Frame-of-Reference Effects on Academic Self-Concept: Addressing Unresolved Issues with New Designs

Dissertation

zur Erlangung des Doktorgrades

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Moritz Fleischmann

Tübingen

2020

1. Betreuer:	Prof. Dr. Ulrich Trautwein
2. Betreuer:	Prof. Dr. Benjamin Nagengast
Tag der mündlichen Prüfung:	10.12.2020
Dekan:	Prof. Dr. Josef Schmid
1. Gutachter:	Prof. Dr. Ulrich Trautwein
2. Gutachter:	Prof. Dr. Steffen Zitzmann

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisors Prof. Dr. Ulrich Trautwein and Prof. Dr. Benjamin Nagengast for giving me the freedom to develop this dissertation project and supporting me throughout the last three years. Many thanks also go to Dr. Nicolas Hübner, who inspired me with his enthusiasm for educational research and stood by my side whenever I needed him. I would also like to thank Prof. Herb Marsh and all staff members from the Institute of Positive Psychology and Education for welcoming me into their community on two lab visits and allowing me to gain insights into the world of research on an international level, while simultaneously improving my ping pong skills considerably. Next, I would like to thank all members of the SIG Motivation and SIG Educational Effectiveness for feedback that sharpened my research ideas, contributing substantially to the present dissertation. Many thanks also go to my (former) officemates Constanze and Fabian, and office neighbors Sven and Thomas, who enriched my everyday working life with coffee and lunch breaks and many academic but also non-academic discussions. Furthermore, I wish to thank Patricia for substantively and emotionally supporting me at the end of my dissertation phase and all other friends and colleagues at the Hector Research Institute for making me feel welcome in your presence. Finally, I thank my friends, family, and Franzi, who have accompanied me over various ups and downs and provided me with unconditional love during my life. Thank you so much!

ABSTRACT

Students' self-perceived competence in academic domains, also referred to as *academic self-concept* (Marsh et al., 2016), is assumed to be a central motivational factor that affects academic effort, achievement, aspirations, and choices (e.g., Guo et al., 2015; Valentine et al., 2004). Consequently, the formation of a positive academic self-concept is regarded as an essential requirement for successful learning processes, and researching its antecedents is of high theoretical and practical relevance. Social comparison processes, in which students evaluate their academic achievement in relation to their classmates (e.g., Huguet et al., 2009; Marsh, Kuyper, et al., 2014), are assumed to be one focal determinant of academic self-concept. Empirical evidence for this assumption stems from the finding that, on average, equally able students have a higher academic self-concept in low-achieving schools and classes than in high-achieving ones. In the former scenario, students feel like big fish in little ponds; thus, this phenomenon has been labeled the *big fish little pond effect* (*BFLPE*). The BFLPE has been extensively investigated in the last three decades. For instance, numerous studies have demonstrated its robustness to individual and contextual level moderators, the fact that frame-of-reference effects also impact other educational outcomes, and its existence across cultures (for an overview, see Marsh & Seaton, 2015).

Despite the massive number of studies investigating the BFLPE, an in-depth understanding of the frame-of-reference effect is still lacking with respect to mechanisms (e.g., To which reference groups do students tend to compare themselves?), implications (e.g., What does the BFLPE mean for the design of educational systems?), and interdisciplinary integration (e.g., How can the BFLPE theory be embedded in other social science disciplines that focus on social comparison processes?). In the present dissertation, I argue that one of the reasons for this lack of in-depth understanding is that previous research on the BFLPE has partly been characterized by homogeneity in terms of the research designs used (Dai et al., 2013; Dai & Rinn, 2008). More specifically, researchers have investigated the BFLPE using education-specific cluster sampling data in which either a random sample of students within schools or a random sample of intact classrooms was drawn. This cross-sectional data was then analyzed using multilevel models in which the school or classroom represented the higher level. The overarching aim of the present dissertation is to address unresolved issues in research on the BFLPE by extending the range of research designs used and consequently providing new insights. Specifically, in the present dissertation, I discuss four unresolved issues in research on the BFLPE: multiple class environments as frames of reference for academic self-concept formation, the association between grading on a curve and the BFLPE, the effects of tracking

on academic self-concept, and neighborhood effects on academic self-concept. For each of these unresolved issues, I will describe design-based challenges of previous research on a conceptual level and clarify what is needed for new designs. Finally, in four empirical studies, I address the presented issues with innovative research designs.

Alongside this overarching aim, this dissertation pursues two subordinate aims that were addressed with two empirical studies each. The first subordinate aim is to use extensive large-scale data (comprehensive educational monitoring data and interdisciplinary large-scale data) for an in-depth investigation of the mechanisms and interdisciplinary integration of the BFLPE. Specifically, Study 1 used Austrian educational monitoring data to investigate multiple class environments as pivotal frames of reference for academic self-concept formation in systems with course-by-course tracking. Study 2 used interdisciplinary large-scale data to explore the neighborhood as a frame of reference for academic self-concept formation. The second subordinate aim is to use natural experiments to investigate the mechanisms and implications of the BFLPE in greater detail. For this purpose, Study 3 used a school reform that abolished formal grades to investigate the association between grading on a curve and the BFLPE. Additionally, Study 4 used two school reforms in which students were detracked to test the BFLPE's predictions concerning tracking.

The first study (*Which Class Matters? Juxtaposing Multiple Class Environments as Frames of Reference for Academic Self-Concept Formation*) was based on the 2012 Austrian Educational Standard Assessment (BIFIE, 2016; Schreiner & Breit, 2012), a comprehensive survey of all Austrian eighth-grade students in the domain of mathematics that contains identifiers for the multiple educational environments students experience. This extensive dataset made it possible to investigate the pivotal frames of reference for academic self-concept formation in school systems with course-by-course tracking. Secondary school students were tracked according to ability in the core subjects (mathematics, German, English) but attended all other subjects in the same mixed-ability class. When regressing math self-concept on math achievement aggregates on all levels in which students were nested, the math class BFLPE was most negative. The regular class BFLPE was less negative, and the school BFLPE was the least negative. These results are in line with local dominance theory, which argues that more local comparative information matters the most for self-evaluations.

The second study (*Living in the Big Pond: How Socioeconomic Neighborhood Composition Predicts Students' Academic Self-Concept*) benefited from data from Starting Cohort 3 of the German National Educational Panel Study (NEPS; Blossfeld et al., 2011), an

interdisciplinary large-scale assessment that contains information on students' educational outcomes as well as neighborhood characteristics. This extensive dataset made it possible to investigate the neighborhood as a frame of reference for academic self-concept formation. Better neighborhood socioeconomic conditions did not or negatively affected students' academic self-concept. Our results stand in supposedly contrast to neighborhood effects research in sociology, which has found that better neighborhood socioeconomic conditions positively impact a broad range of educational outcomes.

The third study (*Can Grades Move the Big Fish? The Consequences of Receiving Report Cards for Frame-of-Reference Effects on Academic Self-Concept*) was based on data from the Swedish Evaluation Through Follow-Up Study (ETF; Härnqvist, 2000). The data were collected during a reform period in which Swedish municipalities were free to decide whether or not to abolish formal grading in elementary school, enabling us to compare non-graded and graded students regarding the BFLPE in a natural experiment. We found no differences between non-graded and graded students regarding the BFLPE. The results are in line with an evolutionary approach to social comparisons, in which social comparison processes present an innate human drive that exists independent of grade provision.

The fourth study (*The Dark Side of Detracking: Mixed-Ability Classrooms Hurt Low Achievers' Math Motivation*) uses data coming from the Austrian National Educational Standard Assessment from 2012 and 2017 (Schreiner et al., 2017; Schreiner & Breit, 2012) as well as the Additional Study Thuringia from the German National Educational Panel Study (NEPS; Blossfeld et al., 2011). These data measured students before and after detracking school reforms, enabling us to investigate such reforms by employing a cohort-control design, in which student cohorts before and after detracking are compared. The BFLPE predicts that detracking decreases low achievers' motivation in terms of academic self-concept, as this group of students is exposed to high-achieving classmates in mixed-ability classrooms. In line with this prediction, we found that low-achieving students' academic self-concept was negatively impacted by detracking, whereas this was not the case for high achievers. Our results stand in contrast to detracking proponents, who argue that abolishing ability grouping will increase low-achieving students' motivation, while highlighting the practical implications of the BFLPE.

Finally, at the end of the dissertation, all four studies' findings are embedded in a broader research context. There is also a final assessment of how successfully the design-based challenges raised have been addressed. Additionally, strengths and limitations are presented, and implications for practice and future research are discussed.

ZUSAMMENFASSUNG

Die Einschätzung von Schülerinnen und Schülern (SuS) über ihre eigenen schulischen Fähigkeiten, welche auch als *akademisches Selbstkonzept* bezeichnet wird (Marsh et al., 2016), stellt einen wichtigen motivationalen Faktor dar. Es wird davon ausgegangen, dass ein positives akademisches Selbstkonzept die schulische Leistungsbereitschaft und Leistungsentwicklung steigert und einen großen Einfluss auf Bildungsaspirationen und Entscheidungen hat (z.B. Guo et al., 2015; Valentine et al., 2004). Daher gilt die Entwicklung eines positiven akademischen Selbstkonzepts als eine wichtige Voraussetzung für erfolgreiche Lernprozesse. Gleichzeitig ist damit die Erforschung der Einflussgrößen des akademischen Selbstkonzepts sowohl aus theoretischer als auch aus praktischer Sicht von hoher Relevanz. Es wird davon ausgegangen, dass das akademische Selbstkonzept in einem entscheidenem Ausmaß durch soziale Vergleichsprozesse beeinflusst wird. In diesen bewerten SuS ihre akademische Leistung im Vergleich mit ihren Mitschülerinnen und Mitschülern (Huguet et al., 2009; Marsh, Kuyper, et al., 2014). Empirische Evidenz für diese Annahme stammt von Studienergebnissen, welche zeigen, dass SuS mit einer bestimmten schulischen Leistung, welche Schulen oder Klassen mit einem niedrigem durchschnittlichen Leistungsniveau besuchen, ein höheres akademisches Selbstkonzept aufweisen als SuS mit der gleichen Leistung, die sich in Schulen und Klassen mit sehr leistungsstarken Mitschülerinnen und Mitschülern befinden. Weil sich in diesem Szenario SuS wie große Fische im kleinen Teich fühlen wird dieser Befund auch *big fish little pond effect* (*BFLPE*) genannt. In den letzten drei Jahrzehnten wurde der BFLPE intensiv untersucht. Zum Beispiel konnten zahlreiche Studien die Unveränderlichkeit des BFLPEs entgegen individueller oder kontextueller Moderatoren zeigen. Weiterhin wurde gezeigt, dass solche Bezugsrahmeneffekte auch andere bildungsbezogene Outcomes beeinflussen und dass der BFLPE auch über verschiedene Kulturkreise hinweg existiert (für einen Überblick, siehe Marsh & Seaton, 2015).

Trotz der großen Anzahl an Studien, die sich mit dem BFLPE befassen, mangelt es immer noch an einem tieferen Verständnis der Mechanismen (z.B. Mit welchen Bezugsgruppen vergleichen sich SuS vorwiegend?), Implikationen (z.B. Was bedeutet der BFLPE für die Gestaltung von Bildungssystemen?) und der interdisziplinären Integration (z.B. Wie kann die BFLPE-Theorie in andere sozialwissenschaftliche Disziplinen eingebettet werden, die sich mit sozialen Vergleichsprozessen befassen?). In der vorliegenden Dissertation arbeite ich heraus, dass einer der Gründe für diesen Mangel an vertieftem Verständnis darin liegt, dass die bisherige Forschung zum BFLPE teilweise durch eine Homogenität hinsichtlich der verwendeten Forschungsdesigns gekennzeichnet war (z.B. Dai, 2004; Dai & Rinn, 2008).

Genauer gesagt wurde der BFLPE oft anhand bildungsspezifischer Cluster-Stichprobendaten untersucht, bei denen entweder eine Zufallsstichprobe von SuS innerhalb von Schulen oder eine Zufallsstichprobe intakter Schulklassen gezogen wurde. Diese Querschnittsdaten wurden dann mithilfe von Mehrebenenmodellen analysiert, bei denen die Schule oder die Schulklasse die höhere Ebene darstellte. Das übergeordnete Ziel der vorliegenden Dissertation ist es, Forschungslücken im Hinblick auf den BFLPE anzugehen, indem die Bandbreite der verwendeten Forschungsdesigns erweitert wird. Konkret bearbeite ich in der vorliegenden Dissertation vier Forschungslücken im Hinblick auf den BFLPE: Mehrere Klassenumgebungen als Bezugsrahmen für die akademische Selbstkonzeptgenese, der Zusammenhang zwischen „grading on a curve“ und dem BFLPE, Trackingeffekte auf das akademische Selbstkonzept und Nachbarschaftseffekte auf das akademische Selbstkonzept. Für jede dieser Forschungslücken werde ich design-basierte Herausforderungen vorheriger Forschung auf konzeptueller Ebene beschreiben und klären, welche Voraussetzungen neue Designs erfüllen müssen.

Neben diesem übergeordneten Ziel verfolgt die vorliegende Dissertation zwei untergeordnete Ziele, die mit jeweils zwei empirischen Studien verfolgt wurden. Das erste untergeordnete Ziel ist die Nutzung umfangreicher Large-Scale Datensätze (Bildungsmonitoring Gesamterhebungsdaten und interdisziplinäre Large-Scale Daten) für eine vertiefte Untersuchung der Mechanismen und der interdisziplinären Integration des BFLPE. Konkret wurden in Studie 1 österreichische Bildungsmonitoringdaten verwendet, um mehrere Klassenumgebungen als Bezugsrahmen für die akademische Selbstkonzeptgenese in Schulsystemen mit Course-by-Course-Tracking zu untersuchen. Studie 2 verwendete interdisziplinäre Large-Scale Daten, um die Nachbarschaft als Bezugsrahmen für die akademische Selbstkonzeptgenese zu untersuchen. Das zweite untergeordnete Forschungsziel besteht darin, natürliche Experimente für eine eingehende Untersuchung der Mechanismen und Implikationen des BFLPE zu nutzen. Zu diesem Zweck verwendete Studie 3 Daten zu einer Notenabschaffungsreform, um den Zusammenhang zwischen der „grading on a curve“ und dem BFLPE zu untersuchen, indem die BFLPEs zwischen nicht benoteten und benoteten SuS verglichen wurden. Darüber hinaus verwendete Studie 4 Daten zu zwei Schulreformen in welchen Leistungsgruppierung abgeschafft wurde, um die BFLPE Vorhersagen bezüglich Tracking zu testen.

Die erste Studie (*Which Class Matters? Juxtaposing Multiple Class Environments as Frames of Reference for Academic Self-Concept Formation*) nutzte Daten der österreichischen Bildungsstandardüberprüfung von 2012 (Schreiner & Breit, 2012), einer Gesamterhebung aller österreichischen SuS der achten Klasse im Bereich Mathematik, welche Informationen über

mehrere Klassenumgebungen von SuS enthält. Dieser umfangreiche Datensatz ermöglichte es, die Bezugsrahmen für die akademische Selbstkonzeptgenese in Schulsystemen mit Course-by-Course-Tracking zu untersuchen. Die SuS wurden in den Kernfächern (Mathematik, Deutsch, Englisch) nach ihren Fähigkeiten gruppiert, besuchten aber in allen anderen Fächern die gleiche Stammklasse mit einem gemischten Leistungsniveau. Nach der Regression des mathematischen Selbstkonzepts auf die Leistungsaggregate in Mathematik auf allen Ebenen war der Matheklassen BFLPE am negativsten. Die BFLPE auf der Stammklassenebene war weniger negativ und der Schul BFLPE war am schwächsten ausgeprägt. Diese Ergebnisse stehen im Einklang mit der Theorie der lokalen Dominanz, die besagt, dass lokale Bezugsrahmen am bedeutsamsten für Selbstevaluationen sind.

Die zweite Studie (*Living in the Big Pond: How Socioeconomic Neighborhood Composition Predicts Students' Academic Self-Concep*t) verwendete Daten der Startkohorte 3 des Nationalen Bildungspanels (NEPS; Blossfeld et al., 2011), einer interdisziplinären Large-Scale Studie, die sowohl Informationen über bildungsbezogene Outcomes von SuS als auch Merkmale bezüglich deren Nachbarschaft enthält. Dieser umfangreiche Datensatz ermöglichte es, die Nachbarschaft als Bezugsrahmen für die akademische Selbstkonzeptgenese zu untersuchen. Vorteilhafte sozioökonomische Nachbarschaftsbedingungen hatten keinen oder einen negativen Einfluss auf das akademische Selbstkonzept der SuS. Die Ergebnisse stehen im vermeintlichen Gegensatz zur soziologischen Nachbarschaftsforschung, die davon ausgeht, das die sozioökonomischen Nachbarschaftsbedingungen ein breites Spektrum von Bildungsoutcomes positiv beeinflussen.

Die dritte Studie (*Can Grades Move the Big Fish? The Consequences of Receiving Report Cards for Frame-of-Reference Effects on Academic Self-Concept*) basierte auf Daten der schwedischen Evaluation Through Follow-Up Studie (ETF; Härnqvist, 2000). Die Daten wurden während einer Reformperiode gesammelt, in der schwedische Kommunen entscheiden konnten, ob sie die Vergabe von Schulnoten in der Grundschule abschaffen oder nicht. Diese Datengrundlage ermöglichte es uns, mit einem natürlichen Experiment nicht benotete und benotete SuS hinsichtlich des BFLPEs zu vergleichen. Wir fanden keine Unterschiede zwischen beiden Schülergruppen bezüglich des BFLPEs. Die Ergebnisse stimmen mit einer evolutionären Perspektive auf soziale Vergleiche überein, welche soziale Vergleichsprozesse als einen unausweichlichen menschlichen Trieb ansieht.

Die vierte Studie (*The Dark Side of Detracking: Mixed-Ability Classrooms Hurt Low-Achievers' Math Motivation*) verwendete Daten der österreichischen

Bildungsstandardüberprüfungen aus den Jahren 2012 und 2017 (Schreiner et al., 2017; Schreiner & Breit, 2012) sowie Daten der Zusatzstudie Thüringen des Nationalen Bildungspanels (NEPS; Blossfeld et al., 2011). In beiden Studien wurden Schülerkohorten vor und nach Detrackingschulreformen gemessen. Somit ermöglichten diese Daten, mithilfe eines Kohortenkontrolldesigns, Schülerkohorten vor und nach der Abschaffung von Leistungsgruppierung zu vergleichen. Der BFLPE prädiziert, dass Detrackingschulreformen das akademische Selbstkonzept von leistungsschwachen SuS verringern, da sich diese Schülergruppe in leistungsheterogenen Klassenverbänden mit leistungsstärkeren SuS in Kontakt kommen. In Übereinstimmung mit dieser Vorhersage fanden wir heraus, dass die Detrackingreformen das akademische Selbstkonzept von leistungsschwachen SuS negativ beeinflussten, während dies bei leistungsstärkeren SuS nicht der Fall war. Unsere Ergebnisse sprechen gegen die Annahme von Detracking Befürwortern, die für die Abschaffung von Leistungsgruppierung plädieren, um die Motivation von leistungsschwachen Schülern zu erhöhen, und weisen auf die praktischen Implikationen des BFLPE hin.

Am Ende der Dissertation werden die Ergebnisse aller vier Studien in einen übergreifenden Forschungskontext eingebettet. Darüber hinaus bewerte ich abschließend, inwiefern die herausgestellten design-basierten Herausforderungen erfolgreich angegangen wurden. Schließlich werden Stärken und Grenzen der vorliegenden Dissertation vorgestellt und Implikationen für die Praxis und die zukünftige Forschung aufgezeigt.

CONTENTS

1 Introduction and Theoretical Background

"I am happy with my life," "I am not satisfied with my wage," "I have a big house," "I have lots of friends," "I am not a good student"—all these statements are evaluations of peoples' selves. These self-evaluations are broadly described as *self-beliefs*[1], namely as the entity of inferences persons have made about themselves (Pajares & Schunk, 2002). Positive self-beliefs are a central construct in positive psychology, which deals with how people can get the best out of life (Seligman & Csikszentmihalyi, 2000). They are referred to "as a hot variable that makes good things happen, facilitating the realization of full human potential in a range of settings" (Marsh & Craven, 2006, p. 137). Empirical evidence on the importance of positive self-beliefs comes from research spanning a variety of disciplines. For instance, studies have found positive self-beliefs to positively predict desirable life outcomes such as physical and mental health (e.g., Orth et al., 2008; Vingilis et al., 1998) and economic prospects (e.g., Trzesniewski et al., 2006). Further research has found that positive self-beliefs contribute to championship performances beyond what can be explained by previous personal best performances (Marsh & Perry, 2005). Other studies have found positive self-beliefs to negatively predict undesirable outcomes such as drug abuse (Richter et al., 1991) and antisocial behavior (Donnellan et al., 2005; Marsh et al.).

A critical point is that self-beliefs typically are in no way based on objective evaluations, but rather on comparisons with a distinct standard, as exemplified by this famous quote by Karl Marx:

> A house may be large or small; as long as the surrounding houses are equally small, it satisfies all social demands for a dwelling. But if a palace rises beside the little house, the little house shrinks into a hut. (as quoted by Lipset, 1960, p. 63)

It is now more than 150 years since Karl Marx made this seemingly paradoxical observation. The social sciences literature has described many other similar paradoxes since then: Why do workers who experience low pay report comparatively high wage satisfaction levels? Why do the more advantaged members of disadvantaged groups engage in protest and rebellion, even though these people are not the most underprivileged members of their group? Why does more wealth not lead to greater happiness? The common answer to all of these questions is that people's reactions to objective circumstances depend on subjective comparisons. With respect

[1] In this section, the term "self-belief" is used as an umbrella term for all constructs that refer to individuals' self-evaluations (e.g., self-esteem, self-concept, self-confidence, self-worth).

to wages, it has been argued that wage satisfaction emerges from comparing one's wages to those of peer workers (e.g., Gardener et al., 2005). Concerning protest and rebellion, relatively advantaged people in disadvantaged groups are most likely to evaluate their situations by making subjective comparisons with even more advantaged individuals (Taylor & Moghaddan, 1994). With respect to wealth and happiness, the evaluation of one's situation heavily depends on comparisons with a social norm. When everyone becomes more prosperous, the comparison norm rises, and people do not become happier, as their self-evaluations with regard to others usually do not change on average (e.g., Easterlin, 1974). Thus, it has been concluded that self-beliefs are determined by relative rank within a distinct reference group rather than an individual's rank within the whole population.

In educational psychology, the relativity of self-beliefs has been the subject of extensive research. Marsh (1987) found equally able students to have lower self-perceived competencies in academic domains—also referred to as academic self-concept—when they are members of high-achieving schools or classrooms. The author interpreted this result by suggesting that students' academic self-concept results from social comparison processes with their school- and classmates and dubbed it the big fish little pond effect (BFLPE). In the last thirty years, the BFLPE has become a widely researched phenomenon (for an overview, see Marsh & Seaton, 2015).

