statistics. As displayed in Table 1, there were considerable correlations between subject-specific grades and achievement on the standardized test in the overall sample within the two states. In Thuringia, standardized math achievement was correlated $r = .48$ ($p < .001$) with the math grade and standardized achievement in English was correlated $r = .62$ ($p < .001$) with the English grade. The correlations in Baden-Württemberg revealed a similar pattern, although math grade and math achievement had a slightly stronger correlation, $r = .64$ ($p < .001$), which might be related to the focus of the mathematical literacy test in the NEPS Thuringia study, which could be judged as less curricularly valid compared with the test used in Baden-Württemberg (e.g., Mullis et al., 1998; Weinert et al., 2011).

Next, we estimated the means and standard deviations of grades and standardized test achievement in mathematics and English for each state (see Table 2). In Thuringia, before the reform, 47.1% of all students were enrolled in advanced courses in math and 32.9% in English. In Baden-Württemberg, 37.5% of all students were enrolled in advanced courses in math, and 47% were enrolled in advanced courses in English.

Regarding the achievement tests, we found quite a comparable pattern across states. Standardized achievement was the lowest in basic courses before the reform. In math in Thuringia, for instance, students in basic courses achieved $M = 43.35$ ($SD = 8.40$) points, whereas students in advanced courses, on average, achieved a statistically significantly higher score of $M = 54.93$ ($SD = 9.27$, $d = -1.09$, $p < .001$). Achievement in the core course after the reform ($M = 50.34$, $SD = 9.91$) was statistically significantly lower compared with the advanced course ($d = -0.47$, $p < .001$) but statistically significantly higher when compared with the basic course ($d = 0.54$, $p < .001$). Similar patterns were found for standardized achievement in English in Thuringia and for both subjects in Baden-Württemberg (see Table 2).

In contrast to the standardized achievement tests, grades revealed a slightly different picture (see Table 2). In Baden-Württemberg, basic course grades in mathematics ($M = 8.01$, $SD = 3.59$) were statistically significantly different from advanced course grades in math ($M = 9.85$, $SD = 3.08$, $d = -0.54$, $p < .001$). Similarly, basic course grades in English ($M = 9.13$, $SD = 2.86$) were statistically significantly different from advanced course grades in English ($M = 9.67$, $SD = 2.70$, $d = -0.19$, $p < .001$). As shown, differences between students' achievement and grades in basic and advanced courses were considerably smaller for grades than for standardized achievement, a finding that supports our assumption about the impact of the reference group on grades.

In addition, we found differences between grades in core courses in mathematics (*M* = 8.27, *SD* = 3.64) and English (*M* = 9.25, *SD* = 2.92) and advanced courses in mathematics (*d* = -0.53, *p* < .001) and English (*d* = -0.15, *p* = .001). However, grades were not statistically significantly different between basic and core courses in mathematics (*d* = -0.08, *p* = .078) or in English (*d* = -0.04, *p* = .342). English grade differences in Thuringia were comparable to the patterns found for Baden-Württemberg. For math in Thuringia, however, average grades were not statistically significantly different between advanced (*M* = 9.24, *SD* = 3.04) and core courses (*M* = 8.94, *SD* = 3.12, *d* = -0.10, *p* = .087), whereas grades from basic courses (*M* = 8.21, *SD* = 3.02) were statistically significantly different from grades in core courses (*d* = -0.24, *p* < .001) and advanced courses (*d* = -0.34, *p* < .001).

Grades and Standardized Achievement

Following these first basic analyses, we estimated multiple-group models for three different groups for each state and subject in order to investigate differences between predicted achievements in the three different courses for an average grade. In line with our assumptions, we found statistically significant differences between all courses (all *p*s < .001) and for both subjects in both states, indicating that for students with an average course grade, their standardized achievement differed between courses (see Table 3).

In order to obtain a more coherent picture, we then specified multiple-group models for the 12 different groups that resulted from the three different course levels and four different grade groups. The results of these analyses are displayed in Table 4 and 5 and Figure 1. As can be seen in Figure 1, we found very comparable patterns for the two states in mathematics, and these findings were in line with our hypothesis. Before the reform, and similar to the findings displayed in Table 3, there was a considerable difference in achievement between students with comparable grades from advanced and basic courses. In Baden-Württemberg, these differences in mathematics were statistically significant across all different grade groups. For instance, before the reform, students in Grade Group B, on average, scored 58.3 points in an advanced course, whereas they scored only 46.4 points in basic courses (*p* < .001). After the reform, where all students had to participate in a mandatory core course in mathematics, the achievement in Grade Group B (55.3 points) was statistically significantly higher when compared with the basic courses (*p* < .001), whereas it was statistically significantly lower when compared with the advanced courses ($p < .001$). As can be seen in Figure 1, this pattern was comparable to the pattern found in Thuringia (see Table 4).

For English, we found a somewhat comparable picture in one state (Baden Württemberg, see Figure 1 and Table 5). All differences between courses were statistically significant here ($p < .001$). However, in Thuringia, achievement differences were statistically significant only for Grade Group C between basic courses ($M = 44.4$) and advanced courses ($M = 48.4$, $p = .001$) and for Grade Group B between core courses ($M = 52.4$) and advanced courses ($M = 55.9$, $p < .001$) and between basic courses ($M = 50.9$) and advanced courses ($p < .001$).

Next, we also checked for interaction effects of Grade Group × Course Level for Grade Groups D and A, and found statistically significant effects in mathematics but not in English. In both states, achievement differences between core courses and basic courses were statistically significantly smaller for Grade Group D compared with Grade Group A in math (Thuringia: 6.6 points, $p = .015$; Baden-Württemberg: 8.9 points, $p < .001$). Furthermore, in Baden-Württemberg, there was also a statistically significantly larger difference between core course achievement and advanced course achievement in Grade Group D compared with Grade Group A (-3.1 points, $p = .016$). These results suggest that differences in students' standardized achievement in math, given a similar grade in core courses, rather resemble the grading of basic courses in low grade groups (Ds), whereas it is closer to the grading in advanced courses in higher grade groups (As).

Finally, we took a closer look at the differentiability of student achievement before and after the reform. To test this, we estimated the explained variance using a model for predicting student achievement before the reform, including grades (in points), a course dummy (basic vs. advanced), as well as the interaction effect. To predict achievement after the reform, the model included only the grades in the core courses as a predictor. These models explained 44% of the variance in students' achievement in Thuringia in English before and 35% after the reform ($|\Delta R^2| = .09$, $p = .136$). A similar pattern was found for math in Thuringia, where 37% of the variance was explained before and 28% was explained after the reform ($|\Delta R^2| = .09$, $p = .084$). In Baden-Württemberg, we found a comparable pattern (mathematics before: 59%, after: 56%, $|\Delta R^2| = .04$, $p = .170$; English before: 44%, after: 40%, $|\Delta R^2| = .04$, $p = .243$). These results indicate that the prediction of standardized achievement from grades, or the distinction between students' grades given their standardized achievement and information on courses, was not statistically significantly different before and after the reform.

Discussion

In the current study, we investigated how the reform of upper secondary school in two German states introduced changes in the meaning of teacher-assigned grades. School grades in upper secondary school are a major criterion for student selection at college and university and are relevant for later employment. In spite of the societal relevance of grades for student allocation and selection, various studies have indicated that school grades and standardized achievement tend to differ, as indicated by their far from perfect correlations. Research has suggested that one central factor that might cause such differences is variations in grading standards (e.g., Guskey, 2006; Hochweber et al., 2014; Willingham et al., 2002), for instance, norm-referenced grading (e.g., Trautwein et al., 2006).

Overall, the majority of the results of this study are in line with our assumptions. We found statistically significant differences between the standardized achievement of students in advanced, basic, and core courses, given similar grades in mathematics. Students in core courses after the reform performed better, on average, compared with basic course students, given a similar grade. Furthermore, in line with our assumptions, comparing standardized achievement between advanced courses (before the reform) and core courses (after the reform) revealed the opposite pattern: Here, average achievement in the advanced courses was statistically significantly higher than in the core courses, given a similar grade.

The differences between standardized math achievement in core and basic courses given similar grades were more pronounced in the high grade groups (those who got As) compared with the low grade groups (those who got Ds). On the basis of this finding, grading in the previous advanced courses more closely resembled the grading in core course for high grade groups (those who got As) after the reform, and grading in the previous basic courses was more similar to the grading in core course for low grade groups (those who got Ds). Finally, differentiation among student achievement, in terms of the variance explained by grades and course level (only before the reform) compared with grades after the reform (all students in core courses) did not differ.

Implications for Research, Policy, and Practice

The findings of this study have several implications for research, policy, and practice. First, the results of this study are in line with an increasing amount of literature on potentially unintended side effects of policy reforms in general and CI reforms more specifically (e.g., Gross, Booker, & Goldhaber, 2009; Hübner, Wille et al., 2017). Similar to Nomi and Raudenbush (2016), our results suggest that reforms, which change the composition of students in classes, can also have an impact on teacher-assigned grades. Our study further expands these previous findings to the end of upper secondary schools, where school grades are an important measure for third parties (e.g., employers or universities) that rely on them for selection purposes as an indicator of students' abilities (e.g., Clinedinst et al., 2015; Koretz et al., 2016). Therefore, in order to interpret grades as an indicator of students' abilities more meaningfully, the introduction of reforms and their effects have to be monitored more rigorously by the different stakeholders who make use of grades as ability indicators. These results are also important for research that focuses on the transition of students to university or vocational training after upper secondary school because if students with lower achievement get similar grades to students with higher achievement, this might have an impact on postsecondary student allocation and success.

Second, the results indicate that differences in students' standardized achievement in math, given a similar grade in core courses, rather resembles the grading of basic courses for low grade groups (Ds), whereas it is closer to grading in advanced courses in higher grade groups (As). This suggests that, after the reform, teachers appeared to adapt their grading to some extent to resemble the full range of grading before the reform when the range of grades was extreme. However, such adaptions seem to occur more often in mathematics than in English, which might be the result of grading standards that are easier to adapt (e.g., points for correct/incorrect steps on a math test).

Third, findings from this study can be further integrated into the discussion on the challenges of constructing effective educational policy reforms in general. As is evident, effects of reforms are hard to anticipate even if they are planned carefully (e.g., Gross et al., 2009), and current research suggests that reforms that have only positive effects and no negative side effects are an exception (e.g., Domina & Saldana, 2012). Our results therefore strengthen claims that there should be a policy that rigorous research should accompany reforms right from the beginning and that funding should be provided for extensive evaluations with formative and summative parts. Investing in rigorous reform monitoring on a national or state level could further contribute to the acceptance of school reforms among teachers as the "ultimate enactors of any change effort" (Porter, Fusarelli, & Fusarelli, 2015, p. 5), reduce normative parts of reform efforts, and increase sustainable knowledge about effective and ineffective reform characteristics.

Finally, our results also suggest that teachers should apply standards-based references in order to judge student achievement. As outlined, the differences that we found are strongly related to grading on the curve, which strongly relies on the composition of the students in a class, independent of the occurrence of educational reforms. Of course, implementing more standardized grading systems would, on a large scale, involve huge efforts (e.g., improving teacher training and training for current teachers).

Limitations and Future Prospects

There are some limitations that should be mentioned before outlining further prospects. First, the data we considered were based on a cross-sectional cohort control design (Shadish et al., 2002). Here, students were assessed right before the policy reform, and a different cohort of students was assessed afterwards. On the basis of these data, which were assessed in upper secondary school, we were not able to consider variables from lower secondary schools, which might have helped us identify potential selection effects. However, in line with previous research using these data, we found no considerable differences between the student cohorts on observed background variables (Hübner, Wagner, Nagengast, & Trautwein, 2017; Hübner, Wille et al., 2017).

Next, we used standardized achievement tests so that we could apply an objective measure of student achievement, but a closer look at the psychometric properties of the instruments indicated that they were not perfectly reliable. We addressed this issue by using IRT models in all analyses to avoid having biased estimates due to unreliability. Related to this, we were not able to consider identical instruments in both assessments (NEPS and the TOSCA study), and this is why results should be compared between states only with caution. However, as the results were comparable between states/instruments, this also points to the generalizability, robustness, and significance of our findings. Finally, in this study, we were not able to empirically identify which component of the “reform package” contributed most strongly to the effects we found, as these were perfectly confounded.

Based on this, two suggestions in particular seem to arise for future research. First, findings for the CI reforms outlined here should also be tested in the context of other reforms. This is important for increasing knowledge about how reforms affect student achievement and related factors of school achievement such as self-concept and interest. It can be assumed that comparable results might arise, particularly if the reforms introduce changes in tracking procedures and the student composition of classes. However, if reforms affect other surface structures of the school system (e.g., accountability structures), grades might remain completely unaffected, and other variables should be evaluated (e.g., standardized achievement or motivation). Therefore, more research is needed to further provide insights into potential channels between isolated characteristics of education reforms and their implementation and school-, teacher-, and student-level variables.

As a final note, it is important to investigate options to further implement standardized grading strategies in class because otherwise, grading will continue to vary unsystematically between students with similar achievement depending on how teachers come up with a reference group and whether teachers consider students' gender, socioeconomic status, and so forth when assigning norm-referenced grades (e.g., Guskey, 2006; Hochweber et al., 2014; Westphal et al., 2016, 2016; Zimmermann et al., 2013).

References

Asparouhov, T. (2005). Sampling Weights in Latent Variable Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 411–434. https://doi.org/10.1207/s15328007sem1203\textunderscore

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

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

Borghans, L., Golsteyn, B. H. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(47), 13354–13359. https://doi.org/10.1073/pnas.1601135113

Brookhart, S. M. (2008). *How to give effective feedback to your students*. *Gale virtual reference library*. Alexandria, Virginia: Association for Supervision and Curriculum Development. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=287379

Brookhart, S. M. (2015). Graded achievement, tested achievement, and validity. *Educational Assessment*, *20*(4), 268–296. https://doi.org/10.1080/10627197.2015.1093928

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

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. A. (1995). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment*, *3*(2), 159–179. https://doi.org/10.1207/s15326977ea0302_3

Clinedinst, M., Koranteng, A.-M., & Nicola, T. (2015). *State of college admission*. Arlington, VA: National Association for College Admission Counseling.

Cortes, K. E., Goodman, J. S., & Nomi, T. (2015). Intensive math instruction and educational attainment. *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, New Jersey: Wiley. https://doi.org/10.1002/9781118963418.childpsy407

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53–72. https://doi.org/10.1207/s15324818ame1201_4

Dickinson, E. R., & Adelson, J. L. (2015). Choosing among multiple achievement measures: Applying multitrait-multimethod confirmatory factor analysis to state assessment, ACT, and student GPA data. *Journal of Advanced Academics, 27*(1), 4–22. https://doi.org/10.1177/1932202X15621905

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/10.3102/0002831211426347

Enders, C. K. (2010). *Applied missing data analysis*. *Methodology in the social sciences*. New York: Guilford Press.

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

Guskey, T. R. (2006). Making high school grades meaningful. *Phi Delta Kappan, 87*(9), 670–675.

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

Hochweber, J., Hosenfeld, I., & Klieme, E. (2014). Classroom composition, classroom management, and the relationship between student attributes and grades. *Journal of Educational Psychology*, *106*(1), 289–300. https://doi.org/10.1037/a0033829

Holland, P. W., & Thayer, D. T. (1985). An alternative definition of the ETS delta scale of item difficulty. *ETS Research Report Series*, *1985*(2), 1-10. https://doi.org/10.1002/j.2330-8516.1985.tb00128.x

Hübner, N., Wagner, W., Nagengast, B., & Trautwein, U. (2017). Putting all students in one basket does not produce equality: Gender-specific effects of curricular intensification in upper secondary school. Manuscript submitted for publication.

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

Jansen, M., Scherer, R., & Schroeders, U. (2015). Students' self-concept and self-efficacy in the sciences: Differential relations to antecedents and educational outcomes. *Contemporary Educational Psychology*, *41*, 13–24. https://doi.org/10.1016/j.cedpsych.2014.11.002

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631–639. https://doi.org/10.1080/01621459.1975.10482485

Kiefer, T., Robitzsch, A., & Wu, M. (2017). *TAM: Test analysis modules. R package version 1.99999-31*. Retrieved from http://CRAN.R-project.org/package=TAM

Köller, O., Watermann, R., Trautwein, U., & Lüdtke, O. (2004). *Wege zur Hochschulreife in Baden-Württemberg: TOSCA - eine Untersuchung an allgemein bildenden und beruflichen Gymnasien* [Ways towards higher education entrance qualification in Baden-Württemberg: TOSCA – an investigation of general and vocational upper secondary schools]. Opladen: Leske und Budrich.

Koretz, D., Yu, C., Mbekeani, P. P., Langi, M., Dhaliwal, T., & Braslow, D. (2016). Predicting freshman grade point average from college admissions test scores and state high school test scores. *AERA Open*, *2*(4), 1-13. https://doi.org/10.1177/2332858416670601

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal*, *125*(588), 397-424. https://doi.org/10.1111/ecoj.12233

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*. (42), 847–862.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

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

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*(3), 280–295. https://doi.org/10.1037/0022-0663.79.3.280

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The Big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review*, *20*(3), 319–350. https://doi.org/10.1007/s10648-008-9075-6

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: reciprocal effects models of causal ordering. *Child Development*, *76*(2), 397–416. https://doi.org/10.1111/j.1467-8624.2005.00853.x

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2016). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000078

Mullis, I. V., Martin, M. O., Beaton, A. E., 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.

Muthen, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267–316.

Muthén, B., & Muthén, L. K. (1998-2012). *Mplus User's Guide. Seventh Edition*. Los Angeles and CA.: Muthén & Muthén.

Nagy, G., Neumann, M., Trautwein, U., & Lüdtke, O. (2010). Voruniversitäre Mathematikleistungen vor und nach der Neuordnung der gymnasialen Oberstufe in Baden-

Württemberg [Advanced math before and after the reform of general upper secondary school in Baden-Württemberg]. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke, & K. Maaz (Eds.), *Schulleistungen von Abiturienten. Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand* (pp. 147–180). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92037-5_6

NEPS. (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

Neumann, M., Trautwein, U., & Nagy, G. (2011). Do central examinations lead to greater grading comparability? A study of frame-of-reference effects on the University entrance qualification in Germany. *Studies in Educational Evaluation*, *37*(4), 206–217. https://doi.org/10.1016/j.stueduc.2012.02.002

Nomi, 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

Nomi, T., & Allensworth, E. (2009). “Double-dose” algebra as an alternative strategy to remediation: Effects on students' academic outcomes. *Journal of Research on Educational Effectiveness*, *2*(2), 111–148. https://doi.org/10.1080/19345740802676739

OECD. (2004). *Learning for tomorrow’s world: First results from PISA 2003*. Paris: OECD.

OECD. (2015). *Education policy outlook 2015*. Paris: OECD Publishing.

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 multicontext study. *Developmental Psychology*, *48*(6), 1629–1642. https://doi.org/10.1037/a0029167

Penfield, R. D., & Camilli, G. (2006). 5: Differential item functioning and item bias. In C. R. Rao & S. Sinharray (Eds.), *Handbook of Statistics 26. Psychometrics* (Vol. 26, pp. 125–167). Elsevier. https://doi.org/10.1016/S0169-7161(06)26005-X

Porter, R. E., Fusarelli, L. D., & Fusarelli, B. C. (2015). Implementing the common core: How educators interpret curriculum reform. *Educational Policy, 29*(1), 111–139. https://doi.org/10.1177/0895904814559248

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

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

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353–387. https://doi.org/10.1037/a0026838

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, *130*(2), 261–288. https://doi.org/10.1037/0033-2909.130.2.261

Robinson, M., & Monks, J. (2005). Making SAT scores optional in selective college admissions: A case study. *Economics of Education Review*, *24*(4), 393–405. https://doi.org/10.1016/j.econedurev.2004.06.006

Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940-2009. *Teachers College Record*. (114), 1–23.

Rupp, A. A., Vock, M., Harsch, C., & Köller, O. (2008). *Developing standards-based assessment tasks for English as a first language: Context, processes and outcomes in Germany*. *Standards-Based Assessment Tasks for English as a First Language: Vol. 1*. Münster: Waxmann.

Schinske, J., & Tanner, K. (2014). Teaching more by grading less (or differently). *CBE life sciences education*, *13*(2), 159–166. https://doi.org/10.1187/cbe.CBE-14-03-0054

Schoon, I., & Eccles, J. S. (Eds.). (2014). *Gender differences in aspirations and attainment: A life course perspective*. Cambridge: Cambridge University Press.

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

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

Sørensen, A. B. (1970). Organizational differentiation of students and educational opportunity. *Sociology of Education*, *43*(4), 355. https://doi.org/10.2307/2111838

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, *35*(5), 401–426. https://doi.org/10.1016/j.intell.2006.09.004

Swanson, C. B., & Stevenson, D. L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. *Educational Evaluation and Policy Analysis*, *24*(1), 1–27. https://doi.org/10.3102/01623737024001001

The National Commission on Excellence. (1983). *A nation at risk: The imperative for educational reform*. Washington: Government Printing Office.

Thorsen, C., & Cliffordson, C. (2012). Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research and Evaluation*, *18*(2), 153–172. https://doi.org/10.1080/13803611.2012.659929

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., & 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* (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: Routledge.

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*(S2), 233–249. https://doi.org/10.1007/s11618-011-0196-1

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). 5: Development of competencies across the life span. In H.-P. Blossfeld, H. G. Rossbach, & J. von Maurice (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67–86). Zeitschrift für Erziehungswissenschaft: 14. https://doi.org/10.1007/s11618-011-0182-7

Westphal, A., Becker, M., Vock, M., Maaz, K., Neumann, M., & McElvany, N. (2016). The link between teacher-assigned grades and classroom socioeconomic composition: The role of classroom behavior, motivation, and teacher characteristics. *Contemporary Educational Psychology*, *46*, 218–227. https://doi.org/10.1016/j.cedpsych.2016.06.004

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, *39*(1), 1–37. https://doi.org/10.1111/j.1745-3984.2002.tb01133.x

Zimmermann, F., Schütte, K., Taskinen, P., & Köller, O. (2013). Reciprocal effects between adolescent externalizing problems and measures of achievement. *Journal of Educational Psychology*, *105*(3), 747–761. https://doi.org/10.1037/a0032793

Table 1

Correlations between Grades and Achievement in the two Federal States

	Math grade (TH)	English grade (TH)	Math grade (BW)	English grade (BW)
Math achievement (TH)	.48	.32	-	-
English achievement (TH)	.32	.62	-	-
Math achievement (BW)	-	-	.64	.31
English achievement (BW)	-	-	.25	.62

Note. Standardized achievement was modeled by latent variables in these models instead of point estimates to avoid biased estimates due to unreliability. TH = Thuringia; BW = Baden-Württemberg.

Table 2

Descriptive Statistics for the Central Outcome Variables

Variable				Thuringia					
	Before the reform						After the reform		
Course	Basic (B)			Advanced (A)			Core (C)		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Math grade	614	8.21_{ac}	3.02	512	9.24_{b}	3.04	959	8.94_{b}	3.12
English grade	719	8.89_{ac}	2.45	369	10.11_{bc}	2.20	924	9.62_{ab}	2.43
Math ACH	614	45.35_{ac}	8.40	512	54.93_{bc}	9.27	959	50.34_{ab}	9.91
English ACH	719	47.50_{ac}	8.81	369	54.10_{bc}	9.79	924	50.30_{ab}	9.32
				Baden-Württemberg					
Math grade	1636	8.01_{a}	3.59	1017	9.85_{bc}	3.08	2800	8.27_{a}	3.64
English grade	1198	9.13_{a}	2.86	1078	9.67_{bc}	2.70	2479	9.25_{a}	2.92
Math ACH	1636	43.75_{ac}	6.96	1017	57.15_{bc}	9.49	2800	51.06_{ab}	9.31
English ACH	1198	46.6_{ac}	9.53	1078	54.20_{bc}	9.77	2479	49.80_{ab}	9.67

Note. Standardized achievement was transformed into scores with M = 50 and SD = 10. ACH = standardized achievement. N = Sample size; M = Mean; SD = Standard deviation. On the basis of results by Hübner, Wagner, Nagengast, and Trautwein (2017), we decided to exclude the specific group of basic course students in English after the reform, as only very few students with a different language as the core subject chose English as a basic course after the reform. Compared with Trautwein et al. (2010), we considered data from only general upper secondary schools. p-values were controlled for the false discovery rate within a subject and country (Benjamini & Hochberg, 1995). Indices indicate two-sided statistically significant group differences (p < .05) between the advanced, basic, or core courses, respectively: Index $_{b}$ = statistically significantly different from basic course; Index $_{a}$ = statistically significantly different from advanced courses; Index $_{c}$ = statistically significantly different from core course.

Table 3

Achievement Predicted by Average Grades in Thuringia and Baden-Württemberg

	Thuringia	
Course / Subject	Mathematics	English
Advanced	54.3_{bc}	52.5_{bc}
Basic	46.0_{ac}	48.7_{ac}
Core	50.3_{ab}	50.1^{ab}
	Baden-Württemberg	
Advanced	54.8_{bc}	53.4_{bc}
Basic	44.3_{ac}	47.0_{ac}
Core	51.6_{ab}	50.0_{ab}

Note. Grades were given on a scale from 0 (*lowest achievement*) to 15 points (*highest achievement*). *p*-values were controlled for the false discovery rate within a subject and state (Benjamini & Hochberg, 1995). The metric of the latent variable was transformed to $M = 50$ and $SD = 10$ on the basis of pooled means and the standard deviation of the latent variable. Indices indicate two-sided statistically significant group differences ($p < .05$) between the advanced, basic, or core courses, respectively: Index $_b$ = statistically significantly different from basic course; Index $_a$ = statistically significantly different from advanced courses; Index $_c$ = statistically significantly different from core course.