Despite the huge amount of research on the BFLPE, there is still a lack of in-depth understanding of the BFLPE's mechanisms (e.g., To which reference groups do students tend to compare themselves?), implications (e.g., What does the BFLPE mean for the design of educational systems?), and interdisciplinary integration (e.g., How can the BFLPE theory be embedded in other social science disciplines that focus on social comparison processes?). Thus, the present dissertation's overarching aim is to make use of new designs to address these unresolved issues in research on frame-of-reference effects on academic self-concept. The first subordinate aim is to use extensive large-scale (comprehensive educational monitoring data and interdisciplinary large-scale data) data to achieve this goal. Using data from a comprehensive national educational monitoring survey, Study 1 investigated multiple class environments as pivotal frames of reference for academic self-concept formation in systems with course-by-course tracking. Using data from an interdisciplinary large-scale cohort study, Study 2 examined neighborhood effects on academic self-concept. The second subordinate aim was to examine natural experiments. Using data from a school reform in Sweden that abolished formal grades, Study 3 investigated the association between grading on a curve and the BFLPE. Using

data from two detracking school reforms, Study 4 examined the BFLPE's predictions concerning tracking.

The present dissertation is structured as follows: Chapter 1 presents the introduction and theoretical background. It introduces the construct of academic self-concept (Section 1.1), reviews research on the BFLPE (Section 1.2), and describes four unresolved issues and corresponding design-based challenges (Section 1.3). Chapter 2 introduces the dissertation's research aims and the research questions of the four empirical studies. Chapters 3 to 6 present the four empirical studies. Chapter 7 provides a general discussion of the present dissertation and its component studies. It describes the dissertation's contribution to the BFLPE literature (Section 7.1) and makes a final evaluation of the subordinate research aims (Section 7.2). This chapter also deals with strengths and limitations (Section 7.3), practical implications (Section 7.4), implications for future research (Section 7.5), and makes a final conclusion (Section 7.6).

1.1 Academic Self-Concept

In this chapter, I introduce the present dissertation's central psychological construct, academic self-concept, which refers to an individual's self-evaluation in academic domains. In addition to discussing the construct's history, definition, and structure, this chapter includes information on the effects and determinants of academic self-concept.

1.1.1 History, Definition, and Structure

From the very beginning of human history, mankind has been interested in the *self*, which "encompasses the direct feeling each person has of privileged access to his or her own thoughts and feelings and sensations" (Baumeister, 1997, p. 681). Research on the self has a long history, starting with the work of ancient Greek philosophers like Plato and Aristotle (Hattie, 2014). Subsequently, the self had long been widely studied by philosophers, politicians, and other thinkers long before psychology existed (Hattie, 2003). It has been assumed that positive evaluations of the self represent one of the deepest human motives and that such positive self-evaluations build the cornerstone of a successful life (Greenwald, 1988). William James, whose magnum opus *The Principles of Psychology* was published in 1890, is credited as the founder of self-concept research within psychology. James (1890) divided the human self into *I* and *Me*. The latter includes people's evaluations of their selves, also known as *self-concept* or *self-esteem*, which was proposed to result from subjective interpretations of successes and failures. Another milestone in the evolution of modern self-concept research was an article by Shavelson et al. (1976). The authors criticized prior self-concept research in various ways. Concretely, their critique concerned the inconsistent definitions of self-concept, the diversity of instruments used to measure it, and the fact that threats to the endeavor of measuring self-concept, like social desirability, had not been investigated in detail. Due to the aforementioned shortcomings of self-concept research at this time, Shavelson et al. (1976) provided a broad definition of self-concept as persons' self-perceptions that are formed through experience with their environment. As can be seen in Figure 1, Shavelson et al. (1976) proposed a model with a general self-concept factor at the apex of the hierarchy, which influences an academic and several non-academic self-concept factors. The model becomes more and more differentiated at lower levels of the hierarchy. This model is known as the *Shavelson model*.

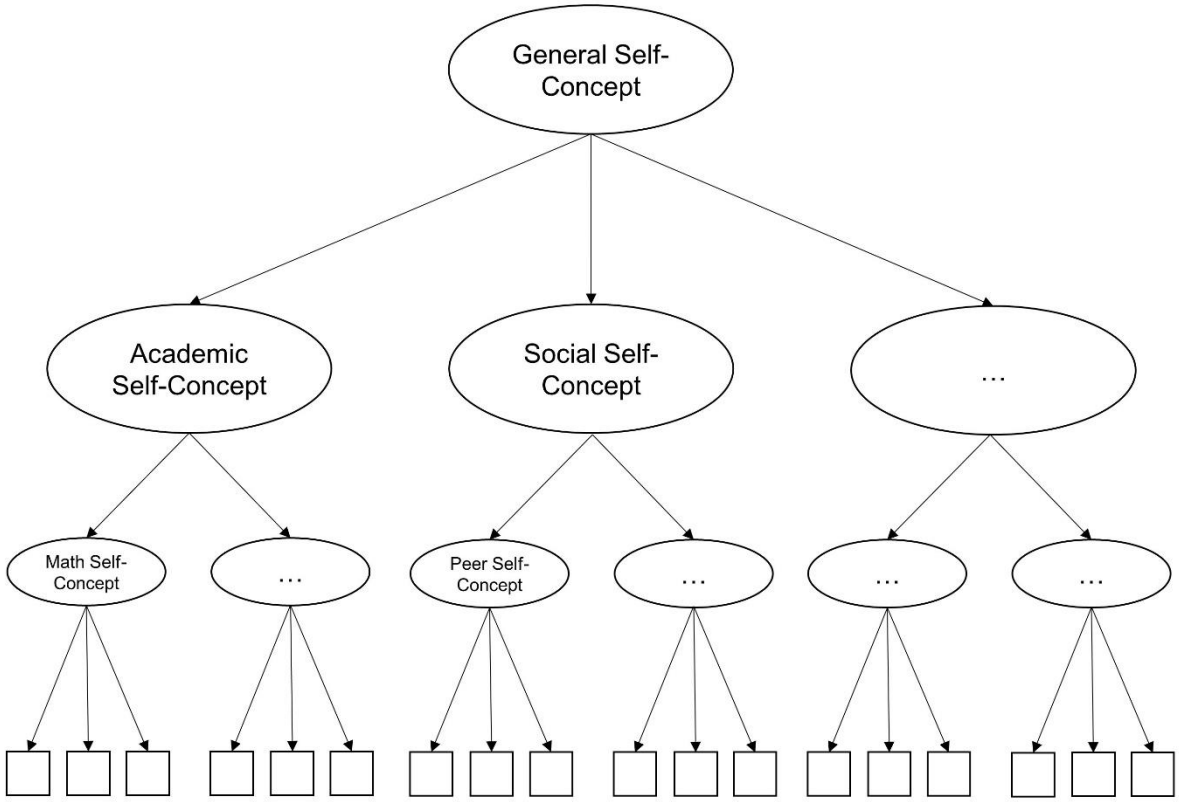

Figure 1. The Shavelson Model.

One of the more differentiated self-concept constructs in the Shavelson model is *academic self-concept*—namely, self-perceived competence in academic domains (Marsh & Martin, 2011). Academic self-concept is domain-specific and can itself be divided into domain-specific facets, such as math self-concept or self-concept in other subjects (Marsh, Martin, et al., 2001). Academic self-concept is typically measured using self-report scales (Trautwein & Möller, 2016). Students are presented with statements such as "Usually, I do well in math" or "I learn things quickly in math" to rate on a Likert scale ranging from *I do not agree* to *I agree* (e.g., Marsh, 1990). From a more general perspective, a vast number of other self-related constructs that also refer to self-perceived competence in academic domains exist (Valentine et al., 2004). First, there is the related construct *expectancies of success* from modern expectancy-value theory (EVT; Eccles et al., 1983), which is one of the major frameworks for studying achievement motivation. EVT divides student motivation into expectancies (*Can I do this?*) and values (*Should I do this?*). Academic self-concept and expectancies of success are typically not empirically distinguishable (Eccles, 2009; Schunk & Pajares, 2005). Thus, academic self-concept has often been used as a proxy measure for expectancies of success in EVT studies (e.g., Simpkins et al., 2012; Wang et al., 2013). Second, there is the related construct of *academic self-efficacy*. Academic self-concept differs from academic self-efficacy in that it is

retrospective rather than prospective (beliefs about *what I have done in the past* vs. *what I can do in the future*) and evaluative rather than descriptive (beliefs about *how well behavior matches personal standards* vs. beliefs about *how well behavior matches external standards*; Marsh et al., 2019). In the present dissertation, I will use the term academic self-belief to represent the broad range of self constructs referring to academic self-perceptions.

Marsh and Shavelson (1985; see also Byrne & Shavelson, 1986; Marsh et al., 1988) proposed dividing Shavelson's academic self-concept factor into two distinct academic self-concept factors, namely mathematical self-concept and verbal self-concept. As can be seen in Figure 2, the *Marsh/Shavelson model* posits that mathematical self-concept influences domain-specific self-concepts in subjects like mathematics, physics, and biology, whereas verbal self-concept influences domain-specific self-concepts in subjects like native language, foreign language, and history.

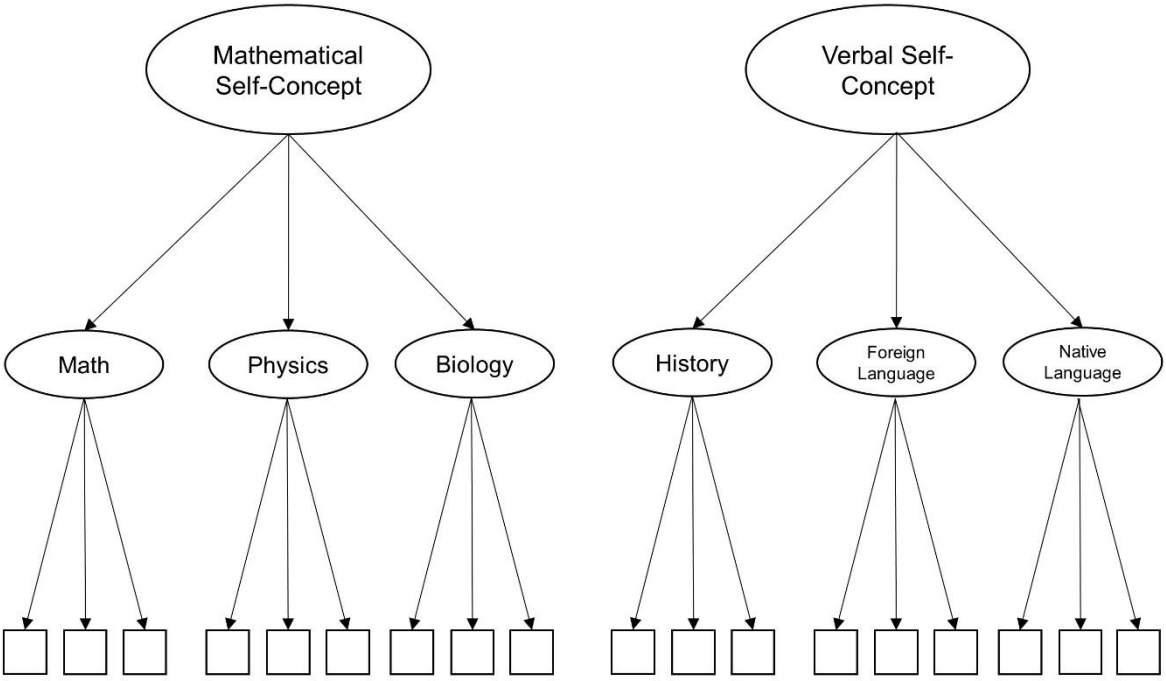

Figure 2. The Marsh/Shavelson Model.

More recently, Brunner et al. (2010) proposed the *nested Marsh/Shavelson model* in which a general self-concept factor influences the manifestations of domain-specific self-concept factors. Note that the discussion on the structure of academic self-concept is not over. For instance, Braun et al. (2020) questioned the reflective nature of the hierarchical model (higher-order self-concept affects lower-order self-concepts) and proposed a formative model in which lower-order self-concepts determine one's higher-order self-concept.

1.1.2 Effects and Determinants

Academic self-concept is regarded as an important motivational factor due to the assumption that self-concept is a “hot variable that makes things happen” (Marsh, 2005, p. 1). One reason for such thinking is the strong association between academic self-concept and academic achievement (Möller et al., 2009). Calsyn and Kenny (1977) distinguish between self-enhancement and skill development approaches to explaining the causal relationship between academic self-concept and academic achievement. Self-enhancement models assume that academic self-concept is the cause of academic achievement, while skill development models assume academic achievement to be the cause of academic self-concept. Marsh and Martin (2011, p. 64) state that "either-or answers to this question are too simplistic" and that empirical evidence from crossed-legged panel models suggests a reciprocal effects model (REM) in which both variables reinforce each other. Recent evidence strongly supports the REM (e.g., Marsh et al., 2018; Preckel et al., 2019). Perhaps the strongest evidence for the validity of the REM stems from Valentine et al. (2004), who reviewed 55 publications investigating the reciprocal effects of academic self-beliefs in their meta-analysis. Valentine et al. (2004) found small effects of academic self-beliefs on overall academic achievement and moderate effects of academic self-beliefs on domain-specific academic achievement, providing strong evidence for the reciprocal effects model. Other research supports these findings (e.g., Huang, 2011; Richardson & Bond, 2012; Steinmayr & Spinath, 2009; Swann et al., 2007). Research investigating the mechanisms behind self-enhancement models is rather scarce (Preckel et al., 2019). In general, it is supposed that the positive effects of academic self-concept on academic achievement are mediated by achievement-related behavior (Marsh et al., 2016). Marsh and Martin (2011) state that effort, persistence, and intrinsic motivation may serve as mediators (see also Trautwein & Möller, 2016). Other studies propose interest (Marsh et al., 2005), academic emotions (Pekrun et al., 2007), or engagement (Wigfield & Eccles, 2000) as potential mediators. Furthermore, the positive effects of academic self-concept on academic achievement may be mediated by performance-enhancing behavior (e.g., increased concentration) in test situations (Eckert et al., 2006). In addition to academic achievement, academic self-concept is assumed to affect academic aspirations and choices. Super (1951) already noted that individuals’ vocational and career choices realize their ideas of themselves—their self-concepts. For instance, adolescents who regard themselves as good writers may decide in favor of a career as a journalist. Moreover, in choosing a vacation or a career, individuals test their self-concepts against reality (Savickas, 2002, 2005). The determinant power of academic self-concept for academic choices is also emphasized by EVT, which links choices to expectancy-

related beliefs. Likewise, self-efficacy theory highlights the determinant power of academic self-beliefs for career choices (Bandura, 1986). In line with these theoretical considerations, Parker et al. (2012) found academic self-concept to be a predictor for university entry. In addition, math self-concept is a predictor to choosing math-related careers in STEM (Parker et al., 2014).

Due to the importance of academic self-concept as a motivational factor, emphasis has been placed on identifying its determinants. One determinant of academic self-concept is gender. Generally, it has been found that girls have a lower academic self-concept in math-related domains and a higher self-concept in language-related domains (e.g., Dai, 2001; Marsh, 1989; Marsh & Yeung, 1998; Nagy et al., 2007). Typically, these gender differences in academic self-concept remain when controlling for domain-specific academic achievement. Thus, it is assumed that gender differences in academic self-concept are to some extent *gender-stereotypical*. Another academic self-concept determinant is age. Multiple studies show that academic self-concept declines over the course of students' educational careers. More specifically, is assumed that academic self-concept declines over the course of elementary education (e.g., Archambault et al., 2010; Spinath & Spinath, 2005; Wigfield & Eccles, 1994) and decreases even further over the course of secondary education (Jacobs et al., 2002; Marsh, 1989). It has been noted that not every student might experience a steady decline in academic self-concept, but that there might rather be several prototypical self-concept trajectories—for instance, students with a moderate decline and students with a rapid decline (Musu-Gillette et al., 2015). Additionally, it is assumed that academic self-concept might become more realistic over time (Harter, 1998). Academic self-concept corresponds to students' actual academic abilities only to a minor extent. For instance, students who fall above the 90th percentile in terms of reading ability might nevertheless evaluate themselves as not very talented readers. Due to this relativity of self-evaluations, a large number of research articles have examined different kinds of comparisons as determinants of academic self-concept (Möller & Trautwein, 2015). One of these comparisons are dimensional comparisons (Marsh, 1986). Dimensional comparisons assume that students compare their achievements in different domains—for instance, from the math and verbal domains—with each other. In other words, students' self-evaluations in one specific domain depend on their achievement in another domain. Thus, students will have a lower math self-concept if they are good in German because they compare their math achievement with their German achievement. Empirical evidence for dimensional comparison processes has been found by regressing academic self-concept in one domain on academic achievement in another (Marsh, Möller, et al., 2014; Möller & Marsh, 2013). Social

comparisons as determinants of academic self-concept are the main topic of this dissertation and will be described in the next section.

1.2 The Big Fish Little Pond Effect (BFLPE)

As mentioned in the previous section, social comparison processes are considered a focal determinant of academic self-concept formation. Evidence for this assumption comes from the finding that equally able students have lower academic self-concepts in high-achieving learning environments—the BFLPE. In this section, I will introduce the foundations of research on the BFLPE and review and systemize previous research into the three research areas *mechanisms, implications,* and *interdisciplinary integration*.

1.2.1 Foundations

Already at the very beginning of psychological research, evaluations and judgments were assumed to be relative. William James, the founder of modern scientific psychology, expressed this notion in his seminal work by stating: “We have the paradox of a man shamed to death because he is only the second pugilist or the second oarsman in the world” (James, 1890, p. 310). Similarly, research on psychophysical judgment emphasizes that evaluations of stimuli—for instance, the size of a square—are relative and strongly depends on other stimuli presented (e.g., very small squares; Parducci, 1965, 1968). Additionally, early social psychologists emphasized the relativity of self-evaluations (Sherif, 1935; Upshaw, 1969). The first empirical evidence for the relativity of self-evaluations likely came from Stouffer et al. (1949), who found that US soldiers in the 1940s were dissatisfied when they felt that other people were unjustly promoted faster than themselves. The idea that humans evaluate themselves via social comparison processes was also noted by Festinger (1957) in his social comparison theory. Festinger (1957, p. 1) stated that “there exists, in the human organism, a drive to evaluate his opinions and his abilities.” Empirical evidence that such frame-of-reference effects also play a role in the educational setting came from sociologists. For example, Davis (1966) found academic achievement to predict the choice of a high-performance career, whereas school-average achievement did not. From this result, he concluded: "It is better to be a big frog in a small pond than a small frog in a big pond" (Davis, 1966, p. 31). Additionally, Soares and Soares (1969) as well as Trowbridge (1972) found disadvantaged students to have a higher self-concept than advantaged students. It was in the 1980s that Marsh and Parker (1984) as well as Marsh (1987) integrated this frame-of-reference logic into educational psychology. The authors found equally able students to have a lower self-concept in high-achieving schools and proposed social comparison processes as driving forces behind this seemingly paradoxical effect. They named this effect the *big fish little pond effect* (*BFLPE*; Figure 3).

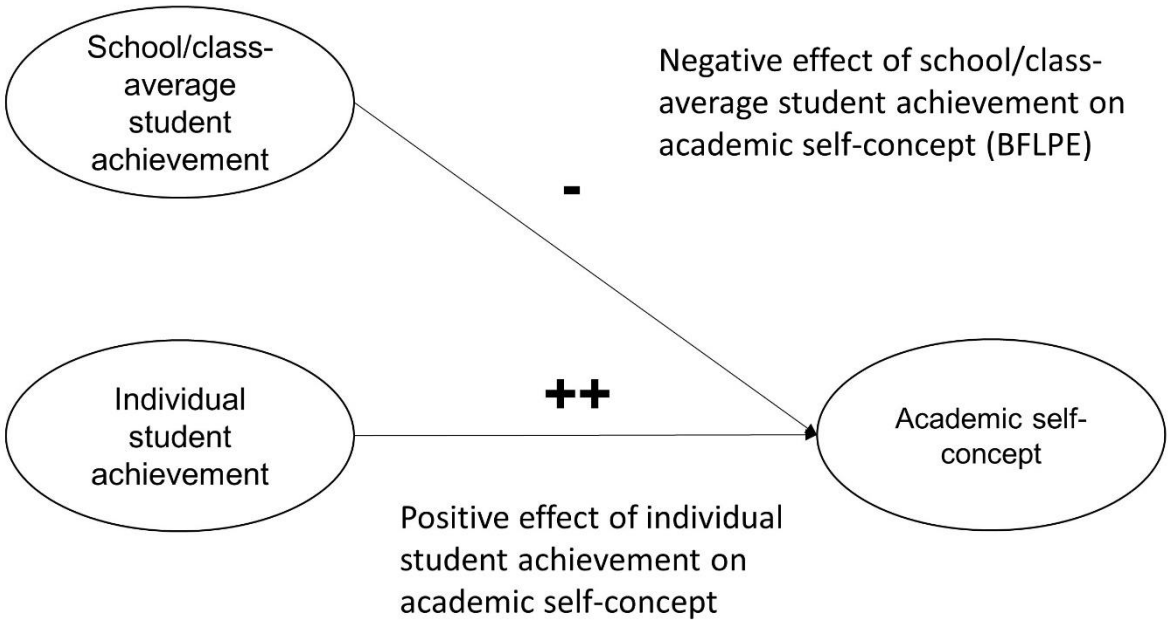

Figure 3. The BFLPE model.

A large share of the more recent research on the BFLPE has investigated the cross-cultural generalizability of the frame-of-reference effect. Using international large-scale assessment data (e.g., PISA or TIMSS), these studies found the BFLPE to generalize remarkably well across cultures, describing the contextual effect as a *panhuman phenomenon* (Seaton et al., 2009; see also Guo et al., 2018; Marsh et al., 2015; Marsh, Abduljabbar, et al., 2014; Marsh & Hau, 2003; Nagengast & Marsh, 2012; Seaton et al., 2010; Wang, 2015; Wang & Bergin, 2017). Educational psychology research has also focused on the question of to what extent frame-of-reference effects also affect other educational outcomes besides academic self-concept. Studies have shown that similar frame-of-reference effects also exist for coursework selection (Marsh, 1991), academic aspirations (Nagengast & Marsh, 2012), academic interest (Trautwein, Lüdtke, Marsh, et al., 2006), school grades (Neumann et al., 2011) and even long-term outcomes such as income (Göllner et al., 2018).

Frame-of-reference effects on academic self-concept—also known as the big fish little pond effect (BFLPE)—have been extensively studied in the past three decades. In the present dissertation, I identify and systemize three major research areas, namely mechanisms, implications (for individual educational pathways and educational systems), and interdisciplinary integration, each of which encompasses several research issues (Figure 4). Generally, it must be noted that this systematization is heuristic in that it aims to structure previous research and that the boundaries between these three research areas are not clear-cut.

Indeed, issues falling within more than one research area may exist. These issues were assigned to the research area with which they were most clearly aligned. By engaging in this systemization, I aim to review research on the BFLPE in a structured way and highlight areas in which more research is needed.

1.2.2 Mechanisms

The fundamental notion of BFLPE theory, which states that students form their academic self-concept through social comparisons with their school- and classmates, is a rather general assumption. Thus, from the beginning of research on the BFLPE, researchers were interested in obtaining a more detailed picture of the mechanisms underlying this contextual effect. Research on mechanisms underlying social science theories is considered important because it makes the higher-level theory “more supple, more accurate, or more general." (Stinchcombe, 1991, p. 367). Specifically, researchers were interested in (a) whether being a member of a high-achieving learning environment also has positive effects (*reflected glory*), (b) whether individuals prefer to compare themselves with certain groups of students (*frames of reference*), and (c) whether there are individual or contextual characteristics that reinforce or attenuate the BFLPE (*moderators*).

Research on the BFLPE argues that membership in a high-achieving learning environment is associated with a lower academic self-concept. The proposed mechanism underlying this finding is social comparison processes in which students evaluate their abilities by using their learning environment as a frame of reference. The respective line of thought is: "There are a lot of students better than I, so I might not be as good a student as I thought". In this line of thinking, students contrast themselves with their peers; thus, the BFLPE has been termed a *contrast effect* (Marsh et al., 2000). On the other hand, psychological research has found that individuals evaluate themselves more positively when they are a member of a high-status group (Felson, 1984; Felson & Reed, 1986). These findings were interpreted as suggesting that people “bask in the reflected glory” of successful group members. Based on these findings, one could theoretically argue that students in selective educational environments might have a higher academic self-concept as a consequence of basking in the reflected glory of others. The respective line of thought is: “I am good enough to be in this selective learning environment, thus I must be a very good student.” Because these proposed mechanisms are in opposite directions, research has sought to evaluate the relative influence of assimilating and contrasting comparison processes. To do so, researchers asked students to rate the prestige of their learning group and added an aggregate measure of these perceptions as an additional predictor to the BFLPE model (e.g., Marsh et al., 2000; Trautwein et al., 2009). These studies

found that high-prestige educational environments positively predict students' academic self-concept while simultaneously creating an even more negative BFLPE. The latter finding was interpreted as indicating that in the conventional BFLPE model—which does not control for prestige—the negative frame-of-reference effect is counterbalanced by a positive assimilation effect. Thus, the authors of these studies concluded that positive assimilation effects exist but are completely absorbed by negative contrast effects. Similar results were found when modeling academic track as an indicator for the prestige of students' learning environments (Chmielewski et al., 2013).

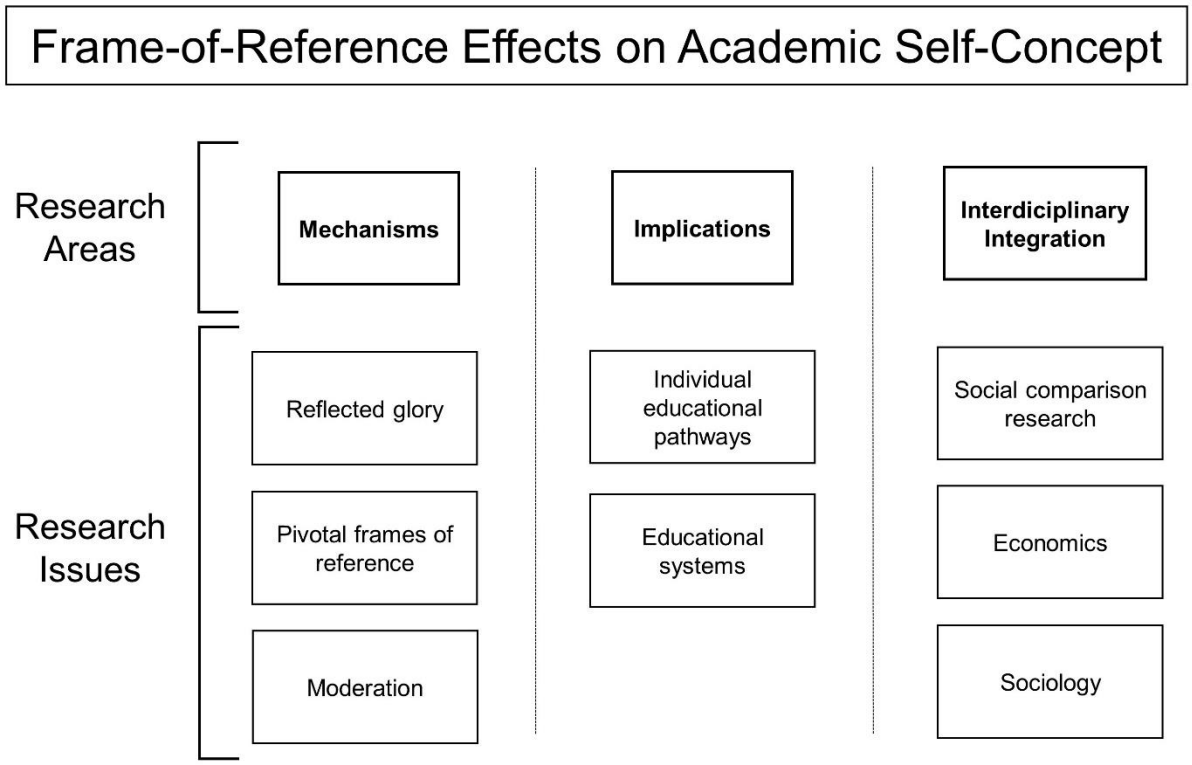

Figure 4. Research Strands on the BFLPE.

Turning to the proposed targets of students' social comparisons, two different approaches have evolved to identifying the pivotal frames of reference for academic self-concept formation. In the first approach, researchers drew upon social comparison theory and the related attribute hypothesis, stating that "given a range of possible persons for comparison, someone who should be close to one's own performance or opinion, given his standing on characteristics related to an predictive of performance or opinion, will be chosen for comparison" (Goethals & Darley, 1977, p. 265). Accordingly, it was proposed that same sex or same ethnicity classmates represent preferred comparison targets. To test this hypothesis, researchers regressed academic self-concept on two distinct achievement aggregates, that of

ingroup classmates (e.g., same sex or same ethnicity) and outgroup ones (different sex or different ethnicity; Fleischmann, 2017; Liem et al., 2013; Thijs et al., 2010). However, in these studies, no sex- or ethnicity-specific frame-of-reference effects on academic self-concept could be found. The second approach to investigating the pivotal frames of reference for academic self-concept formation drew on local dominance theory, which states: "When people have multiple feedback sources, as they often do in their daily lives, the influence of local comparisons dominates and supersedes the influence of general comparisons" (Zell & Alicke, 2010, p. 380). Based on local dominance theory, it was proposed that local scholastic learning environments, such as classrooms, are much more important than distal ones, such as schools. Consequently, academic self-concept was regressed on both school- and class-average achievement (Liem et al., 2013; Marsh, Kuyper, et al., 2014). These studies found that school-average achievement had no effect on academic self-concept when controlling for class achievement and concluded that the classroom is the pivotal frame of reference for academic self-concept formation. Wouters et al. (2012) found that when regressing academic self-concept on friends' and class-average achievement, the latter effect was more pronounced, suggesting that students do not necessarily conduct comparisons with the most proximal environments, but rather the most informative ones.

Third, research in this area has focused on the question of to what extent the BFLPE is moderated by individual and contextual characteristics. Probably the most frequently investigated individual-level moderator variable is individual achievement, addressing the question of whether the BFLPE is equally pronounced for low-achieving and high-achieving students. High-achieving students might experience a smaller BFLPE because they may have less reason to negatively compare themselves to high-achieving classmates, given that they still perform relatively well. Evidence is mixed here. Some studies found the expected positive moderation effect (e.g., Huguet et al., 2009; Marsh & Rowe, 1996; Trautwein et al., 2009), some studies found no interaction effect (e.g., Marsh et al., 1995; Marsh & Hau, 2003), and others even found a negative interaction effect (e.g., Marsh et al., 2007; Marsh, Kuyper, et al., 2014). A second frequently investigated moderator is gender. It has been theorized that females engage in social comparison processes more often than males (Guimond et al., 2007; Wehrens et al., 2010), and thus experience a more pronounced BFLPE. Evidence is also mixed with respect to gender as a potential BFLPE moderator. Some studies found an interaction between gender and the BFLPE (Marsh et al., 2007; Plieninger & Dickhäuser, 2013), whereas others did not (Chanal et al., 2005; Marsh et al., 1995; Marsh, 2016). Jonkmann et al. (2012) investigated whether personality traits moderate the BFLPE and found narcissism to attenuate and

neuroticism to reinforce the BFLPE, whereas the other Big Five characteristics had no effect on the BFLPE. Lüdtke et al. (2005) found teachers' adoption of an individualized frame of reference (i.e., emphasizing improvement upon one's own prior achievement) to not moderate the BFLPE. In addition, Schwabe et al. (2019) reported that students with positive relationships to the teacher experienced smaller BFLPEs than students with negative or average relationships to the teacher. Seaton et al. (2010) investigated 16 individual difference variables (e.g., self-regulation, socioeconomic status) as potential BFLPE moderators using the PISA 2004 dataset. Most interaction effects between these individual characteristics and the contextual effect were small or nonsignificant, and none of the variables was able to eliminate the BFLPE or even change its direction. Thus, the BFLPE has been described as generalizing well across individual and contextual characteristics (Marsh & Seaton, 2015).

1.2.3 Implications

The fundamental BFLPE (i.e., the negative effect of average school or class achievement on academic self-concept when controlling for individual achievement differences) based on large-scale cross-sectional assessment data is a very static phenomenon (e.g., Dai & Rinn, 2008; Wouters et al., 2012). The static BFLPE is important not as a result of immediate practical implications but due to its predictions for educational practice. More specifically, the BFLPE makes predictions about situations in which students are exposed to a change in achievement-related class composition. Such changes in achievement-related class composition can occur, for instance, on the individual level as the result of individual educational pathways (e.g., grade retention, changing academic tracks). However, for the BFLPE model to become practically relevant, these predictions have to be tested in a more dynamic way. For example, the BFLPE predicts that students who transition from elementary schools to more selective secondary schools will experience a drop in their academic self-concept. However, this prediction must be tested with longitudinal designs that follow students across the transition to secondary education. Generally, researching the BFLPE in a static way is much easier due to the availability of large-scale cross-sectional data (e.g., PISA and TIMSS). In contrast, investigating the BFLPE dynamically, e.g., by testing its predictions regarding the transition to selective secondary schools, is much harder because it requires data that is not easily accessible. Generally, the BFLPE makes predictions about the effects of changes in the achievement-related composition of learning environments. These can occur on (a) the individual level as a result of individual educational decisions (*individual educational pathways)* or (b) on the collective level, for instance, due to educational policy reforms (*educational systems*).

With respect to individual educational pathways, the BFLPE predicts that individual educational decisions that lead to changes in the achievement-related composition of students' educational environments—more generally speaking, in students' achievement-related rank—will result in academic self-concept changes. Such individual educational decisions include, for example, age of school entry, grade retention or acceleration, educational transitions, changing academic tracks, course choices, or school transfer. Marsh (2016) found support for a negative year in school effect, in which students who entered school at a relatively young age had a lower self-concept compared to students who entered school relatively late. Similar results were reported by Parker, Marsh, et al. (2019), who also found that early school enrolment decreased academic self-concept. Similarly, grade retention increased academic self-concept, whereas acceleration decreased academic self-concept (Marsh et al., 2017). Dai et al. (2013) measured academic self-concept before and after entering a summer program for gifted students. In contrast to expectations based on the BFLPE, the authors did not find evidence of a decline in academic self-concept (see also Makel et al., 2012). Additionally, Wouters et al. (2012) followed a large sample of students throughout secondary education. In line with the predictions of the static BFLPE model, the authors found that moving to a lower track resulted in a higher academic self-concept. In sum, studies testing the BFLPE dynamically suggest that the BFLPE exerts an effect as a result of changes in individual educational pathways.

With respect to educational systems, the BFLPE predicts that all kinds of decisions that impact the composition of students' environments will also affect students' academic self-concept. Such decisions include all forms of ability grouping practices such as educational tracking in the regular achievement spectrum, gifted education, and special education (mainstreaming/inclusion). The BFLPE model predicts that when students are grouped by ability, low achievers will have a higher academic self-concept because they are exposed to classmates with weaker achievement on average than in the absence of ability grouping. Conversely, the BFLPE model predicts that when students are not ability grouped, low achievers will be exposed to classmates with higher achievement on average, resulting in a decrease in academic self-concept. In other words: "Less tracking means greater heterogeneity in the student body, thus leading to lower-achieving students being confronted with higher-achieving students" (Trautwein & Möller, 2016, p. 206). In line with these predictions of the BFLPE, studies report that low achievers have a higher academic self-concept when tracked as opposed to not being tracked (e.g., Dupriez et al., 2008; Kulik, 1985; Marsh, Köller, et al., 2001). Regarding special education, studies have tested the BFLPE by comparing the academic self-concept of students with intellectual disabilities in special education schools vs.

mainstream schools. In these studies, students with intellectual disabilities had consistently higher academic self-concepts when placed in special education schools as opposed to mainstream schools (Chapman, 1988; Crabtree, 2003; Marsh et al., 2006; Rheinberg & Enstrup, 1978; Tracey et al., 2003). With respect to gifted education, empirical support for the BFLPE stems from research investigating the academic self-concept development of gifted students who joined gifted and talented programs as opposed to matched comparison students. Generally, these studies found that academic self-concept declines among gifted students who enter special gifted and talented programs (Craven et al., 2000; Marsh et al., 1995).

In general, there is much less research on the implications of BFLPE than research on its mechanisms. This is particularly unfortunate given the vital importance of the former research area for decisions regarding individual educational pathways and the design of educational systems.

1.2.4 Interdisciplinary Integration

Educational psychology is not the only discipline to examine how social comparison processes influence individuals' self-evaluations. Other social science disciplines, such as social comparison research, economics, and sociology, have extensively explored this issue. Interdisciplinary integration has been described as "the cognitive process of critically evaluating disciplinary insights and creating common ground among them to construct a more comprehensive understanding" (Repko, 2012, p. 263). Thus, interdisciplinary integration can be considered the final step in scientific theory development that integrates discipline-specific insights into a broader interdisciplinary framework. Interdisciplinary integration of the BFLPE is especially important because other social science disciplines have also investigated social comparison processes, but from different perspectives and with different methodological approaches. Generally speaking, research on the BFLPE can be integrated with (a) social psychological research based on social comparison theory (SCT; *social comparison research*), (b) *sociology*, and (c) *economics*.

The integration of the BFLPE with SCT has a unique history. In two broad essays, Dai (2004) and Dai et al. (2013) critiqued prior research on the BFLPE for neglecting findings from social comparison research and oversimplifying social comparison processes. For instance, the authors argued that social comparisons processes were assumed but not directly observed or measured, causality was not ensured, and effects on other educational outcomes such as achievement were not integrated into the model. In respective responses, Marsh et al. (2004) as well as Marsh, Seaton, et al. (2008) were able to invalidate some of these critiques. However,

they also acknowledged the need to better integrate research on the BFLPE with SCT. Following this discussion, multiple articles were published that were joint ventures between both sides. For instance, Marsh, Trautwein, et al. (2008) simultaneously investigated the importance of generalized (e.g., the classroom) as well as specific others (specific comparison targets, e.g., friends) for academic self-concept formation and found both comparison targets to have substantial negative effects on academic self-concept. Huguet et al. (2009) found that the BFLPE was eliminated after controlling for students' invidious comparisons with their class and interpreted this result as evidence that social comparison processes do indeed drive the BFLPE (see also Marsh, Kuyper, et al., 2014). Marsh et al. (2010) showed that positive assimilation effects, which are typically reported in social comparison research, were overestimated due to uncontrolled measurement error in pretest achievement.

Turning to the intersection of research on the BFLPE and sociology, it is of particular interest that a similar finding to the BFLPE was discovered by sociologists long before educational psychologists began exploring the topic. As previously mentioned, Davis (1966) found that students from high-achieving colleges had lower educational aspirations compared to students from low-achieving colleges in his study *The Campus as a Frog Pond*. As in the BFLPE literature, the author explained this finding with reference to social comparison processes and connected it to the sociological concept of relative deprivation (Stouffer et al., 1949), namely "the judgment that one is worse off compared to some standard accompanied by feelings of anger and resentment" (Smith, Pettigrew, Pippin, & Bialosiewicz, 2012, p. 203). Traditionally, sociologists have noted the negative effects of school-average ability and positive effects of school-average socioeconomic status (Alexander & Eckland, 1975; Becker & Neumann, 2018; Meyer, 1970). Whereas sociological research has examined frame-of-reference effects on a diverse set of outcomes (e.g., aspirations or grades), educational psychology research focuses on the underlying social comparison processes. In an effort to unite the two disciplines, Marsh (1991) proposed academic self-concept as a mediating variable for frame-of-reference effects on academic aspirations. In a similar vein, Nagengast and Marsh (2012) found academic self-concept to mediate frame-of-reference effects on academic aspirations.

The idea that relative position within a given environment affects humans' self-evaluations and thus also behavior also became popular in economics. Duesenberry (1949) used the *relative income hypothesis* to explain the seemingly paradoxical finding that rich people save a higher fraction of their income compared to poor people, but that when economies grow and people become richer, the fraction of national income saved does not change. Duesenberry

(1949) theorized that it is relative but not absolute income that matters (for more information, see McCormick, 2018). People at the bottom of the income distribution will be more often exposed to "better goods" than those they are currently consuming (compared to people at the top), thus leading to higher consumption and less savings. When economies grow and all people become richer, this leads to a stagnation of the fraction of national income saved. Duesenberry's (1949) idea was picked up by Easterlin (1974), who found that the US' positive economic development was not accompanied by gains in happiness. The authors explained their results with the idea that the evaluation of one's situation heavily depends on comparisons with a social norm. When everybody gets richer, the comparison norm rises, and people do not become happier, as their self-evaluations do not change on average. Further studies within economic research on frame-of-reference effects have shown that job satisfaction is negatively related to co-workers' wages (Clark & Oswald, 1996). Brown et al. (2008) also found wage rank to positively affect wage satisfaction and employees' well-being (see also Card et al., 2012). In addition, work in the field of educational economics has investigated the importance of students' achievement-related rank within educational environments. Students' academic rank in elementary school was found to have positive effects on secondary school achievement, high school completion, college enrollment, and even income 19 years later (Denning et al., 2018; Murphy & Weinhardt, 2014). The authors explained their findings with the notion that high-rank students having a higher academic self-concept, which in turn positively affects persistence, effort, and academic achievement. In the tradition of economic research on relative rank, social comparison processes have long been seen as an inevitable reflex within human psychology. Following Darwin, Frank (2011) states that in an evolutionary process, not individuals with genetic mutations that care about absolute rank, but individuals with genetic mutations that care about relative rank are favored. In sum, Frank (2011, p. 26) states:

> To survive and prosper, an individual need not be the strongest, fastest, or smartest animal in the universe. He may be weak, slow, and stupid. What matters is that he be able to compete successfully against members of his species vying for the same resources.

Thus, the Darwinian-economic approach to social comparisons claims that the human brain cares deeply about relative position, as survival in an evolutionary sense depends on relative rather than absolute resources.

Generally, interdisciplinary integration can be considered the most neglected area within research on the BFLPE. Whereas integration with the SCT literature is relatively well developed, integration with economics and sociology is still at an early stage.

1.3 Unresolved Issues and Requirements for New Designs

As outlined in the previous section, an impressive amount of research has investigated the BFLPE, otherwise known as frame-of-reference effects, on academic self-concept with respect to mechanisms, implications, and interdisciplinary integration. As was also clarified above, despite these vital research efforts, a deeper understanding of the BFLPE is still lacking. More specifically, a substantial share of studies in the research area of mechanisms has used data from well-known large-scale assessments that are limited because they rely on one wave of observational data. With respect to implications, a few studies have tested the BFLPE's predictions about individual educational careers. However, less research has tested the BFLPE's predictions concerning educational systems. Concerning interdisciplinary integration, research on the BFLPE has had little contact with the sociological relative deprivation and the economic "rank-order" literature.

One of the reasons why there are still unresolved issues in research on the BFLPE is the homogeneity of research designs (see Dai, 2004; Dai & Rinn, 2008). More specifically, many BFLPE studies are based on education-specific cluster sampling data in which either a random sample of students within schools or a random sample of intact classrooms were drawn. This cross-sectional data was then analyzed with the help of multilevel models in which the school or the classroom represented the higher level. This data is well suited for its primary aim, accurately estimating educational achievement for a diverse set of countries. However, when it comes to research on the BFLPE, such large-scale data is limited in two ways. First, because of cluster sampling procedures, it only contains information on a subset of students' educational environments, and the measures used are typically education-specific. Second, such studies usually rely on one wave of observational data, thus restricting the potential knowledge gains. Therefore, the present dissertation seeks to investigate these unresolved issues with new designs.

In this section, I will introduce four unresolved issues in research on the BFLPE. For each of the unresolved issues, I will describe design-based challenges of previous research on a conceptual level and clarify the requirements for new designs to address these unresolved issues (for an overview, see Table 1).

Table 1

Overview of Unresolved Issues, Design-Based Challenges, and Requirements for New Designs

Unresolved Issues	Design-Based Challenges	Requirements for New Designs
Multiple class environments as frames of reference	High correlation between multiple student environments	Comprehensive data and cross-classified multilevel models
Association between grading on a curve and the BFLPE	Internal validity of traditional mediation models	Random or as-if random variation of grading practices
Tracking effects on academic self-concept	Non-random variation in tracking practices	Random or as-if random variation of tracking practices
Neighborhood effects on academic self-concept	Confounding of neighborhood and school characteristics	Neighborhood-school data and cross-classified multilevel models

1.3.1 Multiple Class Environments as Frames of Reference

Initially, research on social comparison processes from the perspective of *social comparison theory* (*SCT*) addressed the question of to whom students compare themselves when evaluating their academic capabilities by asking students to name classmates whom they preferably compare themselves to. These study designs explicitly instructed students to choose a comparison target (for an overview, see Wood, 1989). This line of research found that students prefer to conduct upward comparisons with students who are similar with respect to basic characteristics such as age, gender, or other attributes potentially related to the outcome under evaluation (for an overview, see Dijkstra et al., 2008). Thus, research based on SCT typically uses the concept of a *specific other* to answer the question of to whom students compare themselves when evaluating their academic abilities. In contrast, educational psychology research on the BFLPE typically addresses this question by regressing academic self-concept on the average achievement of specific reference groups. This approach assumes that students make use of a *generalized other* as an implicit comparison target. Educational psychology studies operationalize the generalized other through the average academic achievement of educational environments such as the school or the classroom. Both approaches have certain drawbacks. The approach used in SCT research depends on correct self-reports of students' comparison targets. The accuracy of these self-reports seems particularly doubtful when considering that social comparison processes typically happen spontaneously and

unconsciously. In the approach used in BFLPE research, social comparison processes are typically hypothesized but not observed.