Table 4

Achievement in Mathematics Predicted in different Grade Groups in Thuringia and Baden-Württemberg

		Thuringia		
Course / Group	D	C	B	A
Advanced	48.2_{bc}	52.7_{bc}	56.8_{bc}	64.2_{b}
Basic	41.8_{ac}	45.9_{a}	46.8_{ac}	50.5_{ac}
Core	44.4_{ab}	47.7_{a}	52.5_{ab}	59.7_{b}
		Baden-Württemberg		
Advanced	49.4_{bc}	52.8_{bc}	58.3_{bc}	65.8_{bc}
Basic	40.0_{ac}	43.0_{ac}	46.4_{ac}	50.6_{ac}
Core	43.8_{ab}	49.7_{ab}	55.3_{ab}	63.3_{ab}

Note. Grades were given on a scale ranging from 0 (*lowest achievement*) to 15 points (*highest achievement*). Grade groups were built as follows: Group D: < 6.5 points, Group C: ≥ 6.5 and < 9.5 points, Group B: ≥ 9.5 and < 12.5 points, Group A: ≥ 12.5 points. *p*-values were controlled for the false discovery rate within a subject, state, and grade group (Benjamini & Hochberg, 1995). When considering only grades from Grade 12 Semester 1 in Thuringia, all differences in grade group results remained comparable. However, all differences in Grade Groups A and C were statistically significant in these analyses, and basic and core courses were not statistically significantly different in Grade Group D. The metric of the latent variables were transformed to *M* = 50 and *SD* = 10 on the basis of pooled means and the standard deviations of the latent variables. Indices indicate two-sided statistically significant group differences ($p < .05$) between the advanced, basic, or core courses, respectively: Index $_b$ = statistically significantly different from basic course; Index $_a$ = statistically significantly different from advanced courses; Index $_c$ = statistically significantly different from core course. All differences between grade groups were statistically significant at $p < .05$ except for the basic course for Grade Groups C and B in Thuringia.

Table 5

Achievement in English Predicted by Different Grade Groups in Thuringia and Baden-Württemberg

		Thuringia		
Course / Group	D	C	B	A
Advanced	41.8	48.4_{b}	55.9_{bc}	63.1
Basic	40.4	44.4_{a}	50.9_{a}	61.4
Core	41.9	46.5	52.4_{a}	60.9
		Baden-Württemberg		
Advanced	44.9_{bc}	50.3_{bc}	56.5_{bc}	64.0_{bc}
Basic	39.6_{ac}	44.4_{ac}	48.9_{ac}	57.3_{ac}
Core	42.0_{ab}	46.8_{ab}	52.4_{ab}	61.1_{ab}

Note. Grades were given on a scale from 0 (*lowest achievement*) to 15 points (*highest achievement*). Grade groups were built as follows: Group D: < 6.5 points, Group C: ≥ 6.5 and < 9.5 points, Group B: ≥ 9.5 and < 12.5 points, Group A: ≥ 12.5 points. *p*-values were controlled for the false discovery rate within a subject, state, and grade group (Benjamini & Hochberg, 1995). When considering only grades from Grade 12 Semester 1 in Thuringia, all differences in grade group results remained comparable. However, in these analyses, advanced and basic courses in Grade Group C were statistically significant different as well as advanced and basic courses in Grade Group A. The metric of the latent variable was transformed to $M = 50$ and $SD = 10$ on the basis of pooled means and the standard deviation of the latent variable. Indices indicate two-sided statistically significant group differences ($p < .05$) between the advanced, basic, or core courses, respectively: Index $_{b}$ = statistically significantly different from basic course; Index $_{a}$ = statistically significantly different from advanced courses; Index $_{c}$ = statistically significantly different from core course. All differences between grade groups were statistically significant at $p < .01$.

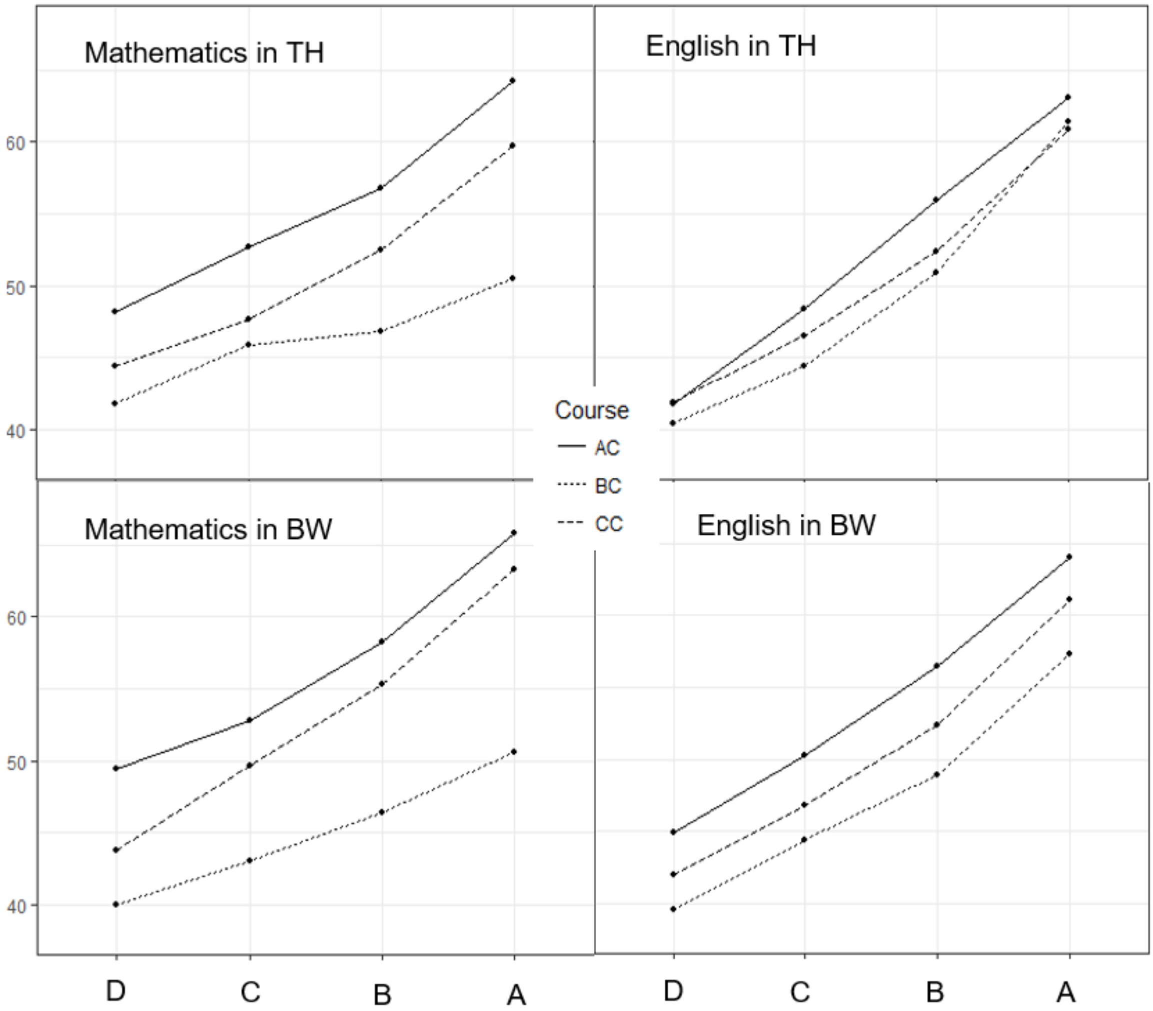

Figure 1. Standardized achievement by grade group and course. In Baden-Württemberg, all differences between courses within a specific grade group were statistically significant for both subjects. Tables 4 and 5 provide exact information about the statistical significance of differences displayed for Thuringia. The grade groups were built as follows: Group D: < 6.5 points, Group C: ≥ 6.5 and < 9.5 points, Group B: ≥ 9.5 and < 12.5 points, Group A: ≥ 12.5 points. The metric of the latent variable was transformed to *M* = 50 and *SD* = 10 on the basis of the pooled means and the standard deviation of the latent variable. TH = Thuringia; BW = Baden-Württemberg. AC = Advanced course; BC = Basic course, CC = Core course.

4.4 Study 4

Hübner, N., Wagner, W., Kramer, J., Nagengast, B., & Trautwein, U. (2017). Die G8-Reform in Baden-Württemberg: Leistungen, Wohlbefinden und Freizeitverhalten vor und nach der Reform (The G8-reform in Baden-Württemberg: Competencies, wellbeing and leisure time before and after the reform). *Zeitschrift für Erziehungswissenschaft*. doi:10.1007/s11618-017-0737-3

Zusammenfassung

Die Konsequenzen der Einführung des achtjährigen Gymnasiums (G8) werden in Politik und Öffentlichkeit kontrovers diskutiert, u.a. weil es lange an belastbaren empirischen Daten mangelte. Der vorliegende Beitrag untersucht die Frage, ob sich Abiturientinnen und Abiturienten aus G8- und G9-Jahrgängen in Baden-Württemberg im Hinblick auf verschiedene Kompetenzbereiche sowie in ihren Selbstberichten zu ihrer schulischen Beanspruchung, ihren gesundheitlichen Beschwerden und in ihrem Freizeitverhalten unterschieden. Die Analysen beruhen auf Daten von vier Kohorten der Zusatzstudie Baden-Württemberg des Nationalen Bildungspanels: der letzte reine G9-Jahrgang (*N* = 1341), der G9-Doppeljahrgang (*N* = 1284), der G8-Doppeljahrgang (*N* = 1293) und der erste reine G8-Jahrgang (*N* = 1292). Im Hinblick auf die fachspezifischen Kompetenzen von Schülerinnen und Schülern zeigten sich zwischen G8- und G9-Jahrgängen in den Bereichen Mathematik und Physik keine Unterschiede, in Biologie geringfügige und in der Englisch-Lesekompetenz substanzielle Unterschiede zugunsten der Schülerinnen und Schüler aus G9-Jahrgängen. Bei der schulischen Beanspruchung und den gesundheitlichen Beschwerden fanden sich in G8-Jahrgängen substanziell höhere Werte. Im Hinblick auf das Freizeitverhalten fanden sich uneinheitliche Ergebnisse. Fragen nach Ursachen der Reformeffekte sowie Implikationen der Befunde für die Schulpolitik werden abschließend diskutiert.

Schlüsselwörter: G8/G9 Reform, Kompetenzen, schulisches Beanspruchungserleben, gesundheitliche Beschwerden, Freizeitverhalten

Abstract

The present study compared students from G8 and G9 cohorts in Baden-Württemberg in regard to cognitive variables such as competence in mathematics, English reading, physics and biology as well as non-cognitive outcomes such as school related stress, health problems and leisure time activities. Based on representative data from the National Educational Panel Study (NEPS; Add-on-Study Baden-Württemberg), students from four cohorts from 2011 to 2013 were compared. In regard to the subject-specific competences we found no differences between students from G8 and G9 cohorts in mathematics and physics, minor disadvantages for G8 students in biology and the largest disadvantage for G8 students in English reading achievement. Concerning stress and health problems we found disadvantages for G8 students, whereas effects for leisure time use remained inconsistent. Interpretations of the findings and possible implications will be discussed.

Keywords: G8-reform, competencies, school related stress, health problems, leisure time

Einleitung

Die Verkürzung der ursprünglich neunjährigen gymnasialen Schulzeit auf acht Jahre bei gleichzeitiger Beibehaltung des Gesamtvolumens von 265 Jahreswochenstunden wurde in zahlreichen westdeutschen Bundesländern in der ersten Dekade des neuen Millenniums umgesetzt (KMK, 2014; Trautwein & Neumann, 2008). Diese flächendeckende Einführung von G8 wurde und wird von Befürwortern und Gegnern kontrovers diskutiert (Jacobsen & Buhse, 2013; *Schul-Volksbegehren in Niedersachsen,* 2011; Tulodetzki & Gohr, 2012; Vieth-Entus, 2014). In Niedersachsen wurde mit Verweis auf die vermuteten negativen Effekte der G8-Reform inzwischen eine landesweite Rückkehr zu G9 zum Schuljahr 2015/2016 (KMK, 2014; Kultusministerium Niedersachsen, 2014) veranlasst, andere Bundesländer haben G9-Optionen eingeführt.

Die intensive öffentliche Diskussion um G8/G9 steht in auffälligem Kontrast zu einem „Schweigen" der Erziehungswissenschaft, der nach Weiler (2003) sowohl bei der Einführung von G8 als auch bei der jetzigen (partiellen) Rückkehr zu G9 keine bedeutsame Rolle zukam (für eine Ausnahme, vgl. Spiewak, 2014). Tatsächlich lässt sich der derzeitige Forschungsstand zu den Reformeffekten der Schulzeitverkürzung als unbefriedigend bezeichnen (Kühn, van Ackeren, Bellenberg, Reintjes & Im Brahm, 2013). Dies drückt sich ebenfalls im Fehlen eines konkreten theoretischen Rahmenmodells aus, welches die Reform z.B. in Bezug auf ihre Entstehung, ihre Ziele und potenziell wirksam werdenden Mechanismen oder Nebenwirkungen auf der Ebene des Unterrichts, der Schule oder unter Rückbezug auf weitere Akteure systematisch fundiert. In dem vorliegenden Beitrag werden beispielhaft für ein Bundesland Daten zu den Effekten von G8 zum Zeitpunkt des Abiturs vorgestellt und dazu genutzt, die Rolle von empirischen Befunden in der politischen Meinungsbildung zu diskutieren.

Diskussionen und Forschungsbefunde zu Schulzeitverkürzungen

Das Gymnasium und seine Weiterentwicklung haben schon immer in besonderer Weise die Aufmerksamkeit von Bildungspolitik und Öffentlichkeit gefunden (Fuchs, 2004; Trautwein & Neumann, 2008). Ein besonders umstrittenes Thema war und ist die Beschulungsdauer auf dem Gymnasium. Für die Einführung bzw. Beibehaltung von G8 (z.B. Herrmann, 2002; Kühn et al., 2013) wurden u.a. ökonomische und demographische Argumente ins Feld geführt; darüber hinaus wurde auf Straffungsmöglichkeiten im Curriculum des G9 sowie eine wahrgenommene Entwicklungsakzeleration von Kindern und Jugendlichen verwiesen, weshalb G8 auch eine Stärkung der Eigenverantwortlichkeit junger Erwachsener ermögliche. Hingegen kritisieren Befürworter des G9 die Argumente für G8 als zu vereinfacht (vgl. Kühn et al., 2013;

siehe auch Herrmann, 2002). Besonders hervorgehoben wird dabei die Qualität gymnasialer Bildung, die durch G9 besser garantiert werden könne als durch G8, wobei neben Aspekten des Kompetenzerwerbs und des interessenorientierten Lernens auch mögliche positive Effekte auf die Persönlichkeitsentwicklung im weiteren Sinne genannt werden. Zusätzliche Argumente, die für G9 angeführt werden, betreffen negative Auswirkungen von G8 auf die Berufs- und Studienorientierung, Auslandsaufenthalte, extracurriculare Aktivitäten, Stresserleben und gesundheitlichen Beschwerden. Zudem werden mögliche negative Effekte von G8 in leistungsheterogenen Klassen sowie in Hinblick auf die Durchlässigkeit des Schulsystems (im Sinne der Aufwärtsmobilität) thematisiert.

Insgesamt ist die empirische Datenlage im Vergleich zur Bedeutung der Thematik und zum Ausmaß der Umsetzung der flächendeckenden Reformmaßnahmen in fast allen Bundesländern eher dünn und fällt sehr viel weniger eindeutig aus als viele Befürworter von G8 oder G9 suggerieren. Man kann in dieser Debatte drei unterschiedliche Datenquellen unterscheiden (vgl. Kühn et al., 2013):

Erstens sind Befunde aus Studien mit begabten und hochbegabten Schülerinnen und Schülern zu nennen (z.B. Heller, 2002). Die oftmals vorgetragenen positiven Befunde aus Studien zu verkürzten Schulzeiten für diese Schülerschaft („Hochbegabtenzüge“) eignen sich jedoch nicht für eine Generalisierung auf breitere Schülergruppen, von methodischen Problemen der entsprechenden Studien ganz abgesehen.

Zweitens werden teilweise internationale Befunde zum Zusammenhang von Beschulungsdauer und Schulleistungen in die Diskussion eingebracht. Inzwischen liegen eine Reihe von Reviews vor, die – bei relativ großer Streuung der Befunde – in der Mehrheit einen eher positiven Zusammenhang zwischen Beschulungsdauer und Schulleistung bzw. anderen kognitiven Kriteriumsmaßen nahelegen (vgl. Ceci, 1991; Patall, Cooper & Allen, 2010; Scheerens, 2014). Allerdings unterscheiden sich die berichteten Studien im Hinblick auf Stichproben, Zeitmaße und Zielkriterien so stark, dass ihre Implikationen für die Situation in Deutschland nur sehr gering sind.

Die dritte Gruppe von Studien, Vergleiche von G8- und G9-Regelgymnasien, sind potenziell besonders aussagekräftig, allerdings ist die Datenlage in Hinblick auf relevante Kriteriumsmaße sehr begrenzt. Die vorliegenden Leistungsvergleiche zwischen Schülerinnen und Schülern aus G8- und G9-Systemen beziehen sich nahezu ausnahmslos auf Schulnoten. Hier zeigten sich überwiegend keine oder kleine Effekte teilweise gegensätzlicher Natur, die nur teilweise statistisch signifikant waren (Büttner & Thomsen, 2015; Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, 2013). Generell ist bei der Interpretation der Effekte der Reform auf Schulnoten kritisch anzumerken, dass sich Schulnoten nur begrenzt dafür eignen, Reformeffekte auf die Leistungsentwicklung adäquat abzubilden, da die Noten starken Referenzgruppeneffekten unterliegen können (Trautwein, Lüdtke, Marsh, Köller & Baumert, 2006; vgl. Trautwein, Lüdtke, Becker, Neumann & Nagy, 2008).

Auch für weitere Kriteriumsmaße wie Lernverhalten und Beanspruchungserleben ist die empirische Befundlage dünn und uneinheitlich. Böhm-Kasper und Weishaupt (2002) untersuchten in einer Studie verschiedene psychosoziale Merkmale wie beispielsweise den Leistungsdruck, das Schulklima, Beanspruchungsgefühle und die Konkurrenz zwischen Schülerinnen und Schülern in der Klassenstufe 8 und in der Kursstufe. Sie fanden uneinheitliche Effekte innerhalb und zwischen den untersuchten Bundesländern und deutliche Geschlechtereffekte. Unabhängig vom Bundesland fühlten sich Schülerinnen höher belastet als Schüler. Auch Milde-Busch et al. (2010) gingen der Frage nach Zusammenhängen einer verkürzten Gymnasialzeit mit dem gesundheitlichen Beschwerden bei Münchener Schülerinnen und Schülern der Klassestufen 10 (G8) und 11 (G9) nach und fanden lediglich im Hinblick auf den Anteil unverplanter Freizeit und in Bezug auf die Einschätzung der Erholung in dieser Zeit substanzielle Unterschiede zuungunsten der G8-Schülerinnen und Schüler. Quis (2015) untersuchte bereits Schülerinnen und Schüler des G8-G9-Doppeljahrgangs in Baden-Württemberg hinsichtlich möglicher Unterschiede im Wohlbefinden, ebenfalls auf Basis der Daten des Nationalen Bildungspanels, jedoch ohne den ersten reinen G8-Jahrgang. Es zeigte sich ein Unterschied von rund 30% einer Standardabweichung beim Beanspruchungserleben und 10% einer Standardabweichung bei den gesundheitlichen Beschwerden zuungunsten der G8-Kohorte. Trotz des Fehlens eines klaren konzeptuellen pädagogischen Rahmens für die Reform lassen sich natürlich mögliche Wirkfaktoren aus der wissenschaftlichen Literatur heranziehen. Im vorliegenden Falle liegen die Wirkfaktoren (Änderungen im Curriculum, Beibehaltung von Gesamtstundenzahl, Veränderung der Gesamt-Schulzeit, Alter beim Abitur, etc.) in einer komplexen Mischung vor, deren Gesamteffekt nur schwer zu antizipieren ist. Aus diesem Grund wollen wir unseren Artikel nicht im Sinne einer experimentellen Studie als Prüfung einer expliziten Theorie (z.B. in Bezug auf die Lernzeit; Bloom, 1968; Carroll, 1963, 1989), bzw. eines bestimmten Wirkfaktors unter Kontrolle aller anderen verstehen, sondern wir untersuchen ein Maßnahmenpaket. Eine theoretische Einbettung der G8-Reform (auch über den Bezug zur Lernzeit hinaus) ist daher notwendig, kann von uns aber in diesem Artikel, der zunächst Ergebnisse der Reform darstellt, nicht vollumfänglich geleistet werden. Gleichwohl sei darauf verwiesen, dass aktuelle Studien mit Schülerinnen und Schüler aus G8- und G9-Jahrgängen auf Basis von PISA-Daten in Klassenstufe 9 kleine Vorsprünge der G8-Kohorte nahelegen, schwache Schüler nicht von der Reform zu profitieren scheinen und Leistungsunterschiede zwischen starken und schwachen Schülerinnen und Schüler sich verstärken (Huebener, Kuger & Marcus, 2016)

Umsetzung der G8-Reform in Baden-Württemberg

In der öffentlichen Wahrnehmung wird häufig nicht beachtet, dass die G8-Reformen in den einzelnen Bundesländern unterschiedlich implementiert wurden. Neben der „reinen“ Schulzeitverkürzung sollten deshalb immer auch weitere Faktoren identifiziert werden, die einen Effekt auf die Kriteriumsmaße haben können.

In Baden-Württemberg wurden im Zuge der Umsetzung der G8-Reform die durchschnittlichen Wochenstunden am allgemeinbildenden Gymnasium (Trautwein & Neumann, 2008) erhöht, um die Vorgabe der Kultusministerkonferenz von 265 Jahreswochenstunden bis zum Abitur einzuhalten. Darüber sah der gemeinsam mit G8 eingeführte neue Bildungsplan für das Gymnasium die Einführung von Bildungsstandards mit Kerncurricula, die Verpflichtung zur Erstellung eines Schulcurriculums, das Erlernen einer zweiten Fremdsprache ab Klassenstufe 5 und die Einführung des Faches Naturwissenschaft und Technik (NwT) vor (Kultusministerium Baden-Württemberg, 2004a, Kultusministerium Baden-Württemberg, 2004b).

Ein Vergleich der Stundentafeln zeigt, dass sich als bedeutsamer Unterschied in Bezug auf die G8- und der G9-Systeme z.B. die Stundenreduktionen im Fach Mathematik in der Sekundarstufe I (G8: 24 Stunden; G9: 28 Stunden) nennen lässt. Im Fach Biologie erfolgte eine Reduktion um durchschnittlich 2 Stunden in der Sekundarstufe I. Für die erste und zweite Fremdsprache kam es zu einer Stundenreduktion in der Sekundarstufe I, die durch die Einführung von acht Jahreswochenstunden Grundschulenglisch für alle Jahrgänge ab dem Einschulungsjahr 2004/2005 kompensiert wurde. Im Fach Physik blieb das Stundenvolumen gleich (Kultusministerium Baden-Württemberg, 2004b; Landesinstitut für Schulentwicklung, 1999). Die hier berücksichtigten G8-Jahrgänge hatten in der Grundschule also noch keinen Englischunterricht, wie von der Reform für aktuelle G8-Jahrgänge vorgesehen. Dies sollte bei einem Vergleich der Englischleistung von G8- und G9- Jahrgängen stets berücksichtigt werden.

Fragestellung

Welche Effekte die G8-Reformen in den einzelnen Bundesländern hatten, ist höchst umstritten und empirisch weitgehend ungeklärt. Für das Bundesland Baden-Württemberg werden in dieser Studie auf der Basis belastbarer Daten nun erstmals zentrale Kriteriumsmaße untersucht. Dabei ist zu beachten, dass – wie in den anderen Bundesländern auch – die G8-Reform in Baden-Württemberg von weiteren Maßnahmen begleitet wurde.

In der hier vorgestellten Studie wird die Veränderung in den Kompetenzen in vier Domänen (Mathematik, Englisch-Lesekompetenz, Physik und Biologie) untersucht. Hierbei stellt sich als zentrale Frage, ob und in welchem Maße sich die G8-Reform in geringeren Kompetenzen niederschlug. Im Hinblick auf das Freizeitverhalten wurde die Befürchtung geäußert, dass Abiturienten in G8 weniger Zeit für außerunterrichtliche Aktivitäten wie Sport und Musik haben könnten (z.B. Greiner & Himmelrath, 2014; Laging, Böcker & Dirks, 2014). In der vorliegenden Studie konnten insgesamt elf Freizeitaktivitäten herangezogen werden, um etwaige Effekte zu prüfen. Schließlich wurde in Bezug auf das Beanspruchungserleben und die selbst eingeschätzten gesundheitlichen Beschwerden untersucht, ob sich diese zwischen den G8- und G9-Abiturienten unterscheiden.

Methode

Stichprobe

Es wurden Daten aus drei Erhebungswellen (Studiennummern: A72, A73 und A74) aus dem Scientific Use File Version 3.0.0.[1] der NEPS Zusatzstudie Baden-Württemberg (Blossfeld, Rossbach & Maurice, 2011) herangezogen (vgl. Tabelle 1). Konkret wurden der G9-Abschlussjahrgang 2011 (Welle I), der „Doppeljahrgang" 2012 (Welle II) sowie der erste reine G8-Abschlussjahrgang 2013 (Welle III) erfasst. Es handelt sich also um ein Kohorten-Kontroll-Design (Shadish, Cook & Campbell, 2002), welches hier die Grundlage für ein natürliches Experiment bildet (Murnane & Willett, 2011).

Insgesamt nahmen 48 zufällig gezogene Schulen aus Baden-Württemberg (zwei dieser Schulen konnten aus organisatorischen Gründen in der erste Welle nicht berücksichtigt werden) mit insgesamt rund 5000 Abiturienten (Welle 1: N = 1341; Welle 2: N = 2577; Welle 3: N = 1292) an der Untersuchung teil.[2]

Instrumente

[1] *Diese Arbeit nutzt Daten des Nationalen Bildungspanels (NEPS) Zusatzstudie Baden-Württemberg, doi:10.5157/NEPS:BW:3.0.0. Die Daten des NEPS wurden von 2008 bis 2013 als Teil des Rahmenprogramms zur Förderung der empirischen Bildungsforschung erhoben, welches vom Bundesministerium für Bildung und Forschung (BMBF) finanziert wurde. Seit 2014 wird NEPS vom Leibniz-Institut für Bildungsverläufe e.V. (LIfBi) an der Otto-Friedrich-Universität Bamberg in Kooperation mit einem deutschlandweiten Netzwerk weitergeführt.*

[2] In der ersten Welle liegen zusätzlich zu den Daten der G9-Schülerinnen und Schüler auch Daten von 52 Schülerinnen und Schülern aus dem G8-Schnelläufer Jahrgang vor. Diese wurden als Teil des ursprünglichen G9-Systems betrachtet (das sogenannte G8-Schnelläuferklassen umfasste) und bei den Analysen entsprechend als G9-Schüler kodiert. Der Ausschluss der G8-Schnelläufer hatte keinen Einfluss auf die Ergebnisse.