Design based-challenge. The major design-based challenge of previous studies applying the generalized other approach to answering the question of to whom students compare themselves when evaluating their academic capabilities is that average achievement across several potential frames of reference is highly correlated. Specifically, a student who attends a school composed of high-achieving students typically also attends a classroom in which average achievement is high. Likewise, students from high-ability schools have friends that are academically high-achieving. Due to the high correlations between potential frames of reference for academic self-concept formation, contextual effects in an ordinary two-level model (e.g., students within schools) might result from a *noisy reflection* of some other frame of reference. For instance, a negative effect of school-average achievement on academic self-concept (when controlling for individual achievement) does not necessarily mean that students form their academic self-concept in comparison with their schoolmates. Because school-average achievement is highly correlated with class-average achievement, the comparison might just as well happen with classmates or friends. Empirical evidence on the high correlation of potential student environments stems from Marsh, Kuyper, et al. (2014), for instance. The authors report correlations between school-average achievement and class-average achievement of $r = .81$ for Dutch, $r = .83$ for math, $r = .78$ for English. Similarly, Wouters et al. (2012) report a correlation between friends' average achievement and average class achievement of $r = .74$. Fleischmann (2017) reports a correlation between male class achievement and class achievement of $r = .85$ and female class achievement and class achievement of $r = .83$. On a deeper level, the design-based challenge resulting from the high correlations between average achievement within several potential frames of reference is one of internal validity. Whereas internal validity problems can often be solved using experimental designs, this is practically and ethically impossible when investigating the pivotal frames of reference for academic self-concept formation. For instance, to juxtapose the school and the class as potential frames of reference, one would have to assign students to schools and classes with different average achievement levels. Likewise, juxtaposing class and friends as pivotal frames of reference is impossible because one cannot assign students to different groups of friends. Thus, controlling for other confounding variables seems to be the only feasible way to disentangle which frames of reference are pivotal for academic self-concept formation. On a general level, the design-based challenge posed by the high correlations between several potential frames of reference for academic self-concept formation has been recognized by

previous studies. For instance, Marsh, Kuyper, et al. (2014) conducted a study in which they juxtaposed the school versus classroom as pivotal frames of reference for academic self-concept formation by regressing academic self-concept on both school and class-average achievement (see also Liem et al., 2013). The authors found that the class BFLPE completely absorbed the school BFLPE in that school-average achievement had no predictive power for academic self-concept once class achievement was controlled for. However, a closer investigation of which frames of reference are pivotal for academic self-concept formation is still pending. For instance, in secondary education systems worldwide, students are tracked on a course-by-course basis, and are thus exposed to several class environments (Chmielewski, 2014; Loveless, 2013). To date, no study has investigated the pivotal frames of reference for academic self-concept formation in school systems with course-by-course tracking, in which students are members of multiple class environments. Answering this unresolved issue is of great theoretical and practical relevance, as it contributes to the theory of academic self-concept formation in systems with course-by-course tracking.

Requirements for new designs. Addressing the design-based challenge posed by the high correlation between multiple student environments and investigating multiple class environments as frames of reference for academic self-concept formation in systems with course-by-course tracking is an issue that can only be addressed with new designs. Three requirements have to be met in order to adequately investigate this issue. First, the data must include information on students' multiple educational environments, such as the school and multiple classrooms. Many conventional large-scale data sets only contain school identifiers (e.g., PISA) or school identifiers and classroom identifiers in one specific domain (e.g., TIMSS). Second, ideally, all students from the target student population should be tested to build reliable achievement aggregates on all levels of the data hierarchy. Conventional large-scale studies (e.g., PISA or TIMSS) typically draw random student samples within schools. In such data, measurement error on the individual level will lead to higher-level aggregates presenting more reliable values for lower ones, resulting in biased contextual effects estimates. Additionally, the dataset must be large enough to estimate contextual effects on the higher levels of the hierarchy precisely. Third, the juxtaposition of multiple class environments depends on applying *cross-classified multilevel models* (e.g., Goldstein, 2016) because multiple class environments are not hierarchically nested in systems with course-by-course tracking.

1.3.2 The Association Between Grading on a Curve and the BFLPE

Not just academic self-concept but also teacher-assigned grades are assumed to be subject to frame-of-reference effects (e.g., Hübner et al., 2020; Südkamp & Möller, 2009; Zeidner, 1992). More specifically, it is assumed that teachers assign the best grades to the best students, the worst grades to the worst students, and place the others somewhere in-between, thereby implicitly adjusting school grades to average achievement within a given educational environment (Cizek et al., 1995). This grading practice is referred to as *grading on a curve* or *class-referenced grading* (these terms are used interchangeably in this dissertation). In class-referenced grading, it is rather unlikely for all students to get a good grade because teachers stretch even small differences in achievement between individuals across the grade continuum. Thus, in class-referenced grading, grades are strongly associated with individuals' relative positions in the respective learning environment. Empirical evidence for teachers' tendency to grade on a curve comes from qualitative work (e.g., McMillan et al., 2002) but also empirical studies showing that students' standardized achievement varied across educational environments, whereas this was not the case for grades (e.g., Dompnier et al., 2006). Moreover, regressing teacher-assigned grades on individual and context achievement typically reveals a negative contextual effect in that equally able students have lower grades in high-achieving educational environments (e.g., Neumann et al., 2011; Trautwein, Lüdtke, Köller, et al., 2006). Class-referenced grading is assumed to be one of the determinants of grade provision in most grading systems (Cizek, Fitzgerald, & Rachor, 1995; Dompnier, Pansu, & Bressoux, 2006; Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007). This ubiquity of grading on a curve was illustrated by Neumann et al. (2011), who showed that teachers provided class-referenced grades even in highly standardized central examinations. The association between these two frame-of-reference effects—on academic self-concept and teacher-assigned grades—is a controversial topic. Two contrasting assumptions concerning this association exist. First, the two frame-of-reference effects may be related in that grading on a curve causes or reinforces the BFLPE by providing students with worse grades in high-achieving classes, which in turn negatively affect their academic self-concept. The basic assumption here is that teacher-assigned grades represent easily accessible class-rank information that students base their self-evaluations on. Second, the two frame-of-reference effects might be independent phenomena that emerge from students comparing themselves with each other (for academic self-concept) or teachers comparing their students with each other (for teacher-assigned grades). One approach to clarifying the association between frame-of-reference effects on academic self-concept (BFLPE) and frame-of-reference effects on grades (grading on a curve) is to investigate

the extent to which the BFLPE is mediated by teacher-assigned grades. This was done by applying traditional mediation analysis (e.g., Baron & Kenny, 1986; MacKinnon, 2012) and adding teacher-assigned grades to the BFLPE model as an additional predictor variable, thus controlling the contextual effect for school grades. In this model, the effect of aggregate achievement can be interpreted as the effect of placing equally able students given equal grades in high-achieving classes. The idea behind this mediation approach is that if grading on a curve contributes to the BFLPE, the BFLPE should decline when additionally controlling for teacher-assigned grades. Multiple studies took this approach and found controlling for grades to substantially reduce the BFLPE (e.g., Marsh, 1987; Marsh et al., 2007; Marsh & Rowe, 1996; Trautwein, Lüdtke, Marsh, et al., 2006.). The overall interpretation of this result is that grading on a curve is separate from, but contributes to, the BFLPE.

Design based-challenge. The major design-based challenge of previous studies investigating the association between grading on a curve and the BFLPE is the low internal validity of traditional mediation models. Thus, previous research has not been able to determine whether grading on a curve reinforces the BFLPE of whether the two frame-of-reference effects coexist without being (causally) related to each other. Indeed, the authors of previous studies acknowledged that the results of traditional mediation models were only weak evidence for a causal relationship between the two frame-of-reference effects in the sense that grading on a curve reinforces the BFLPE (e.g., Marsh et al., 2007; Marsh, Kuyper, et al., 2014). One must also critically examine the ambiguous relationship between the two frame-of-reference effects because studies have shown that the BFLPE shrinks in a similar way when controlling for a measure of class rank (Dijkstra et al., 2008; Huguet et al., 2009). This means that the traditional mediation model might be "controlling within-class social comparison processes rather than class marks per se that is the reason why BFLPEs are substantially reduced when class marks are controlled" (Marsh, Kuyper, et al., 2014, p. 61). On a general level, the design-based challenge posed by the traditional mediation approach has been recognized by previous studies, which argued for stronger designs to investigate the association between grading on a curve and the BFLPE (e.g., Marsh, Kuyper, et al., 2014). These studies always called for the disentanglement of the confounding effects of these two processes as a fruitful direction for further research. However, a closer investigation of the association between grading on a curve and the BFLPE is still pending. To date, no study has investigated the association between grading on the curve and the BFLPE with designs other than the traditional mediation approach. Answering this unresolved issue is of great theoretical and practical relevance, as it contributes

to the theory of academic self-concept formation. In addition, it is also of great practical relevance, as grading practices may be one factor that can be used to manipulate the BFLPE.

Requirements for new designs. Addressing the design-based challenge posed by the weak internal validity of traditional mediation models, and thus investigating the association between grading on a curve and the BFLPE, is an issue that can only be addressed with new designs. In addition to new approaches to studying causal mediation (e.g., Imai et al., 2010), one alternative approach would be to investigate the association between grading on a curve and the BFLPE by examining not whether grades mediate the BFLPE but whether the provision of class-referenced grades moderates the BFLPE. Two requirements have to be met in order to adequately investigate the association between grading on a curve and the BFLPE with such a moderation approach. First, feedback practices often vary between (national) educational systems. However, comparing the BFLPEs of non-graded students from Country A with those of graded students from Country B does not help, because one can never rule out the possibility that any potential BFLPE differences result from cultural differences. Thus, the moderation approach depends on variation in grading practices within (national) educational systems. Second, if there are non-graded and graded students within a country, it is most likely that those groups of students are exposed to fundamentally different teaching styles, making it hard to identify the effect grading specifically has on the BFLPE. Thus, this approach requires variation in grading practices within systems and variation that is random or at least *as-if random*[2].

1.3.3 Tracking Effects on Academic Self-Concept

As already mentioned above, the main implication of the BLFPE for educational systems refers to educational tracking[3] practices (Trautwein & Möller, 2016). The BFLPE predicts that low achievers will have a higher self-concept in segregated systems than in comprehensive ones because they are surrounded by relatively low-achieving students. Conversely, the BFLPE indicates that low achievers will have a lower self-concept in comprehensive systems than segregated systems because they are surrounded by relatively high-achieving students. Generally, it can be said that, according to the BFLPE, classroom composition will not impact academic self-concept for the overall student population. For example, the BFLPE predicts that comprehensive grouping will increase low achievers'

[2] According to Dunning (2012), the term *as-if random* variation means that the stimulus under investigation was manipulated is such a way that one can plausibly argue that variation is "as good as random". The term as-if random variation is identical to the term *arguably exogenous* variation used by Murnane and Willett (2010).

[3] In this section, the term "tracking" is broadly defined as grouping students by achievement levels into separate educational environments (Chmielewski, 2014).

academic self-concept and decrease high achievers' self-concept, resulting in no changes overall. Therefore, it has been argued that the social comparison processes underlying the BFLPE are a *zero-sum game* (e.g., Marsh, 1984; Trautwein & Möller, 2016). Investigating the topic of the BFLPE and tracking is not trivial for a variety of reasons. First, in the real world, tracking refers not only to the composition of learning environments and thus social comparison processes but also involves providing students with different curricula (Domina et al., 2019). More specifically, this means that when an educational system is tracked, students are not only grouped according to ability but also provided with curricula with different performance requirements. The fact that tracking involves more than student body composition makes clear that the BFLPE might be an insufficient basis for making precise predictions about how tracking affects academic self-concept. Second, academic self-concept is only one of several desirable educational outcomes. Generally, it is assumed that tracking negatively affects low achievers' academic achievement (Rui, 2009). The fact that academic self-concept is considered a determinant of academic achievement (Valentine et al., 2004) leads to a kind of paradox. How can tracking positively impact low achievers' self-concept and negatively impact students' achievement? This paradox has been the target of intensive debate but seems to have not yet been resolved (Dicke et al., 2018; Stäbler et al., 2017). The fact that academic self-concept is just one out of several desirable educational outcomes means that a comprehensive evaluation of educational tracking practices (e.g., "Should educational systems track or not?", "Is tracking good?") must take a multifaceted approach in which the BFLPE represents only one element. In this context, it has also been argued that educational stratification raises low achievers' academic self-concept. However, low achievers might not be able to translate this higher academic self-concept into educational success because membership in a low track limits their educational pathways (e.g., concerning university entry; Parker, Dicke, et al., 2019). Third, tracking practices vary considerably with regard to the level on which ability grouping is conducted. For example, Chmielewski (2014) differentiates between-school tracking (students from different tracks attend different learning institutions), within-school streaming (students are assigned to a single track for all subjects), and course-by-course tracking (students are tracked separately in one or more subjects). The existence of these different tracking types makes the topic of tracking and academic self-concept even more complicated because the results may differ from one tracking system to another (Chmielewski et al., 2013).

To date, two approaches have been taken to test the BFLPE's predictions regarding tracking. First, studies have compared the academic self-concept of students in higher and lower tracks after controlling for individual achievement differences. The BFLPE predicts that equally

able students have a lower academic self-concept in higher tracks because the average achievement of their educational environments is higher here. In line with these predictions, studies have found that equally able students have a lower self-concept in higher tracks (e.g., Liem et al., 2013; Trautwein, Lüdtke, Marsh, et al., 2006). Generally, it has been suggested that such negative track level effects depend on the specific tracking type. For instance, Chmielewski et al. (2013) found that high-track students in between-school tracking and within-school streaming had a lower self-concept, but high-track students in course-by-course tracking did not. Additionally, these negative track-level effects are assumed to change over time. Liu et al. (2005) found that low-track students had a lower academic self-concept than high-track students directly after being streamed. But these low-track students had a higher academic self-concept compared to high-track students after three years. On a general level, it is essential to note that studies investigating track-level effects on academic self-concept by comparing groups of students from higher and lower tracks can only indirectly test the hypothesis that tracking increases the academic self-concept of low achievers, as these studies did not compare non-tracked with tracked students. Thus, a second approach to testing the BFLPE's predictions regarding tracking is to compare the academic self-concept of non-tracked with that of tracked students. Based on the BFLPE's predictions, one would assume that non-tracked and tracked students should not differ in their academic self-concept overall. However, one would predict that low-achieving non-tracked students have lower self-concept than low-achieving tracked students. Kulik (1985) provided the first data to support the BFLPE's predictions concerning tracking. In their meta-analysis, the author found tracking to positively affect self-evaluations among low achievers, whereas negative effects were found for high achievers (see also Marsh, Köller, et al., 2001). In a more recent study, Dupriez et al. (2008) compared the academic self-concept of students in non-tracked vs. tracked systems. They found that self-concept differences between high and low achievers were especially pronounced in non-tracked systems and interpreted this finding as resulting from contrastive social comparisons within educational environments. Further evidence stems from research comparing the academic self-concept of students with intellectual disabilities placed in special education schools compared to mainstream schools. In these investigations, students with intellectual disabilities consistently had a higher academic self-concept when placed in special education schools than mainstreamed schools (e.g., Chapman, 1988; Crabtree, 2003; Marsh et al., 2006; Rheinberg & Enstrup, 1978; Tracey et al., 2003).

Design based-challenge. The major design-based challenge of previous studies investigating the BFLPE's predictions concerning tracking is the non-random variation in

tracking practices. Generally, both approaches for testing the BFLPE's predictions concerning tracking (comparing the academic self-concepts of high-track vs. low-track students and comparing the academic self-concepts of non-tracked vs. tracked students) are based on non-random variation in tracking practices. In such studies, low-achieving non-tracked students might, for example, report a lower self-concept than low-achieving tracked students not because of tracking but because of individual characteristics related to tracking. Existing research has addressed this design-based challenge only to a minor extent to date. For instance, Hübner et al. (2017) used a cohort-control design that compared two consecutive student cohorts, one before and one after a detracking school reform. They found that girls had a lower academic self-concept after detracking and assumed the BFLPE to be responsible for this finding, as girls are more likely to attend lower tracks. However, Hübner et al. (2017) did not explicitly test the BFLPE's assumption that low achievers have a lower self-concept after detracking. To my knowledge, no study to date has investigated the BFLPE's predictions concerning tracking using random or at least as-if random variation in tracking practices. Consequently, there is a lack of evidence on whether tracking positively impacts low achievers' academic self-concept as predicted by the BFLPE. Addressing this unresolved issue on a theoretical level would test the BFLPE in an internally valid setting. On a practical level, it would have major implications for educational practice.

Requirements for new designs. Addressing the design-based challenge posed by non-random variation of tracking practices, and thus investigating tracking effects on academic self-concept, is an issue that can only be addressed with new designs. One basic requirement has to be met in order to adequately investigate the BFLPE's predictions concerning tracking. Variation in tracking practices has to be random, or at least as-if random. Ideally, this means that research designs should randomly assign students to different groups with respect to educational policy. However, a randomized controlled trial regarding this issue is not feasible for ethical and practical reasons (for an exception, see Duflo et al., 2011). Another way of addressing the design-based challenge posed by the non-random variation of tracking practices is to make use of natural experiments resulting in random variation in tracking practices. For example, such natural experiments can result from policy interventions (Dunning, 2012).

1.3.4 Neighborhood Effects on Academic Self-Concept

A vast amount of research is dedicated to the question of how the neighborhood as a social environment influences people's behavior. Many different disciplines, including economics, sociology, geography and other social sciences, study such *neighborhood effects*

(Dietz, 2002). Generally, a neighborhood effect has been described as "a social interaction that influences the behavior or socioeconomic outcome of an individual" (Dietz, 2002, p. 450). The neighborhood effects literature examines how neighborhoods affect various outcomes, such as emotional problems, sexuality and childbearing, and educational outcomes (Leventhal & Brooks-Gunn, 2000). The typical methodological approach in the neighborhood effects literature is to regress these outcomes on indicators of neighborhood socioeconomic composition, such as average income, socioeconomic status, or employment, while simultaneously controlling for possible confounding variables (e.g., individual social status; Galster, 2008). The debate about the magnitude or even existence of neighborhood effects is highly controversial (Sharkey & Faber, 2014). Generally, the neighborhood effects literature suggests that "good" neighborhoods—in terms of advantageous socioeconomic conditions—positively affect a broad range of outcomes (Wilson, 1987). The proposed mechanisms for these positive effects include social contagion (neighborhood peers change behaviors or attitudes) or social networks (Galster, 2012).

Design based-challenge. One design-based challenge of the neighborhood effects literature is the confounding of neighborhoods and school characteristics (e.g., Jargowsky & Komi, 2011). Due to specific catchment areas, schools' student bodies are typically composed according to residential criteria. This means students in certain schools typically live in particular neighborhoods and students in certain neighborhoods attend particular schools. This design-based challenge implies that studies focusing on only one context are not able to identify the relative importance or overlap between these two contextual effects. Thus, any analysis that omits one of these contexts runs the risk of overstating or misstating the effect of the other (Jargowsky & Komi, 2011). Theoretically, the relation between neighborhood and school contextual effects has been expressed by viewing schools as a mediating factor of neighborhood effects (Arum, 2000; Ferryman et al., 2008; Jencks & Mayer, 1990; Johnson, 2012; Mayer & Jencks, 1989; Sanbonmatsu et al., 2006; Wilson, 1987). In this sense, schools are considered an important pathway or institutional mechanism influencing children and youth. Additionally, schools are viewed as the place where youth interact with their neighborhood peers (Sykes & Musterd, 2010). Despite these assumptions, only limited empirical work has been devoted to understanding the intersection of these two contexts, also known as the school–neighborhood mesosystem (Gaias et al., 2018). Due to the confounding of school and neighborhood characteristics, the joint investigation of school and neighborhood processes has been acknowledged as a research area of particular interest (e.g., Arum, 2000; Johnson, 2012; Sampson et al., 2002). Consequently, an increasing number of studies simultaneously model

students' schools and neighborhoods to disentangle contextual effects on both levels. Many of these studies found neighborhood effects to decrease substantially when controlling for school characteristics (e.g., Dunn, Milliren, et al., 2015; Kauppinen, 2008; Sykes & Musterd, 2010; Wicht & Ludwig-Mayerhofer, 2014). However, contrary findings have also been published (Wodtke & Parbst, 2017). While there is a growing awareness that neighborhood effects that do not control for school characteristics might actually represent school contextual effects, it is also possible that school effects that do not control for the neighborhood are actually neighborhood effects. Concerning frame-of-reference effects on academic self-concept, the design-based challenge posed by the confounding of school and neighborhood characteristics means that to date, there is no clear understanding of how schools and neighborhoods influence academic self-concept. A joint investigation of school and neighborhood composition on academic self-concept is of particular interest, as one would expect negative neighborhood effects, but the neighborhood effects literature only reports positive ones. To my knowledge, no study to date has simultaneously analyzed the effects of specific educational environments and specific neighborhoods on students' academic self-concept. Consequently, there is a lack of evidence on the relation between school and neighborhood effects with respect to academic self-concept formation. Answering this unresolved issue is of great theoretical and practical relevance. On a theoretical level, it can contribute to the integration of BFLPE research and neighborhood effects research. On a practical level, it provides information on whether and how students' motivation relates to neighborhood composition.

Requirements for new designs. Addressing the design-based challenge posed by the confounding of neighborhood and school characteristics, and thus simultaneously analyzing the effects of educational environments and neighborhoods on students' academic self-concept, is an issue that can only be addressed with new designs. Two requirements have to be met in order to adequately investigate this issue. First, data must include information on students' educational environments, such as schools or classrooms, as well as students' neighborhoods. The fact that many studies investigate the effects of each context in isolation (Dunn, Richmond, et al., 2015) stems from the problem that educational assessments which include a broad array of education-related variables typically do not contain neighborhood information (Jargowsky & Komi, 2011). Conversely, neighborhood studies often rely on census data that usually do not contain educational variables. Second, addressing the design-based challenge posed by the confounding of neighborhood and school characteristics requires applying cross-classified multilevel models (Beretvas, 2011; Goldstein, 2016). In contrast to schools and classrooms (all students in a classroom attend the same school), schools/classrooms and neighborhoods are not

hierarchically nested because usually, students from a particular neighborhood do not all attend the same school/classroom and students from a certain school/classroom do not all live in the same neighborhood. Failure to model the cross-classification of the data can lead to bias in estimating standard errors and variance components (e.g., Meyers & Beretvas, 2006).

2 Aims and Research Questions

The previous chapter discussed how academic self-concept is impacted by frame-of-reference effects, thus outlining the importance of educational environments' achievement-related composition for academic self-concept formation. Despite a considerable number of studies investigating the BFLPE in the areas of mechanisms (e.g., To which reference groups do students tend to compare themselves?), implications (e.g., What does the BFLPE mean for the design of educational systems?), and interdisciplinary integration (e.g., How can the BFLPE theory be embedded in other social science disciplines that focus on social comparison processes?), I demonstrated that a deeper understanding of the BFLPE is still lacking. One reason for this is that research on the BFLPE is partly characterized by a lack of heterogeneity in terms of the research designs used. Thus, in the preceding chapter, I introduced unresolved issues in research on the BFLPE, described the underlying design-based challenges of previous research on a conceptual level, and clarified the requirements for new designs. Building upon this foundation, the present dissertation's overarching aim is to tackle these unresolved issues by addressing design-based challenges of research on frame-of-reference effects on academic self-concept using new research designs.

Thereby, the first subordinate aim of this dissertation is to use extensive large-scale data, including comprehensive educational monitoring data and interdisciplinary data, to address unresolved issues in research on the BFLPE. Comprehensive educational monitoring data that seeks to capture all students within a certain population while simultaneously obtaining information on students' multiple class environments allows for tackling the design-based challenge posed by the high correlation between multiple student environments, thus investigating multiple class environments as frames of reference for academic self-concept formation. Previous studies have simultaneously investigated the effects of school- and class-average achievement on academic self-concept. However, in school systems with course-by-course tracking, students are exposed to several non-hierarchically nested classrooms. Interdisciplinary data with information on students' educational outcomes as well as neighborhood characteristics allows for tackling the confounding of neighborhood and school characteristics, thus allowing for the investigation of neighborhood effects on academic self-concept. Previous studies have not simultaneously modeled school and neighborhood characteristics as predictors of academic self-concept formation.

The second subordinate aim of this dissertation is to use natural experiments, namely a school reform abolishing grades and two detracking school reform, to address unresolved issues

in research on the BFLPE. The school reform abolishing grades allows for tackling the design-based challenge posed by the weak internal validity of traditional mediation models by comparing the BFLPEs of non-graded and graded students, thus investigating the association between grading on a curve and the BFLPE. Previous research has found teacher-assigned grades to mediate the BFLPE, suggesting that grading on a curve is associated with the frame-of-reference effect. However, because of traditional mediation models' inability to demonstrate causality, it is also possible that the two frame-of-reference effects on grades and academic self-concept coexist. Two detracking school reforms made it possible to tackle the design-based challenge posed by the non-random variation of tracking practices by comparing student cohorts before and after detracking school reforms, thus investigating tracking effects on academic self-concept. Previous research has found low-achieving students to have lower self-concepts in tracked systems. However, due to the correlational nature of the existing data, this is not very strong evidence.

Figure 4 shows how the two subordinate aims and four studies conducted are embedded in the classification of research on frame-of-reference effects on academic self-concept presented above. Below, the four dissertation studies are described in greater detail.

Study 1 (*Which Class Matters? Juxtaposing Multiple Class Environments as Frames of Reference for Academic Self-Concept Formation*) investigated multiple class environments as potentially pivotal frames of references for academic self-concept formation in school systems with course-by-course tracking. Prior researchers usually focused on the school and one specific class environment as frames of reference for academic self-concept formation. However, many school systems worldwide employ forms of course-by-course tracking, thus exposing students to multiple class environments. Due to the high correlation between multiple student environments, the frame of reference actually used for academic self-concept formation in systems with course-by-course tracking is unclear to date. To address this design-based challenge and tackle this unresolved issue, we used the 2012 Austrian Educational Standard Assessment (BIFIE, 2016; Schreiner & Breit, 2012), a comprehensive survey of all Austrian eight-grade students, and also collected information on multiple class environments. This data enabled us to investigate the pivotal frames of reference for academic self-concept formation in school systems with course-by-course tracking using cross-classified multilevel models. Study 1 tackles the design-based challenge posed by the high correlation between multiple student environments and contributes to the first subordinate aim of using extensive large-scale data to address this design-based challenge.

Study 2 (*Living in the Big Pond: How Socioeconomic Neighborhood Composition Predicts Students' Academic Self-Concept*) separately and simultaneously analyzes the effects of educational environments and neighborhoods on students' academic self-concept. Prior researchers have theorized academic self-concept as affected by educational environments such as the school or the class. However, a large body of research within sociology considers advantageous neighborhood socioeconomic conditions to positively affect students' educational outcomes. Due to the confounding of neighborhood and school characteristics and the fact that neighborhood effects research did not consider academic self-concept as an outcome, it is not clear how socioeconomic neighborhood composition affects academic self-concept. To address this design-based challenge and tackle this unresolved issue, we used Starting Cohort 3 of the German National Educational Panel Study (NEPS; Blossfeld et al., 2011), a longitudinal multi-cohort study that includes measures of students' academic self-concept as well as neighborhood information. Study 3 tackles the design-based challenge of the confounding of neighborhood and school characteristics and contributes to the first subordinate dissertation aim of using interdisciplinary data to address design-based challenges.

Study 3 (*Can Grades Move the Big Fish? The Consequences of Receiving Report Cards for Frame-of-Reference Effects on Academic Self-Concept*) investigates the association between grading on a curve and the BFLPE. Prior researchers have theorized the BFLPE as reinforced by class-referenced grading, as teacher-assigned grades mediate the BFLPE in traditional mediation models. However, as grades are highly correlated with a hypothetical measure of students' class rank, this comes as no surprise. To address this design-based challenge, we used a quasi-experimental design (Shadish et al., 2002) and compared the BFLPEs of non-graded and graded students during a school reform that abolished grades using data from the Swedish Evaluation Through Follow-Up study (ETF; Härnqvist, 2000). Study 3 tackles the design-based challenge posed by the weak internal validity of traditional mediation models and contributes to the second subordinate aim of using natural experiments to address this design-based challenge.

Study 4 (*The Dark Side of Detracking: Mixed-Ability Classrooms Hurt Low-Achievers' Math Motivation*) investigates how detracking—i.e., the abolishment of ability grouping—affects students' academic self-concept. In detracked school systems, low-achieving students are typically exposed to more high-achieving classmates than in tracked ones. Thus, the BFLPE predicts that detracking decreases low achievers' academic self-concept. However, previous research could not directly test this prediction as suitable data with variation in tracking practices was not available. To address this design-based challenge, we made use of cohort-

control designs (Shadish et al., 2002) and compared cohorts before and after detracking school reforms using data from the Austrian National Educational Standard Assessments in 2012 and 2017 (Schreiner et al., 2017; Schreiner & Breit, 2012) as well as the Additional Study Thuringia from the German National Educational Panel Study (NEPS; Blossfeld et al., 2011). Study 4 tackles the design-based challenge posed by the non-random variation of tracking practices and contributes to the second subordinate aim of using natural experiments to address this design-based challenge.

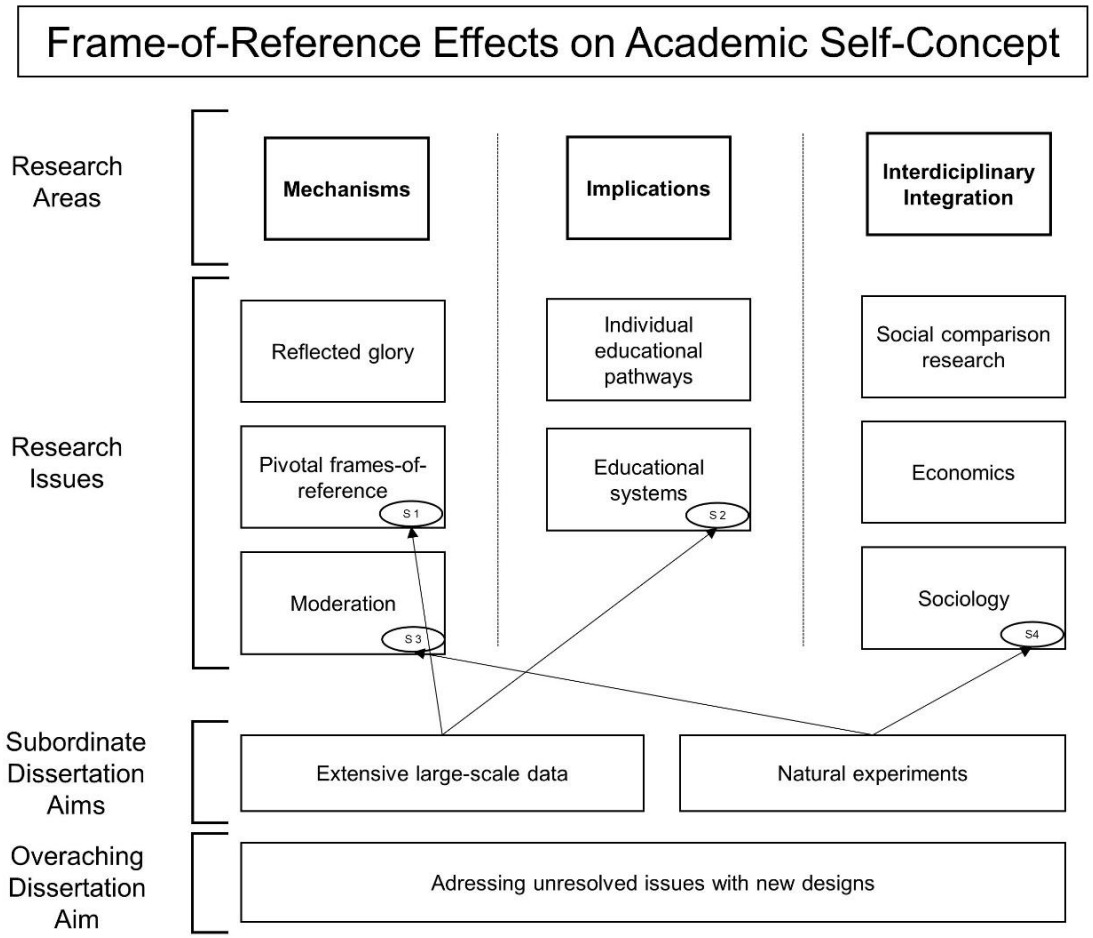

Figure 3. Embedding of the two subordinate aims and the four dissertation studies in the systematization of research on frame-of-reference effects on academic self-concept

3 Study 1: Which Class Matters? Juxtaposing Multiple Class Environments as Frames of Reference for Academic Self-Concept Formation

Fleischmann, M., Hübner, N., Marsh, H. W., Guo, J., Trautwein, U., Nagengast, B. (2020). Which Class Matters? Juxtaposing Multiple Class Environments as Frames of Reference for Academic Self-Concept Formation. Manuscript in revision.

Abstract

Equally able students have lower academic self-concept in high achieving schools or classes, a phenomenon known as the big fish little pond effect (BFLPE). The class (more so than the school) has been shown to be the pivotal frame-of-reference for academic self-concept formation—a local dominance effect. Course-by-course tracked school systems group students according to ability in one or more subjects. However, students remain in the same regular class for the other ones, thus being exposed to several class environments. We evaluated the effects of these multiple frames-of-reference with data from a comprehensive survey that measured the entire population of Austrian eighth-grade students without special educational needs in the domain of mathematics in 2012. General secondary school students (N = 50,208, 48% female, M_{age}= 14.44 years) were in the core subjects (i.e., mathematics, German, and English) grouped according to ability, whereas regular class composition was the same in all other subjects. Using cross-classified multilevel models, we regressed math self-concept on average math achievement of students' school, math class, and regular class. Consistent with the local dominance effect we found the BFLPE on the school level to be weak after controlling for the class levels. We found a stronger BFLPE on the regular class level and the strongest BFLPE on the math class level. Additionally, the math class BFLPE was reduced by controlling math grades, whereas this was not the case for the regular class BFLPE. Our study demonstrates the importance of multiple class environments as frames-of-reference for academic self-concept formation.

Which Class Matters? Juxtaposing Multiple Class Environments as Frames-of-Reference for Academic Self-Concept Formation

Academic self-concept—that is, students' perceptions of their academic abilities (Marsh, Martin, Yeung, & Craven, 2016)—is predicted by the average academic achievement of educational environments (i.e., school or class) when controlling for individual achievement differences (Marsh, 1987; Marsh & Parker, 1984). In particular, equally able students have lower academic self-concept in high achieving schools or classes. This frame-of-reference effect—which has been labeled big fish little pond effect (BFLPE; for an overview, see Marsh & Seaton, 2015)—is assumed to be induced by social comparison processes in which students compare their academic achievement with that of their schoolmates or classmates (Huguet et al., 2009; Marsh, Kuyper, Morin, Parker, & Seaton, 2014).

Typically, research on the BFLPE regresses academic self-concept on aggregated achievement of either the school or the class level. However, average achievement of educational environments is highly correlated, making it difficult to identify the relative strength of both frames-of-reference. To overcome this, Marsh, Kuyper, et al. (2014) employed a three-level approach and found the class to be the pivotal frame-of-reference for academic self-concept formation. In line with the local dominance effect (see Zell & Alicke, 2010), they concluded that local comparison information matters the most for ability self-evaluations.

However, many school systems around the world—for instance, those of many Anglo-Saxon countries—group students according to ability in one or more subjects (often in core subjects like math) while allowing them to remain in the same regular class for the other (untracked) ones. Such an educational practice is referred to as course-by-course tracking (Chmielewski, 2014). Students in course-by-course tracked systems are members of at least two class environments. In such a situation, the question arises to what extent academic self-concept is impacted by the average achievement of multiple class environments. Juxtaposing domain-specific and regular classes as pivotal frames-of-reference for academic self-concept formation is especially interesting as both educational environments are equal regarding their local proximity but differ concerning their domain-specific proximity.

Previous research was not able to juxtapose multiple class environments as frames-of-reference for academic self-concept formation because educational large-scale datasets (e.g., PISA, TIMSS) typically do not include information on multiple class environments. And even if information on multiple class environments were included, the fact that typically only a subsample of students from one school is tested would not allow for calculating reliable

achievement aggregates on all levels of student nesting. In the present study, we were able to overcome this limitation by making use of data coming from the Austrian national educational standard assessment from 2012 (BIFIE, 2016; Schreiner & Breit, 2012), a comprehensive survey that tested the entire population of Austrian eighth-grade students without special educational needs in the domain of mathematics. Austrian general secondary school students were grouped according to ability in math, German, and English classes and attended all other subjects in the same (untracked) regular class. As the complete student population was tested and information on students' math and regular classes was available, this dataset provided us with an unprecedented opportunity for juxtaposing multiple class environments as frames-of-reference for academic self-concept formation.

The Big Fish Little Pond Effect and Its Proposed Mechanisms

The BFLPE, namely the finding that academic self-concept is negatively affected by school- or class-average achievement is supposed to be the result of social comparison processes. Based on classical social comparison theory (Festinger, 1957), there is a human drive for self-evaluation that results in students comparing with school- and classmates, consequently building their academic self-concept based on these comparisons. Thus, an average-ability student would develop a positive self-concept in low achieving educational environments, whereas the opposite would occur in high achieving environments. Several studies support the idea that the BFLPE is driven by social comparison processes (Huguet et al., 2009; Marsh, Kuyper, et al., 2014). Overall, the BFLPE has received strong empirical support. First, the effect is generalizable across cultures (e.g., Marsh, Abduljabbar, et al., 2014; Marsh & Hau, 2003; Nagengast & Marsh, 2012). Additionally, the BFLPE generalizes well over individual characteristics as well as characteristics of educational environments (e.g., Lüdtke, Köller, Marsh, & Trautwein, 2005; Seaton, Marsh, & Craven, 2010; Seaton, Marsh, Yeung, & Craven, 2011). Finally, frame-of-reference effects affect other desirable outcomes, even though the effects are smaller in size, such as academic effort, interest, participation in physical education, and even long-term income (Göllner, Damian, Nagengast, Roberts, & Trautwein, 2018; Marsh, 1991; Trautwein, Gerlach, & Lüdtke, 2008; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006)

Building on work from social psychology (e.g., Cialdini & Richardson, 1980; Snyder, Lassegard, & Ford, 1986), there has been speculation that membership in a high achieving educational environment might also come with benefits in terms of academic self-concept because students "bask in the reflected glory" of successful others. To put this idea to an empirical test, the BFLPE model was extended by including a measure capturing the prestige

of students' learning environments (Marsh, Kong, & Hau, 2000; Trautwein, Lüdtke, Marsh, & Nagy, 2009). Students' perception of their educational environment's status positively affected academic self-concept, and controlling for the prestige of students' educational environment led to an even more negative BFLPE. Membership in a high achieving learning group seems to lead to positive assimilation effects which are counterbalanced by contrastive frame-of-reference effects. Similarly, high within-school track membership (in contrast to low within-school track membership) positively predicts academic self-concept, after controlling for individual and aggregated achievement (Chmielewski, Dumont, & Trautwein, 2013; Trautwein et al., 2006), suggesting assimilation effects. Again, accounting for within-school track led to a more negative BFLPE. In other words, equally able students experience a much stronger academic self-concept decline in high achieving learning environments when track level is kept constant. Generally, there are two different interpretations of such assimilation effects. First, these effects might result from track-level assignment, thus being assimilative track-branding effects (Chmielewski et al., 2013). Second, track level might be an indicator of students' prior academic achievement that is not captured by the standardized achievement measure, thus positively predicting academic self-concept (Marsh et al., 2018). As track level and prior achievement are always correlated, correlational analyses cannot clarify the interpretation of track-level effects on academic self-concept.

Early on, researchers speculated that the BFLPE is driven by grading on a curve or class-referenced grading, which is the tendency of teachers to give the best grades to the best students, the worst grades to the worst students, and place the others somewhere in-between (Neumann, Trautwein, & Nagy, 2011). For instance, Marsh (1987) theorized that the BFLPE might be the consequence of equally able students getting worse grades in high achieving classes, subsequently leading to lower academic self-concept. This idea has been tested by considering teacher-assigned grades as an additional predictor variable in the BFLPE model. In this model, controlling for grades typically leads to a substantial decline in the size of the BFLPE (e.g., Marsh, 1987; Trautwein et al., 2006). Statistically speaking, equally able students with equal grades have only a slightly lower academic self-concept in high achieving learning environments. It is important to note that these results do not ensure a causal relationship between grades and the BFLPE. There is still the possibility that students compare to each other independent of grades.

Juxtaposing the School and the Class as Frames-of-Reference for Academic Self-Concept Formation

Early studies on the BFLPE (e.g., Marsh, 1987, 1991; Marsh & Parker, 1984) most typically used some measure of school-average achievement to predict academic self-concept. By contrast, recent studies more often investigated the effects of class-average achievement on academic self-concept (e.g., Marsh, Abduljabbar, et al., 2014; Marsh, Köller, & Baumert, 2001; Trautwein et al., 2006). The decision of choosing the school or the class as students' learning environment was typically guided by the properties of the data to be analyzed. The school was chosen when data with a student sample from schools were available. The class was chosen when whole classes were drawn. Because these two-level studies modeled either the school or the class (but not both) and because school- and class-average achievement are typically highly correlated, these examinations were not able to investigate what the pivotal frame-of-reference for academic self-concept formation is.

From a social comparison literature perspective, clear expectations exist regarding the relative importance of the school and the class as frames-of-reference for academic self-concept formation. In several experiments, it was shown that local comparison information supersedes the influence of distal comparison information on ability self-evaluations (Alicke, Zell, & Bloom, 2010; Buckingham & Alicke, 2002; Zell & Alicke, 2009). Based on their experimental work, Zell and Alicke (2010) hypothesized the local dominance effect in self-evaluation stating that "when multiple comparison standards are available for self-evaluation, people rely on the most local comparison information while deemphasizing more general, and typically more diagnostic, forms of comparison feedback" (Zell & Alicke, 2010, p. 369).

Marsh, Kuyper, et al. (2014) conducted a study with 15,356 Dutch students nested in 651 classes and 95 schools and juxtaposed the school and the class as frames-of-reference for academic self-concept formation. When modeled separately, school-average achievement, as well as class-average achievement, negatively predicted academic self-concept, controlling for individual achievement. However, when juxtaposed in a joint model, class achievement negatively predicted academic self-concept, whereas school achievement had no effect. These results led Marsh, Kuyper, et al. (2014) to conclude, "This might even suggest that school context really has no effect and its apparent effect is merely a reflection that schools with high school average achievement are made up of classes with high class-average achievement" (p. 58). Similarly, Liem, Marsh, Martin, McInerney, and Yeung (2013), using a sample of 4,461

Singaporean students from 136 classes and 9 schools, found significant class effects but no school effects in a joint model.

Juxtaposing Multiple Class Environments as Frames-of-Reference for Academic Self-Concept Formation

In school systems around the world, including those in the United States, the United Kingdom, Australia, Canada, and New Zealand, many schools track students on a course-by-course basis (Chmielewski, 2014). In contrast to between-school tracking (students from different ability levels attend different schools) and within-school streaming (students from different ability levels attend the same school but are then assigned to different streams for all subjects), course-by-course tracking is defined as "offering courses at varying levels of difficulty in one or more subjects within a school" (Chmielewski, 2014, p. 293). In the following, we will sometimes—for reasons of simplicity—refer to course-by-course tracking as "tracking".

In course-by-course tracked systems students are, depending on the subject, assigned to different ability tracks that in turn are taught in separate classrooms. These systems usually do not assign students to ability tracks in all of the subjects. For instance, Loveless (2013) showed for the United States that course-by-course tracking in math occurs much more frequently compared with language, science, or history. The fact that students in course-by-course tracked systems are not ability tracked in all subjects typically leads to students attending non-tracked subjects in the same regular class.

In course-by-course tracked school systems, in which students belong to several class environments, students can form their academic self-concept in a certain domain, for example, in math, by comparisons with classmates from their domain-specific class (e.g., math class) and their regular class. In relation to the local dominance effect, both classes are local as students are directly exposed to classmates from both classes in daily teaching lessons. However, they differ concerning their domain-specific proximity. Thus—according to the local dominance theory—one would expect the domain-specific class to be the pivotal frame-of-reference for academic self-concept formation in that domain.

In course-by-course tracked systems, the question is not only to which class environments students compare but also how respective comparison processes might differ regarding assimilation effects. Based on previous research, one would expect that controlling for domain-specific track level should increase the BFLPE on the domain-specific class level because it controls for assimilation (Chmielewski et al., 2013). For instance, students in high

achieving math classes are more likely to be high math track members, resulting in a confounding of contrast and assimilation. However, one would not expect that controlling domain-specific track level will increase a BFLPE on the regular class level as students in high achieving regular classes are not expected to be more likely to be high math track members.

In course-by-course tracked systems, the question is not only to which class environments students compare themselves but also how respective comparison processes might differ regarding grading on a curve. Whereas previous research found frame-of-reference effects of domain-specific class environments to be mediated by grades and interpreted this result as a consequence of class-referenced grading stimulating the BFLPE (Marsh, 1987; Trautwein et al., 2006), no study exists that has investigated if this is the case for domain-unrelated class environments. Investigating this question is especially important as it contributes to a better understanding of the mechanisms of frame-of-reference effects on different class levels. Based on previous research, one would expect that controlling for domain-specific grades will decrease the BFLPE on the domain-specific class level because it controls for the teachers' tendency to conduct class-referenced grading. For instance, students in high achieving math classes are provided with worse math grades resulting in confounding of the BFLPE and grading on a curve. However, one would not expect that controlling domain-specific grades will increase the BFLPE on the regular class level as students in high achieving regular classes are not expected to be provided with worse grades.

To date, research focused on only one class environment—in most cases the domain-specific class environment—as the frame-of-reference for academic self-concept formation. To our knowledge, no study has juxtaposed several class environments as frames-of-reference for academic self-concept formation. One reason for that is that educational large-scale datasets usually do not contain information about multiple class memberships. Another reason for the scarce research on this issue is that it relies on survey designs that test all students within sampled schools. Not testing complete schools will lead to differential sampling rates for different classes that will in turn result in differences in the reliability of aggregates and biased estimates.

The juxtaposition of multiple class environments as frames-of-reference for academic self-concept formation has high theoretical and practical relevance. Regarding the former, it captures the full complexity of academic self-concept formation in course-by-course tracked systems—an issue that previous research neglected. Regarding the latter, disentangling

contextual effects of multiple class environments comes with implications for the composition of learning environments.