Mathematische Kompetenz. Aufgaben zur Messung der mathematischen Kompetenz basierten auf dem Konzept der *Mathematical Literacy,* das auch in PISA und den Nationalen Bildungsstandards verwendet wird (NEPS, 2011). Hierbei werden vier Inhaltsbereiche unterschieden: *Quantität, Raum und Form, Veränderung und Beziehungen* sowie *Daten und Zufall,* die sich wiederum in sechs Komponenten mathematischer Denkprozesse unterscheiden lassen: technische Fertigkeiten einsetzen, modellieren, argumentieren, kommunizieren, repräsentieren und Probleme lösen. Im Mathematiktest wurden jeweils vier Items in den Bereichen *Quantität* und *Raum und Form* sowie jeweils sechs Items in den Bereichen *Veränderung und Beziehung* und *Daten und Zufall* administriert (Duchhardt, 2015). Insgesamt wurden in der NEPS Zusatzstudie 21 Mathematikitems im Multiple Choice oder offenen Antwortformat administriert, für deren Bearbeitung 30 Minuten Zeit zur Verfügung standen. Die Aufgaben orientieren sich in der Mehrzahl an den Inhalten der Mittelstufe.[3]

Englisch-Lesekompetenz. Zur Erfassung der Englisch-Lesekompetenz wurde auf am Institut zur Qualitätsentwicklung im Bildungswesen (IQB) entwickelte Aufgaben zurückgegriffen (Rupp, Vock, Harsch & Köller, 2008). Diese Aufgaben berücksichtigen einerseits die Bildungsstandards für das Fach Englisch, auf der anderen Seite orientieren sie sich am Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER; Europarat, 2001). Im Englischtest wurden insgesamt fünf Items auf dem Niveau B1, vier Items auf dem Niveau B1/B2 und 16 Items auf dem Niveau B2 administriert. Darüber hinaus lagen acht Items auf dem C1 Niveau des GER vor. Insgesamt wurden 33 Aufgaben, die die Niveaustufen B1 bis C1 (selbständige bis kompetente Sprachverwendung) abdecken, administriert (21 Items pro Testheft). Die Bearbeitungszeit lag bei 30 Minuten (Hübner, Rieger & Wagner, 2016b).

Biologische Kompetenz. Die Erfassung der – in der NEPS-Studie so bezeichneten – „biologischen Kompetenz“ erfolgte anhand eines im Rahmen der EVAMAR II-Studie (Eberle et al., 2008) entwickelten Instruments. Ähnlich wie bei der mathematischen Kompetenz wurde zunächst eine Unterteilung des Konstrukts in Inhaltsbereiche und drei Klassen kognitiver Anforderungsbereiche vorgenommen. Im Biologietest wurden mit 27 Items die Bereiche *Cytologie, Anatomie und Soffwechsel,* mit 10 Items die Bereiche *Informationsverarbeitung, Verhalten und Immunbiologie* und mit 7 Items die Bereiche *Genetik und Entwicklungsbiologie* erfasst. Darüber hinaus wurden 11 Items zum Thema *Ökologie* sowie 5 Items im Bereich *Systematik und Evaluation* administriert.

[3] Ein Item wurde im Zuge der Skalierung ausgeschlossen (vgl. Duchhardt, 2015).

Bei den kognitiven Anforderungsbereichen handelt es sich zunächst um die Stufe I, die sich mit dem Reproduzieren und Anwenden von Eingeübtem beschäftigt, und um Stufe II, die kognitive Operationen erfordert, die auf das Umstrukturieren und Übertragen von Inhalten abzielen. Die letzte Stufe III nimmt schließlich Operationen des Beurteilens und Problemlösens in den Fokus (vgl. NEPS, 2011). In der NEPS-Zusatzstudie Baden-Württemberg wurden Biologische Kompetenzen mit insgesamt 60 Items gemessen. Jede Schülerin und jedem Schüler sollte im Rahmen des Booklet-Designs dabei ein Ausschnitt von 36 Items bearbeiten. Die vorgegebene Bearbeitungszeit betrug insgesamt 45 Minuten. Die Items wurden in Multiple Choice Format oder in offenen Antwortformaten präsentiert (NEPS, 2011). Die Aufgaben orientieren sich primär an den Inhalten der Kursstufe (Hübner, Rieger & Wagner, 2016a).

Physikalische Kompetenz. Die physikalische Kompetenz wurde mit 41 Items erfasst, die zum Teil aus vorhandenen Instrumenten (z.B. TIMSS; Baumert et al., 1999) übernommen wurden, zum Teil speziell für die beiden NEPS-Zusatzstudien (Thüringen, Baden-Württemberg) entwickelt wurden (NEPS, 2011)[4]. Hierbei sollte jede Schülerin und jeder Schüler einen Ausschnitt aller Items (19 bis 21 Items pro Testheft) bearbeiten. Im Physiktest wurden drei Items aus dem Bereich Elektrische Felder und Wechselwirkung, sechs Items aus dem Bereich Magnetische Felder und Elektromagnetische Induktion und zwei Items aus dem Bereich Spezielle Relativitätstheorie administriert. Darüber hinaus beinhaltete der Test jeweils vier Items aus den Bereichen Wellen, Quantenphysik: Quanten und Materie, Dynamik: Schwingungen und Dynamik: Mechanik des starren Körpers. Zuletzt wurden für die Bereiche Optik und Thermodynamik jeweils sieben Items administriert. Die Bearbeitungszeit für den Test lag ebenfalls bei insgesamt 45 Minuten. Die Items waren im Multiple Choice, Forced Choice sowie im offenen Antwortformat formuliert. Die Konstruktion dieser Items orientiert sich an den Einheitlichen Prüfungsanforderungen für die Abiturprüfung (EPA) in Physik. Die Aufgaben orientieren sich primär an den Inhalten der Kursstufe (Hübner, Rieger & Wagner, 2016c).

Die Analyse aller Kompetenzen erfolgte simultan, unter Verwendung eines vierdimensionalen Mehrgruppen-1PL-IRT-Modells. Für alle Tests zeigten sich substantielle Zusammenhänge zwischen der jeweiligen Note im Fach am Ende der Sekundarstufe II und der latenten Variable der Testleistung, die für Mathematik bei r =.59 lag, für Englisch bei r =.57, für Biologie bei r = .49 und für Physik bei r =.51. Die Kodierung der Items aller Kompetenztests

[4] Wir bedanken uns im Namen der Etappe 5 (Gymnasiale Oberstufe und Übergänge in (Fach-)Hochschule, Ausbildung oder Arbeitsmarkt) des NEPS für die Unterstützung bei der Erstellung des Physiktests bei Knut Neumann, Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN).

in „korrekt“ und „falsch“ liegt im aktuellen Scientific Use File 3.0.0 (Blossfeld, Rossbach & Maurice, 2011) bereits vor, sodass dies nicht im Rahmen der Analysen der vorliegenden Studie erfolgte. Offene Antworten wurden entweder als „falsch“ (0) oder „korrekt“ (1) kodiert. Bei Items, bei denen sowohl ein numerischer Wert als auch eine Maßeinheit angegeben werden musste, wurden Antworten nur als korrekt gewertet, wenn beide Angaben richtig waren. Fehlende Angaben wurden gemäß der NEPS Standards (Pohl & Carstensen, 2012) mit dem speziellen Missing Code „nicht bearbeitet“ gekennzeichnet, unleserlichen Angaben wurden als „nicht valide“ kodiert. Die Kodierung dieser offenen Items erfolgte computerbasiert per Syntax nach den Vorgaben des Auswertungsmanuals.

Schulbezogenes Beanspruchungserleben. Das schulbezogene Beanspruchungserleben wurde mit einer im Rahmen der NEPS-Zusatzstudie konzipierten Skala mit 15 Items erfasst. Dabei sollten die Abiturientinnen und Abiturienten die Zustimmung zu verschiedenen schulbezogenen Aussagen von 1 (stimme gar nicht zu) bis 4 (stimme völlig zu) beurteilen (Itembeispiele: „Wenn ich von der Schule nach Hause komme, bin ich angespannt“ und „Manchmal kann ich schwer einschlafen, weil mir Probleme aus der Schule durch den Kopf gehen“). Die interne Konsistenz der Skala (Cronbachs α) lag bei .91.

Gesundheitliche Beschwerden. Selbstberichte über gesundheitliche Beschwerden wurden mit insgesamt 26 Items erfasst. Hierbei handelt es sich um eine Skala, die bereits im Rahmen von PISA 2003 eingesetzt wurde (Bergmüller, 2003) Schülerinnen und Schüler sollten jeweils die Häufigkeit des Auftretens verschiedener physischer und psychischer Symptome in den letzten sechs Wochen auf einer Skala von 1 (nie) bis 4 (öfter als sechsmal) angeben. Gefragt wurde hierbei beispielsweise nach „starkem Herzklopfen“, „Angst, dass alles zu viel wird“ oder „Erbrechen“. Die Skala wies eine interne Konsistenz (Cronbachs α) von .93 auf. Die Auswertungen der gesundheitlichen Beschwerden erfolgte simultan mit dem Beanspruchungserleben unter Anwendung von Graded Response Modellen (Embretson & Reise, 2000; Samejima, 1997).

Freizeitverhalten. Das Freizeitverhalten wurde in Stunden pro Woche für insgesamt elf Bereiche erfasst (Trautwein, Neumann, Nagy, Lüdtke & Maaz, 2010). Diese sind „Freizeitangebote der Schule (z.B. Sport-, Hobby-, Arbeitsgruppen)“, „Computer spielen, chatten etc.“, „Freundinnen und Freunde treffen“, „Fernsehen“, „Lesen“, „etwas mit der Familie unternehmen“, „Sport treiben (alleine, mit Freundinnen oder Freunden, im Verein)“, „zum Orchester, Kirchengruppen oder anderen Gruppen (außer Sport) gehen“, „Zeit mit anderen Hobbys verbringen (z.B. Instrumente, Basteln)“, „Nebenjob“ und „Sonstiges“. Diese Items sollten in Bezug auf die wöchentliche Beschäftigung und in Stunden beantwortet werden.

Da es bei dieser Skala keine Antwortmöglichkeit für „keine Betätigung“ im jeweiligen Freizeitbereich gab, gilt zu beachten, dass fehlende Werte und „keine Betätigung“ bei diesen Skalen nicht eindeutig unterscheidbar sind und lediglich Aussagen über die relative Betätigungszeit von Schülerinnen und Schülern möglich ist, die eine konkrete Betätigungszeit angaben.

Sozialer und kultureller Hintergrund. Die Erfassung des sozialen Status der Familie der Schülerinnen und Schüler erfolgte auf Basis des International Socio-Economic Index of Occupational Status 2008 (ISEI-08; Ganzeboom & Treiman, 2003). Aus dem ISEI-08 wurde in den vorliegenden Analysen der höchste ISEI (HISEI) aus dem jeweils höchsten ISEI der beiden Eltern gebildet. Der häusliche Buchbestand diente als Indikator des kulturellen Kapitals. Der familiäre Migrationsstatus wurde auf Basis des Geburtslands der Eltern bestimmt. Dabei wurde als Kriterium die Geburt mindestens eines Elternteils im Ausland festgelegt.

Kognitive Grundfähigkeiten. In der Zusatzstudie Baden-Württemberg wurden als nonverbale kognitive Grundfähigkeiten einerseits die Wahrnehmungsgeschwindigkeit und andererseits das schlussfolgernde Denken der Schülerinnen und Schüler erfasst (NEPS, 2011). Konkret wurde die Wahrnehmungsgeschwindigkeit über den Bilder-Zeichen-Test (NEPS-BZT) erfasst, einem Test mit insgesamt 93 Items, für die in jeweils drei Itemblöcken mit 31 Items eine Bearbeitungszeit von jeweils 30 Sekunden pro Block vorgesehen war. Das schlussfolgernde Denken wurde mit Hilfe eines Matrizentests erfasst (NEPS-MAT), bei dem insgesamt 12 Items verwendet wurden. Der Test misst figurale kognitive Fähigkeiten (Brunner, Lang & Lüdtke, 2014). Die Bearbeitung dieser Items erfolgte in drei Blöcken mit jeweils vier Items; hierfür standen jeweils drei Minuten pro Block Zeit zur Verfügung.

Kursbelegung. Im Rahmen des Schülerfragebogens wurde erfasst, ob die Schülerinnen und Schüler die Fächer Englisch, Biologie und Physik in der Oberstufe abgewählt bzw. als Kernfach gewählt hatten.

Statistische Analyse

Zunächst wurden Unterschiede in den Kursbelegungsquoten der Verschiedenen Gruppen für die Bereiche Englisch-Lesekompetenz, Biologie und Physik mittels multinomialer logistischer Regressionen mit der Gruppenzugehörigkeit (G9W1, G9W2, G8W2, G8W3) als abhängiger Variable und dummy-kodierten Prädiktoren für Kurswahl (Kernfach bzw. Abwahl) anhand von Wald-Tests geprüft (Annahme: sämtliche Regressionskoeffizienten sind gleich Null). Zur adäquaten Untersuchung möglicher Unterschiede von Schülerinnen und Schülern aus G8- und G9-Jahrgängen wurde ein mehrstufiges Vorgehen gewählt. Die einzelnen

Kompetenzmaße wurden zunächst mit einem eindimensionalen Rasch-Modell, bzw. Partial-Credit-Modell skaliert, um die psychometrische Qualität des Tests und der einzelnen Items zu überprüfen. Es wurden DIF-Analysen für den HISEI, den Migrationshintergrund, das Geschlecht, das Kursniveau (bei Biologie- bzw. Physiktest) und die Erhebungswelle durchgeführt. Hierbei zeigte sich insgesamt nur auf wenigen Items starker DIF >= 0.60 Logits. Der Ausschluss dieser Items aus den Analysen führte zu keiner substantiellen Veränderung der Ergebnisse. DIF bedeutet nicht zwangsläufig, dass Items „unfair" sind (item bias), sondern können auch ein Hinweis auf valide Unterschiede zwischen Subgruppen darstellen (Zumbo, 1999). In verschiedenen Studien konnte darüber hinaus gezeigt werden, dass IRT-basierte Modellinferenzen bei moderaten Verletzungen der Messinvarianzannahme relativ robust sind (Rupp & Zumbo, 2006). Anschließend wurden die Kompetenzdaten, ebenso wie das schulische Beanspruchungserleben und die gesundheitlichen Beschwerden unter Verwendung von mehrdimensionalen Mehrgruppen-IRT-Modellen ausgewertet. Hierbei wurden zunächst vier latente Variablen (eine latente Dimension pro Kompetenzbereich) spezifiziert und deren Kovarianz frei geschätzt. Die Schätzung der latenten Mittelwerte der jeweiligen Kompetenzdimension erfolgte unter Verwendung des MLR-Schätzers. Indikatoren der latenten Dimensionen wurden als kategorial definiert. Der Vorteil einer mehrdimensionalen Skalierung gegenüber einer eindimensionalen Skalierung liegt in der theoretisch plausiblen und komplexen Abhängigkeit der Kompetenzen untereinander, die in einem mehrdimensionalen Modell explizit berücksichtigt werden kann (vgl. Reckase, 2009) und mit einer höheren Teststärke einhergehen sollte. Die Analysen zum Freizeitverhalten erfolgten schließlich unter Verwendung von Mehrgruppen-Analysen für metrische Daten. Sämtliche Analysen wurden in Mplus 7.4 durchgeführt (Muthén & Muthén, 1998-2012). Zu berücksichtigen ist, dass die von uns spezifizierten latenten Variablenmodelle Parameterschätzungen bezogen auf „messfehlerfrei" erfasste Konstrukte ermöglichen (sofern die Modelle angemessen spezifiziert wurden), wobei geringere Reliabilität der Instrumente sich lediglich in (etwas) größeren Standardfehlern der Schätzungen niederschlägt.

Aus den jeweiligen Analysen resultierten Parameterschätzungen getrennt für vier Kohorten: Dem letzten reinen G9-Jahrgang (G9W1), dem G9-Doppeljahrgang (G9W2), dem G8-Doppeljahrgang (G8W2) und dem ersten reinen G8-Jahrgang (G8W3). Die Analyse der Reformeffekte erfolgte auf Basis verschiedener möglicher Kohortenvergleiche. Hierbei erfolgte sowohl ein Vergleich des Doppeljahrgangs, der beiden reinen Jahrgänge als auch ein Vergleich der gesamten G8- versus G9-Schülerinnen und Schüler. Diese Vergleiche bieten sich an, da sich Schülerinnen und Schüler im Doppeljahrgang zumindest theoretisch von Schülerinnen und

Schülern in den beiden reinen Jahrgängen unterscheiden können. Hierbei gilt zu berücksichtigen, dass es einerseits keine Unterschiede zwischen Schülerinnen und Schülerinnen im Doppeljahrgang in Bezug auf den Lernstoff in der Oberstufe gab, während sich andererseits Schülerinnen und Schüler der reinen Jahrgänge diesbezüglich theoretisch unterscheiden können. Gleichzeitig führt die gemeinsame Unterrichtung von G8- und G9-Schülerinnen im Doppeljahrgang möglicherweise auch zu Referenzgruppeneffekten, die bei einem Vergleich der beiden reinen Jahrgänge praktisch auszuschließen sind. Aufgrund des neuen Curriculums in der Oberstufe, das beim ersten reinen G8-Jahrgang erstmals implementiert war, sollten sich die diesbezüglichen Befunde auch eher auf aktuelle G8-Jahrgänge generalisieren lassen.

Eine zentrale Herausforderung in quasi-experimentellen Designs besteht in der Trennung von Selektions- und Behandlungseffekten (Morgan & Winship, 2007; Murnane & Willett, 2011). Dazu wurden, unter Berücksichtigung zusätzlicher Daten des statistischen Landesamts, mögliche Selektionsunterschiede (z.B. in Übergangsquoten und Nichtversetztenquoten) geprüft. Anschließend wurde die Vergleichbarkeit der Schülerinnen und Schüler aus den G8- bzw. G9-Kohorten im Abschlussjahr bezüglich relevanter Hintergrundmerkmale untersucht. In einem letzten Schritt erfolgte schließlich die Untersuchung von Unterschieden der G8- und G9-Schülerinnen und Schüler auf den Kriteriumsmaßen unter Kontrolle von Hintergrundmerkmalen.

Alle Analysen erfolgten zunächst im Rahmen eines unadjustierten Modells (ohne Kovariaten) und anschließend mit Adjustierung. Bei den berücksichtigten Kovariaten handelte es sich um das Geschlecht, den Migrationshintergrund, den häuslichen Buchbestand, kognitive Grundfähigkeiten[5] und Informationen zu Klassenwiederholungen in der gesamten Sekundarsufe. Die im Rahmen der unadjustierten Modellschätzungen aufgeführten Werte spiegeln die Mittelwerte der Variablen für die jeweiligen Kohorten ohne Adjustierung wider. Bei den Analysen der Kompetenzen wurde in einem zusätzlichen Modell das Kursniveau statistisch kontrolliert. Die adjustierten Modelle bieten zusätzlich zu den Modellen ohne Kovariaten die Möglichkeit einer Betrachtung von Unterschieden zwischen den jeweiligen Kohorten, unter statistischer Kontrolle möglicher bestehender Gruppenunterschiede.

In den adjustierten Modellen wurden die Kovariaten vor der Analyse am Gesamtmittelwert über die drei Erhebungswellen zentriert. Unterscheiden sich die

[5] Die kognitiven Grundfähigkeiten hatten in allen Modellen einen nicht signifikanten oder einen geringen Einfluss auf die Ergebnisse. Die adjustierten Modelle mit und ohne Kontrolle der kognitiven Fähigkeiten unterschieden sich hinsichtlich ihrer Ergebnisse nicht bedeutsam voneinander.

gruppenspezifischen Mittelwerte für die Kovariaten bei Regressionsgewichten ungleich Null, repräsentieren die Intercepts aus diesen Modellen adjustierte Gruppenmittelwerte für die „typische“ Schülerkomposition (als durchschnittliche Zusammensetzung in allen drei Erhebungswellen). Um Unterschiede zwischen Schülerinnen und Schülern aus den verschiedenen Kohorten auf den abhängigen Variablen zu untersuchen, wurden entsprechend Mittelwert- (unadjustierte Modelle) bzw. Intercept-Differenzwerte (adjustierte Modelle) inklusive Standardfehler geschätzt. Zur besseren Interpretierbarkeit möglicher Unterschiede wurden die aus den Analysen resultierenden Parameter linear transformiert. Für die Kompetenzen erfolgte eine Transformation auf eine Metrik mit *M* = 500 und *SD* = 100. Die Ergebnisse zum Wohlbefinden wurden in die T-Metrik überführt (*M* = 50 und *SD* = 10).

Die mit der Adjustierung verbundenen Annahmen sind zwar plausibel, müssen aber keinesfalls zwingend zu korrekten (oder wenigstens korrekteren) Schätzungen führen. So könnten etwa Kurswahlunterschiede auch als Reformeffekte interpretiert werden, sodass eine Adjustierung für das Kursniveau – je nach Blickwinkel – auch als ungerechtfertigt betrachtet werden kann. Auch eine Adjustierung auf Basis von im Abschlussjahrgang erhobener Maße der allgemeinen kognitiven Fähigkeiten kann prinzipiell zu Verzerrungen führen, da diese möglicherweise durch die Reform beeinflusst wurden. Aufgrund dieser Einschränkungen lässt sich kein klar zu favorisierendes Analysemodell formulieren, wenngleich der Einbezug von Hintergrundmerkmalen für die Schätzung unverzerrter Reformeffekte sinnvoll erscheint. Unterscheiden sich die Schätzungen aus verschiedenen Modellen (unterschiedliche Adjustierungen oder ohne Adjustierung) nur geringfügig, so kann dies im Sinne einer hohen „Robustheit“ der Befunde interpretiert werden.

Die Besonderheiten des Sampling Designs (Ziehung von Schulen, Surveygewichte; Schönberger & Aßmann, 2014) wurden anhand entsprechender Mplus-Optionen berücksichtigt (Type = Complex; Weight-Option). Fehlende Werte wurden in den vorliegenden Analysen mithilfe der Full Information Maximum Likelihood-Methode (FIML) berücksichtigt.

Ergebnisse

Selektivitätsanalysen

Um zu prüfen, ob die Schülerschaft der vier berücksichtigten Kohorten vergleichbar war oder sich von vornherein (z.B. durch Klassenwiederholungen oder Schulwechsel) unterschied, wurden zunächst Hinweise auf unterschiedliche Selektionsprozesse näher untersucht. Auf Basis von Daten des Statistischen Landesamts Baden-Württemberg (2014b) wurden zunächst gymnasiale Übergangsquoten untersucht. Hierbei zeigte sich für die Jahre

2003, 2004 und 2005 ein geringfügiger Anstieg (2003: 35,3 %; 2004: 36,1 %; 2005: 37,8 %), der vor dem Hintergrund eines allgemein zunehmenden gymnasialen Übergangsverhaltens interpretiert werden kann.

Neben dem Übergangsverhalten sind auch die Anteile der Nichtversetzten und der Klassenwiederholer zentral für die Vergleichbarkeit von Schülerinnen und Schülern unterschiedlicher Kohorten. Die Nichtversetztenquote variierte zwischen Klassenstufe 5 und 11 nur geringfügig zwischen G8- und G9-Jahrgängen (G9: 0,4 % - 3,1 %; G8: 0,4 % - 3,7 %;Schwarz-Jung, 2008; Statistisches Landesamt Baden-Württemberg, 2014a). Die Gruppe der G8-Schülerinnen und Schüler aus dem Doppeljahrgang wies allerdings einen besonders geringen Anteil an Klassenwiederholern auf (Statistische Ämter des Bundes und der Länder, 2015). Da sich die Nichtversetztenquote aus Klassenwiederholern und Abgängern zusammensetzt, lässt sich schließen, dass ein größerer Anteil der nichtversetzten Schülerinnen und Schüler aus dem letzten G9-Jahrgang eher auf eine andere Schulform wechselte anstatt eine Klasse zu wiederholen. In der zweiten G8-Kohorte zeigte sich dann wieder eine Wiederholerquote vergleichbar mit der vor der Reform. Zur Vergleichbarkeit der Kohorten wurden Klassenwiederholungen daher in den adjustierten Analysen statistisch kontrolliert. Bei der Überprüfung möglicher Unterschiede in den Belegungsquoten zeigten sich für die Bereiche Physik ($\chi^2(6) = 5{,}68$, $p = {,}46$) und Biologie ($\chi^2(6) = 9{,}62$ $p = {,}14$) keine Unterschiede. Für das Fach Englisch fand sich ein statistisch bedeutsamer Unterschied ($\chi^2(6) = 27{,}57$, $p < {,}001$). So wählten G9W1-Schülerinnen und Schüler Englisch weniger häufig als Kernfach (91%; in den nachfolgenden Kohorten jeweils mindestens 94%) und häufiger als Grundkurs (4%; in den nachfolgenden Kohorten jeweils weniger als 1%). Die Abwählerquote lag in sämtlichen Kohorten relativ konstant im Bereich von 5% bis 6%.

Bei der deskriptiven Statistik (vgl. Tabelle 2) zeigten sich in Bezug auf die meisten Variablen lediglich geringfügige Unterschiede zwischen den untersuchten Kohorten. Schülerinnen und Schüler aus G8-Kohorten waren im Mittel erwartungsgemäß ein Jahr jünger als Schülerinnen und Schüler aus G9-Kohorten.