The Present Study

The present study is based on data from the Austrian national educational standard assessment in 2012 (BIFIE, 2016; Schreiner & Breit, 2012), which measured all Austrian eighth-grade students in the domain of math. Austrian general secondary school students were assigned to one of three tracks (low, medium, high) in the core subjects of mathematics, German, and English, based on teachers' subjective impression of students' achievement. Students from the different tracks were usually taught in separate classrooms according to curricula that differed in performance requirements. As there might have been students who were good in all three core subjects, the class composition of core subjects might have been more or less similar. Secondary school students attended all other subjects (e.g., history, geography, biology, chemistry, physics, music, domestic education) in the same regular class that was not grouped according to ability. Thus, in our multilevel data, students (level 1) were nested in the cross-classification between math classes (level 2a) and regular classes (level 2b) that were nested within schools (level 3; see Figure 1 for a graphical description of the data structure; a more detailed explanation of the complex data structure can be found in the Data section).

As the Austrian national educational standard assessment in 2012 (BIFIE, 2016; Schreiner & Breit, 2012) identified students' school, math class, and regular class and additionally measured all students—what enabled us to build reliable math achievement aggregates on all levels of the data hierarchy—these data were perfectly suited for juxtaposing multiple class environments as frames-of-reference for academic self-concept formation, thus filling the research gap concerning the pivotal frames-of-reference for academic self-concept formation in course-by-course tracked systems. In order to take a closer look at the different mechanisms of the level-specific frame-of-reference effects, we were also interested in how additionally modeling math track and math grades affected the different contextual effects. In detail, we hypothesized the following:

Hypothesis 1 (H1; see also Figure 2a): When considered separately, each of the three math achievement aggregates (school, math class, and regular class math achievement) is expected to have a negative effect on math self-concept, controlling for individual math achievement. Thus, we expected to find a school, a math class, and a regular class BFLPE.

Hypothesis 2 (H2; see also Figure 2b): When all three math achievement aggregates are considered together, controlling for individual math achievement, we expected the math class BFLPE to be more negative than the regular class BFLPE which we in turn expected to be more negative than the school BFLPE.

Hypothesis 3 (H3; see also Figure 2c): When additionally modeling track level, we expected it to contribute positively to math self-concept and result in a more negative math class BFLPE. However, we did not expect it to substantially change the regular class BFLPE. In an exploratory endeavor, we were also interested in whether effects differed between pure and mixed math classes (i.e., math classes with students from the same math track [pure] vs. math classes with students from different math tracks [mixed]).

Hypothesis 4 (H4; see also Figure 2d): In a preliminary analysis, we were interested in whether grades are impacted by frame-of-reference effects. When additionally modeling grades in the BFLPE model, we expected it to contribute positively to math self-concept and result in a less negative math class BFLPE. However, we did not expect it to substantially change the regular class BFLPE.

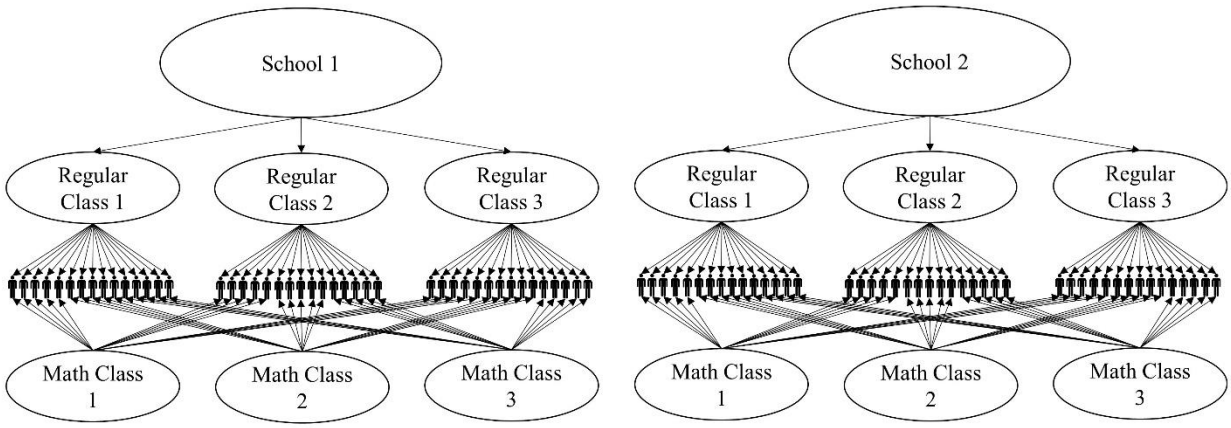

Figure 1. Graphical illustration of the cross-classified data structure (exemplary for two schools).

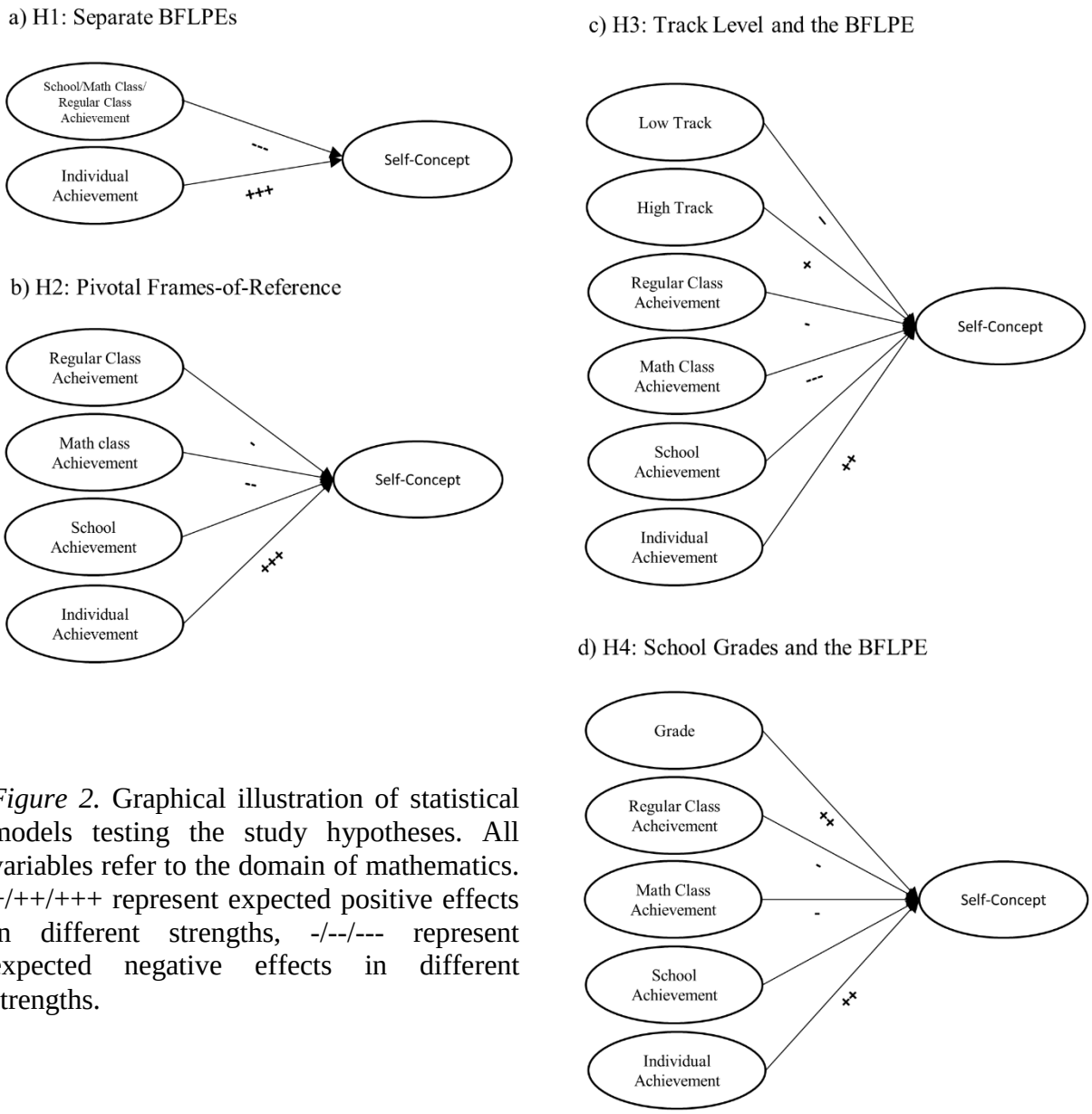

Figure 2. Graphical illustration of statistical models testing the study hypotheses. All variables refer to the domain of mathematics. +/++/+++ represent expected positive effects in different strengths, -/--/--- represent expected negative effects in different strengths.

Method

The Austrian Educational System

In Austria, children attend primary school from Grades 1 to 4, then attend secondary school from Grade 5 onward (for a detailed description of the Austrian school system, see Bruneforth, Chabera, Vogtenhuber, & Lassnigg, 2016). Depending on their primary school achievement, students attend either (a) academic or (b) general secondary school. Academic secondary school provides students with deepened general knowledge and requirements for a transition to university. In the school year 2011-2012—in which eighth-grade student data for the present study were collected—about 33% of all Austrian eighth-grade students attended this school type. In contrast, general secondary school prepares students for vocational training or the transition to higher education. In the school year 2011-2012, about 67% of all Austrian eighth-grade students attended this school type (Schreiner & Breit, 2012). In the present study, we focus on general secondary school students only.

Students attending general secondary school were assigned to one of three tracks (low, medium, high) in three core subjects including mathematics. Generally, students from different tracks were provided with different curricula that differed concerning the topics to be addressed as well as the depth with which the topics were treated. However, in the end, it was left to the teacher to decide how to design the curriculum. In math classes with students from different math tracks, it was also left to the teacher to decide how to deal with math class heterogeneity in terms of track level. Some teachers differentiated their teaching by instructing students from one track while students from the other tracks worked on their own. Some teachers provided very much the same classroom instruction to students from different tracks, however, provided tests with varying degrees of difficulty according to students' track level. Other teachers provided the same tests to students from varying tracks but applied different grading schemes. Note that beginning in the school year 2012-2013, course-by-course tracking was successively abolished and does not exist anymore today (Eder, Altrichter, Hofmann, & Weber, 2015).

Data

In Austria, the Federal Institute for Educational Research, Innovation, and Development of the Austrian School System (BIFIE) conducts national educational monitoring. The examinations of Austria's educational standards are conducted as comprehensive surveys, aiming at measuring all Austrian students attending the fourth or the eighth grade without special educational needs. The national educational standard assessment from 2012 (BIFIE, 2016; Schreiner & Breit, 2012), which is the database for the present study, was conducted in

May 2012. The assessment was aimed at testing all Austrian eighth-grade students without special educational needs in the domain of mathematics. About 4% of the students could not be tested, mostly due to absence at the main and alternative testing dates. The Austrian national educational standard assessment is prescribed by law and does not require the consent of students or parents. Data access was approved by the BIFIE and required consent to data protection regulations.

Our sample included 50,208 students from 1,078 general secondary schools, 3,449 math classes, and 2,729 regular classes. On average, there were M = 3.20 (SD = 1.11) math classes and M = 2.53 (SD = 0.96) regular classes per school with M = 14.56 (SD = 5.32) students per math class and M = 18.40 (SD = 4.10) students per regular class. The math classes were on average smaller than regular classes because schools that contained only two regular classes split the student body into three math classes according to the three track levels. As noted above, math classes typically contained students from one and the same math track. However, in small schools, math classes might also have contained students from different math tracks. Generally, 73% of all math classes were composed of students from one and the same math track. In the subsample of students from mixed math classes, we found that every student attends his math class with M = 12.54 (SD = 6.54) other students from his regular class, indicating a moderate overlap between both class environments for students in mixed math classes. Generally, students spent about 15% of the weekly lesson time in each of the three core subjects (in total 45%) and the other 55% in their regular class.

Instruments

Math self-concept. Math self-concept (MSC) was assessed using four items (i.e., *Usually I am good in mathematics; Mathematics is harder for me than for many of my classmates; I am just not good in mathematics; I learn quickly in mathematics*), which were answered on a 4-point Likert scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*; BIFIE, 2012). For subsequent analyses, a mean score comprising these items was constructed (at least two items had to be completed for mean score calculation; α = .85. Table 1 contains descriptives for all model variables. Average MSC in our sample was M = 2.97 (SD = 0.76). Most of the MSC variation was located on the individual level ($\hat{\rho}_{ind}$ = .95). Math class variation in MSC was lower ($\hat{\rho}_{mcl}$ = .05), and variability on the regular class level ($\hat{\rho}_{rcl}$ < .01) and the school level ($\hat{\rho}_{sch}$ = < .01) was even lower. Table S1 in the online supplemental material

presents the descriptives for each math track separately. Students in the high track (M = 3.16, SD = 0.70) had higher MSC than those in the medium track (M = 2.91, SD = 0.73) and those in the low track (M = 2.55, SD = 0.79).

Math achievement. Math achievement (MACH) was measured using a math competencies test that was based on the competency model of the Austrian educational standards (Schreiner & Breit, 2012). The competency model of the Austrian educational standards—similar to the PISA concept of mathematical literacy (OECD, 2017a)—focuses on the mastery of processes, the understanding of concepts, and the ability to deal with different everyday situations and problems within a competence area on the basis of sustainably networked knowledge. The test lasted about 90 minutes and was delivered by means of a multi-matrix design that contained several test booklets. Students completed approximately 48 items, mostly being presented in a multiple-choice format. There also existed a limited amount of half-open and open item formats.

The BIFIE provides ten plausible values (PVs) that represent the likely distribution of a person's ability (von Davier, Gonzalez, & Mislevy, 2009; Wu, 2005). Large-scale assessment studies typically use PVs because such a procedure allows taking into account the uncertainty of person parameter estimation, thus allowing for correctly estimating associations with other variables. However, due to the multi-matrix design, it was not possible to calculate marginal reliabilities for the PVs. Thus, we calculated an alternative reliability coefficient as it is used in PISA, deducting the within-person PV variance proportion from one. A reliability coefficient close to one indicates that PVs vary within individuals only to a small extent, thus pointing to high measurement accuracy (Adams, 2005; OECD, 2017b). This reliability coefficient was 0.91. MACH showed high variation on the individual level ($\hat{\rho}_{ind}$ = .41) and the math class level ($\hat{\rho}_{mcl}$ = .52), whereas variability on the regular class level ($\hat{\rho}_{rcl}$ = .01) and the school level ($\hat{\rho}_{sch}$ = .06) was lower. This finding empirically underlines the group assignment mechanism. Average MACH was M = 504.24 (SD = 86.85). MACH correlated with MSC by r = .39. Students from the high math track (M = 564.77, SD = 69.50) had higher MACH as compared to students from the medium track (M = 477.38, SD = 59.10) and the low track (M = 414.02, SD = 56.06).

Math grade. Students self-reported their math grade (MGRA) in the last half-year report card. As students were measured in May 2012 and report cards were provided in February 2012, this kind of performance feedback was still relatively current. Generally, it can be assumed that self-reported grades provide a reliable measure of actual grades (Sticca et al.,

2017). In Austria, grades are given on a scale from 1 to 5 with 1 representing the best grade. For subsequent analyses, we inverted the grade variable so that higher grades reflect higher achievement. MGRA mainly varied on the individual level ($\hat{\rho}_{ind}$ = .88) with variance proportions on the math class level ($\hat{\rho}_{mcl}$ = .08), the regular class level ($\hat{\rho}_{rcla}$ = .01), and the school level ($\hat{\rho}_{sch}$ = .02) being substantially smaller. Average MGRA was *M* = 3.19 (*SD* = 0.94). The correlation between MGRA and MACH was *r* = .39.

Math track. Students self-reported the math track they were associated with. Math track can be regarded as a level 1 variable as there where some math classes that contained students from several math tracks (also see above). Generally, track assignment was mainly based on teachers' subjective impression of their students' achievement. We created two dummy variables for the low and the high math tracks with medium math track representing the reference category.

Statistical Analyses

We applied multilevel linear regression analyses (Hox, Moerbeek, & van de Schoot, 2017; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) using the statistical computing software R (R Core Team, 2008) and the package lme4 (Bates, Mächler, Bolker, & Walker, 2015). Generally, all analyses were run with 10 datasets that differed concerning the achievement variables (10 plausible values are provided by the BIFIE). Results then were pooled by Rubin's (1987) rules using the lmer_pool function drawn from the package miceadds (Robitzsch, Grund, & Henke, 2018).

We addressed our research question by calculating multilevel models in which we regressed math self-concept on math achievement aggregates on all levels of student nesting. In these models, level 1 variables were standardized and all three achievement aggregates (math achievement for level 3 school, level 2a math class, and level 2b regular class) were calculated based on the standardized measure, but not re-standardized. As a result, all math achievement variables are in the same metric, namely standard deviations of individual math achievement, making coefficients comparable across levels and models. By grand mean centering of level 1 variables, respective higher level effects can be interpreted as effects of the higher level aggregates, controlling for individual variables, also referred to as contextual effects (Enders & Tofighi, 2007).

For juxtaposing multiple class environments as frames-of-reference for academic self-concept formation, we regressed math self-concept on individual math achievement and math achievement aggregates at the school, the math class, and the regular class levels:

$$Self-concept_{i(j,k)l} = \gamma_{000} + \gamma_{100} \cdot achievement_{i(j,k)l} + \gamma_{001} \cdot school\ achievement_l + \gamma_{010} \cdot math\ class\ \ achievement_j + \gamma_{020} \cdot regular\ class\ achievement_k + w_{0l} + v_{0j} + u_{0k} + e_{i(j,k)l}$$

In this model γ_{000} is the predicted self-concept value for a student with average achievement in educational environments (school, regular class, math class) with average achievement. γ_{001}, γ_{010}, and γ_{020} can be interpreted as the BFLPEs on the school, the math class, and the regular class levels, respectively. w_{0l}, v_{0j}, u_{0k}, are random school, regular class, and math class effects and $e_{i(j,k)l}$ is the residual term.

We ran all our statistical models using a complete case analysis approach (also known as "listwise deletion"). Thus, cases that had missing values on at least one model variable were excluded. The procedure resulted in exclusion rates between 1 and 8%, depending on the statistical model. Research on missing data suggests that when the loss of cases is small—like it was in our study—a complete case analysis will result in negligible parameter bias (Graham, 2009).

Results

Separate BFLPEs (H1)

H1 was "When considered separately, each of the three math achievement aggregates (school, math class, and regular class math achievement) is expected to have a negative effect on math self-concept, controlling for individual math achievement". As shown in Table 2 and consistent with our hypothesis, we found BFLPEs on the school (Model 1; b = -.42, p < .001, 95% CI = [-.45, -.39]), the math class (Model 2: b = -.43, p < .001, CI = [-.45, -.41]), and regular class levels (Model 3; b = -37, p < .001, CI = [-.39, -.35]) when modeled separately. However, as the math achievement aggregates on the different levels are highly correlated, these models do not provide a good basis for evaluating the relative importance of the different frames-of-reference for academic self-concept formation.

Pivotal Frames-of-Reference (H2)

In order to reveal the pivotal frames-of-reference for academic self-concept formation, we examined math achievement aggregates on all levels of student nesting together in the same model (Model 4; see Table 2). H2 was "When all three math achievement aggregates are considered together, controlling for individual math achievement, we expected the math class BFLPE to be more negative than the regular class BFLPE which we in turn expected to be more negative than the school BFLPE". In line with H2 we found the math class BFLPE to be b = -

.37, $p < .001$, 95% CI = [-.39, -.35], whereas the regular class BFLPE was $b = -.11$, $p < .001$, CI = [-.15, -.08] and the school BFLPE was $b = -.06$, $p = .004$, CI = [-.10, -.02]. Thus, equally able students in equally able schools and regular classes had much lower self-concept in high achieving math classes. Additionally, these results indicate that the regular class BFLPE was more negative than the school BFLPE. Equally able students in equally able schools and math classes had lower self-concept in high achieving regular classes, but equally able students in equally able math and regular classes had only a little lower self-concept in high achieving schools. These results are in line with local dominance theory as the size of the level-specific BFLPEs differs as a function of the proximity of respective learning environments.

Track Level and the BFLPE (H3)

Next, we modeled math track as an additional predictor variable (Model 5; see Table 3). We found high math track students to have more positive self-concept as opposed to medium track students ($b = .19$, $p < .001$, 95% CI = [.17, .22]). Conversely low track students had lower self-concept ($b = -.32$, $p < .001$, CI = [-.35, -.30]). In line with H3, the math class BFLPE changed (from $b = -.37$ in Model 4) to $b = -.53$, $p < .001$, CI = [-.56, .51]. Thus, equally able students in equal math tracks experienced a more severe math class BFLPE. Additionally we found no such substantial changes for BFLPEs associated with either the school level ($b = .05$, $p = .024$, CI = [.01, .09]) or the regular class level ($b = -.09$, $p < .001$, CI = [-.13, -.06]).

To check if effects differed between students from pure and mixed math classes, we calculated an additional set of analyses in which we included all interactions between a “mixed” dummy variable and the model variables (see Table 4). Mixed math classes included classes with students from different math tracks whereas pure math classes included classes with students from the same math track. Generally, we did not find differences in the frame-of-reference effects between students from mixed and pure math classes. However, we indeed found differences in the track-level effects. Students in mixed math classes experienced more negative track-level effects from the low track as indicated by the negative interaction between the low track and the pureness dummy ($b = -.19$, $p < .001$). Additionally, students in mixed math classes experienced a more positive track-level effect from the high track as indicated by the positive interaction between the high track and the pureness dummy ($b = .14$, $p < .001$).

School Grades and the BFLPE (H4)

To investigate the frames-of-reference for grade provision, we regressed grades on achievement aggregates on all levels of student nesting (see Table 5). As expected, we found the math class average achievement effect to be most pronounced ($b = -.30$, $p < .001$). We found

the school average achievement effect to be less negative (b = -.11, p < .001) and the regular class effect to be positive (b =.07, p < .001).

Following this, we modeled math grades as an additional predictor of academic self-concept (Model 6; see Table 3). We found math grades to have a strong positive effect on math self-concept (b = .39, p < .001, 95% CI = [.38, .40]). Students with better grades had higher self-concept. In line with H4, we found the math class BFLPE to be substantially reduced (from b = -.37 in Model 4) to b = -.26, p < .001, CI = [-.28, -.24]. Additionally we found that the school (b = -.02, p = .297, CI = [-.06, .02]; b = -.06 in Model 4) and the regular class BFLPEs (b = -.14, p < .001, CI = [-.18, -.11]; b = -.11 in Model 4) were not substantially changed by the inclusion of school grades. This suggests that equally able students with equal grades did not experience a more or less severe school or regular class BFLPE.