Kompetenzen der Abiturientinnen und Abiturienten vor und nach der Oberstufenreform

Für die Mathematik ergaben sich in den adjustierten Modellen (ohne bzw. mit Kontrolle des Kursniveaus) keine statistisch signifikanten Unterschiede zwischen beiden G9- und G8-Kohorten (adjustiert ohne Kursniveau: M_{G9ges}-M_{G8ges}: -3, $p = {,}54$; adjustiert mit Kursniveau: M_{G9ges}-M_{G8ges}: -4, $p = {,}25$, siehe Tabelle 3). Auch die übrigen Gruppenvergleiche in den Modellen mit Adjustierung waren nicht statistisch signifikant. Das Ergebnismuster des adjustierten Modells zeigte sich auch in den Modellen ohne Berücksichtigung weiterer Kovariaten.

Bei der Englisch-Lesekompetenz fanden sich statistisch signifikante Unterschiede zwischen beiden G9- und G8-Kohorten sowohl im adjustierten Modell ohne und unter Kontrolle des Kursniveaus. Im Mittel schnitten hier Schülerinnen und Schüler aus G9-Jahrgängen rund 18 bzw. 20 Punkte besser ab als Schülerinnen und Schüler aus G8-Jahrgängen. Gleiches gilt für die Unterschiede zwischen den Kohorten des Doppeljahrgangs und den beiden reinen G8- bzw. G9-Jahrgängen, bei denen ebenfalls jeweils die G9-Jahrgänge höhere Werte aufwiesen (vgl. Tabelle 3).

Für die Biologische Kompetenz ergab sich ein Unterschied zwischen Schülerinnen und Schülern aus G9- und G8-Jahrgängen, der jedoch nur im adjustierten Modell mit Kursniveau statistisch signifikant war. Darüber hinaus unterschieden sich Schülerinnen und Schüler aus dem Doppeljahrgang in ihrer Biologiekompetenz nicht voneinander. Der Vergleich der reinen G9- bzw. G8-Jahrgänge ergab einen statistisch signifikanten Unterschied zugunsten der Schülerinnen und Schüler im letzten reinen G9-Jahrgang. Für die Physikkompetenz fanden sich ähnlich wie bei der Mathematikkompetenz keine Unterschiede zwischen Schülerinnen und Schülern aus G9- und G8-Jahrgängen (vgl. Tabelle 3).[6]

Schulisches Beanspruchungserleben und gesundheitliche Beschwerden

Beim schulischen Beanspruchungserleben zeigte sich zunächst ein signifikanter Effekt für den Unterschied zwischen Schülerinnen und Schülern aus G9- und G8-Jahrgängen (M_{G9ges}-M_{G8ges}: -4,0, $p < ,001$), bei dem G8-Schülerinenn und Schüler angaben, sich im Mittel höher beansprucht zu fühlen. Darüber hinaus fanden sich Unterschiede zwischen dem G8-G9-Doppeljahrgang (M_{G9W2}-M_{G8W2}: -3,1, $p < ,001$) und bei einem Vergleich des letzten reinen G9-Jahrgangs mit dem ersten reinen G8-Jahrgang (M_{G9W1}-M_{G8W3}: -4,9, $p < ,001$). Diese Ergebnisse waren äquivalent zu den Ergebnissen im unadjustierten Modell (vgl. Tabelle 4).

In Bezug auf die gesundheitlichen Beschwerden zeigten sich im Mittel ebenfalls höhere Werte bei Schülerinnen und Schülern aus G8-Jahrgängen. Der Unterschied zwischen den Kohorten innerhalb des G8-G9-Doppeljahrgangs wurde nicht signifikant, wohingegen der Unterschied zwischen den beiden reinen Jahrgängen statistisch signifikant war. Es fanden sich ebenfalls keine Unterschiede zwischen diesen Ergebnissen und den Ergebnissen im unadjustierten Modell (vgl. Tabelle 4).

[6] Die adjustierten Modelle mit und ohne Kontrolle der kognitiven Fähigkeiten unterschieden sich hinsichtlich ihrer Ergebnisse nicht bedeutsam voneinander.

Freizeitverhalten

Bei der Analyse der Angaben zu Zeitinvestitionen für Freizeitbereiche zeigten sich in vier der elf untersuchten Bereiche signifikante Unterschiede im adjustierten und im unadjustierten Modell zwischen G9- und G8-Jahrgängen (vgl. Tabelle 5). Zu beachten gilt, dass diese Analysen lediglich die Informationen von Schülerinnen und Schülern berücksichtigen, die Angaben zur durchschnittlichen wöchentlichen Dauer der Aktivitäten in einem Freizeitbereich gemacht haben. Nicht berücksichtigt werden konnte hierbei die relative Betätigungshäufigkeit, da „keine Betätigung“ nicht als Antwortoption vorgesehen war und somit nicht von fehlenden Werten („nicht bearbeitet“) unterschieden werden konnte. Für den Bereich „Freunde treffen“ lag der Unterschiede zwischen allen G8- und G9-Schülerinnen und Schülern bei 96 Minuten (M_{G9ges}-M_{G8ges}: 95,9, p < ,001). Hierbei gaben Schülerinnen und Schüler im G9-Jahrgang durchschnittlich eine längere Beschäftigungsdauer in diesem Freizeitbereich an. Der Unterschied zwischen den beiden reinen G8- und G9-Jahrgängen belief sich hier auf rund 171 Minuten (M_{G9W1}-M_{G8W3}: 170,5, p < ,001), ebenfalls mit höheren Angaben der G9-Schülerinnen und Schüler. Im Freizeitbereich „Nebenjob“ gaben die Schülerinnen und Schüler aus G9-Jahrgängen im Mittel eine höhere zeitliche Investition an als Schülerinnen und Schüler aus G8-Jahrgängen (M_{G9ges}-M_{G8ges}: 75,3, p < ,001). Weiterhin fanden sich signifikante Unterschiede für die Bereiche „Sport treiben“ und „Fernsehen“, die sich auf 18,2 Minuten und 22,3 Minuten beliefen und bei denen jeweils G9-Schülerinnen und Schüler eine längere Beschäftigungsdauer angaben.

Diskussion

Die G8-Reform gilt als *die* zentrale Reform des Gymnasiums des ersten Jahrzehnts im neuen Jahrtausend (Trautwein & Neumann, 2008). Mit der vorliegenden Studie konnten nun erstmals – zumindest für ein Bundesland – Befunde vorgestellt werden, die auch standardisierte Kompetenzmaße umfassen sowie auf einer repräsentativen Stichprobe beruhen. Im Folgenden werden zunächst die Ergebnisse zusammengefasst und mögliche Erklärungsansätze vorgestellt, bevor auf Implikationen für die Bildungspolitik in Baden-Württemberg sowie dem Bundesgebiet eingegangen wird. Abschließend wird die Rolle der Bildungsforschung bei Bildungsreformen kritisch hinterfragt.

Zentrale Ergebnisse und Erklärungsansätze

In Bezug auf die Kompetenzen fand sich ein bemerkenswertes Ergebnismuster: Während sich in Mathematik und Physik keinerlei Leistungseinbußen durch G8 fanden, zeigten sich für die Lesekompetenz in Englisch substanzielle sowie für Biologie tendenzielle Unterschiede zugunsten der G9-Absolventen. Eine mögliche Erklärung ist, dass die Umstellung auf G8 in den einzelnen Fächern unterschiedlich gut gelang. Im Fach Englisch kam es in Baden-Württemberg wegen der gleichzeitig zur Umstellung auf G8 erfolgten Einführung des Grundschulenglisch und der parallelen Reduktion des Unterrichtsvolumens in Englisch in der Sekundarstufe I um insgesamt acht Wochenstunden zu einer vorübergehenden Reduktion der Gesamtstundenzahl bis zum Abitur; auch war die Wertigkeit des Faches Englisch wegen des parallelen – inzwischen wieder aufgehobenen – Starts mehrerer Fremdsprachen in Klassenstufe 5 für die Schülerinnen und Schüler ggf. etwas in Frage gestellt. Vielleicht spielt es aber auch eine Rolle, dass Englisch nicht nur in der Schule gelernt wird, sondern auch im Freizeitbereich (Fernsehserien, Musik, Reisen, Alltagskultur) eine Rolle spielt und im G9 also auch im nichtschulischen Bereich mehr gelernt werden konnte. Zu beachten ist, dass die Unterschiede im Fach Englisch durchaus substanziell ausfielen; sollten die Absolventen jedoch in nennenswerter Zahl das „gewonnene“ Jahr für einen Aufenthalt im englischsprachigen Ausland nutzen, könnte dies den Unterschied rasch wettmachen.

Prononcierte Unterschiede zugunsten von G9 fanden sich beim schulischen Beanspruchungserleben und den gesundheitlichen Beschwerden. Dies war auch der Fall im Doppeljahrgang. Dieser Befund mag etwas überraschen, da die Schülerinnen und Schüler aus G8 und G9 im Doppeljahrgang gemeinsam die Kurse besuchten und mit exakt denselben schulischen Anforderungen konfrontiert waren. Darüber hinaus zeigten sich Unterschiede zwischen Schülerinnen und Schülern, die ähnlich groß oder leicht größer als die Kohortenunterschiede zwischen G8- und G9-Schülerinnen und Schülern ausfielen. Als Erklärungsansätze kommen deshalb in Frage, dass (1) die Absolventen aus G8 jünger sind und deshalb für dieselben Anforderungen mehr Energie aufwenden müssen, (2) die G8-Absolventen in der Mittelstufe Defizite aufbauten, die in der Oberstufe korrigiert werden, oder dass (3) die Selbstberichte der G8-Absolventen z.T. auch die öffentliche Diskussion um die erwarteten negativen Folgen der Schulzeitverkürzung widerspiegeln. Leider standen für die Auswertungen keine objektiven Markiervariablen für die Gesundheit zur Verfügung, so dass offen bleiben muss, wie sehr die genannten möglichen Ursachen zu dem Ergebnismuster beigetragen haben. Hinweisen sollte man allerdings für die Einordnung der Bedeutsamkeit der Befunde auch darauf, dass die Kohortenunterschiede geringer ausfielen als die (in jeder Kohorte auftretenden) Unterschiede zwischen männlichen und weiblichen Abiturienten.

Hinsichtlich der Freizeitaktivitäten bestätigen die vorliegenden Daten die Befürchtungen, wonach es zu einem Einbruch bei „wertvollen“ Freizeitaktivitäten bei G8-

Absolventen käme, nur sehr bedingt; zum Zeitpunkt des Abiturs fanden sich in der Mehrzahl der berücksichtigten Bereiche keine signifikanten Unterschiede.

Die dokumentierten Befunde entsprechen somit nur teilweise den oftmals vorgebrachten Sorgen in Hinblick auf G8. Die Datenbasis für die hier vorgestellten Analysen darf hierbei als gut gelten. So wurde im Nationalen Bildungspanel ein Kohorten-Kontroll-Design umgesetzt, bei dem unmittelbar aufeinanderfolgende Kohorten untersucht wurden. Zur Absicherung der Befunde wurde eine Serie von unterschiedlichen Modellen berechnet, die sich in den berücksichtigten Kontrollvariablen unterschieden. Insgesamt zeigten sich hierbei keine oder nur sehr geringe Unterschiede zwischen den adjustierten und den unadjustierten Modellen. Wichtig für eine adäquate Interpretation und Einordnung der Ergebnisse bezüglich der Leistungsunterschiede der Schülerinnen und Schüler ist die Qualität der Messinstrumente. Wie bereits oben angeführt, wurden die Leistungstest latent modelliert, sodass aufgrund der teilweise unbefriedigenden Score-Reliabilität einzelner Instrumente keine Verzerrungen der Effektstärken zu erwarten sind. Darüber hinaus zeigten unsere Analysen durchaus substantielle Zusammenhänge zwischen den Fachnoten am Ende der Sekundarstufe II und den Leistungstests sowie insgesamt geringes DIF (auch in Bezug auf den Kohortenvergleich) und einen moderaten Itemfit. Diese Ergebnisse legen keine aufgabenspezifischen Unterschiede nahe, sondern lassen eher vergleichbare Ergebnisse bei einem größeren Itempool erwarten. Gleichwohl zeigte sich auch, dass das *test targeting* noch nicht vollständig befriedigend war. So ist der Englischtest tendenziell eher leicht für die Schülerinnen und Schüler, während der Physiktest viele schwierige Items enthielt. Diese Tendenz zeigte sich jedoch in gleicher Weise sowohl für Schülerinnen und Schüler aus den G8- als auch den G9-Kohorten. Bezogen auf die Validität der eingesetzten Leistungstests ist zu bemerken, dass diese in unterschiedlichem Ausmaß das Curriculum repräsentieren. Besonders deutlich ist die unvollständige Abdeckung des Curriculums beim Englischtest, der lediglich Lesekompetenz (auf insgesamt eher niedrigem Niveau) erfasst, womit in der vorliegenden Studie beispielsweise der Bereich der produktiven Teilkompetenzen im Englischen nicht berücksichtigt wurde. Wenn man aber davon ausgeht, dass die Leistungstests die kohortenspezifischen Curricula jeweils in vergleichbarer Weise abdecken, dann lassen sich die gefundenen Unterschiede (weitgehend) im Sinne von Effekten der Schulzeitverkürzung auf die jeweils erfasste Kompetenz interpretieren.

Bildungspolitische Implikationen

Welche Implikationen haben die Ergebnisse in Hinblick auf bildungspolitische Entscheidungen? Sind sie ein Beleg für das Funktionieren von G8 in Baden-Württemberg oder lassen sie sich als Basis für eine Forderung nach Rückkehr zu G9 verwenden? Grundsätzlich ist festzuhalten, dass (1) die Ergebnisse nur einen Teil der Wirkungen von G8 reflektieren und (2) erst durch eine subjektive Gewichtung von Zielen und durch den Vergleich mit Erreichtem mit bildungspolitischen Implikationen angereichert werden (vgl. Bromme, Prenzel & Jäger, 2014). Im vorliegenden Fall dürfte es für eine Abschätzung des „Erfolgs" der Reform wesentlich darauf ankommen, (1) als wie bedeutsam man die „Kosten" (also beispielsweise die Kompetenzunterschiede in Englisch und beim Wohlbefinden) von G8 bewertet, (2) ob man annimmt, dass inzwischen vorgenommene Nachregulierungen bei G8 (u.a. Grundschulenglisch sowie Unterstützungsangebote in der Oberstufe) die identifizierten Schwachstellen überwinden und (3) wie positiv man das in G8 „gesparte" Lebensjahr betrachtet.

Darüber hinaus müssen bei Forderungen nach Wiedereinführungen von G9 nach dem Vorbild von Niedersachsen auch potenzielle ungewollte Nebenwirkungen bedacht werden. So würde eine erneute Reform erstens Ressourcen binden, die – so implizieren es viele empirische Studien – vielleicht effizienter in die Unterrichtsentwicklung investiert werden könnten (z.B. Hattie, 2008). Zweitens würde eine Rückkehr zu G9 dafür sorgen, dass es in absehbarer Zeit einen Jahrgang gäbe, bei dem kein Abiturient das allgemeinbildende Gymnasium verlassen würde, was wiederum massive negative Konsequenzen für die Hochschulen des Bundeslandes haben dürfte (ein „Nullerjahrgang" anstatt des „Doppeljahrgangs"). Drittens lässt sich auch spekulieren, ob eine Rückkehr zu G9 angesichts der kürzlich aufgehobenen Verbindlichkeit der Grundschulempfehlungen in Baden-Württemberg eine Veränderung des Schulwahlverhaltens zur Folge haben könnte, was wiederum im Konflikt mit der anvisierten Architektur der Schulformen stehen könnte.

Die Implikationen der vorgelegten Studie beschränken sich nicht auf nur ein Bundesland. Natürlich ist zu berücksichtigen, dass es bundesweit nicht *die* G8-Reform gab – vielmehr kam G8 immer gemeinsam mit bestimmten Veränderungen in der Organisation der Mittelstufe und bestimmten curricularen Veränderungen. In empirischen Studien lassen sich diese zwei Faktoren nur schwer trennen, so dass sich in den Befunden zum Zeitpunkt des Abiturs immer zwei Komponenten, nämlich „G8 plus landesspezifische Regelungen", niederschlagen und die spezifischen Wirkungen nicht generalisierbar sind. Trotzdem hat unsere Studie Implikationen jenseits des lokalen Kontextes eines Bundeslands. So ist festzuhalten, dass es – siehe beispielsweise die Mathematik – sehr wohl möglich ist, auch unter den Bedingungen von G8 das Abitur ohne Qualitätsverlust abzulegen. Zweitens können die identifizierten Unterschiede zwischen den Fächern als (erneuter) Beleg dafür herangezogen werden (vgl. Hattie, 2008), dass „äußeren" Faktoren, zu denen auch die Frage von G8 vs. G9 gehört, im

Vergleich zur Umsetzung von Qualität im Unterricht eine geringere Rolle spielen. Zuletzt könnten die Befunde ein erster Hinweis darauf sein, dass Zeitkompression in Bezug auf die Leistungen (z.B. in Mathematik und Physik) weniger problematisch sind als eine geringere schulische oder außerschulische Lernzeit (z.B. in Englisch).

Implikationen für Evaluationen bei Reformen

Auf einer abstrakteren Ebene kann die G8-Reform als ein Beleg für die Bedeutungslosigkeit der Erziehungswissenschaft bzw. Bildungsforschung betrachtet werden: Bei der Konzeption der Reform war sie kaum einbezogen und auf begleitende Evaluationsmaßnahmen durch die Wissenschaft, die von Anfang an mit eingeplant hätten werden können, wurde gänzlich verzichtet (vgl.). Umgekehrt lässt sich aber auch argumentieren, dass der Verzicht auf die Mitarbeit und Begleitung durch die Erziehungswissenschaft/Bildungsforschung zeigt, wie wichtig diese sein könnte.

So sollten Evaluationen von vornherein mitgeplant werden. Hierbei kann man sowohl an formative (reformbegleitende Erhebungen, die zu unmittelbaren Veränderungen führen können) und summative (die Gesamtwirkung der Reform auf unterschiedliche Kriteriumsmaße prüfende) Elemente denken. Anhand der von uns vorgestellten Studie lässt sich auch aufzeigen, wie das Studiendesign für die summativen Elemente noch aussagekräftiger hätte werden können, wenn die Studie von vornherein als Teil der Reform mitgeplant wird: So wäre es möglich gewesen, Daten auch in der Sekundarstufe I zu sammeln, in der die Beanspruchung durch G8 möglicherweise besonders deutlich ausfällt. Zudem hätten sich in Zusammenarbeit mit den Verantwortlichen im Land zusätzliche Kriteriumsmaße identifizieren und einsetzen lassen, die für (positive und negative) Reformeffekte besonders sensitiv sein könnten.

Natürlich kann und soll eine solche Begleitforschung nicht die politischen Entscheidungen ersetzen oder öffentliche Debatten überflüssig machen. Die Frage beispielsweise, ob der „Gewinn“ eines schulfreien Lebensjahres bei G8 es ggf. auch rechtfertigen würde, dass im Abitur gewisse Leistungseinbußen zu verzeichnen sind, und die Frage danach, welcher Zeitaufwand für die Schule gefordert wird und welches Maß an Belastung „akzeptabel“ ist, sind normative Entscheidungen, die als Ergebnisse von Aushandlungsprozessen in bildungspolitische Entscheidungen münden. Sie werden nicht von der Bildungsforschung gesteuert – aber diese könnte (wenn man sie denn lässt) mithelfen, die Diskussionsprozesse durch empirische Befunde zu verbreitern (vgl. Bromme, Prenzel & Jäger, 2014).

Literaturverzeichnis

Baumert, J., Bos, W., Klieme, E., Lehmann, R. H., Lehrke, M., Hosenfeld, I. et al. (Hrsg.). (1999). *Testaufgaben zu TIMSS/III. Mathematisch-naturwissenschaftliche Grundbildung und voruniversitäre Mathematik und Physik der Abschlussklassen der Sekundarstufe II (Population 3)*. Berlin: Max-Planck-Institut für Bildungsforschung.

Bergmüller, S. (2003). Schulische Belastung und gesundheitliche Beschwerden. In G. Haider & C. Reiter (Hrsg.), *PISA 2003. Internationaler Vergleich von Schülerleistungen.*. Graz: Leykam.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1* (2), 1–11.

Blossfeld, H.-P., Rossbach, H. G. & Maurice, J. von (Hrsg.). (2011). *Education as a lifelong process. The German National Educational Panel Study (NEPS)*. Zeitschrift für Erziehungswissenschaft: Sonderheft 14.

Böhm-Kasper, O. & Weishaupt, H. (2002). Belastung und Beanspruchung von Lehrern und Schülern am Gymnasium. *Zeitschrift für Erziehungswissenschaft, 5* (3), 472–499. https://doi.org/10.1007/s11618-002-0062-2

Brunner, M., Lang, F. R. & Lüdtke, O. (2014). *Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen der National Educational Panel Study: Expertise* (NEPS Working Paper No. 42). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Büttner, B. & Thomsen, S. L. (2015). Are We Spending Too Many Years in School? Causal Evidence of the Impact of Shortening Secondary School Duration. *German Economic Review, 16* (1), 65–86. https://doi.org/10.1111/geer.12038

Carroll, J. B. (1963). A model for 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

Duchhardt, C. (2015). *NEPS Technical Report for Mathematics: Scaling Results for the Additional Study Baden-Wuerttemberg*. Bamberg: Leibniz-Institut für Bildungsverläufe e.V. (LIfBi).

Eberle, F., Gehrer, K., Jaggi, B., Kottonau, J., Oepke, M. & Pflüger, M. (2008). *Evaluation der Maturitätsreform 1995. Schlussbericht zur Phase II.* Bern: Staatssekretariat für Bildung und Forschung SFB.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists* (Multivariate applications book series). Mahwah, N.J.: Lawrence Erlbaum Associates.

Europarat. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin: Langenscheid.

Fuchs, H.-W. (2004). *Gymnasialbildung im Widerstreit. Die Entwicklung des Gymnasiums seit 1945 und die Role der Kultusministerkonferenz*. Frankfurt: Peter Lang.

Ganzeboom, H. B. G. & Treiman, D. J. (2003). Three Internationally Standardised Measures for Comparative Research on Occupational Status. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Hrsg.), *Advances in Cross-National Comparison* (S. 159–193). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-9186-7_9

Greiner, L. & Himmelrath, A. (2014, 18. August). Studie zum Turbo-Abi: G8-Stress gibt es gar nicht. *Spiegel Online*. Verfügbar unter http://www.spiegel.de/schulspiegel/abi/studie-zu-turbo-abi-kaum-unterschiede-bei-abiturienten-mit-g8-und-g9-a-986159.html

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Heller, K. (2002). *Begabtenförderung im Gymnasium. Ergebnisse einer zehnjährigen Längsschnittstudie*. Opladen: Leske und Budrich.

Herrmann, U. (2002). Achtjähriges Gymnasium? Thesen Pro und Contra. Paralleltitel: 8 graded college track high school (Gymnasium)? Theses pro and con. *Die deutsche Schule, 94* (4), 471–484. Verfügbar unter http://www.digizeitschriften.de/dms/img/?PPN=PPN509092632_0094&DMDID=dmdlog104

Hübner, N., Rieger, S. & Wagner, W. (2016a). *NEPS Technical Report for Biological Competence: Scaling Results for the Additional Study Baden-Württemberg (NEPS Survey Paper No. 9)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Hübner, N., Rieger, S. & Wagner, W. (2016b). *NEPS Technical Report for English Reading: Scaling Results for the Additional Study Baden-Württemberg (NEPS Survey Paper No. 10)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Hübner, N., Rieger, S. & Wagner, W. (2016c). *NEPS Technical Report for Physics Competence: Scaling Results for the Additional Study Baden-Württemberg (NEPS Survey Paper No. 11)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Huebener, M., Kuger, S. & Marcus, J. (2016). Increased instruction hours and the widening gap in student performance. *DIW Discussion Paper, 1561*, 1–42.

IEA DPC. (2013). *Methodenbericht: NEPS Zusatzstudie zur G8-Reform in Baden-Württemberg. Haupterhebung - Frühjahr 2011 (A72).* Zugriff am 17.02.17. Verfügbar unter https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/BW/Methodenbericht_A72.pdf

IEA DPC. (2014a). *Methodenbericht: NEPS Zusatzstudie zur G8-Reform in Baden-Württemberg. Haupterhebung - Frühjahr 2012 (A73).* Zugriff am 17.02.17. Verfügbar unter https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/BW/2-0-0/Methodenbericht_A73.pdf

IEA DPC. (2014b). *Methodenbericht: NEPS Zusatzstudie zur G8-Reform in Baden-Württemberg. Haupterhebung - Frühjahr 2013 (A74).* Zugriff am 17.02.17. Verfügbar unter https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/BW/3-0-0/Methodenbericht_A74.pdf

Jacobsen, L. & Buhse, M. (2013, 22. März). Schulreform: G8 oder G9? Wer will, bleibt länger. *Die Zeit*. Verfügbar unter http://www.zeit.de/2013/12/G8-Entscheidung-Eltern

KMK. (2014). *Die gymnasiale Oberstufe.* Zugriff am 30.09.2014. Verfügbar unter http://www.kmk.org/bildung-schule/allgemeine-bildung/sekundarstufe-ii-gymnasiale-oberstufe.html

Kühn, S. M., van Ackeren, I., Bellenberg, G., Reintjes, C. & Im Brahm, G. (2013). Wie viele Schuljahre bis zum Abitur? Eine multiperspektivische Standortbestimmung im Kontext der aktuellen Schulzeitdebatte. *Zeitschrift für Erziehungswissenschaft, 16* (1), 115–136. https://doi.org/10.1007/s11618-013-0339-7

Kultusministerium Baden-Württemberg. (2004a). *Bildungsplan 2004. Allgemein bildendes Gymnasium*. Verfügbar unter http://www.bildung-staerkt-menschen.de/service/downloads/Bildungsplaene/Gymnasium/Gymnasium_Bildungsplan_Gesamt.pdf

Kultusministerium Baden-Württemberg. (2004b). *Gymnasium 2004. Das pädagogische Konzept*. Verfügbar unter http://www.kultusportal-bw.de/site/pbs-bw/get/documents/

KULTUS.Dachmandant/KULTUS/import/pb5start/pdf/ Gymnasium%202004%20Das%20pdagogische%20Konzept%20G8_Sept2004_klein.pdf

Kultusministerium Niedersachsen. (2014). *Fragen und Antworten zum modernen Abitur nach 13 Jahren,* Kultusministerium Niedersachsen. Verfügbar unter www.mk.niedersachsen.de/download/85662/Fragen_und_Antworten_zum_modernen_Abitur_nach_13_Jahren_hier_herunterladen.pdf

Laging, R., Böcker, P. & Dirks, F. (2014). Zum Einfluss der Schulzeitverkürzung (G8) auf Bewegungs- und Sportaktivitäten von Jugendlichen. *Sportunterricht, 63* (3), 66–72.