Robustness Checks and Additional Analyses

To check the robustness of our results, we ran additional sets of analyses. First, as the academic self-concept item "*Mathematics is harder for me than for many of my classmates*" directly referred to social comparisons, we reran all models with a self-concept score in which this item was excluded. This did not change the results (Table S2). Second, we reran all models with the inclusion of covariates, namely sex, age, SES, and migration background. Also, this did not substantially change the results (Table S3). Third, we ran additional analyses with math class average track level instead of track level. We did this because prestige measures in the assimilation effects literature often represent class-level variables. In our main analyses, this was not true for students from math classes that contain students from several math tracks. Modeling average track level instead of track level did not substantially change the results (Table S4). In additional exploratory analyses, we also calculated the interactions between the track-level dummies as well as the BFLPEs on the different levels. We found that low-track students experienced a more positive school-level BFLPE as opposed to medium-track students, whereas the opposite was true for high-track students (Table S5). Furthermore, high-track students experienced a more negative math class level BFLPE as opposed to medium-track students but a more positive regular class level BFLPE. We also calculated all models with only the social comparison item as the dependent variable. The results were the same as for the complete scale (Table S6). Moreover, we calculated all models with the help of multiply imputed data (Table S7). The results were the same as for the complete case analysis approach. Additionally, we ran all analyses for the subsample of students from math classes that contain only students from the same track (Table S8). This did not substantially change the results.

Finally, in an exploratory endeavor, we also conducted moderation analyses in which we specified the interactions between the achievement aggregates and sex, age, migration, and SES (Table S9). None of these interactions were statistically significantly different from zero.

Table 1

Descriptive Statistics of Model Variables

	Mis	*M*	*SD*	VP_{ind}	VP_{mcl}	VP_{rcl}	VP_{sch}	1	2	3	4	5
1. Self-concept	.01	2.97	0.76	.95	.05	<.01	<.01					
2. Individual achievement	.00	504.24	86.85	.41	.52	.01	.06	.39				
3. Math class achievement	.00	504.24	67.58					.16	.78			
4. Regular class achievement	.00	504.24	49.74					.09	.57	.70		
5. School achievement	.00	504.24	44.09					.06	.51	.65	.89	
6. Grade	.02	3.19	0.94	.88	.08	.01	.02	.51	.39	.20	.15	.12

Note. All variables refer to the domain of mathematics. Variables 1 to 5 are in their original metric. Grade is reverse coded in that higher values indicate better grades. Descriptive statistics were calculated using a complete case analysis approach. Variance proportions were estimated using random intercept models that modeled all levels of student nesting: students (*VPind*), math class (*VPmcl*), regular class (*VPrcl*), and school (*VPsch*). Mis = percent missing.

Table 2

Pivotal Frames-of-Reference for Academic Self-Concept Formation

	Model 1				Model 2				Model 3				Model 4			
	B	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*
Individual achievement	.55	.01	[.54, .56]	<.001	.67	.01	[.66, .68]	<.001	.56	.01	[.55, .57]	<.001	.67	.01	[.66, .69]	<.001
School achievement	-.42	.01	[-.45, -.39]	<.001									-.06	.02	[-.10, -.02]	.004
Math class achievement					-.43	.01	[-.45, -.41]	<.001					-.37	.01	[-.39, -.35]	<.001
Regular class achievement									-.37	.01	[-.39, -.35]	<.001	-.11	.02	[-.15, -.08]	<.001

Note. N = 49,625. All variables refer to the domain of mathematics. Level 1 variables are standardized, and manifest level 2 aggregates are composed of standardized level 1 variables. CI = 95% confidence interval.

Table 3

Pivotal Frames-of-Reference for Academic Self-Concept Formation With Math Grade and Track Level

	Model 5 (N = 46,078)				Model 6 (N = 48,978)			
	B	*SE*	CI	*p*	*B*	*SE*	CI	*p*
Individual achievement	.61	.01	[.60, .63]	<.001	.45	.01	[.44, .46]	<.001
School achievement	.05	.02	[.01, .09]	.024	-.02	.02	[-.06, .02]	.297
Math class achievement	-.53	.01	[-.56, -.51]	<.001	-.26	.01	[-.28, -.24]	<.001
Regular class achievement	-.09	.02	[-.13, -.06]	<.001	-.14	.02	[-.18, -.11]	<.001
Low track	-.32	.01	[-.35, -.30]	<.001				
High track	.19	.01	[.17, .22]	<.001				
Grade					.39	.00	[.38, .40]	<.001

Note. All variables refer to the domain of mathematics. Level 1 variables are standardized, and manifest level 2 aggregates are composed of standardized level 1 variables. The track variables are dummy variables with reference category medium track. Because of the complete case analysis approach, *N*s differed slightly for the statistical models. CI = 95% confidence interval.

Table 4

Pivotal Frames-of-Reference for Academic Self-Concept Formation: Differences Between Students from Pure and Mixed Math Classes

	B	*SE*	CI	*p*
Mixed	-.04	.02	[-.08, .00]	.082
Achievement	.69	.01	[.67, .71]	<.001
School achievement	.00	.04	[-.07, .08]	.941
Math class achievement	-.52	.03	[-.58, -.45]	<.001
Regular class achievement	-.08	.03	[-.13, -.02]	.006
Low track	-.19	.03	[-.26, -.13]	<.001
High track	.09	.04	[.02, .16]	.009
Mixed x achievement	-.14	.01	[-.17, -.11]	<.001
Mixed x school achievement	.06	.05	[-.03, .16]	.199
Mixed x math class achievement	.03	.04	[-.05, .10]	.499
Mixed x regular class achievement	-.05	.04	[-.12, .03]	.220
Mixed x low track	-.19	.04	[-.26, -.11]	<.001
Mixed x high track	.14	.04	[.06, .22]	<.001

Note. All variables refer to the domain of mathematics. Level 1 variables are standardized, and manifest level 2 aggregates are composed of standardized level 1 variables. “Mixed” is a dummy variable indicating if students belong to pure (students from one math track; value 0) or mixed (students from several math tracks; value 1) math classes. The track-level variables are dummy variables with reference category medium track. CI = 95% confidence interval.

Table 5

Pivotal Frames-of-Reference for Grade Provision

	Model 1				Model 2				Model 3				Model 4			
	B	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*
Individual achievement	.50	.01	[.49, .51]	<.001	.58	.01	[.57, .59]	<.001	.49	.01	.48 .50	<.001	.58	.01	[.56, .59]	<.001
School achievement	-.25	.02	[-.28, -.22]	<.001									-.11	.02	[-.15, -.06]	<.001
Math class achievement					-.31	.01	[-.32, -.29]	<.001					-.30	.01	[-.32, -.28]	<.001
Regular class achievement									-.15	.01	-.18 -.13	<.001	.07	.02	[.04, .11]	<.001

Note. All variables refer to the domain of mathematics. Level 1 variables are standardized, and manifest level 2 aggregates are composed of standardized level 1 variables. CI = 95% confidence interval

.

Discussion

When regressing math self-concept on math achievement on all levels of student nesting, math class achievement had the strongest negative effect (math class BFLPE), regular class achievement had a less negative effect (regular class BFLPE), and school achievement had the least negative effect (school BFLPE). Additionally controlling for track level increased the math class BFLPE but did not substantially change the regular class BFLPE. Additionally controlling for grades decreased the math class BFLPE but did not substantially change the regular class BFLPE. In sum, our study suggests that in course-by-course tracked systems, multiple class environments may act as frames-of-reference for academic self-concept formation and that mechanisms of respective social comparison processes might differ from each other.

Our paper offers several unique contributions to the BFLPE literature and more broadly to the literature on academic self-concept formation. Our study's overall contribution is the investigation of the BFLPE and its potential mechanisms in course-by-course tracked systems in which students are members of not one but multiple class environments. More specifically, our study is the first to (a) juxtapose multiple class environments as frames-of-reference for academic self-concept formation and (b) investigate assimilation and grading on a curve as a potential mechanism for frame-of-reference effects of these multiple class environments. Regarding (a) we found the (domain-specific) math class achievement and to a weaker extent the (domain-unrelated) regular class achievement to negatively predict domain-specific academic self-concept. This finding suggests that students in course-by-course tracked systems evaluate their abilities against students from not one but multiple class environments. Regarding (b) we found BFLPEs of multiple class environments to differentially react to controlling for track level and grades. The math class BFLPE increased when controlling for track level and decreased when controlling for grades; in contrast, there was no substantial change to the regular class BFLPE. One interpretation of our results is that math class BFLPE is counterbalanced assimilation and associated with grading on a curve, whereas this is not the case for the regular class BFLPE. Additionally, our study contributes to the educational psychological literature by investigating differential track-level effects for students from math classes that contain students from the same math track (pure math classes) and students from math classes that contain students from different math tracks (mixed math classes). We found more pronounced track-level effects in mixed math classes, suggesting track-level saliency to amplify this prestige effect. Also in additional analyses, we investigated frame-of-reference effects on grades in course-by-course tracked school systems. We found grades to be negatively

predicted by math class and to a lower extent also by school achievement, suggesting that teachers conduct class- and school-referenced grading.

Pivotal Frames-of-Reference for Academic Self-Concept Formation

Generally, we found all math achievement aggregates on all levels of student nesting (math class, regular class, and school) to negatively predict math self-concept when modeled separately. However, when conjointly modeling all these predictors, the math class BFLPE was dominant. These results provide renewed evidence that the use of traditional large-scale datasets that do not allow for modeling all levels of student nesting is most likely to result in biased estimates of level-specific BFLPEs.

When conjointly modeling math achievement aggregates on all levels of student nesting, we found a small school BFLPE. This result is somewhat in contrast to that of Marsh, Kuyper, et al. (2014) who did not find a school BFLPE when class achievement was taken into account. However, Marsh, Kuyper, et al. (2014) conducted their study with a Dutch student sample in which students were tracked in relation to all classes. We also note that the very large sample size in our study meant that even a small BFLPE at the school level was highly significant.

When conjointly modeling math achievement aggregates on all levels of student nesting, we also found a regular class BFLPE that was smaller than the math class BFLPE. As both educational environments might be considered to be similar concerning their local proximity but differ concerning their domain-specific proximity, these results are in line with—but also clarify and extend—local dominance theory. But how can average math achievement of regular classes affect students' math self-concept? Our explanation is that it is likely that students had a relatively accurate perception of the math achievement of regular classes because track membership was highly salient. Thus, students might have had lower math self-concept in regular classes with high math achievement as a consequence of being surrounded by lots of students from the high math track. Conversely, students might have had lower math self-concept in regular classes with low math achievement as a consequence of being surrounded by lots of students from the low math track.

Track Level and the BFLPE

When additionally modeling track level, we found students from higher math tracks to have higher math self-concept. One interpretation of this finding is that students experienced assimilative track branding effects (e.g., "I am in a high math track, thus I am good at math"). However, as already noted in the introduction, information on track level is confounded with

students' prior achievement, as track designation is based on prior achievement. Unfortunately, we cannot resolve the issue of these opposing interpretations with data available in the present investigation. Thus, the disentanglement of positive track branding effects and effects of prior achievement is a fruitful direction for future research.

When additionally modeling track level, the math class BFLPE increased, whereas this did not change the regular class BFLPE. We interpret this finding as a consequence of the math class BFLPE being counterbalanced by assimilation, whereas this was not the case for the regular class BFLPE. This result suggests that frame-of-reference effects of multiple class environments might differ in their mechanisms.

In addition, we found students from mixed math classes to experience more pronounced track-level effects on academic self-concept. We interpret this as a consequence of increased salience of track level in mixed math classes.

School Grades and the BFLPE

We found school grades to be negatively predicted by math class achievement and school achievement, whereas the effect of regular class achievement was slightly positive. This result suggests that teachers—next to providing grades on a class-referenced basis—additionally grade on a school-referenced basis. The rather unexpected frame-of-reference effect on the school level may result from two aspects. First, several math classes from one school might be taught by the same teacher. These teachers might evaluate students in their classes on the same scale, for instance with the same tests, which induced the frame-of-reference effect on the school level. Unfortunately, no teacher ID is provided in the data so we are not able to empirically test our assumption. Another explanation for our finding might be that math teachers use common testing standards. For instance, they might use identical test materials and standardized result protocols, which are comparable across classes within schools. We also found that regular-class achievement positively affected grades when controlling for achievement on all other levels of student nesting. This finding is somewhat surprising as we would not have expected any associations between regular class achievement and grades. In other words, why should teachers provide better grades for students that come from a high achieving regular class? We can only speculate about possible mechanisms. For instance, teachers might perceive students from high achieving regular classes to be more competent, thus providing them with better school grades.

When math grades were included in the model, there was a substantial positive effect of math grades on math self-concept. Additionally, the math class BFLPE decreased

substantially whereas this was not the case for the regular class BFLPE. We interpret this finding in that the math class BFLPE was associated with grading on a curve. This result suggests that in frame-of-reference effects of multiple class environments might differ in their mechanisms. As already noted in the introduction, previous research interpreted this to mean that the BFLPE was caused, at least in part, by grading-on-a-curve driving the BFLPE (e.g., Marsh, 1987; Trautwein et al., 2006). However, hypothesized causal effects are difficult to test with correlational data. Indeed, recent discussion suggests that there is a strong evolutionary basis for social comparison processes (Frank, 2011). Marsh et al. (2018) argued that this explains why the BFLPE is so cross-culturally robust, and this supports claims that social comparison processes underpinning the BFLPE are pan-human and universal (Marsh & Seaton, 2015). From this perspective, it might be possible that the social comparison processes underlying the BFLPE are so strong that they are independent of the provision of class-referenced grades because students socially compare themselves in relation to other students in a similar fashion whether or not they are assigned with school grades. Thus, for example, would the size of the BFLPE decrease if students were not assigned grades at all or were assigned grades in relation to a common metric rather than grading on a curve? Although beyond the scope of the present investigation, we note that more research is needed to determine whether grading-on-a-curve is a causal contributor to the BFLPE or merely an effect that is correlated with the BFLPE.

Limitations and Directions for Future Research

Although our study is based on strong data, some potential limitations should be addressed in the future. First, students were tracked not only in relation to math but also in German and English. However, we had no information about German and English class membership. Thus, every student was associated with two more class environments that were not included in our analysis. Future research should aim at juxtaposing all class environments as frames-of-reference for academic self-concept formation. However, such an endeavor requires a comprehensive dataset with complete information on students' multiple course memberships.

A second potential limitation of the present investigation is that it is based on cross-sectional population data, thus we cannot provide firm causal inference. For two reasons, we argue that our correlational approach, which is of course not perfect, still provides a rather strong design to investigate the desired research questions. First, an internally valid juxtaposition of multiple class environments as frames-of-reference for academic self-concept formation would require the random assignment to multiple class environments that differ in

their average achievement. More specifically, it would require randomizing students to schools with different achievement levels, while simultaneously keeping class achievement constant. Likewise, it would require randomizing students to classes with different achievement levels, while simultaneously keeping school achievement constant. For ethical, organizational, and political reasons there is no chance to conduct such a study. Second, our study shows that achievement aggregates from different student environments are highly correlated and that controlling for all student environments results in a completely different picture. For instance, the school BFLPE shrinks by about 85% (from -.42 to -.06). Thus, we argue that our study which controls for achievement aggregates of different student environments has substantially improved in internal validity in contrast to previous studies that considered only one student environment (e.g., the school).

Third—although data from the Austrian national educational standard assessment represents a comprehensive survey, thus providing nearly perfect external validity for the Austrian context—it remains unclear to what extent our results are generalizable to other countries and educational systems. Due to differences in teacher communication or grading policies, it might be the case that pivotal frames-of-reference for academic self-concept formation in other student populations deviate from those we found. Thus, future studies should replicate our findings in different cultural contexts. Additionally, prior research has produced evidence that the BFLPE is stronger in math as opposed to verbal domains (e.g., Guo, Marsh, Parker, & Dicke, 2018). In this paper, we focused on mathematics, as this was the central domain of the national educational standard assessments in 2012. Future research is needed to test the generalizability of our results to other domains (e.g., language). Limitations of external validity also concern the transferability of the results to other age groups. For instance, the local dominance effect might be stronger in younger age groups that evaluate their abilities primarily concerning very proximal environments, whereas older age groups might take into account also less proximal comparison information. Limitations of external validity also concern the transferability of the results to other educational systems.

Finally, track level, as well as grades, were self-reported by students. Thus self-report bias, such as social desirability, might have impacted the reliability and validity of our measures. As it is rather unlikely that self-report bias differentially occurred for different groups of students, we think that it did not affect the relationship between the variables. If this would have been the case, however, the grade- and track-level estimates that we found would have been conservative estimates. Concerning grades, there is also empirical evidence that self-

reported measures provide reliable indicators of actual grades (Sticca et al., 2017). Additionally, neither grades nor track level were the central constructs in Hypothesis 3 and Hypothesis 4.

Practical Implications

Generally, the very basic BFLPE finding—equally able students have lower self-concept in high achieving educational environments—has a variety of practical implications. These implications can be divided into (a) implications for individual educational careers and (b) implications for educational systems. Regarding (a), the BFLPE predicts that individual educational careers that will result in changes in the average achievement of a student's educational environment will be accompanied by changes in the student's academic self-concept. In this context, the BFLPE predicts that school transfers, educational transitions, course choices, track changes, or grade retention of a student may be accompanied by changes in his academic self-concept (e.g., Wouters, Fraine, Colpin, van Damme, & Verschueren, 2012). Regarding (b), the BFLPE predicts that changing educational systems concerning the composition of educational environments will result in changes in students' academic self-concept. Specifically, this means that every form of ability segregation (e.g., different forms of tracking) should increase the academic self-concept of low achievers because it decreases the average achievement of these students' educational environments (Hübner et al., 2017; Hübner, Wagner, Hochweber, Neumann, & Nagengast, 2020). Note that the opposite is true for high achieving students. Vice versa the BFLPE predicts that ability desegregation (e.g., detracking) will decrease the academic self-concept of low achievers because it increases the average achievement of these students' educational environments.

Given these predictions of the BFLPE, the question arises on how educational policymakers should shape their school systems to reduce the negative consequences of the frame-of-reference effect. First of all, it has to be noted that the BFLPE is a "zero-sum game" (Trautwein & Möller, 2016). This means that a low achieving student that encounters a high achieving classroom will have lower academic self-concept but he will also lower the class average achievement of that class, increasing the academic self-concept of other students. Similarly, detracking will result in an academic self-concept decline of low achievers but an academic self-concept increase of high achievers. Additionally, the BFLPE applies to student motivation in terms of academic self-concept, however not necessarily to other educational outcomes such as academic achievement (Dicke et al., 2018; Stäbler, Dumont, Becker, & Baumert, 2017). Nevertheless, suggestions have been made for counteracting the negative consequences of the BFLPE. For instance, Marsh and Seaton (2015) suggest avoiding a competitive environment, enhancing students' feeling of connection, or valuing students'

unique accomplishments as potential measures to reduce the negative consequences of BFLPE. Unfortunately, there is no empirical evidence for the effectiveness of such endeavors. Studies show that the size of the BFLPE seems not to be affected by feedback practices (Lüdtke et al., 2005) or motivational climate (Wouters, Colpin, van Damme, & Verschueren, 2013). Thus, the BFLPE has been described as an unavoidable aspect of human nature (Marsh, Parker, Guo, Pekrun, & Basarkod, 2020) .

Our study investigated BFLPE and its proposed mechanisms within course-by-course tracked school systems in which students are members of multiple class environments. Accordingly, our findings allow for a refinement of BFLPE predictions presented above. Regarding individual educational careers, our study results suggest that a student's academic self-concept in a certain domain may be strongly hurt when placing him in high achieving domain-specific classes and may also be hurt, though to a much lesser extent, when placing him in high achieving domain-unrelated classrooms or schools.

Our study also comes with further implications for educational practice. For example, we found track-level effects to be more pronounced in math classes that contain students from more than one math track. As track-level effects on academic self-concept negatively affect low achievers and have the opposite effect for high achievers, these results remind practitioners to carefully think about the arrangement of learning environments. We also found frame-of-reference effects on grades on the math class level and to a weaker extent on the school level. This finding can be interpreted to mean that grades might not only be class-referenced but also be school-referenced. Thus, making grades a more valid instrument for student assessment requires better coordination not only between teachers but also between respective schools.

Conclusion

The present study was aimed at testing predictions from local dominance theory by taking a closer look at the pivotal frames-of-reference for academic self-concept formation in course-by-course tracked school systems. More specifically, we were interested in juxtaposing multiple class environments as frames-of-reference for academic self-concept formation. Data from a comprehensive survey that measured the entire population of Austrian eighth-grade students without special educational needs were well-suited for addressing our research question as general secondary school students were tracked in the core subjects (i.e., mathematics, German, and English) according to ability, whereas regular class composition was the same in all other (non-tracked) subjects. We found math class achievement and to a weaker extent regular class achievement to negatively affect math self-concept, when controlling for

achievement on all levels of student nesting. Our finding is in line with local dominance theory and suggests the more proximal domain-specific and to a lower extent the domain-unrelated environments to be frames-of-reference for academic self-concept formation.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008

Alicke, M. D., Zell, E., & Bloom, D. L. (2010). Mere categorization and the frog-pond effect. *Psychological Science, 21*, 174–177. https://doi.org/10.1177/0956797609357718

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

BIFIE. (2012). *Schülerfragebogen Standardüberprüfung 8. Schulstufe 2012. Gedruckte Version.* Salzburg: BIFIE.

BIFIE. (2016). *Datensatz zur Standardüberprüfung 2012 Mathematik, 8. Schulstufe. Schülerebene, M812I. Forschungsdatenbibliothek (FDB). Nicht-imputierter Datensatz, v2.0.*: BIFIE.

Bruneforth, M., Chabera, B., Vogtenhuber, S., & Lassnigg, L. (2016). *ECD review of policies to improve the effectiveness of resource use in schools. Country background report for Austria*. Wien: Bundesministerium für Bildung und Frauen.

Buckingham, J. T., & Alicke, M. D. (2002). The influence of individual versus aggregate social comparison and the presence of others on self-evaluations. *Journal of Personality and Social Psychology*, *83*, 1117–1130. https://doi.org/10.1037/0022-3514.83.5.1117

Chmielewski, A. K. (2014). An international comparison of achievement inequality in within- and between-school tracking systems. *American Journal of Education, 120*, 293–324. https://doi.org/10.1086/675529

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type. *American Educational Research Journal*, *50*, 925–957. https://doi.org/10.3102/0002831213489843

Cialdini, R. B., & Richardson, K. D. (1980). Two indirect tactics of image management: Basking and blasting. *Journal of Personality and Social Psychology*, *39*, 406–415. https://doi.org/10.1037/0022-3514.39.3.406

Dicke, T., Marsh, H. W., Parker, P. D., Pekrun, R., Guo, J., & Televantou, I. (2018). Effects of school-average achievement on individual self-concept and achievement: Unmasking phantom effects masquerading as true compositional effects. *Journal of Educational Psychology*, *110*, 1112–1126. https://doi.org/10.1037/edu0000259

Eder, F., Altrichter, H., Hofmann, F., & Weber, C. (Eds.). (2015). *Evaluation der Neuen Mittelschule (NMS). Befunde aus den Anfangskohorten.* Graz: Leykam.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*, 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Festinger, L. (1957). A theory of social comparison processes. *Human Relations, 7*, 117–140. https://doi.org/10.1177/001872675400700202

Frank, R. H. (2011). *The Darwin economy: Liberty, competition, and the common good*. Princeton: Princeton University Press.

Göllner, R., Damian, R. I., Nagengast, B., Roberts, B. W., & Trautwein, U. (2018). It's not only who you are but who you are with: High school composition and individuals' attainment over the life course. *Psychological Science*, 1-12. https://doi.org/10.1177/0956797618794454

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

Guo, J., Marsh, H. W., Parker, P. D., & Dicke, T. (2018). Cross-cultural generalizability of social and dimensional comparison effects on reading, math, and science self-concepts for primary school students using the combined PIRLS and TIMSS data. *Learning and Instruction, 58*, 210–219.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. London: Routledge Taylor & Francis Group.