Landesinstitut für Schulentwicklung. (1999). *Stundentafel für die Klassen 5 bis 11 des allgemein bildenden Gymnasiums. Sprachliches Profil und naturwissenschaftliches Profil*. Verfügbar unter http://www.ls-bw.de/bildungsplaene/allgbilschulen/lp1994/stgysn.htm

Milde-Busch, A., Blaschek, A., Borggräfe, I., Kries, R. von, Straube, A. & Heinen, F. (2010). Besteht ein Zusammenhang zwischen der verkürzten Gymnasialzeit und Kopfschmerzen und gesundheitlichen Belastungen bei Schülern im Jugendalter? *Klinische Pädiatrie, 222* (4), 255–260. https://doi.org/10.1055/s-0030-1252012

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2013). *Ministerin Löhrmann: NRW hat Abitur mit Doppeljahrgang erfolgreich bewältigt: Ergebnisse des Zentralabiturs 2013.* Zugriff am 17.02.17. Verfügbar unter https:// www.schulministerium.nrw.de/docs/bp/Ministerium/Presse/Pressekonferenzen/2013/ 130822Zentralabitur/Zentralabitur_Sprechzettel-1.pdf

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

Murnane, R. J. & Willett, J. B. (2011). *Methods matter. Improving causal inference in educational and social science research*. Oxford: Oxford University Press.

Muthén, B. & Muthén, L. K. (1998-2012). *Mplus U ser's Guide. Seventh Edition.* Los Angeles: Muthén & Muthén.

NEPS. (2011). *G8-Reform in Baden-Württemberg: Haupterhebung 2010/11 (A72) Schüler/innen, Klasse 12/13 Informationen zum Kompetenztest*. Bamberg. Verfügbar unter https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/BW/2-0-0/C_ A72_de.pdf

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

Pohl, S. & Carstensen, C. H. (2012). *NEPS Technical Report - Scaling the Data of the Competence Tests: (NEPS Working Paper No. 14)*. Bamberg: Leibniz-Institute for Educaitonal Trajectories, National Educational Panel Study.

Quis, J. S. (2015). *Does higher learning intensity affect student well-being? Evidence from the National Educational Panel Study* (BERG working paper series on government and growth, vol. 94, neue Ausg). Bamberg: BERG.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-89976-3

Rupp, A. A. & Zumbo, B. D. (2006). Understanding Parameter Invariance in Unidimensional IRT Models. *Educational and Psychological Measurement, 66* (1), 63–84. https://doi.org/10.1177/0013164404273942

Rupp, A. A., Vock, M., Harsch, C. & Köller, O. (2008). *Developing standards-based assessment tasks for english as a first language. Context, processes and outcomes in Germany* (Standards-Based Assessment Tasks for English as a First Language, Bd. 1). Münster: Waxmann.

Samejima, F. (1997). Graded Response Model. In van der Linden,Wim J & R. K. Hambleton (Hrsg.), *Handbook of Modern Item Response Theory* (S. 85–100). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-2691-6_5

Scheerens, J. (Hrsg.). (2014). *Effectiveness of Time Investments in Education*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00924-7

Schönberger, B. & Aßmann, C. (2014). *Weighting the Additional Study in Baden-Wuerttemberg of the National Educational Panel Study,* National Educational Panel Study. Zugriff am 17.02.17. Verfügbar unter https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/BW/2-0-0/BW_2-0-0_Weighting.pdf

Schul-Volksbegehren in Niedersachsen. Turbo-Abitur wird Wahlkampfthema (2011, 30. April). *Die Tageszeitung*. Verfügbar unter http://www.taz.de/!5121786/

Schwarz-Jung, S. (2008). Allgemeinbildende Gymnasien in Baden-Württemberg flächendeckend fünf Jahrgänge im „G8“. *Statistisches Monatsheft Baden-Württemberg* (10), 3-10;. Verfügbar unter http://www.statistik-portal.de/Veroeffentl/Monatshefte/PDF/ Beitrag08_10_01.pdf

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

Spiewak, M. (2014). Glaubenskrieg um vier Stunden. Das Hickhack um G8 oder G9 ist eine Armutserklärung für die Bildungspolitik. Nun rufen Schulforscher Halt. Zugriff am 13.02.2017. Verfügbar unter http://www.zeit.de/2014/25/g8-debatte-bildungspolitik

Statistische Ämter des Bundes und der Länder. (2015). *D13.1 und D13.2: Anzahl und Anteile der Klassenwiederholungen*. Verfügbar unter http://www.statistikportal.de/Statistik-Portal/

Statistisches Landesamt Baden-Württemberg. (2014a). *Sommer 2013: Rund 15 500 Schülerinnen und Schüler an Werkreal-/Hauptschulen, Realschulen und Gymnasien verfehlen das Klassenziel,* Statistisches Landesamt Baden-Württemberg. Verfügbar unter http://www.statistik-bw.de/Pressemitt/2014257.asp

Statistisches Landesamt Baden-Württemberg. (2014b). *Übergänge aus Klassenstufe 4 an Grundschulen auf weiterführende Schulen (öffentliche und private Schulen),* Statistisches Landesamt Baden-Württemberg. Verfügbar unter http://www.statistik-portal.de/ BildungKultur/Landesdaten/uebergaenger_0000.asp?y=2003

Trautwein, U., Neumann, M., Nagy, G., Lüdtke, O. & Maaz, K. (Hrsg.). (2010). *Schulleistungen von Abiturienten. Die neugeordnete gymnasiale Oberstufe auf dem Prüfstand* (1. Aufl.). Wiesbaden: VS, Verl. für Sozialwiss.

Trautwein, U., Lüdtke, O., Becker, M., Neumann, M. & Nagy, G. (2008). Die Sekundarstufe I im Spiegel der empirischen Bildungsforschung: Schulleistungsentwicklung, Kompetenzniveaus und die Aussagekraft von Schulnoten. In E. Schlemmer & H. Gerstberger (Hrsg.), *Ausbildungsfähigkeit im Spannungsfeld zwischen Wissenschaft, Politik und Praxis* (S. 91–107). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90839-7_5

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. & Neumann, M. (2008). Das Gymnasium. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer & L. Trommer (Hrsg.), *Das Bildungswesen in der Bundesrepublik Deutschland* (Rororo Sachbuch, Bd. 62339, S. 467–501). Reinbek bei Hamburg: Rowohlt.

Tulodetzki, P. & Gohr, L. (2012, 1. Februar). Turbo-Abi. „Die Schule wird uns stressiggeredet“. *FOCUS*. Verfügbar unter http://www.focus.de/familie/schule/turbo-abi/die-schule-wird-uns-stressiggeredet-turbo-abi_id_2454177.html

Vieth-Entus, S. (2014, 5. Mai). G8, G9 und Turboabitur. Das größte Problem sind die Kultusminister. *Der Tagesspiegel*. Verfügbar unter http://www.tagesspiegel.de/meinung/g8-g9-und-turboabitur-das-groesste-problem-sind-die-kultusminister/9840140.html

Weiler, H. N. (2003). Bildungsforschung und Bildungsreform – Von den Defiziten der deutschen Erziehungswissenschaft. In G. I. & T. R. (Hrsg.), *Innovation durch Bildung. Beiträge zum 18. Kongress der Deutschen Gesellschaft für Erziehungswissenschaft* (S. 181–203). Opladen: Leske und Budrich.

Zumbo, B. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Tabelle 1

Stichprobengrößen der drei Wellen differenziert nach G8- und G9-Anteilen

Kohorte	**2011**		**2012**		**2013**	**Gesamt**	
Jahrgang	**G8***	**G9**	**G8**	**G9**	**G8**	**G8**	**G9**
Teilnahme Schüler	52	1289	1284	1293	1292	2628	2582
Teilnahme Schulen	46	46	48	48	48	48	48
Teilnahmequote	95,7		90,0		94,0	92,5	
Gesamt	1341		2577		1292	5210	

Anmerkung. Die Erhebung der ersten Kohorte fand im Zeitraum vom 03. bis zum 18. Mai 2011 statt. Die zweite Erhebung erfolgte vom 23. April bis zum 22. Mai 2012. Die Erhebung der dritten Kohorte wurde schließlich zwischen dem 13. Mai und dem 12. Juni 2015 durchgeführt (IEA DPC, 2013; 2014a, 2014b). G8*: G8-Schnelläufer Jahrgang.

Tabelle 2

Deskriptive Statistik

	G9W1	**G9W2**	**G8W2**	**G8W3**
Durchschnittsnote Abitur	2,33 (0,62)	2,35 (0,61)	2,38 (0,61)	2,36 (0,64)
Alter	19,02 (0,60)	19,05 (0,51)	17,97 (0,39)	18,07 (0,62)
Weiblich	705 (55,1%)	645 (54,7%)	673 (55,6%)	670 (55,4%)
Migrationshintergrund	287 (22,5%)	274 (23,4%)	248 (20,1%)	277 (22,9%)
Sozioökonomischer Status	61,94 (19,31)	61,10 (19,20)	61,45 (19,55)	63,09 (18,17)
Schlussfolgerndes Denken	10,80 (1,23)	10,81 (1,28)	10,72 (1,26)	10,70 (1,30)
Wahrnehmungsgeschwindigkeit	65,85 (11,52)	64,87 (11,21)	64,55 (12,06)	65,37 (11,94)
Klassenwiederholung	131 (10,2%)	112 (9,5%)	19 (1,6%)	130 (10,7%)

Anmerkung. Aufgeführt sind die Mittelwerte und Standardabweichungen bzw. für „Klassenwiederholung“ die absoluten und relativen Häufigkeiten; der sozioökonomische Status wurde mit dem höchsten ISEI gemessen.

Tabelle 3

Adjustierte und unadjustierte Mittelwerte der fachspezifischen Kompetenzen Mathematik, Englisch-Lesekompetenz, Biologie und Physik für die jeweiligen Kohorten

	G9W1 $_a$	G9W2 $_b$	G8W2 $_c$	G8W3 $_d$	$G9_{ges}$-$G8_{ges}$	*p*
unadjustierte Ergebnisse						
Mathematik	502	495 (5,50)	501 (5,08)	502 (6,02)	-2	,54
Englisch	511$_d$	509 (5,18)$_c$	488 (6,59)$_b$	493 (5,27)$_a$	20	< ,001
Biologie	507$_d$	500 (4,71)	499 (5,01)	494 (6,07)$_a$	7	,09
Physik	501	495 (5,79)	501 (6,05)	503 (5,26)	-4	,25
adjustierte Ergebnisse ohne Kursniveau						
Mathematik	502	496 (4,77)	498 (4,77)	504 (5,19)	-2	,54
Englisch	511$_d$	509 (4,83)$_c$	486 (6,23)$_b$	493 (5,18)$_a$	21	< ,001
Biologie	507$_d$	500 (4,25)	498 (5,01)	495 (5,47)$_a$	7	,07
Physik	501	496 (5,13)	497 (5,00)	506 (4,73)	-3	,31
adjustiert Ergebnisse mit Kursniveau						
Mathematik	500	496 (4,05)	498 (5,08)	506 (5,08)	-4	,25
Englisch	510$_d$	509 (4,83) $_c$	489 (5,97)$_d$	493 (5,01)$_a$	18	< ,001
Biologie	508$_d$	500 (4,40)	494 (5,31)	498 (4,86)$_a$	8	,04
Physik	500	499 (5,13)	498 (5,13)	503 (4,47)	-0,7	,84

Anmerkungen. G9W1: Schülerinnen und Schüler aus G9-Jahrgängen der ersten Erhebungswelle; G9W2: Schülerinnen und Schüler aus G9-Jahrgängen der zweiten Erhebungswelle; G8W2: Schülerinnen und Schüler aus G8-Jahrgängen der zweiten Erhebungswelle; G8W3: Schülerinnen und Schüler aus G8-Jahrgängen der dritten Erhebungswelle; G9ges: Schülerinnen und Schüler aller G9-Kohorten; G8ges: Schülerinnen und Schüler aller G8-Kohorten; $G9_{ges}$-$G8_{ges}$: Mittelwertdifferenz aller G9- und G8-Jahrgänge. Die Metrik der latenten Variablen wurde transformiert auf M = 500 und SD = 100 auf Basis der gepoolten Mittelwerte bzw. Standardabweichungen. Die berichteten p-Werte beziehen sich auf zweiseitige Tests. Subskripte indizieren Unterschiede zwischen den jeweiligen Gruppen. Für Englisch waren alle Unterschiede signifikant mit $p \leq ,001$, für Biologie mit $p < ,05$. Es werden nur Unterschiede zwischen den Doppeljahrgängen, den reinen Jahrgängen und Unterschiede innerhalb der G8- und G9-Jahrgänge dargestellt. Die finale Analysestichprobe basiert auf Daten von N = 4893 Schülerinnen und Schülern, die entweder am Schülerfragebogen oder an einem Leistungstest teilgenommen haben.

Tabelle 4

Adjustierte und unadjustierte Mittelwerte des schulischen Beanspruchungserleben und der wahrgenommenen gesundheitlichen Beschwerden nach Kohorte

	unadjustierte Ergebnisse					
	G9W1$_{a}$	G9W2$_{b}$	G8W2$_{c}$	G8W3$_{d}$	G9$_{ges}$-G8$_{ges}$	p
Beanspruchungserleben	47,2$_{bd}$	49,0 (0,48)$_{ac}$	51,7 (0,55)$_{bd}$	52,1 (0,50)$_{ac}$	-3,9	< ,001
Gesundheitliche Beschwerden	48,2$_{bd}$	49,7 (0,43)$_{a}$	50,5 (0,42)$_{d}$	51,6 (0,44)$_{ac}$	-2,1	< ,001
	adjustierte Ergebnisse ohne Kursniveau					
Beanspruchungserleben	47,2$_{bd}$	48,9 (0,52)$_{ac}$	51,9 (0,45)$_{bd}$	52,0 (0,47)$_{ac}$	-4,0	< ,001
Gesundheitliche Beschwerden	48,3$_{bd}$	49,6 (0,57)$_{a}$	50,6 (0,64)$_{d}$	51,5 (0,67)$_{ac}$	-2,1	< ,001

Anmerkungen. G9W1: Schülerinnen und Schüler aus G9-Jahrgängen der ersten Erhebungswelle; G9W2: Schülerinnen und Schüler aus G9-Jahrgängen der zweiten Erhebungswelle; G8W2: Schülerinnen und Schüler aus G8-Jahrgängen der zweiten Erhebungswelle; G8W3: Schülerinnen und Schüler aus G8-Jahrgängen der dritten Erhebungswelle; G9ges: Schülerinnen und Schüler aller G9-Kohorten; G8ges: Schülerinnen und Schüler aller G8-Kohorten; G9$_{ges}$-G8$_{ges}$: Mittelwertdifferenz aller G9- und G8-Jahrgänge. Die Metrik der latenten Variablen wurde transformiert auf M = 50 und SD = 10 auf Basis der gepoolten Mittelwerte bzw. Standardabweichungen. Die berichteten p-Werte beziehen sich auf zweiseitige Tests. Subskripte indizieren Unterschiede zwischen den jeweiligen Gruppen. Alle Unterschiede waren signifikant mit $p \leq$,001. Es werden nur Unterschiede zwischen den Doppeljahrgängen, den reinen Jahrgängen und Unterschiede innerhalb der G8- und G9-Jahrgänge dargestellt. Die finale Analysestichprobe basiert auf Daten von N = 4887 Schülerinnen und Schülern, die am Schülerfragebogen teilgenommen haben.

Tabelle 5

Bereiche der Freizeitbeschäftigung in Stunden

Bereich	unadjustierte Ergebnisse					
	G9$_{W1a}$	G9$_{W2b}$	G8$_{W2c}$	G8$_{W3d}$	G9$_{ges}$	G8$_{ges}$
Freunde treffen	15,12$_{bd}$	13,27$_{a}$	12,62	12,09$_{a}$	14.20	12.36***
Nebenjob	7,92$_{d}$	7,70$_{c}$	6,54$_{b}$	6,29$_{a}$	7.81	6.42***
Sport	6,68$_{d}$	6,45	6,26	6,13$_{a}$	6.57	6.20**
Fernsehen	8,84$_{d}$	8,69	8,37	8,04$_{a}$	8.77	8.20**
Angebote in Schule	3,16	3,08	3,01	3,05	3.12	3.03
Computer	11,07	10,63	10,80	10,81	10.85	10.80
Lesen	4,64	4,57	4,60	4,37	4.60	4.48
Unternehmungen mit Familie	5,53	5,35	5,59	5,59	5.44	5.59
Orchester	3,04$_{b}$	3,49$_{a}$	3,39	3,21	3.27	3.30
weitere Hobbies	4,46	4,21	3,97$_{d}$	4,38$_{c}$	4.34	4.18
	adjustierte Ergebnisse					
Freunde treffen	14,92$_{bd}$	13,28$_{a}$	12,93$_{d}$	12,08$_{ac}$	14,10	12,50***
Nebenjob	7,80$_{d}$	7,54$_{c}$	6,68$_{b}$	6,15$_{a}$	7,67	6,42***
Sport	6,62$_{d}$	6,43	6,37	6,08$_{a}$	6,53	6,22*
Fernsehen	8,83$_{d}$	8,68	8,69	8,83$_{a}$	8,75	8,38*
Angebote in Schule	3,10	3,09	3,07	3,02	3,10	3,04
Computer	11,01	10,58	10,89	10,82	10,79	10,86
Lesen	4,56	4,47	4,63	4,28	4,52	4,45
Unternehmungen mit Familie	5,51	5,34	5,78	5,57	5,42	5,67
Orchester	3,05$_{b}$	3,48$_{a}$	3,29	3,20	3,26	3,25
weitere Hobbies	4,53	4,19	4,00	4,38	4,36	4,19

Anmerkungen: G9W1: Schülerinnen und Schüler aus G9-Jahrgängen der ersten Erhebungswelle; G9W2: Schülerinnen und Schüler aus G9-Jahrgängen der zweiten Erhebungswelle; G8W2: Schülerinnen und Schüler aus G8-Jahrgängen der zweiten Erhebungswelle; G8W3: Schülerinnen und Schüler aus G8-Jahrgängen der dritten Erhebungswelle; G9ges: Schülerinnen und Schüler aller G9-Kohorten; G8ges: Schülerinnen und Schüler aller G8-Kohorten. Unadjustierte Modelle wurden hier nicht mit FIML, sondern listwise deletion spezifiziert, da fehlende Werte und „keine Betätigung" im entsprechenden Freizeitbereich nicht getrennt erfasst wurden und somit nicht voneinander trennbar sind. In den adjustierten Modellen wurden lediglich fehlende Werte auf unabhängigen Variablen mittels FIML berücksichtigt. Bezüglich der „Rest"-Kategorie „Sonstiges" ergaben sich für keinen der Gruppenvergleiche statistisch signifikante Unterschiede. Indizes stellen statistisch signifikante Gruppenunterschiede p < ,05 dar. Dargestellt wurden lediglich Unterschiede zwischen den beiden reinen Jahrgängen, den Doppeljahrgängen und Unterschiede innerhalb der G9-, bzw, G8-Jahrgänge. Unterschiede zwischen den gesamten G9- und G8-Jahrgängen wurden mit * gekennzeichnet.

*** p < ,001; ** p < ,01; * $p \leq$,05.

5 General Discussion

Educational reforms are a central part of educational governance and can influence educational practice; however, rigorous evaluations of such reforms are scarce (OECD, 2015). As outlined in this dissertation, there are several reasons to invest in and conduct scientific policy evaluations in the field of education: From a perspective of sustainability, evaluations and deepening investigations of reforms and specific programs can be important for justifying and potentially for revising or reforming policy decisions. Furthermore, from a perspective of accountability, they provide important tools for reducing trial-and-error policy implementations and increasing knowledge about how, when, and for whom interventions work or do not work (Black & Wiliam, 2009; McConnell, 2010; Schaffer et al., 1997). In four studies, this dissertation therefore investigated effects of two major German educational policy reforms at the end of upper secondary school, using advanced methods and large representative data sets. The two reforms in the focus of the studies were the upper secondary school reforms in Thuringia and Baden-Württemberg and the G8 reform in Baden-Württemberg. The following discussion will focus on four different issues: (a) a summary of results, (b) strengths and limitations of the present dissertation, (c) implications for future research on educational policy reforms, and (d) implications for policy and practice.

5.1 Summary of Results

The summary of results will focus on (a) a general localization of results in the framework of knowledge on educational policy reforms, (b) the relevance of allocated time, course composition, and curricular demands for student outcomes, and (c) the comparability of reform effects across the two different states.

General Knowledge on Educational Policy Reforms

The results of this dissertation can be integrated into the larger framework of knowledge on effects of educational policy reforms. As can be seen from all of the studies, the reforms analyzed in further detail in this dissertation all had effects on at least some student outcomes. The CI reforms showed no effects on the average performance in Thuringia but a slight increase in Baden-Württemberg (Trautwein et al., 2010). They also showed the higher performance of young women in math after the reform in Baden-Württemberg but not in Thuringia. However, we found differential gender effects for math self-concepts in the two states. Furthermore, we found effects on average course-specific achievement (see Studies 2 and 3). These results therefore provide new and important findings regarding effects of very similar reforms implemented in two different states, which can in practice lead to somewhat adverse effects. Furthermore, a closer look at the results of the G8 reform showed that reforms that introduced curricular compression (a reduction in the total number of years spent in school but an increase in the average number of hours per week) in lower secondary school had negative effects on student achievement in English reading and biology but not in mathematics and physics as well as negative effects on outcomes such as stress or subjective health at the end of upper secondary school. However, which aspects of the reform these effects could be attributed to remained unclear (e.g., curriculum-specific changes or the compression of overall school time).

From a larger perspective of educational effectiveness, this indicates that CI reforms and reforms that compress school time can indeed result in intended but also in unintended effects on student outcomes. Moreover, as indicated in the studies in this dissertation, CI reforms that introduce mandatory enrollment in core subjects and that increase the allocated time and the curricular level might provide a reasonable way to increase achievement in some states and for some subgroups but not for others (see Studies 1 and 2). Results of the dissertation therefore support research that underscores the importance of considering the educational environment in which they are implemented (e.g., Stein et al., 2004).

However, also visible from all studies, these reform effects are, in most cases, not very large in a traditional metric (e.g., Cohen, 1988). However, when interpreting such effects, it is important to keep in mind that these effects are usually not reported for a small sample of students in terms of a randomized experiment or randomized controlled trial study but reflect average effects on the level of states. Therefore, it can be argued that even small effects of policy reforms can have a huge practical relevance (e.g., Konstantopoulos & Hedges, 2008). Based on this, and as outlined by Coe (2002), effect sizes can also be translated into percentages, which might reflect a different perspective on effects of large-scale policy reforms. From this perspective, a decrease in achievement of approximately 0.2 standard deviations would be equal to an increase in which 8% of the students in one group would perform worse than the average of the other group. These percentages might in turn provide a different metric by which to judge the sizes of effects. However, judgments of these effects as large or small will still remain normative if not integrated in and compared with findings from previous comparable studies. In this regard, Konstantopoulos and Hedges (2008) suggested a very reasonable framework from which to judge the size of reform effects by judging the reform effects in terms of the national or state distribution of school effects (e.g., adjusted and unadjusted between school effects of achievement). However, this framework strongly depends on the availability of appropriate data on a national or state level.

Finally, and aspect that was also shown in the present dissertation and that might be somewhat alarming is related to the theoretical foundation of education reforms and the anticipation of reform effects. As outlined in Chapter 3, still only scarce knowledge exists on how components of educational reforms that affect surface structures of the education system specifically interact with characteristics of the school, the teacher, the class, and the individual student (e.g., Elmore, 1995; OECD, 2015).

Although a large number of theoretical models on educational effectiveness exist (e.g., Creemers, 1994; Scheerens, 1990), of which indeed most provide reasonable assumptions, and some have even been tested and have shown empirical proficiency, most reforms are not based on such models. This, of course, can also be related to the complex nature of developing reforms in a policy context, where reforms rather display results of long processes of negotiations between different stakeholders (e.g., Jann & Wegrich, 2007). Furthermore, as shown in Chapters 2 and 3, developments and societal dispositions are of major relevance for interpreting reform-related action by politicians, and from this perspective, it is oftentimes not easy to integrate scientific evidence into these decision processes. As was shown by Dedering (2016), politicians therefore oftentimes use scientific evidence just to legitimize previously made decisions and objectives post hoc. As outlined by Sabatier (2007):

> In short, understanding the policy process requires knowledge of the goals and perceptions of hundreds of actors throughout the country involving possibly very technical scientific and legal issues over periods of a decade or more while most of those actors are actively seeking to propagate their specific 'spin' on events (p. 4).

As shown by example, this complex process oftentimes results in very specific reform packages that consist of many different, possibly very specific, components. General empirical quantitative knowledge about "how reform works" is therefore hard to obtain from this perspective because effects of single aspects of reform packages are methodologically difficult to isolate and therefore difficult to quantify. Neglecting knowledge about educational effectiveness models and the relevance of rigorous educational evaluations can lead to both (a) reforms that are difficult to justify from the perspective of educational research and (b) challenges when anticipating specific reform effects. When reforms are instead built on strong theoretical foundations, uncertainty regarding potential mechanisms and related effects can decrease to some extent (e.g., Porter et al., 2015; Rogers, 2003).

Effects of Allocated Time, Course Composition, and the Curricular Level

As I outlined in Chapter 1, student achievement is assumed to be directly linked to economic growth, and one line of argumentation for educational reform is related to the assumption that reforms can increase student achievement (e.g., Hanushek & Woessmann, 2012). However, as is visible from the theoretical models introduced in Chapter 3.4, policy reforms oftentimes introduce changes only on surface structures of the school system (e.g., Elmore, 1995), and important structures that have consistently been part of scientific debates in this regard are allocated time and standards regarding the curricular level of specific courses (e.g., Ceci, 1991; Domina & Saldana, 2012; Scheerens, 2014b). As outlined by Scheerens and Hendriks (2014), a distinction should be made between the time that is formally given, for instance, in a schedule (allocated time); exposure to instruction in terms of net teaching time (instructional time); and the time spent actively engaged in learning tasks (time-on-task). The CI reforms further investigated in this dissertation introduced core courses that were mandatory for enrolling in upper secondary school. Therefore, the time that students were required to allocate to math courses in upper secondary school in Baden-Württemberg increased by 1 hr for basic courses (from 3 hr before the reform to 4 hr afterwards). By contrast, achievement in Thuringia was limited to a minimum of 4 hr even before the reform, and therefore, for basic courses, the curricular level of the course officially changed because core courses had to be taught on an advanced course level after the reform. Along these same lines, advanced courses that were previously offered for 5 or 6 hr per week were eliminated. As outlined, the group of students who would potentially take basic courses was larger in comparison with the group of advanced-course students. Therefore, changes to the basic courses had a stronger impact on the overall averages of student achievement.

It is interesting that the results of Studies 1 and 2 suggest a somewhat different pattern. Whereas effects of the reform in Baden-Württemberg indicate the better achievement of young women after the reform, this was not the case in Thuringia, where there was no statistically significant Reform × Gender interaction on achievement. This might suggest that the allocated time, which can closely resemble increased instructional time and time-on-task, had a larger effect (e.g., Carroll, 1989) compared with the curricular level. On the other hand, this might also point to potential differential implementation patterns of the reform, related to the curricular level of the core course. Teachers might be more strongly oriented toward the advanced course level from before the reform in one state, whereas they might not be in the other. However, as outlined above, these two reform components were perfectly confounded in our studies and could not be methodologically disentangled at this stage. Furthermore, as different standardized tests were used to analyze student achievement, comparisons between states should be implemented with caution.

Another aspect went along with the introduction of mandatory core courses. These were related to the composition of students within classes. Before the reform, students were tracked into basic or advanced courses. But afterwards, students were detracked into one core course. This resulted in a change in the performance-related student composition within a class (see Study 3). Student composition has been found to have an important impact on teacher-assigned grading as grades depend on the sorting of students along the performance continuum if they are built on a norm-reference basis (e.g., Trautwein et al., 2006). Therefore, not only did students who were likely to take basic courses have a more demanding curriculum and more weekly hours (in Baden-Württemberg), but they were also faced with an, on average, stronger reference group, compared with before the reform (in both states). As shown in Study 3, the CI reform changed the relation of standardized achievement and student grades. According to Marsh et al. (2007), student grades are of central importance for students' academic self-concepts. Therefore, Study 3 offers a deeper explanation for findings from Studies 1 and 2 in mathematics, namely, that there were changes in teacher-assigned grades due to the reforms that introduced changes in the reference group, which in turn could lead to changes in students' self-concepts.

Study 4 focused on the G8 reform, which introduced a slightly different change in time, namely, an increase in the average number of weekly hours followed by a reduction in overall school time for 1 year in the highest school track. It is also interesting that, following debates introduced by the A Nation at Risk report (The National Commission on Excellence, 1983), Henry Levin (1986) had already suggested caution in the 1980s when the length of school days was increased. By contrast, lengthening the school year was suggested to be less problematic at that time (as cited in Carroll, 1989). Results from our study suggest a similar picture. We found that before the G8 reform, students showed a similar performance compared with G8-students in mathematics and physics; however, achievement in English reading and biology was lower afterwards (see Study 4). Furthermore, G8-students reported higher stress levels and more subjective health problems. Results on stress are in line with results from other research (e.g., Böhm-Kasper & Weishaupt, 2002; Meyer & Thomsen, 2015; Quis, 2015). However, effects on student achievement have been inconsistent (e.g., Huebener et al., 2017; Ivanov, Nikolova, & Vieluf, 2016).

Comparability of Reform Effects Among States

This chapter might come as a bit of a surprise because none of the studies in this dissertation explicitly tested for differences between reform effects in different states. However, as outlined, Studies 1 to 3 were all concerned with a similar reform in two different states, and therefore, at least some findings can be used in this regard.

First, what becomes evident from a comparison of the results of Studies 1 and 2 is that the two different states showed a broad variety of findings. Whereas in Baden-Württemberg, no effects were found on average achievement (Trautwein et al., 2010), and differential effects were found to work to the advantage of young women in math, we did not detect these effects in Thuringia. Here, the average achievement between students before and after the reform did not differ on any of the standardized tests (mathematics, English reading, biology, and physics), and the difference between young women's math achievement before and after the reform did not change. However, we found quite a comparable pattern regarding young women's math self-concept, which was statistically significantly lower after the reform in both states. Beside these findings, we also found an increase in the self-concept of young men in English in Thuringia (see Study 2).[43] As outlined above, as we were not able to clearly disentangle different dimensions of the CI reform from each other (e.g., increase in instructional time, increase in curricular standards, detracking), we were not able to determine which changes might have truly caused the effects we found. What seems to be suggested by the pattern of

[35] Note that subject-specific self-concept in English was tested in the TOSCA 2006 study only but not in the TOSCA 2002 study.

results found in our studies is that for math, the allocated time “does the trick” (as done in Baden-Württemberg) but not the increase in the curricular level only (as done in Thuringia). However, these suggestions would need to be tested in further studies.

Second, when comparing changes in grades in math between Thuringia and Baden-Württemberg, quite a comparable pattern was found, indicating that given the same grade in math, achievement in core courses was lower when compared with advanced courses and was higher when compared with basic courses. The achievement in core courses more closely resembled achievement in advanced courses when grades were high (As), whereas it more closely resembles basic course achievement when grades were low (Ds). In contrast to this pattern, which we found to be comparable for math in Thuringia and Baden-Württemberg as well as for English in Baden-Württemberg, the CI reform was not found to have a comparable effect on English in Thuringia. Comparing the results between states therefore indicated that, although the reform characteristics were very comparable between the states, the effects were found to be comparable in some regards but different in others. As outlined, there are at least some aspects related to differences between the two systems, which might provide the first post hoc explanations for these differences. In Thuringia before the reform, students were enrolled for 6 hr per week in advanced courses, for 4 hr per week in the basic math course, and for 3 hr in the second basic course (e.g., language). Therefore, mathematics was taught for at least 4 hr before and after the reform. Compared with this, in Baden-Württemberg, advanced courses were taught for 5 hr per week and basic courses for 3 hr per week. On the basis of this, it is obvious that although the CI reform was very comparable between states, the starting points of the two state systems differed in some relevant regards, and this might have caused differences in the results between states.[44]

In sum, these findings also highlights the limitations that occur when reforms from other states are used as a blueprint, as sometimes suggested by major stakeholders in the field. If very comparable reforms are implemented in two states within a country and do not lead to comparable effects, using other countries education system as a blueprint seems to be even more challenging to achieve sustainable improvements.

5.2 Strengths and Limitations of the Present Dissertation

The present dissertation has several strengths and limitations that should be kept in mind when interpreting the results of the different studies.

[36] For research on effects of regional disparities, see, for instance, Kemper and Weishaupt (2011).

Generally, all studies benefitted from the use of large representative data sets for the two states Thuringia and Baden-Württemberg. This is especially related to the external validity of the findings. As outlined by Briggs (2008), external validity might be especially important for studies in the context of education.

Furthermore, we applied state-of-the-art methods in order to analyze the data, such as multidimensional multiple-group IRT models or structural equation models with continuous indicators. These models not only allowed us to test for measurement invariance across different groups of students, but they also offered a suitable option to treat unreliability in the constructs across the different studies. In all studies, we controlled for the cluster structure of the data by using robust standard errors and considered missing values by using multiple imputation (MI) methods or full information maximum likelihood (FIML) estimation.

In addition, as we made use of cohort control design secondary data (Shadish et al., 2002), most of our data sets were based on a natural experiment (e.g., Murnane & Willett, 2011), whereby the reform resembled exogenous variation introduced as discontinuity in the system (Schlotter, Schwerdt, & Woessmann, 2011), thus strengthening the internal validity of our studies. In this regard, we also conducted a variety of selection analysis to compare students on observed covariates, and we also considered additional data from the federal bureau of statistics to check for potential differences on the population level whenever possible, and we indeed found these differences to be small to nonexistent.

Obviously, however, we were not able to formally check the selection for nonobserved covariates, and this may have posed a potential threat to internal validity. This was especially the case for Study 1, as after the reform, students were not assessed immediately after it was implemented but rather 2 years later. Furthermore, related to this point, we had only one cohort before and one cohort after the reform in in Studies 1 through 3, so the potential power of the cohort control design, which would result in very similar successive cohorts, could not be formally tested to consider additional subsequent and previous cohorts. Because of this, we were also not able to distinguish between short- and long-term impacts of policy reforms. This might be problematic because several researchers have suggested that reform effects need some time to show up and develop (e.g., Kyriakides, Charalambous, Philippou, & Campbell, 2006; Rogers, 2003). Related to this point, it is not possible to distinguish between primary side effects (e.g., stress increases with increased instructional time), resulting from structural efforts to implement change and the rather consistent long-term effects of a reform. Therefore, it would be necessary to conduct reform studies in a multicohort sequence design in future research, where multiple groups before and especially after the reform are followed longitudinally.

Besides these aspects, as already mentioned, we were not able to test individual, isolated aspects of the reforms in the focus of this dissertation but instead dealt with “reform packages,” where multiple components were perfectly confounded as these reform packages were implemented simultaneously. Regarding the CI reform, it was therefore not possible to clearly disentangle effects of changes in curricular standards from effects of allocated time or effects of changing course composition. Regarding the G8 reform, we were not able to explicitly isolate the effect of time compression from curricular changes or age.[45] The inability to identify effects of specific components of educational policy reforms is especially unsatisfying from the perspective of generalization of results and restrictions related to suggestions for policy makers. Furthermore, the identification of specific dimensions of the reform could provide a promising way to relate them to aspects of effective teaching (e.g., classroom management, cognitive activation, and supportive climate), which are assumed to be of major importance for learning outcomes (e.g., Good et al., 2009).

Furthermore, all of the analyses focused on individual student outcomes and did not explicitly integrate the perspectives of teachers or school leaders. Future research should additionally focus on this aspect, as recent research has shown that reforms greatly depend on school leaders and teachers working effectively. However, adequate large-scale analyses of teacher and leadership effects during periods of change are still scarce (Bogotch et al., 2010; Fullan, 2016; Porter et al., 2015).

There are also limitations that stem from the use of secondary data. As the data were not collected by the author of the dissertation, the accuracy of the questionnaire data in this data set is not known. In this regard, it has to be noted that many results presented in this dissertation were based on self-reports.[46] Self-reports might not provide the best possible measure for assessing stress or problems related to students’ health (see Study 4). There is already some research that has investigated a similar question using physiological measures such as students’ cortisol levels (Minkley, Rest, Terstegen, Kirchner, & Wolf, 2015). However, such investigations are difficult to implement on a large scale because they are very extensive and therefore usually lack statistical power.

[37] However, in an analysis using an instrumental variable approach suggested by Puhani and Weber (2008), age did not have a meaningful effect on the results found. For the sake of clarity, however, these additional analyses were not published in Study 4.

[38] There are also several variables that were based on school data, such as grades, gender, and information on the track or cohort. Furthermore, additional data from the Federal Statistical Office of Germany and regional offices were used.

Finally, from a more general perspective, the implementation fidelity of the reforms was not assessed. This can be argued to be due to the characteristics of the reforms, which were legally binding and were implemented by means of a top-down policy strategy. However, considerable variety might still exist in how schools implemented the reforms and how teachers dealt with the reform-related alterations. This might have been the case, for instance, during the CI reforms in the advanced courses after the reform, where teachers were told to teach on an advanced course level, but they were simultaneously confronted with a larger number of potentially low-performing students. Furthermore, an increasing amount of research suggests that reforms need time to fully develop (e.g., Kyriakides et al., 2006; Rogers, 2003) and that teachers cycle through different stages of concern during the implementation of reforms (e.g., George, Hall, & Stiegelbauer, 2008). The studies presented in this dissertation did not consider the status of the implementation due to limitations in the data set. Therefore, they relied on the assumption that implementation and implementation status were fairly similar across different teachers due to the legally binding characteristics of the reform.

5.3 Implications for Future Research on Educational Policy Reforms

Among others, three implications in particular arise from research on educational policy reforms conducted in this dissertation. These are related to (a) evaluations of educational policy reforms, (b) a stronger integration of findings into theoretical models of educational reform and educational effectiveness, and (c) the provision of information that is adequate to inform policy decisions on the basis of rigorous educational research.

Evaluation of Educational Policy Reforms

As outlined above, although a broad variety of different educational policy reforms occur, still, in general, most reforms are currently not part of rigorous evaluations (e.g., OECD, 2015). This might seem to be an individual disadvantage of the states that do not invest in such evaluations because they will, in the best case scenario, be able to recognize problems and start finding solutions after they have already implemented the reform. In line with this, in the best case scenario, a few years of schooling will pass before problems are adequately identified and resolved, if they are identified at all. During this period of identifying and resolving side effects of previous education reforms, school goes on, and students, teachers, and/or principals might continuously suffer from problems, and this will in some respects decrease the effectiveness of the whole school system. These threats, which are also strongly related to ideas of accountability and sustainability of policy decisions for the public, should be made clear whenever scientists address policy or the public.

Furthermore, what can be generally useful for decreasing the likelihood that such scenarios will occur are rigorous educational evaluations and scientific policy consulting right from the beginning of the policy process. Such evaluations should generally consider both the implementation process and characteristics of the reform itself in order to obtain a holistic picture of the reform. As outlined by Gräsel (2010), there are different aspects that have an impact on the implementation process. These are related to the characteristics of the reform itself (e.g., Does the reform match the values of the institutional environment?), teacher characteristics (e.g., Does the reform "matter" from the perspective of teachers?), characteristics of the school (e.g., Do teachers cooperate?), and environmental support (e.g., Is there enough support for the implementation of the reform?). Along these same lines, Fullan (2016) identified the teacher, the principal, the student, parents, and the community as well as the district administrator as central for successful educational change. Therefore, also considering the limitations outlined above, when evaluating reforms, relevant stakeholders should be included, at least to some extent.

Scientists with a focus on research related to education and educational policy have a long tradition of investigating effects of reforms and the process of reform implementation in the education system (e.g., Fullan, 1983, 2016; Gross et al., 2009; Porter et al., 2015), all of which are related to what was defined as descriptive and change knowledge by Bromme et al. (2014). However, knowledge that can explain which components of educational policy reforms work given specific governmental structures is more available in terms of heterogeneous case studies and complex theoretical approaches than in a broader more general empirical framework. Reform evaluations should therefore generally go beyond the strong focus on the individual student level and closely investigate processes at the level of the class, the school, and the administration.

Moreover, in line with Stein et al. (2004), considering further research for both the upper secondary school reform and the G8 reform (Studies 1 to 4) underlines what has already been visible when comparing the results from upper secondary schools in Baden-Württemberg and Thuringia (Studies 1 and 2): Reforms often result in quite different effects between states even when they seem quite similar on the surface (e.g., Huebener et al., 2017; Wagner et al., 2014). If the anticipation of intended and unintended reform effects is difficult, and if the results of studies show quite different patterns when similar reforms are implemented in different states, this clearly points to the importance of rigorous evaluations of reforms to shine light into the black box to address specific mechanisms of educational policy reforms as well as the specificities of the different states where the reforms are introduced. However, this also highlights the limitations that occur when reforms from other states are used as a blueprint.

Finally, evaluations should follow more rigorous evaluation standards. First, analyses of reform effects should generally be based on strong multicohort sequence designs, where it is possible to continuously monitor multiple subsequent cohorts before and after the reform. As it usually takes some time for reform effects to become apparent (e.g., Rogers, 2003), this is especially important for a continuous monitoring and identification of short- and long-term impacts and in order to adequately consider differences between cohorts between states on all relevant layers (e.g., the administration, the school, the teacher, and the student).

Next, what is evident from the information outlined above is that no universal standards exists on how to adequately investigate effects of educational policy reforms in educational research. Therefore, rather general knowledge about research methods is usually applied when analyzing effects of reforms, with a strong dependency on the individual researcher's field of research (e.g., Goldthorpe, 2001). This often goes along with a focus on a research-field-specific subset of variables (e.g., variables related to GDP vs. variables related to student learning), whereas other important variables are neglected, respectively. Furthermore, this leads to a research-field-specific perspective of the analysis of a reform (e.g., national vs. state-specific vs. district-specific).[47] In sum, and from a research perspective, this might present a major opportunity because knowledge about different aspects of reforms is increasing. However, this also provides an open gateway for politicians to choose among a variety of potentially contradictory results. Therefore, what is needed are more general discussions about variables that are important to investigate and especially discussions about how to investigate these variables, possibly including researchers from different scientific disciplines. Similar to the field of medical science, reform research is very closely related to what is actually happening in school and education policy, and this accountability should be considered in terms of continuous discussions and a general search for the best evaluation standards.

[39] Potential differences can be identified, for instance, from taking a closer look at research on the G8-reform, conducted by researchers in the field of economy as opposed to education/psychology (e.g., Huebener, Kuger, & Marcus, 2017; Hübner, Wagner, Kramer, Nagengast, & Trautwein, 2017).

Integrating Findings into Theoretical Models of Educational Reform and Educational Effectiveness

Although models and theories of and related to educational effectiveness have shown proficiency in explaining student achievement in recent decades (e.g., Creemers, 1994; Eccles & Wigfield, 2002; Helmke, 2006; Helmke & Weinert, 1997), less attention has been paid to linking reform effects more closely with such models. In this dissertation, some examples were outlined in Chapter 3.4 in this regard. However, this can be taken as only a very first step toward a stronger empirical theory of educational reform and educational change that explicitly considers individual-, class-, and school-level variables more closely.

What seems to be especially important, although also especially challenging, is the need to identify specific objectives, dimensions, and structures of educational policy reforms and to test their effects on core components of models of or related to educational effectiveness. On this basis, previous research could be used to provide an evidence base of average effects of individual reform characteristics on specific aspects of educational effectiveness (e.g., McLaughlin, 1987; Young & Lewis, 2015).

This could also go along with a promotion of stepwise implementation procedures of reform-specific components rather than the implementation of whole reform packages that do not allow researchers to isolate and test individual reform components for its effects. In this regard, reforms have been shown to differ along a broad variety of different dimensions (e.g., Cuban, 1990; Fullan, 1983), and these dimensions could be disentangled by means of a stepwise implementation and therefore investigated one at a time in educational research to gain knowledge about the effectiveness of individual reform-specific components. Furthermore, introducing reforms in a stepwise manner could also increase acceptance among teachers because only small changes would be introduced continuously and could therefore be integrated more accurately into teachers' daily routines.

Besides integrating previous research into such theoretical models, future research should be based more strongly on such models. This could also lead to a change from rather exploratory evaluations to confirmatory procedures for testing specific hypotheses that are more strongly based on and integrated into previous research. A stronger dedication to theoretical models would also help to identify potentially relevant research from related disciplines and increase the effectiveness of formative evaluations in terms of identifying and assessing the most relevant variables that are likely to be affected by a specific reform. As can be seen from research related to educational effectiveness (e.g., Reynolds et al., 2014; Scheerens, 1990), there is already a large amount of research on variables affecting student achievement and other student outcomes. However, this research stands somewhat apart from the current research on effects of educational policy reforms or is integrated in a rather isolated way into the context of individual studies. In this regard, profound reviews could increase seemingly little knowledge and can strongly contribute to the anticipation and interpretation of specific reform effects. As outlined in Chapter 3.3, educational policy reforms oftentimes follow the assumption that affecting surface structures of the education system (e.g., tracking, allocated time) will have a positive effect on individual student characteristics or teaching. Future research should place a greater focus on identifying and testing these links using prior theoretical work.

Providing Adequate Information to Inform Policy Decisions

As is evident, evaluations of policy reforms have a very strong link and great potential to inform policy (e.g., Briggs, 2008; Campbell, 1969; Slavin, 2002). Not only can rigorous evaluations of educational policy reforms provide information about whether or not a policy reform worked out well, in terms of a summative evaluation, but results of evaluations can also be integrated into actual policy decisions in a formative manner. They can be used to learn from potential side effects when reforms are conducted, or they can be used to legitimize policy decisions. However, informing policy also requires educational policy research to comply with specific requirements: As outlined by Hazle Bussey, Welch, and Mohammed (2014), consultants need to have expertise regarding not only the specific contents of the topic but also regarding the process. Furthermore, they need to have specific interpersonal abilities to establish a trustworthy relationship with the client and need to be open to incorporating adjustments to previous plans over the course of consulting in order to maximize the fit to the client's objectives. As I will outline in the following, from a perspective of accountability, for research, this could also mean working together in groups of expert researchers rather than conducting policy consulting individually.

From a more general perspective, results of scientific evaluations are often conducted so that they can be presented to a scientific community rather than to the public audience. Also, they often focus more on theoretical scientific success rather than having the goal to contribute to improving the educational system in practice. Therefore, in the case of promoting the rigorous monitoring of education reforms, it is important to prepare information from evaluations for both the scientific and the political community, and this is especially important when promoting evidence-based or evidence-informed policy (e.g., Hedges & Waddington, 1993; Slavin, 2008). There are two strands in particular that seem promising for contributing to this demand in the future.

First, and related to the discussion of evidence from EER, outlined in Chapter 3.2, there are two major challenges linked to the current problematic fit between research, policy, and practice. As has been outlined, the definition and comprehension of what is meant to be “a causal effect” varies between and within different scientific disciplines (Goldthorpe, 2001). For some researchers, correlations still seem to provide a reasonable method for providing implications for policy and practice, whereas others make use of only more advanced methods that are explicitly concerned with the identification of causal effects (Murnane & Willett, 2011; Reinhart et al., 2013). Although these researchers differ in their methodological preferences, they are oftentimes concerned with similar research questions and based on different designs, methods, and results they might provide different answers to similar questions, shown for instance in medical research (e.g., Haddad, 2016). However, even when using comparable gold standard research designs (e.g., randomized experiments), the recently proclaimed replication crises in the field of psychology points to great variability in findings (Open Science Collaboration, 2015).

This, in turn, can lead to a communication problem, resulting from the broad variety of possibly contradictory research available, of which politicians are free to choose, for instance, to legitimize their decisions (e.g., Dedering, 2016). Moreover, this might lead to a reduction in the informative value of scientific evidence for policy in general. As politicians have to integrate a broad variety of information from different stakeholders during the process of public policy making (e.g., Jann & Wegrich, 2007; Sabatier, 2007), providing a variety of different answers from science will decrease the likelihood that politicians will be able to identify the best available research knowledge for evidence-informed policy, even if they are very interested in doing so. Therefore, if the scientific community is interested in informing policy decisions, it has to target such threats. It might be promising, for instance, for scientists to form special interest groups (SIGs) that are connected to specific topics. Within these groups, the scientific community could then review and judge existing literature according to predefined standards (e.g., Konstantopoulos & Hedges, 2008). Even if different SIGs develop to address similar specific topics (e.g., with an economic focus or a psychological focus), this would lead to an important reduction in single-study researchers informing policy.[48]

[40] Of course, the idea of special interest groups is not new, as shown by 27 SIGs from The European Association for Research on Learning and Instruction (EARLI) or by a total of 180 SIGs from the American Educational Research Association (AERA). However, at least in Germany, these SIGs are not very engaged in informing

Second, in a more general sense, research needs to increase awareness of how policy works (e.g., Black & Donald, 2001), and research has to focus more on questions about how the best evidence on educational policy reforms can be made more accessible for politicians and the educational administration. This would not only include information for stakeholders after a specific scientific article is published, but also the summarizing of results for a nonscientific community in the presence of the potential limitations of scientific research (e.g., Baumert, 2016). Increasing awareness and knowledge to move toward the use of research for policy decisions in the field of education could also be more strongly oriented toward research from the field of healthcare, which has been concerned with such questions for a long time (e.g., Hughes, 2008). The focus of research on policy might pose a huge challenge for scientists, but doing research on effects of educational reforms should be linked more closely to accountability in informing policy decisions. There are already some initiatives that generally foster this movement, such as the WWC (WWC, 2015), but these institutions focus on only a specific amount of predefined high-quality standards and therefore might ignore other potentially useful research (e.g., Briggs, 2008). Furthermore, until now, such movements have rarely been seen in Germany.

Finally, the relevance of the media as a central mediator between research and the public (e.g., Baumert, 2016) has to be acknowledged for its potential power to put their specific stamp on results of scientific evaluations. Therefore, educational research has to work more strongly on outlining specific strategies for how to work with journalists in the field of education. This is again related to the aspect of accessibility and comprehensibility of research results in order to avoid erroneous press releases and to increase awareness about the limitations of research.

5.4 Implications for Policy and Practice

There are several implications of this dissertation that are relevant for policy and practice, some of which have already been outlined in Studies 1 to 4. However, from a more general perspective, the results of all studies in this dissertation can be used as descriptive information about effects of specific reforms on student outcomes (e.g., Bromme et al., 2014). In this regard, the studies relating to reforms of CI suggest that comparable reforms do not automatically lead to similar results, which emphasizes the idea that supporting and funding rigorous evaluations of large reforms should be the standard rather than an exception.

educational policy decisions, and decisions about educational policy consulting are oftentimes based on individual preferences related to specific aspects of successful policy consulting (e.g., trustworthy relationships with a researcher) rather than on other relevant aspects such as the profound expertise of the researcher in the field of policy (e.g., Hazle Bussey, Welch, & Mohammed, 2014).

Furthermore, as shown, CI reforms do not generally result in increases in student achievement, and this might depend on different dimensions of CIs such as detracking, allocated time, or curricular standards. However, Studies 1 and 2 both suggested that detracking seems to be particularly related to self-concepts in math. This finding was also supported by Study 3, which suggested that CI reforms can change the meaning of teacher-assigned grades. Finally, including Study 4 in this dissertation allowed us to take a closer look at another reform in which a different component of time was changed, namely, the number of weekly hours in lower secondary school. The results of this reform suggest that after the reform, students showed lower performance in biology and English reading, increased stress, and more health-related problems. All of these results should not be used in isolation but should instead be integrated into the broader framework of evidence from the specific reforms and their effects.

Based on this, the overarching conclusion is that these reforms, which are constituted at the highest layer of the education system, can make a difference in practice, and the ideas of educational governance about how to impact educational practice can work well in specific cases (e.g., Schaffer et al., 1997; Swanson & Stevenson, 2002). However, a general uncertainty about potential mechanisms and effects of reforms have to be acknowledged, and this uncertainty can be meaningfully targeted by considering knowledge from educational research, when reforms are conceptualized and implemented.

Furthermore, the results of this dissertation strongly suggest that individual state-specific characteristics have to be accounted for when implementing policy reforms. As seen in Study 2, which further investigated effects of the upper secondary school reform in Thuringia, no main effects on achievement were found. However, the reform introduced great changes to the whole upper secondary school system, which, to some extent, might have temporarily impeded other core concepts of educational effectiveness that are related to learning and teaching because more time is spent on organizational issues. Furthermore, not only can reforms impede learning and teaching at school, but they usually also lead to huge investments to offer additional teacher training to implement the reform and adapt learning materials. This underscores the importance of considering all aspects of Wortman’s (1983) taxonomy, which includes efficacy, effectiveness, and efficiency, before enacting large programs. Differences between states should be acknowledged and considered whenever reforms are implemented. Unfortunately, instead of recognizing the advantages of the federal system and the enthusiasm that some states have for reforms from a secure distance before they are adopted, policy decisions are oftentimes immediately copied across many states, for instance, based on the worry that the other states will fall apart. In the presence of all actions to make education comparable across the states, this dissertation underscores the idea that similar reforms do not necessarily lead to comparable effects, and this aspect should be acknowledged more strongly by policy before new programs are implemented. Along the same lines, this also highlights the great number of challenges that have to be overcome in order to adequately make use of the best strategies for practice (e.g., OECD, 2014), whereby successful nations work as blueprints for others (see Chapter 3).

Results of this dissertation further contribute to an increase in the amount of literature suggesting that rigorous educational policy evaluations that are focused on psychological factors such as standardized student achievement or achievement motivation are vital for profound evidence-informed policy. This, of course, strongly points to the importance of considering rigorous evaluations of reforms also in the financial planning of the implementation of a reform. Whenever large-scale policy reforms are to be implemented, for instance, as the recently introduced large education reform package in Bavaria, which will cost 870 million euro (e.g., Günther & Wittl, 2017, April 6), rigorous evaluations of such huge investments, possibly including different educational researchers with expertise in different fields, should be a given—if not for the interest in knowing whether a reform will work the way it should work, then at least to justify such a huge investment to the tax payer. Moreover, if research results are already available, not only should educational policy decisions rely on specific norms, values, and traditions of political parties or individuals, but they can also be based and legitimized by employing educational research (e.g., Dedering, 2016). This aspect is also part of recent discussions in Germany (see Study 4) related to the introduction of the two reforms that this dissertation focused on. Oftentimes, different stakeholders promote specific features of educational systems using emotional statements or based on individual judgments. However, what should matter for major policy decisions is what works best for the majority of students. From this perspective, it seems especially irrational to make an argument for different reforms instead of a general argument for more profound knowledge prior to reforming at all (e.g., in terms of small-scale tests). Maybe even worse, if reforms were already enacted on the basis of vague knowledge and by means of normative arguments, this can foster reactions at a similar, normative level, which can be orthogonally related to what is right from an objective, scientific perspective.

As outlined above, however, of course evidence-informed policy needs research to point objectively to specific policy solutions, rather than to an unclear compound of different solutions to a policy problem. This might include the need for the scientific community to work together more effectively (e.g., in scientific-expert groups) to provide specific recommendations rather than by relying on trusted individual experts who support policy decisions under the “cloak of science.”

To sum up, this dissertation systematically showed that there are still many open research questions, and much more has to be investigated regarding educational policy reforms and their evaluation, integration, and standards. Strongly underlined by the findings of this dissertation, large-scale policy reforms can only be implemented in an accountable manner with serious consideration of recent educational developments, educational research, and knowledge on or related to educational effectiveness. Especially in the face of the recent developments of opportunities of educational research in recent decades, but also when considering the frequent use of reforms introduced to change educational practices, knowledge about what such reforms theoretically and empirically (intentionally and unintentionally) impact seems to be scarce. Therefore, in order to increase the likelihood of implementing “good reforms,” educational policy reforms must be linked to the results of scientific research and rigorous educational evaluations, and the importance of the psychological factors that are related to student achievement must be acknowledged. Such practices must become the standard rather than the exception.

6 References

Aebli, H. (1961). *Grundformen des Lehrens. Eine Allgemeine Didaktik auf kognitionspsychologischer Grundlage* [Foundations of teaching. General didactics based on cognitive psychology]. Stuttgart: Klett.

Altrichter, H., & Maag Merki, K. (Eds.). (2016). *Handbuch Neue Steuerung im Schulsystem* [Handbook of new governance in the education system]. Wiesbaden: VS Verlag für Sozialwissenschaften.

Altrichter, H., & Wiesinger, S. (2005). Implementation von Schulinnovationen – aktuelle Hoffnungen und Forschungswissen. *Journal für Schulentwicklung, 9*(4), 28–36.

Angrist, J. D., & Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics, 114*(2), 533–575. https://doi.org/10.1162/003355399556061

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review, 92*(5), 1535–1558. https://doi.org/10.1257/000282802762024629

Au, W. (2007). High-stakes testing and curricular control: A qualitative meta-synthesis. *Educational Researcher*, *36*(5), 258–267. https://doi.org/10.3102/0013189X07306523

Autorengruppe Bildungsberichterstattung. (2014). *Bildung in Deutschland 2014: Ein indikatorengestützter Bericht mit einer Analyse zur Bildung von Menschen mit Behinderungen* [Education in Germany 2014: An indicator-based report including an analysis of the situation of people with special educational needs and disabilities]. Bielefeld: Bertelsmann.

Autorengruppe Bildungsberichterstattung. (2016). *Bildung in Deutschland 2016: Ein indikatorengestützter Bericht mit einer Analyse zu Bildung und Migration* [Education in Germany 2016: An indicator-based report including an analysis on education and migration]. Bielefeld: Bertelsmann.

Baird, J.-A., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T. L., & Daugherty, R. (2011). Policy effects of PISA. Retrieved from http://oucea.education.ox.ac.uk/wordpress/wp-content/uploads/2011/10/Policy-Effects-of-PISA-OUCEA.pdf

Baker, E. L., & O´Neil, H. F. (2016). The United States: 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. 122–139). New York: Routledge.

Bandura, A. (1997). *Self-efficacy: The exercise of control* (1. print). New York, NY: Freeman.

Baumert, J., Stanat, P., & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie [PISA 2000: Object of investigation, theoretical foundations and implementation]. In Deutsches PISA-Konsortium (Ed.), *PISA 2000* (pp. 15–65). Opladen: Leske + Budrich.

Baumert, J. (2016). Leistungen, Leistungsfähigkeit und Leistungsgrenzen der empirischen Bildungsforschung [Large-scale assessment studies between science and politics]. *Zeitschrift für Erziehungswissenschaft*, *19*(1), 215–253. https://doi.org/10.1007/s11618-016-0704-4

Bellmann, J., & Weiß, M. (2009). Risiken und Nebenwirkungen Neuer Steuerung im Schulsystem. Theoretische Konzeptualisierung und Erklärungsmodelle [Risks and side effects of new control strategies in the educational system. Theoretical conceptualization and explanatory models]. *Zeitschrift für Pädagogik*, *55*(2), 286–308.

Bennett, C. J., & Howlett, M. (1992). The lessons of learning: Reconciling theories of policy learning and policy change. *Policy Sciences*, *25*(3), 275–294. https://doi.org/10.1007/BF00138786

Benz, A. (Ed.). (2010). *Governance - Regieren in komplexen Regelsystemen: Eine Einführung* [Governance - Governance in complex systems] (2. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Berkemeyer, N. (2010). *Die Steuerung des Schulystems: Theoretische und praktische Explorationen* [Governance in the school system. Theoretical and practical explorations] (1. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Berkemeyer, N., Manitius, V., Müthing, K., & Bos, W. (2009). Ergebnisse nationaler und internationaler Forschung zu schulischen Innovationsnetzwerken [Results of national and international research on school innovation networks]. *Zeitschrift für Erziehungswissenschaft*, *12*(4), 667–689. https://doi.org/10.1007/s11618-009-0102-2

Berliner, D. (1990). What’s all the fuss about instructional time? In M. Ben-Peretz & R. Bromme (Eds.), *The nature of time in schools: Theoretical concepts, practitioner perceptions.* New York: Teachers College Press.

Bieber, T., Martens, K., Niemann, D., & Windzio, M. (2014). Grenzenlose Bildungspolitik?: Empirische Evidenz für PISA als weltweites Leitbild für nationale Bildungsreformen [Boundary-less education policy? Empirical evidence for PISA as a world-wide model for national education reforms]. *Zeitschrift für Erziehungswissenschaft*, *17*(S4), 141–166. https://doi.org/10.1007/s11618-014-0513-6

Biehl, J., Hopmann, S., & Ohlhaver, F. (1996). Wie wirken Lehrpläne? Modelle, Strategien, Widersprüche [How do curriula work? Models, strategies, contradictions]. *Pädagogik*, *48*(5), 32–35.

Black, N., & Donald, A. (2001). Evidence based policy: Proceed with care Commentary: research must be taken seriously. *BMJ*, *323*(7307), 275–279. https://doi.org/10.1136/bmj.323.7307.275

Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning*. London: Sage.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

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

Bogotch, I., Townsend, T., & Acker-Hocevar, M. (2010). Leadership in the Implementation of Innovations. In E. Baker, B. McGaw, & P. Peterson (Eds.), *International encyclopedia of education* (pp. 128–134). Amsterdam: Elsevier Academic.

Böhme, K., & Hoffmann, L. (2016). Mittelwerte und Streuungen der im Fach Deutsch erreichten Kompetenzen [Means and variances of German competencies]. In P. Stanat, K. Böhme, S. Schipolowski, & N. Haag (Eds.), *IQB-Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich*. Münster: Waxmann.

Böhm-Kasper, O., & Weishaupt, H. (2002). Belastung und Beanspruchung von Lehrern und Schülern am Gymnasium [Burden and Strain on Teachers and Pupils in Gymnasia]. *Zeitschrift für Erziehungswissenschaft*, *5*(3), 472–499. https://doi.org/10.1007/s11618-002-0062-2

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive School Reform and Achievement: A Meta-Analysis. *Review of Educational Research*, *73*(2), 125–230. https://doi.org/10.3102/00346543073002125

Bowen, S., & Zwi, A. B. (2005). Pathways to "evidence-informed" policy and practice: A framework for action. *PLOS Medicine*, *2*(7), 600-605. https://doi.org/10.1371/journal.pmed.0020166

Briggs, D. C. (2008). Comments on Slavin: Synthesizing causal inferences. *Educational Researcher*, *37*(1), 15–22. https://doi.org/10.3102/0013189X08314286

Bromme, R., Prenzel, M., & Jäger, M. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik [Educational research and evidence-based educational policy]. *Zeitschrift für Erziehungswissenschaft*, *17*(4), 3–54. https://doi.org/10.1007/s11618-014-0514-5

Brühwiler, C., & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and Instruction*, *21*(1), 95–108. https://doi.org/10.1016/j.learninstruc.2009.11.004

Brunner, E. J., Imazeki, J., & Ross, S. L. (2010). Universal vouchers and racial and ethnic segregation. *The Review of Economics and Statistics*, *92*(4), 912–927. https://doi.org/10.1162/REST_a_00037

Brunsson, N. (2009). *Reform as routine: Organizational change in the modern world* (1. publ). Oxford: Oxford University Press.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, *24*(4), 409–429. https://doi.org/10.1037/h0027982

Campbell, R. J., Kyriakides, L., Muijs, R. D., & Robinson, W. (2003). Differential teacher effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education*, *29*(3), 347–362. https://doi.org/10.1080/03054980307440

Caplan, N., Morrison, A., & Stambaugh, R. J. (1975). *The use of social science knowledge in policy decisions at the national level: A report to respondents*. Ann Arbor: Institute for Social Research, University of Michigan.

Carroll, J. B. (1963). A model for 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

Cattaneo, M. A., Oggenfuss, C., & Wolter, S. C. (2017). The more, the better? The impact of instructional time on student performance. *Education Economics*, *142*(1), 1–13. https://doi.org/10.1080/09645292.2017.1315055

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

Cerna, L. (2013). *The nature of policy change and implementation: A review of different theoretical approaches*. Paris: OECD.

Chen, H.-T., & Rossi, P. H. (1987). The theory-driven approach to validity. *Evaluation and Program Planning, 10*(1), 95–103. https://doi.org/10.1016/0149-7189(87)90025-5

Chin, R., & Benne, K. D. (1969). General strategies for effecting changes in human systems. In W. G. Bennis, K. D. Benne, & R. Chin (Eds.), *The planning of change* (pp. 32–59). New York: Holt, Rinehart & Winston.

Coburn, C. E. (2005). Shaping Teacher Sensemaking: School Leaders and the Enactment of Reading Policy. *Educational Policy*, *19*(3), 476–509. https://doi.org/10.1177/0895904805276143

Coe, R. (2002). *It's the effect size, stupid. What effect size is and why it is important. Paper presented at the annual conference of the British Educational Research Association*. University of Exeter, England. Retrieved from http://www.cem.org/attachments/ebe/ESguide.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington: National Center for Educational Statistics.

Conley, D. T. (1994). *Roadmap to restructuring: Policies, practices and the emerging visions of schooling*. Eugene, Oregon: ERIC Clearinghouse on Educational Management, University of Oregon.

Cooper, A., Levin, B., & Campbell, C. (2009). The growing (but still limited) importance of evidence in education policy and practice. *Journal of Educational Change*, *10*(2-3), 159–171. https://doi.org/10.1007/s10833-009-9107-0

Creemers, B. P. (1994). *The effective classroom*. *School development series: 1994: 1*. London: Cassel.

Creemers, B. P., & Reezigt, G. J. (1996). School level conditions affecting the effectiveness of instruction. *School Effectiveness and School Improvement*, *7*(3), 197–228. https://doi.org/10.1080/0924345960070301

Cronbach, L. J. (1980). *Toward reform of program evaluation: aims, methods and institutional arrangements*. San Francisco: Jossey-Bass.

Cuban, L. (1990). A fundamental puzzle of school reform. In A. Lieberman (Ed.), *School development and the management of change series: Vol. 3. Schools as collaborative cultures. Creating the future now* (pp. 71–78). London: Falmer.

Darling-Hammond, L. (2000). Teacher quality and student achievement. *Education Policy Analysis Archives*, *8*(1), 1–44. https://doi.org/10.14507/epaa.v8n1.2000

Datnow, A. (2005). The sustainability of comprehensive school reform models in changing district and state contexts. *Educational Administration Quarterly*, *41*(1), 121–153. https://doi.org/10.1177/0013161X04269578

Davies, P. (2000). The relevance of systematic reviews to educational policy and practice. *Oxford Review of Education*, *26*(3-4), 365–378. https://doi.org/10.1080/713688543

Deaton, A., & Cartwright, N. (2016). Understanding and misunderstanding randomized controlled trials. *NBER Working Paper*. (22595). Retrieved from https://www.princeton.edu/~deaton/downloads/Deaton_Cartwright_RCTs_with_ABSTRACT_August_25.pdf

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*(6), 627-68; discussion 692-700. https://doi.org/10.1037/0033-2909.125.6.627

Decristan, J., & Klieme, E. (2016). Bildungsqualität und Wirkung von Angeboten in der Ganztagsschule. Einführung in den Thementeil [Quality and effectiveness of extra-curricular activities in German all-day schools]. *Zeitschrift für Pädagogik*, *62*(6), 757-759.

Dedering, K. (2016). Entscheidungsfindung in Bildungspolitik und Bildungsverwaltung [Decision making in education policy and educational administration]. In H. Altrichter & K. Maag Merki (Eds.), *Handbuch Neue Steuerung im Schulsystem* (pp. 53–73). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-18942-0_3

Deutscher Bildungsrat. (1970). *Empfehlungen der Bildungskommission. Strukturplan für das Bildungswesen* [Suggestions of the educational comission. Structural plan for the education system]. Stuttgart: Klett.

Dewey, J. (1916). *Democracy and education: An introduction to the philosophy of education*. New York: The Macmillan Company.

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/10.3102/0002831211426347

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3-4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

Easterly, W. (2001). The lost decades: Developing countries' stagnation in spite of policy reform 1980–1998. *Journal of Economic Growth*, *6*(2), 135–157. https://doi.org/10.1023/A:1011378507540

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: 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

Elmore, R. L. (1995). Structural reform and educational practice. *Educational Researcher, 24*(9), 23–26.

European Commission. (2016). Communication from the Commission to the European Parliament, The European Council and The Council: Better regulation: Delivering better results for a stronger union. Retrieved from https://ec.europa.eu/info/sites/info/files/better-regulation-delivering-better-results-stronger-union_sept_2016_en.pdf

Interinstitutional Agreement between the European Parliament, the Council of the European Union and the European Commission on Better Law-Making, European Parliament; Council of the European Union; European Commission 2016.

Federal Statistical Office of Germany. (2017). *Bildung und Kultur: Allgemeinbildende Schulen* [Education and Culture: General Education Schools]. Wiesbaden: Federal Statistical Office of Germany.

Fend, H. (2004). Was stimmt mit den deutschen Bildungssystemen nicht? Wege zur Erklärung von Leistungsunterschieden zwischen Bildungssystemen. [What is wrong with the German education system? Ways to explain performance differences in the education system]. In G. Schümer (Ed.), *Die Institution Schule und die Lebenswelt der Schüler. Vertiefende Analysen der PISA-2000-Daten zum Kontext von Schülerleistungen* (pp. 15–38). Wiesbaden: VS Verlag für Sozialwissenschaften.

Fend, H. (2009). *Neue Theorie der Schule: Einführung in das Verstehen von Bildungssystemen* [New school theory: Introduction to an understanding of educational systems]. Wiesbaden: VS Verlag für Sozialwissenschaften.

Fischer, N., Kuhn, H. P., & Tillack, C. (Eds.). (2016). *Was sind gute Schulen? Theorie, Forschung und Praxis zur Qualität von Ganztagsschulen*. *Theorie und Praxis der Schulpädagogik: Band 38*. Immenhausen: Prolog-Verlag.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature*. Tampa, FL: University of South Florida: Louis de la Parte Florida Mental Health Institute.

Friedman, M. (1955). The role of government in education. In R. A. Solo (Ed.), *Economics and the Public Interest* (pp. 123–144). New Jersey: Rutgers University Press.

Fullan, M. (1983). Evaluating Program Implementation: What Can Be Learned from Follow Through. *Curriculum Inquiry*, *13*(2), 215. https://doi.org/10.2307/1179640

Fullan, M. (2000). The return of large-scale reform. *Journal of Educational Change*, *1*(1), 5–27. https://doi.org/10.1023/A:1010068703786

Fullan, M. (2016). *The new meaning of educational change* (Fifth edition). New York, NY: Teachers College Press.

Fusarelli, L. D. (2002). Tightly coupled policy in loosely coupled systems: Institutional capacity and organizational change. *Journal of Educational Administration*, *40*(6), 561–575. https://doi.org/10.1108/09578230210446045

Fussangel, K., & Gräsel, C. (2012). Lehrerkooperation aus der Sicht der Bildungsforschung [Teacher cooperation from the perspective of educational research]. In E. Baum (Ed.), *Schule und Gesellschaft: Vol. 51. Kollegialität und Kooperation in der Schule. Theoretische Konzepte und empirische Befunde* (pp. 29–40). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-94284-1_2

Füssel, H.-P., & Leschinsky, A. (2008). Der institutionelle Rahmen des Bildungswesens. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer, & L. Trommer (Eds.), *Das Bildungswesen in der Bundesrepublik Deutschland.* Reinbek bei Hamburg: Rowohlt.

George, A. A., Hall, G. E., & Stiegelbauer, S. (2008). *Measuring implementation in schools: The stages of concern questionnaire*. Austin: Southwest Educational Development Laboratory.

Glasgow, R. E., & Emmons, K. M. (2007). How can we increase translation of research into practice? Types of evidence needed. *Annual Review of Public Health*, *28*, 413–433. https://doi.org/10.1146/annurev.publhealth.28.021406.144145

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 319–330. https://doi.org/10.1080/0969594042000304618

Goldstein, H. (2014). Responses to Andreas Schleicher's reply to open letter. *Policy Futures in Education*, *12*(7), 880–882. https://doi.org/10.2304/pfie.2014.12.7.880

Goldthorpe, J. H. (2001). Causation, statistics, and sociology. *European Sociological Review*, *17*(1), 1–20. https://doi.org/10.1093/esr/17.1.1

Good, T. L., Wiley, C. R. H., & Florez, I. R. (2009). Effective Teaching: an Emerging Synthesis. In L. J. Saha & A. G. Dworkin (Eds.), *Springer international handbooks of education: Vol. 21. International Handbook of Research on Teachers and Teaching* (pp. 803–816). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-73317-3_51

Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438.

Gräsel, C., Fussangel, K., & Pröbstel, C. (2006). Lehrkräfte zur Kooperation anregen - eine Aufgabe für Sisyphos? [Encourage teacher to cooperate - a task for Sisyphus]. *Zeitschrift für Pädagogik*, *52*(2), 205–219.

Gräsel, C. (2010). Stichwort: Transfer und Transferforschung im Bildungsbereich [Keyword: transfer and transfer research in education science]. *Zeitschrift für Erziehungswissenschaft*, *13*(1), 7–20. https://doi.org/10.1007/s11618-010-0109-8

Gräsel, C. (2011). Was ist Empirische Bildungsforschung? [What is educational research?]. In H. Reinders, H. Ditton, C. Gräsel, & B. Gniewosz (Eds.), *Empirische Bildungsforschung* (pp. 13–27). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-93015-2_1

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

Günther, A., & Wittl, W. (2017, April 6). Bayern will 870 Millionen Euro in Schulen investieren [Bavaria wants to invest 870 million euro in schools]. *Süddeutsche Zeitung*. Retrieved from http://www.sueddeutsche.de/bayern/schulpolitik-bayern-will-millionen-euro-in-schulen-investieren-1.3453649

Guo, J., Nagengast, B., Marsh, H. W., Kelava, A., Gaspard, H., Brandt, H.,. . . Trautwein, U. (2016). Probing the Unique Contributions of Self-Concept, Task Values, and Their Interactions Using Multiple Value Facets and Multiple Academic Outcomes. *AERA Open, 2*(1), 1-20. https://doi.org/10.1177/2332858415626884

Gustafsson, J.-E., & Yang Hansen, K. (2017). Changes in the Impact of Family Education on Student Educational Achievement in Sweden 1988?: 2014. *Scandinavian Journal of Educational Research*, *10*(2), 1–18. https://doi.org/10.1080/00313831.2017.1306799

Haddad, F. S. (2016). Similar questions, different answers. *The Bone & Joint Journal*, *98-B*(9), 1153–1154. https://doi.org/10.1302/0301-620X.98B9.38077

Haddad, W. D., & Demsky, T. (1995). *Education policy-planning process: An applied framework*. *Fundamentals of educational planning: Vol. 51*. Paris: UNESCO Internat. Inst. for Educational Planning.

Hall, G. E., & Hord, S. M. (2000). *Implementing change: Patterns, principles, and potholes*. Boston: Allyn and Bacon.

Hameyer, U., Frey, K., & Haft, H. (Eds.). (1983). *Handbuch der Curriculumforschung: Übersichten zur Forschung 1970 - 1981* [Handbook of curriculum research] (1. Ausg). Weinheim: Beltz.

Hamilton, L. S., Stecher, B. M., & Yuan, K. (2009). *Standards-based reform in the United States: History, research, and future directions.* California: RAND Cooperation. Retrieved from http://www.rand.org/pubs/reprints/RP1384.html.

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

Hanushek, E. A., & Woessmann, L. (2010). Education and economic growth. In E. Baker, B. McGaw, & P. Peterson (Eds.), *International encyclopedia of education* (pp. 245–252). Amsterdam: Elsevier Academic.

Hanushek, E. A., & Woessmann, L. (2012). The economic benefit of educational reform in the European Union. *CESifo Economic Studies*, *58*(1), 73–109. https://doi.org/10.1093/cesifo/ifr032

Hanushek, E. A., & Woessmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal*, *116*(510), 63-76. https://doi.org/10.1111/j.1468-0297.2006.01076.x

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hazle Bussey, L., Welch, J. C., & Mohammed, M. B. (2014). Effective consultants: A conceptual framework for helping school systems achieve systemic reform. *School Leadership & Management*, *34*(2), 156–178. https://doi.org/10.1080/13632434.2013.849684

Heck, R. H., & Hallinger, P. (2009). Assessing the contribution of distributed leadership to school improvement and growth in math achievement. *American Educational Research Journal*, *46*(3), 659–689. https://doi.org/10.3102/0002831209340042

Heck, R. H., & Hallinger, P. (2010). Collaborative leadership effects on school improvement: Integrating unidirectional- and reciprocal-effects models. *The Elementary School Journal*, *111*(2), 226–252. https://doi.org/10.1086/656299

Hedges, L. V., & Waddington, T. (1993). From evidence to knowledge to policy: Research synthesis for policy formation. *Review of Educational Research*, *63*(3), 345–352. https://doi.org/10.3102/00346543063003345

Helmke, A., & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. [Determinants of school achievement]. In F. E. Weinert (Ed.), *Enzyklopädie der Psychologie* (pp. 71–176). Göttingen: Hogrefe-Verlag.

Helmke, A. (2006). Was wissen wir über guten Unterricht? Über die Notwendigkeit einer Rückbesinnung auf den Unterricht als dem "Kerngeschäft" der Schule [What do we know about good teaching? On the need to return to teaching as the "core activity" of schools]. *Pädagogik*, *58*, 42–45.

Hendriks, M., Luyten, H., Scheerens, J., & Sleegers, P. (2014). Meta-Analyses. In J. Scheerens (Ed.), *Effectiveness of time investments in education* (pp. 55–142). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00924-7_4

Hitchcock, J. H., Kratochwill, T. R., & Chezan, L. C. (2015). What Works Clearinghouse standards and generalization of single-case design evidence. *Journal of Behavioral Education*, *24*(4), 459–469. https://doi.org/10.1007/s10864-015-9224-1

Hübner, N., Wagner, W., Nagengast, B., & Trautwein, U. (2017). Putting all students in one basket does not produce equality: Gender-specific effects of curricular intensification in upper secondary school. Manuscript submitted for publication.

Hübner, N., Wagner, W., Hochweber, J., Neumann, M., & Nagengast, B. (2017). Comparing apples and oranges: Curricular intensification reforms can change the meaning of students' grades! Manuscript submitted for publication.

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

Hübner, N., Wagner, W., Kramer, J., Nagengast, B., & Trautwein, U. (2017). Die G8-Reform in Baden-Württemberg: Kompetenzen, Wohlbefinden und Freizeitverhalten vor und nach der Reform [The G8 reform in Baden-Württemberg: Competencies, well-being, and leisure time cefore and after the reform]. *Zeitschrift für Erziehungswissenschaft*, *1*(2), 1–24. https://doi.org/10.1007/s11618-017-0737-3

Huebener, M., Kuger, S., & Marcus, J. (2017). Increased instruction hours and the widening gap in student performance. *Labour Economics.* Advance online publication. https://doi.org/10.1016/j.labeco.2017.04.007

Hughes, R. G. (Ed.). (2008). *Patient Safety and Quality: An Evidence-Based Handbook for Nurses. AHRQ Publication No. 08-0043*. Rockville: Agency for Healthcare Research and Quality.

Ivanov, S., Nikolova, R., & Vieluf, U. (2016). G8 vs. G9 im Kohortenvergleich [G8 vs. G9 as a comparison of cohorts]. In J. Kramer, M. Neumann, & U. Trautwein (Eds.), *Edition ZfE: Band 2. Abitur und Matura im Wandel. Historische Entwicklungslinien, aktuelle Reformen und ihre Effekte* (pp. 81–106). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-658-11693-4_4

Jann, W., & Wegrich, K. (2007). Theories of the policy cycle. In F. Fischer, G. Miller, & M. S. Sidney (Eds.), *Public administration and public policy: Vol. 125. Handbook of public policy analysis: Theory, politics, and methods* (Vol. 125, pp. 43–62). Boca Raton: CRC/Taylor & Francis.

Kemper, T., & Weishaupt, H. (2011). Region und soziale Ungleichheit [Regional and social disparity]. In H. Reinders, H. Ditton, C. Gräsel, & B. Gniewosz (Eds.), *Empirische Bildungsforschung* (pp. 209–219). Wiesbaden: VS Verlag für Sozialwissenschaften.

Klieme, E. (2006). Empirische Unterrichtsforschung: aktuelle Entwicklungen, theoretische Grundlagen und fachspezifische Befunde. Einführung in den Thementeil [Empirical classroom teaching research: Current developments, theoretical foundations and subject-specific findings. Introduction to the topic]. *Zeitschrift für Pädagogik*, *52*(6), 765–773.

KMBW. (2013). Bildungswege in Baden-Württemberg: Abschlüsse und Anschlüsse [School careers in Baden-Württemberg: Graduation and access]. Retrieved from https://www.baden-wuerttemberg.de/fileadmin/redaktion/dateien/PDF/Bildungswege-BW-2014.pdf

KMBW. (2015). Die Gemeinschaftsschule in Baden-Württemberg [The community school in Baden-Württemberg]. Retrieved from https://www.baden-wuerttemberg.de/fileadmin/redaktion/dateien/PDF/Gemeinschaftschule_Broschuere_neu.pdf

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

KMK. (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring* [Overall strategy of the KMK on educational monitoring] (1. Aufl.). Neuwied: Luchterhand in Wolters Kluwer Deutschland.

KMK. (2016). *KMK Bildungsmonitoring (II): Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring* [Educational monitoring strategy of the KMK]. Kronach: Carl Link.

KMK. (2017). The Standing Conference of the Ministers of Education and Cultural Affairs (KMK). Retrieved from https://www.kmk.org/kmk/information-in-english/standing-conference.html

Köller, O., Watermann, R., Trautwein, U., & Lüdtke, O. (2004). *Wege zur Hochschulreife in Baden-Württemberg: TOSCA - eine Untersuchung an allgemein bildenden und beruflichen*

Gymnasien [Ways towards higher education entrance qualification in Baden-Württemberg: TOSCA – an investigation of general and vocational upper secondary schools]. Opladen: Leske und Budrich.

Konstantopoulos, S., & Hedges, L. V. (2008). How large an effect can we expect from school reforms? *Teachers College Record*, *110*(8), 1611–1638.

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Cambridge, Mass.: Harvard University Press.

Kühn, S. M., van Ackeren, I., Bellenberg, G., Reintjes, C., & Im Brahm, G. (2013). Wie viele Schuljahre bis zum Abitur? Eine multiperspektivische Standortbestimmung im Kontext der aktuellen Schulzeitdebatte [How many years until abitur in German upper secondary schooling? – Taking stock in the context of current school duration debates]. *Zeitschrift für Erziehungswissenschaft, 16*(1), 115–136. https://doi.org/10.1007/s11618-013-0339-7

Kunter, M., & Trautwein, U. (2013). *Psychologie des Unterrichts* [Psychology of teaching]. Paderborn: Schöningh.

Künzli, R., Fries, A.-V., Hürlimann, W., & Rosenmund, M. (2013). *Der Lehrplan - Programm der Schule* [The curriculum - Program of the school]. *Institutionenforschung im Bildungsbereich*. Weinheim u.a.: Beltz Juventa.

Kyriakides, L., Charalambous, C., Philippou, G., & Campbell, R. J. (2006). Illuminating reform evaluation studies through incorporating teacher effectiveness research: A case study in mathematics. *School Effectiveness and School Improvement*, *17*(1), 3–32. https://doi.org/10.1080/09243450500404293

Landesinstitut für Schulentwicklung. (2016). *Beiträge zur Bildungsberichterstattung: VERA 8 2015* [Contributions to the education report: VERA 8 2015]. Stuttgart: Landesinstitut für Schulentwicklung.

Lasswell, H. D. (1956). *The decision process: Seven categories of functional analysis*. College Park: University of Maryland Press.

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal*, *125*(588), 397-424. https://doi.org/10.1111/ecoj.12233

Lee, J. (2015). Educational testing: Measuring and remedying achievement gaps. In R. A. Scott & S. M. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences. An* *interdisciplinary, searchable, and linkable resource* (pp. 1–14). Hoboken, N.J.: John Wiley & Sons.

Levin, B. (2000). Putting students at the centre in education reform. *Journal of Educational Change*, *1*(2), 155–172. https://doi.org/10.1023/A:1010024225888

Levin, H. (1986). Are longer school sessions a good investment? *Contemporary Economic Policy, 4*(3), 63–75. https://doi.org/10.1111/j.1465-7287.1986.tb00851.x

Lohr, K. N. (2004). Rating the strength of scientific evidence: relevance for quality improvement programs. *International journal for quality in health care : journal of the International Society for Quality in Health Care, 16*(1), 9–18. https://doi.org/10.1093/intqhc/mzh005

Lossen, K., Tillmann, K., Holtappels, H. G., Rollett, W., & Hannemann, J. (2016). Entwicklung der naturwissenschaftlichen Kompetenzen und des sachunterrichtsbezogenen Selbstkonzepts bei Schüler/-innen in Ganztagsgrundschulen. Ergebnisse der Längsschnittstudie StEG-P zu Effekten der Schülerteilnahme und der Angebotsqualität [Development of competencies in science and subject-related self-concept of students in all-day primary schools. Results from the StEG-P longitudinal study of the effects of student participation in and quality of extra-curricular activities]. *Zeitschrift für Pädagogik, 62*(6), 760–779. Retrieved from http://www.beltz.de/fachmedien/erziehungs_und_sozialwissenschaften/zeitschriften/zeitschrift_fuer_paedagogik/article/Journal.html?tx_beltz_journal%5Barticle%5D=34655&cHash=5f3a6f4041c0d6bb6cc6cbdfb392c28e

Ma, X., & Johnson, W. (2008). Mathematics as the critical filter: Curricular effects on gendered career choices. In H. M. G. Watt & J. S. Eccles (Eds.), *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences* (pp. 55–83). Washington: American Psychological Association. https://doi.org/10.1037/11706-002

Maier, U. (2008). Vergleichsarbeiten im Vergleich – Akzeptanz und wahrgenommener Nutzen standardbasierter Leistungsmessungen in Baden-Württemberg und Thüringen [Comparative tests in comparison – Acceptance and perceived use of standardized student assessment in Baden-Württemberg and Thuringia]. *Zeitschrift für Erziehungswissenschaft, 11*(3), 453–474. https://doi.org/10.1007/s11618-008-0036-0

Malouf, D. B., & Taymans, J. M. (2016). Anatomy of an evidence base. *Educational Researcher, 45*(8), 454–459. https://doi.org/10.3102/0013189X16678417

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/10.3102/00028312023001129

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

Marsh, H. W., & Yeung, A. S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and English constructs. *American Educational Research Journal, 35*(4), 705–738. https://doi.org/10.3102/00028312035004705

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The Big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review, 20*(3), 319–350. https://doi.org/10.1007/s10648-008-9075-6

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: reciprocal effects models of causal ordering. *Child Development, 76*(2), 397–416. https://doi.org/10.1111/j.1467-8624.2005.00853.x

Martin, M. O., Mullis, I. V. S., Foy, P., & Olson, J. F. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: IEA and TIMSS & PIRLS International Study Center, Boston College.

Mayntz, R. (1977). Die Implementation politischer Programme: Theoretische Überlegungen zu einem neuen Forschungsgebiet [The implementation of political programs: Theoretical considerations on a new field of research]. *Die Verwaltung : Zeitschrift für Verwaltungsrecht und Verwaltungswissenschaften, 10*(1), 51–66.

McConnell, A. (2010). Policy success, policy failure and grey areas in-between. *Journal of Public Policy, 30*(03), 345–362. https://doi.org/10.1017/S0143814X10000152

McDermott, A. M., Fitzgerald, L., & Buchanan, D. A. (2013). Beyond acceptance and resistance: Entrepreneurial change agency responses in policy implementation. *British Journal of Management, 24*(S1), 93-115. https://doi.org/10.1111/1467-8551.12012

McLaughlin, M. W. (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis*, *9*(2), 171–178. https://doi.org/10.3102/01623737009002171

Meyer, H.-D., & Zahedi, K. (2014). An open letter: To Andreas Schleicher, OECD, Paris. Retrieved from http://www.globalpolicyjournal.com/blog/05/05/2014/open-letter-andreas-schleicher-oecd-paris#

Meyer, T., & Thomsen, S. L. (2015). Schneller fertig, aber weniger Freizeit? Eine Evaluation der Wirkungen der verkürzten Gymnasialschulzeit auf die außerschulischen Aktivitäten der Schülerinnen und Schüler [Finished faster, but less leisure time? – An evaluation on the effects of shortening time at the Gymnasium on students extracurricular activities]. *Schmollers Jahrbuch*, *135*(3), 249–277. https://doi.org/10.3790/schm.135.3.249

Minkley, N., Rest, M., Terstegen, S., Kirchner, W. H., & Wolf, O. T. (2015). Mehr Stress durch G8? Stressbelastung von Abiturienten mit regulärer und verkürzter Gymnasialzeit in NRW [Stress levels of secondary school learners exposed to a lower and regular number of years in school in NRW]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *47*(4), 188–198. https://doi.org/10.1026/0049-8637/a000133

MSW NRW. (2016). *Das Schulwesen in Nordrhein-Westfalen aus quantitativer Sicht: 2015/2016* [The education system in North Rhine-Westphalia from a quantitative perspective]. Düsseldorf: Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art - teacher effectiveness and professional learning. *School Effectiveness and School Improvement*, *25*(2), 231–256. https://doi.org/10.1080/09243453.2014.885451

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016a). TIMSS 2015 international results in mathematics. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/timss-2015/mathematics/student-achievement/

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016b). TIMSS 2015 international results in science. Retrieved from Boston College, TIMSS & PIRLS International Study

Center website: http://timssandpirls.bc.edu/timss2015/international-results/timss-2015/science/student-achievement/

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. Oxford: Oxford University Press.

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: American Psychological Association. https://doi.org/10.1037/11706-004

Neumann, M., Becker, M., Baumert, J., Maaz, K., & Köller, O. (Eds.). (2017). *Zweigliedrigkeit im deutschen Schulsystem: Potenziale und Herausforderungen in Berlin* [The bipartide German school system: Potentials and challenges in Berlin]. Münster: Waxmann.

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* (19–36). New York: Routledge.

Nomi, 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

Nomi, T., & Allensworth, E. (2009). “Double-dose” algebra as an alternative strategy to remediation: Effects on students' academic outcomes. *Journal of Research on Educational Effectiveness*, *2*(2), 111–148. https://doi.org/10.1080/19345740802676739

OECD. (2007). *PISA 2006 science competencies for tomorrow's world*. Paris: OECD Publishing.

OECD. (2010). *PISA 2009 results: What students know and can do:: Student performance in reading, mathematics and science*. Paris: OECD Publishing.

OECD. (2013). *PISA 2012 results: What students know and can do: Student performance in reading, mathematics, and science*. Paris: OECD Publishing.

OECD. (2014). *PISA 2012 Results: What students know and can do: Student performance in mathematics, reading and science*. *PISA*. Paris: OECD.

OECD. (2015). *Education policy outlook 2015*. Paris: OECD Publishing.

OECD. (2016a). *Education at a glance 2016: OECD indicators*. Paris: OECD Publishing.

OECD. (2016b). *PISA 2015 Results (Volume I): Excellence and equity in education*. *PISA*. Paris: OECD Publishing.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science (New York, N.Y.), 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

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 multicontext study. *Developmental Psychology*, *48*(6), 1629–1642. https://doi.org/10.1037/a0029167

Peterson, P. E., Howell, W. G., Wolf, P. J., & Campbell, D. E. (2003). School vouchers: Results from randomized experiments. In C. M. Hoxby (Ed.), *NBER-Conference Report. The Economics of School Choice* (pp. 107–144). Chicago: University of Chicago Press.

Pont, B., Nusche, D., Moorman, H., & Hopkins, D. (2008). *Improving school leadership*. Paris: OECD.

Porter, R. E., Fusarelli, L. D., & Fusarelli, B. C. (2015). Implementing the common core: How educators interpret curriculum reform. *Educational Policy*, *29*(1), 111–139. https://doi.org/10.1177/0895904814559248

Puhani, P. A., & Weber, A. M. (2008). Does the early bird catch the worm? In C. Dustmann, B. Fitzenberger, & S. Machin (Eds.), *The Economics of Education and Training* (pp. 105–132). Heidelberg: Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-2022-5\textunderscore

Qi, J., & Levin, B. (2013). Assessing organizational efforts to mobilize research knowledge in education. *Education Policy Analysis Archives*. Advance online publication. https://doi.org/10.14507/epaa.v21n2.2013

Quinn, R. E., & Sonenshein, S. (2008). Four general strategies for changing human systems. In T. G. Cummings (Ed.), *Handbook of organization development* (69-78;). Thousand Oaks Calif. u.a.: Sage.

Quis, J. S. (2015). *Does higher learning intensity affect student well-being?: Evidence from the National Educational Panel Study* (neue Ausg). *BERG working paper series on government and growth: Vol. 94*. Bamberg: BERG.

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

Reezigt, G. J., Guldemond, H., & Creemers, B. P. (1999). Empirical validity for a comprehensive model on educational effectiveness. *School Effectiveness and School Improvement*, *10*(2), 193–216. https://doi.org/10.1076/sesi.10.2.193.3503

Reinders, H., Ditton, H., Gräsel, C., & Gniewosz, B. (2011). *Empirische Bildungsforschung: Strukturen und Methoden* [Empirical Educational Research: Structures and Methods] (1. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften. Retrieved from http://dx.doi.org/10.1007/978-3-531-93015-2

Reinhart, A. L., Haring, S. H., Levin, J. R., Patall, E. A., & Robinson, D. H. (2013). Models of not-so-good behavior: Yet another way to squeeze causality and recommendations for practice out of correlational data. *Journal of Educational Psychology*, *105*(1), 241–247. https://doi.org/10.1037/a0030368

Reynolds, D., Sammons, P., de Fraine, 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

Richardson, V., & Placier, P. (2001). Teacher change. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., 905-947;). Washington DC: American Educational Research Assoc.

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

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J.-H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien [Challenges in estimations of trends in large-scale assessments: A calibration of the German PISA data]. *Diagnostica*, *63*(2), 148–165. https://doi.org/10.1026/0012-1924/a000177

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

Rolff, H.-G. (1970). *Bildungsplanung als rollende Reform*. Frankfurt: Diesterweg.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (2004). *Evaluation: A systematic approach* (7. ed.). London: Sage.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Sabatier, P. A. (2007). *Theories of the policy process* (2nd ed.). Boulder, Colo.: Westview Press.

Schaffer, E., Nesselrodt, P., & Stringfield, S. (1997). *Impediments to reform: An analysis of destabilizing issues in ten promising programs*. Arlington, VA: Educational Research Science.

Scheerens, J. (1990). School effectiveness research and the development of process indicators of school functioning. *School Effectiveness and School Improvement*, *1*(1), 61–80. https://doi.org/10.1080/0924345900010106

Scheerens, J. (2014a). Introduction. In J. Scheerens (Ed.), *Effectiveness of time investments in education* (pp. 1–5). Cham: Springer International Publishing.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness* (1st ed.). Oxford, OX and New York, N.Y.: Pergamon.

Scheerens, J. (Ed.). (2014b). *Effectiveness of time investments in education*. Cham: Springer International Publishing.

Scheerens, J., & Hendriks, M. (2014). State of the Art of Time Effectiveness. In J. Scheerens (Ed.), *Effectiveness of time investments in education* (pp. 7–29). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00924-7\textunderscore

Schipolowski, S., & Sachse, K. A. (2016). Mittelwerte und Streuungen der im Fach Englisch erreichten Kompetenzen [Means and variances of English competencies]. In P. Stanat, K. Böhme, S. Schipolowski, & N. Haag (Eds.), *IQB-Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich.* Münster: Waxmann.

Schlotter, M., Schwerdt, G., & Woessmann, L. (2011). Econometric methods for causal evaluation of education policies and practices: A non-technical guide. *Education Economics*, *19*(2), 109–137. https://doi.org/10.1080/09645292.2010.511821

Seidel, T. (2015). Performance Assessment and Professional Development in University Teaching. In I. M. Welpe, J. Wollersheim, S. Ringelhan, & M. Osterloh (Eds.), *Incentives and Performance. Governance of Knowledge-Intensive Organizations* (pp. 465–480). Cham: Springer International Publishing.

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

Shakeel, M. D., Anderson, K. P., & Wolf, P. J. (2016). The participant effects of private school vouchers across the globe: A meta-analytic and systematic review. *EDRE Working Paper, 07*. Retrieved from https://ssrn.com/abstract=https://ssrn.com/abstract=2777633

Slavin, R. E. (2008). Perspectives on evidence-based research in education: What works?: Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 5–14. https://doi.org/10.3102/0013189X08314117

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15–21. https://doi.org/10.3102/0013189X031007015

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., Hubbard, L., & Mehan, H. (2004). Reform ideas that travel far afield: The two cultures of reform in New York City's District #2 and San Diego. *Journal of Educational Change*, *5*(2), 161–197. https://doi.org/10.1023/B:JEDU.0000033053.99363.e4

Steinert, B., Klieme, E., Maag Merki, K., Döbrich, P., Halbheer, U., & Kunz, A. (2006). Lehrerkooperation in der Schule: Konzeption, Erfassung, Ergebnisse [Teacher cooperation in school: Conceptualization, Assessment, Results]. *Zeitschrift für Pädagogik*, *52*(2), 185–204.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, *35*(5), 401–426. https://doi.org/10.1016/j.intell.2006.09.004

Swanson, C. B., & Stevenson, D. L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. *Educational Evaluation and Policy Analysis*, *24*(1), 1–27. https://doi.org/10.3102/01623737024001001

Terhart, E. (1983). Curriculumforschung aufgrund interpretativer Methoden [Curriculum research based on interpretative methods]. In U. Hameyer, K. Frey, & H. Haft (Eds.), *Handbuch der Curriculumforschung. Übersichten zur Forschung 1970 - 1981* (1st ed., pp. 533–542). Weinheim: Beltz.

Terhart, E. (2011). Has John Hattie really found the holy grail of research on teaching?: An extended review of visible learning. *Journal of Curriculum Studies*, *43*(3), 425–438. https://doi.org/10.1080/00220272.2011.576774

The National Commission on Excellence. (1983). *A nation at risk: The imperative for educational reform*. Washington: Government Printing Office.

Thiel, F. (2014). Evidenzbasierte Bildungspolitik. Generierung und Nutzung wissenschaftlichen Wissens. [Evidence-based educational policy: Generation and usage of scientific knowledge]. In BMBF (Ed.), *Bildungsforschung 2020. Herausforderungen und Perspektiven. Dokumentation der Tagung des Bundesministeriums für Bildung und Forschung vom 29.-30. März 2012* (pp. 116–127). Berlin: BMBF.

Thomas, S. M., Gana, Y., & Muñoz-Chereau, B. (2016). England: 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* (37–57;). New York: Routledge.

TMBJS. (2016). Schullaufbahnen in Thüringen: Schuljahr 2016/2017 [School careers in Thuringia: School year 2016/2017]. Retrieved from http://apps.thueringen.de/de/publikationen/pic/pubdownload584.pdf

Torgerson, C. J., & Torgerson, D. J. (2001). The need for randomised controlled trials in educational research. *British Journal of Educational Studies*, *49*(3), 316–328. https://doi.org/10.1111/1467-8527.t01-1-00178

Trautwein, U., Neumann, M., Nagy, G., Lüdtke, O., & Maaz, K. (Eds.). (2010). *Schulleistungen von Abiturienten: Die neugeordnete gymnasiale Oberstufe auf dem Prüfstand* [School achivement of upper secondary school students. The rearrangend upper secondary school on the trial] (1. Auflage). Wiesbaden: VS Verlag für Sozialwissenschaften.

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., & 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* (pp. 467–501). Reinbek bei Hamburg: Rowohlt.

United Nations. (2015). *Millennium development goals report 2015*. New York: United Nations.

United Nations General Assembly. (1948). Universal declaration of human rights. Retrieved from http://www.un.org/en/universal-declaration-human-rights/

van Ackeren, I., Klemm, K., & Kühn, S. M. (2015). *Entstehung, Struktur und Steuerung des deutschen Schulsystems: Eine Einführung* [Development, structure and governance of the German education system: An introduction]. Wiesbaden: Springer Fachmedien Wiesbaden. Retrieved from http://dx.doi.org/10.1007/978-3-531-20000-2

Volante, L. (2016). International achivement testing, education policy and large-sclae reform. In L. Volante (Ed.), *The intersection of international achievement testing and educational policy. Global Perspectives on Large-Scale Reform* (pp. 3–16). New York: Routledge.

Wacker, A., & Kramer, J. (2012). Vergleichsarbeiten in Baden-Württemberg [Standard comparative testing in Baden-Württemberg]. *Zeitschrift für Erziehungswissenschaft*, *15*(4), 683–706. https://doi.org/10.1007/s11618-012-0326-4

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

Wang, J., Odell, S. J., Klecka, C. L., Spalding, E., & Lin, E. (2010). Understanding teacher education reform. *Journal of Teacher Education*, *61*(5), 395–402. https://doi.org/10.1177/0022487110384219

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

Weber, M. (1919). Wissenschaft als Beruf [Science as a profession]. Retrieved from http://verlag.ub.uni-potsdam.de/html/494/html/WL.pdf

Wecker, C., Vogel, F., & Hetmanek, A. (2017). Visionär und imposant – aber auch belastbar? [Visionary and impressive – but also reliable?]. *Zeitschrift für Erziehungswissenschaft*, *20*(1), 21–40. https://doi.org/10.1007/s11618-016-0696-0

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, *21*(1), 1. https://doi.org/10.2307/2391875

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 173–195). San Diego, CA: Academic Press.

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, *12*(3), 265–310. https://doi.org/10.1016/0273-2297(92)90011-P

Wollmann, H. (2014). Zur (Nicht-) Verwendung von Evaluationsergebnissen in Politik und Verwaltung: Eine vernachlässigte Fragestellung der Evaluationsforschung [On (not-) using evaluation results in policy and administration: A neglected question in evaluation research]. In S. Kropp & S. Kuhlmann (Eds.), *Wissen und Expertise in Politik und Verwaltung* (Vol. 1, 87-102;). Opladen: Budrich.

Wortman, P. M. (1983). Evaluation research: A methodological perspective. *Annual Review of Psychology*, *34*(1), 223–260. https://doi.org/10.1146/annurev.ps.34.020183.001255

WWC. (2015). *About the WWC*. Retrieved from http://ies.ed.gov/ncee/wwc/aboutus.aspx

Young, T., & Lewis, W. D. (2015). Educational policy implementation revisited. *Educational Policy*, *29*(1), 3–17. https://doi.org/10.1177/0895904815568936

7 List of Abbreviations

ACT	American College Test
BEE	Best Evidence Encyclopedia
BW	Baden-Württemberg
CI	Curricular Intensification (note: Confidence interval with %)
EER	Educational Effectiveness Research
EVT	Expectancy-Value Theory
GDP	Gross Domestic Product
KMBW	Ministerium für Kultus, Jungen und Sport Baden-Württemberg
LSA	Large-scale assessments
MSW	Ministerium für Schule und Weiterbildung
NAEP	National Assessment of Educational Progress
NRW	North Rhine-Westphalia
OECD	Organization for Economic Co-operation and Development
SAT	Scholastic Assessment Test
SES	Socioeconomic Status
TH	Thuringia
TMBJS	Thüringer Ministerium für Bildung, Jungend und Sport
WWC	What Works Clearinghouse