Hübner, N., Wagner, W., Hochweber, J., Neumann, M., & Nagengast, B. (2020). Comparing apples and oranges: Curricular intensification reforms can change the meaning of students' grades! *Journal of Educational Psychology*, *112*, 204–220. https://doi.org/10.1037/edu0000351

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

Huguet, P., Dumas, F., Marsh, H. W., Wheeler, L., Seaton, M., Nezlek, J., . . . Régner, I. (2009). Clarifying the role of social comparison in the big-fish-little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology*, *97*, 156–170. https://doi.org/10.1037/a0015558

Liem, G. A. D., Marsh, H. W., Martin, A. J., McInerney, D. M., & Yeung, A. S. (2013). The big-fish-little-pond effect and a national policy of within-school ability streaming. *American Educational Research Journal, 50,* 326–370. https://doi.org/10.3102/0002831212464511

Loveless, T. (2013). *The 2013 Brown Center report on American education: How well are American students learning?* Washington: Brookings Institution.

Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish-little-pond effect. *Contemporary Educational Psychology*, *30,* 263–285. https://doi.org/10.1016/j.cedpsych.2004.10.002

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

Marsh, H. W. (1991). Failure of high-ability high schools to deliver academic benefits commensurate with their students' ability levels. *American Educational Research Journal, 28,* 445–480. https://doi.org/10.2307/1162948

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J., Abdelfattah, F., & Nagengast, B. (2014). The big-fish-little-pond effect in mathematics. *Journal of Cross-Cultural Psychology*, *45*, 777–804. https://doi.org/10.1177/0022022113519858

Marsh, H. W., & Hau, K.-T. (2003). Big-fish-little-pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist, 58,* 364–376. https://doi.org/10.1037/0003-066X.58.5.364

Marsh, H. W., Köller, O., & Baumert, J. (2001). Reunification of east and west german school systems: Longitudinal multilevel modeling study of the big-fish-little-pond effect on academic self-concept. *American Educational Research Journal, 38,* 321–350. https://doi.org/10.3102/00028312038002321

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology*, *78*, 337–349. https://doi.org/10.1037//0022-3514.78.2.337

Marsh, H. W., Kuyper, H., Morin, A. J., Parker, P. D., & Seaton, M. (2014). Big-fish-little-pond social comparison and local dominance effects: Integrating new statistical models, methodology, design, theory and substantive implications. *Learning and Instruction, 33,* 50–66. https://doi.org/10.1016/j.learninstruc.2014.04.002

Marsh, H. W., Martin, A. J., Yeung, A. S., & Craven, R. (2016). Competence self-perceptions. In C. Dweck & D. Yaeger (Eds.), *Handbook of competence and motivation.* New York: Guilford Press.

Marsh, H. W., & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, *47*, 213–231. https://doi.org/10.1037/0022-3514.47.1.213

Marsh, H. W., Parker, P. D., Guo, J., Pekrun, R., & Basarkod, G. (2020). Psychological comparison processes and self-concept in relation to five distinct frame-of-reference effects: Pan-human cross-cultural generalizability over 68 countries. *European Journal of Personality, 3*, 180–202. https://doi.org/10.1002/per.2232

Marsh, H. W., Pekrun, R., Murayama, K., Arens, A. K., Parker, P. D., Guo, J., & Dicke, T. (2018). An integrated model of academic self-concept development: Academic self-concept, grades, test scores, and tracking over 6 years. *Developmental Psychology*, *54*, 263–280. https://doi.org/10.1037/dev0000393

Marsh, H. W., & Seaton, M. (2015). The big-fish–little-pond effect, competence self-perceptions, and relativity: Substantive advances and methodological innovation. *Advances in Motivation Science, 2*, 127–184.

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

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*, 206–217. https://doi.org/10.1016/j.stueduc.2012.02.002

OECD. (2017a). *PISA 2015 assessment and analytical framework. Science, reading, mathematic, financial literacy and collaborative problem solving*. Paris: OECD.

OECD. (2017b). *PISA 2015 Technical Report*. Paris: OECD.

R Core Team. (2008). *R: A Language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: Sage.

Robitzsch, A., Grund, S., & Henke, T. (2018). *Miceadds: Some additional multiple imputation functions, especially for mice.* R package version 3.8.9.

Schreiner, C., & Breit, S. (2012). *Standardüberprüfung 2012 Mathematik, 8. Schulstufe. Bundesergebnisbericht*. Salzburg: BIFIE.

Seaton, M., Marsh, H. W., & Craven, R. G. (2010). Big-fish-little-pond effect: Generalizability and moderation - Two sides of the same coin. *American Educational Research Journal*, *47*, 390–433. https://doi.org/10.3102/0002831209350493

Seaton, M., Marsh, H. W., Yeung, A. S., & Craven, R. (2011). The big fish down under: Examining moderators of the 'big-fish little-pond' effect for Australia's high achievers. *Australian Journal of Education, 55*, 93–114. https://doi.org/10.1177/000494411105500202

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Snyder, C. R., Lassegard, M., & Ford, C. E. (1986). Distancing after group success and failure: Basking in reflected glory and cutting off reflected failure. *Journal of Personality and Social Psychology, 51*, 382–388. https://doi.org/10.1037/0022-3514.51.2.382

Stäbler, F., Dumont, H., Becker, M., & Baumert, J. (2017). What happens to the fish's achievement in a little pond? A simultaneous analysis of class-average achievement effects on achievement and academic self-concept. *Journal of Educational Psychology, 109*, 191–207. https://doi.org/10.1037/edu0000135

Sticca, F., Goetz, T., Bieg, M., Hall, N. C., Eberle, F., & Haag, L. (2017). Examining the accuracy of students' self-reported academic grades from a correlational and a discrepancy perspective: Evidence from a longitudinal study: Evidence from a longitudinal study. *PloS One*, *12*, 1-13. https://doi.org/10.1371/journal.pone.0187367

Trautwein, U., Gerlach, E., & Lüdtke, O. (2008). Athletic classmates, physical self-concept, and free-time physical activity: A longitudinal study of frame of reference effects. *Journal of Educational Psychology*, *100*, 988–1001. https://doi.org/10.1037/0022-0663.100.4.988

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, *98*, 788–806. 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*, 853–866. https://doi.org/10.1037/a0016306

Trautwein, U., & Möller, J. (2016). Self-concept: Determinants and consequences of academic self-concept in school contexts. In A. A. Lipnevich, F. Preckel, & R. D. Roberts (Eds.), *Psychosocial Skills and School Systems in the 21st Century* (pp. 187–214). Cham: Springer International Publishing.

Von Davier, M. von, Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Vol. 2. Issues and methodologies in large scale assessments* (pp. 9–36). Princeton, NJ: IEA-ETS Research Institute.

Wouters, S., Colpin, H., van Damme, J., & Verschueren, K. (2013). Endorsing achievement goals exacerbates the big-fish-little-pond effect on academic self-concept. *Educational Psychology, 35*, 252–270. https://doi.org/10.1080/01443410.2013.822963

Wouters, S., Fraine, B. de, Colpin, H., van Damme, J., & Verschueren, K. (2012). The effect of track changes on the development of academic self-concept in high school: A dynamic test of the big-fish–little-pond effect. *Journal of Educational Psychology*, *104*, 793–805. https://doi.org/10.1037/a0027732

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*, 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Zell, E., & Alicke, M. D. (2009). Contextual neglect, self-evaluation, and the frog-pond effect. *Journal of Personality and Social Psychology, 97*, 467–482. https://doi.org/10.1037/a0015453

Zell, E., & Alicke, M. D. (2010). The local dominance effect in self-evaluation: Evidence and explanations. *Personality and Social Psychology Review*, *14*, 368–384. https://doi.org/10.1177/1088868310366144

Supplemental Material

Table S1

Descriptive Statistics for Different Math Tracks

	Low track				Medium track				High track			
	M	*SD*	1	2	*M*	*SD*	1	2	*M*	*SD*	1	2
1. Self-concept	2.55	0.79			2.91	0.73			3.16	0.70		
2. Achievement	414.02	56.06	.25		477.38	59.10	.24		564.77	69.50	.32	
3. Grade	2.69	0.88	.41	.19	3.06	0.81	.41	.17	3.46	0.96	.51	.32

Note. All variables refer to the domain of math. Variables 1 to 2 are in their original metric. Grade is reverse coded in that higher values indicate better grades. Descriptive statistics were calculated using a complete case analysis approach.

Table S2

Pivotal Frames-of-Reference for Academic Self-Concept Formation Without Self-Concept Item That Refers to Social Comparison

	Model 1				Model 2				Model 3				Model 4				Model 5				Model 6			
	B	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*
Individual achievement	.53	.01	[.52, .54]	<.001	.64	.01	[.63, .66]	<.001	.54	.01	[.53, .55]	<.001	.65	.01	[.63, .66]	<.001	.59	.01	[.57 .60]	<.001	.43	.01	[.41, .44]	<.001
School achievement	-.41	.01	[-.43, -.38]	<.001									-.08	.02	[-.13, -.04]	<.001	.03	.02	[-.01 .08]	.162	-.04	.02	[-.08, .00]	.042
Math class achievement					-.41	.01	[-.43, -.39]	<.001					-.35	.01	[-.37, -.33]	<.001	-.52	.01	[-.54 -.49]	<.001	-.24	.01	[-.26, -.22]	<.001
Regular class achievement									-.35	.01	[-.38, -.33]	<.001	-.10	.02	[-.13, -.06]	<.001	-.07	.02	[-.11 -.04]	<.001	-.13	.02	[-.16, -.09]	<.001
Low track																	-.31	.01	[-.33 -.28]	<.001				
High track																	.21	.01	[.18 .24]	<.001				
Grade																					.39	.00	[.38, .40]	<.001

Note. The self-concept item that refers to social comparison is *Mathematics is harder for me than for many of my classmates*. All variables refer to the domain of math. Level 1 variables are standardized, and manifest level 2 aggregates are composed of standardized level 1 variables. The track variables are dummy variables with reference category medium track. CI = 95% confidence interval.

Table S3

Pivotal Frames-of-Reference for Academic Self-Concept Formation Controlling for Covariates

	Model 1				Model 2				Model 3				Model 4				Model 5				Model 6			
	B	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*
Sex	-.35	.01	[-.36, -.33]	<.001	-.33	.01	[-.35, -.32]	<.001	-.35	.01	[-.36, -.33]	<.001	-.33	.01	[-.35, -.32]	<.001	-.34	.01	[-.36, -.32]	<.001	-.41	.01	[-.42, -.39]	<.001
Age	.00	.00	[-.01, .01]	.552	-.01	.00	[-.02, .00]	.008	.00	.00	[-.01, .00]	.353	-.01	.00	[-.02, .00]	.006	.00	.00	[-.01, .01]	.617	.01	.00	[.00, .01]	.133
SES	-.01	.00	[-.02, .00]	.045	.00	.00	[-.01, .01]	.605	-.01	.00	[-.01, .00]	.188	.00	.00	[-.01, .01]	.837	-.01	.00	[-.02, .00]	.045	-.01	.00	[-.02, -.01]	<.001
Migration	.20	.01	[.17, .22]	<.001	.21	.01	[.18, .23]	<.001	.20	.01	[.18, .23]	<.001	.19	.01	[.16, .21]	<.001	.20	.01	[.18, .23]	<.001	.15	.01	[.13, .17]	<.001
Individual achievement	.54	.01	[.53, .55]	<.001	.65	.01	[.64, .67]	<.001	.55	.01	[.54, .56]	<.001	.65	.01	[.64, .67]	<.001	.59	.01	[.58, .61]	<.001	.41	.01	[.40, .43]	<.001
School achievement	-.34	.01	[-.37, -.31]	<.001									-.04	.02	[-.08, .01]	.101	.09	.02	[.05, .14]	<.001	.00	.02	[-.04, .04]	.990
Math class achievement					-.39	.01	[-.41, -.37]	<.001					-.35	.01	[-.37, -.33]	<.001	-.52	.01	[-.55, -.49]	<.001	-.21	.01	[-.23, -.19]	<.001
Regular class achievement									-.30	.01	[-.33, -.28]	<.001	-.09	.02	[-.13, -.05]	<.001	-.06	.02	[-.10, -.02]	.002	-.12	.02	[-.15, -.08]	<.001
Low track																	-.34	.01	[-.37, -.31]	<.001				
High track																	.22	.01	[.19, .24]	<.001				
Grade																					.41	.00	[.41, .42]	<.001

Note. All variables refer to the domain of math. Level 1 variables are standardized, and manifest level 2 aggregates are composed of standardized level 1 variables. Sex is a dummy variable with 0 for male and 1 for female. Migration is a dummy variable with 0 for no migration background and 1 for migration background. The track variables are dummy variables with reference category medium track. CI = 95% confidence interval.

Table S4

Pivotal Frames-of-Reference for Academic Self-Concept Formation With Math Class Average Track Level

	B	*SE*	CI	*p*
Individual achievement	.67	.01	[.66, .68]	<.001
School achievement	.02	.02	[-.02, .07]	.375
Math class achievement	-.53	.02	[-.57, -.49]	<.001
Regular class achievement	-.09	.02	[-.13, -.06]	<.001
Math class average track level	.17	.02	[.13, .21]	<.001

Note. All variables refer to the domain of mathematics. Level 1 variables are standardized, and manifest level 2 aggregates are composed of standardized level 1 variables. CI = 95% confidence interval.

Table S5

Track Interaction Model

	B	*SE*	CI	*p*
Achievement	.55	.01	[.53, .58]	<.001
School achievement	.08	.03	[.01, .15]	.022
Math class achievement	-.44	.02	[-.49, -.39]	<.001
Regular class achievement	-.15	.03	[-.21, -.09]	<.001
Low track	-.34	.03	[-.40, -.29]	<.001
High track	.15	.02	[.12, .18]	<.001
Achievement x Low track	.00	.02	[-.05, .04]	.848
Achievement x High track	.11	.02	[.08, .15]	<.001
School achievement x Low track	.17	.05	[.06, .27]	.002
School achievement x High track	-.20	.05	[-.29, -.11]	<.001
Math class achievement x Low track	-.06	.04	[-.13, .02]	.126
Math class achievement x High track	-.09	.03	[-.15, -.02]	.010
Regular class achievement x Low track	.01	.05	[-.09, .11]	.885
Regular class achievement x High track	.11	.04	[.02, .19]	.010

Note. All variables refer to the domain of mathematics. Level 1 variables are standardized, and manifest level 2 aggregates are composed of standardized level 1 variables. The track variables are dummy variables with reference category medium track. CI = 95% confidence interval.

Table S6

Pivotal Frames-of-Reference for Academic Self-Concept Formation with Only the Social Comparison Item as the Outcome

	Model 1				Model 2				Model 3				Model 4				Model 5				Model 6			
	B	SE	CI	p	B	SE	CI	p	B	SE	CI	p	B	SE	CI	p	B	SE	CI	p	B	SE	CI	p
Individual achievement	.39	.01	[.38, .41]	<.001	.51	.01	[.49, .52]	<.001	.41	.01	[.40, .42]	<.001	.51	.01	[.50, .53]	<.001	.47	.01	[.46, .49]	<.001	.36	.01	[.34, .37]	<.001
School achievement	-.30	.01	[-.32, -.28]	<.001									.00	.02	[-.04, .04]	.983	.07	.02	[.03, .11]	.001	.03	.02	-[.01, .07]	.105
Math class achievement					-.35	.01	[-.37, -.33]	<.001					-.30	.01	[-.32, -.28]	<.001	-.40	.01	[-.42, -.37]	<.001	-.22	.01	[.24, -.20]	<.001
Regular class achievement									-.29	.01	[-.31, -.27]	<.001	-.12	.02	[-.15, -.08]	<.001	-.11	.02	[-.14, -.07]	<.001	-.14	.02	[-.17, -.11]	<.001
Low track																	-.25	.01	[-.28, -.22]	<.001				
High track																	.09	.01	[.06, .12]	<.001				
Grade																					.27	.00	[.26, .28]	<.001

Note. All variables refer to the domain of mathematics. Level 1 variables are standardized, and manifest level 2 aggregates are composed of standardized level 1 variables. The track-level variables are dummy variables with reference category medium track. CI = 95% confidence interval.

Table S7

Pivotal Frames-of-Reference for Academic Self-Concept Formation with Multiple Imputation

	Model 1				Model 2				Model 3				Model 4				Model 5				Model 6			
	B	SE	CI	p	B	SE	CI	p	B	SE	CI	p	B	SE	CI	p	B	SE	CI	p	B	SE	CI	p
Individual achievement	.55	.01	[.54, .56]	<.001	.67	.01	[.65, .68]	<.001	.56	.01	[.55, .57]	<.001	.67	.01	[.66, .69]	<.001	.62	.01	[.61, .64]	<.001	.48	.01	[.46, .49]	<.001
School achievement	-.42	.01	[-.45, -.40]	<.001									-.06	.02	[-.11, -.02]	.004	.01	.02	[-.03, .06]	.564	-.03	.02	[-.07, .01]	.204
Math class achievement					-.43	.01	[-.45, -.41]	<.001					-.37	.01	[-.40, -.35]	<.001	-.49	.01	[-.52, -.47]	<.001	-.27	.01	[-.29, -.25]	<.001
Regular class achievement									-.37	.01	[-.40, -.35]	<.001	-.12	.02	[-.15, -.08]	<.001	-.09	.02	[-.13, -.05]	<.001	-.14	.02	[-.17, -.11]	<.001
Low track																	-.27	.01	[-.29, -.24]	<.001				
High track																	.14	.01	[.11, .17]	<.001				
Grade																					.36	.00	[.36, .37]	<.001

Note. All variables refer to the domain of mathematics. Level 1 variables are standardized, and manifest level 2 aggregates are composed of standardized level 1 variables. The track-level variables are dummy variables with reference category medium track. CI = 95% confidence interval.

Table S8

Pivotal Frames-of-Reference for Academic Self-Concept Formation with the Pure Student Sample

	Model 1				Model 2				Model 3				Model 4				Model 5				Model 6			
	B	SE	CI	p	B	SE	CI	p	B	SE	CI	p	B	SE	CI	p	B	SE	CI	p	B	SE	CI	p
Individual achievement	.53	.01	[.51, .54]	<.001	.68	.01	[.66, .70]	<.001	.53	.01	[.52, .55]	<.001	.68	.01	[.66, .70]	<.001	.69	.01	[.67, .71]	<.001	.42	.01	[.40, .44]	<.001
School achievement	-.40	.03	[-.45, -.35]	<.001									-.07	.03	[-.14, -.01]	.033	.00	.04	[-.08, .08]	.968	-.03	.03	[-.09, .03]	.362
Math class achievement					-.43	.01	[-.46, -.40]	<.001					-.39	.02	[-.42, -.36]	<.001	-.51	.03	[-.58, -.44]	.000	-.25	.01	[-.28, -.22]	<.001
Regular class achievement									-.30	.02	[-.34, -.27]	<.001	-.08	.03	[-.13, -.03]	.002	-.08	.03	[-.13, -.03]	.004	-.11	.02	[-.16, -.06]	<.001
Low track																	-.19	.03	[-.25, -.12]	<.001				
High track																	.09	.04	[.01, .16]	.019				
Grade																					.41	.01	[.39, .42]	<.001

Note. All variables refer to the domain of mathematics. Level 1 variables are standardized, and manifest level 2 aggregates are composed of standardized level 1 variables. The track-level variables are dummy variables with reference category medium track. CI = 95% confidence interval.

Table S9

Pivotal Frames-of-Reference for Academic Self-Concept Formation with Moderators

	Sex				Age				Migration				SES			
	B	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*	*B*	*SE*	CI	*p*
Moderator	-.33	.01	[-.35, -.32]	<.001	.01	.00	[.00, .02]	.008	.19	.01	[.16, .21]	<.001	-.01	.00	[-.02, .00]	.094
Achievement	.65	.01	[.63, .66]	<.001	.67	.01	[.66, .69]	<.001	.68	.01	[.67, .69]	<.001	.68	.01	[.66, .69]	<.001
School achievement	-.06	.03	[-.11, -.01]	.031	-.07	.02	[-.11, -.02]	.002	-.02	.02	[-.07, .03]	.426	-.07	.02	[-.11, -.02]	.003
Math class achievement	-.34	.01	[-.37, -.32]	<.001	-.37	.01	[-.39, -.35]	<.001	-.37	.01	[-.39, -.34]	<.001	-.38	.01	[-.40, -.35]	<.001
Regular class achievement	-.12	.02	[-.17, -.08]	<.001	-.11	.02	[-.15, -.08]	<.001	-.11	.02	[-.15, -.07]	<.001	-.11	.02	[-.14, -.07]	<.001
Moderator x school achievement	-.02	.04	[-.09, .05]	.529	.03	.02	[-.01, .06]	.142	-.02	.05	[-.12, .07]	.612	-.01	.02	[-.05, .02]	.507
Moderator x math class achievement	-.01	.01	[-.04, .02]	.475	-.01	.01	[-.02, .01]	.272	.00	.02	[-.05, .04]	.905	.01	.01	[.00, .03]	.070
Moderator x regular class achievement	.04	.03	[-.03, .11]	.249	.00	.02	[-.03, .03]	.900	.01	.05	[-.08, .10]	.871	.02	.02	[-.02, .05]	.312

Note. All variables refer to the domain of mathematics. Level 1 variables are standardized, and manifest level 2 aggregates are composed of standardized level 1 variables. Sex is a dummy variable with 0 for male and 1 for female. Migration is a dummy variable with 0 for no migration background and 1 for migration background. The column names indicate the moderator under investigation. CI = 95% confidence interval.

4 Study 2: Living in the Big Pond: How Socioeconomic Neighborhood Composition Predicts Students' Academic Self-Concept

Fleischmann, M., Becker, D., Wessling, K., Nagengast, B., Trautwein, U. (2020). Living in the Big Pond: How Socioeconomic Neighborhood Composition Predicts Students' Academic Self-Concept. Manuscript submitted.

Abstract

Research on the big fish little pond effect (BFLPE) suggests that selective learning environments, for example, classrooms composed of high-achieving classmates, harm students' academic self-concept as a consequence of social comparison processes. Sociological studies stress the importance of neighborhoods as non-institutional student environments. In supposed contrast to the BFLPE, this line of sociological research emphasizes the beneficial effects of selective neighborhoods on students' academic development via mechanisms of collective socialization. To test predictions based on these seemingly conflicting research traditions, we individually and jointly analyzed the effects of classroom and neighborhood composition on students' academic self-concept in several domains. Using cross-classified multilevel models and controlling for possible confounding variables, we did not find any positive effects of advantageous socioeconomic neighborhood conditions on students' academic self-concept. Quite the contrary, among German fifth-grade students (N = 3,906), a higher neighborhood social status and a higher employment rate negatively predicted general and math self-concept. The negative effects of neighborhood social status remained even after controlling for class-average achievement. Both neighborhood characteristics also negatively predicted math self-concept of ninth-grade students (N = 3,277); however, the effects vanished after controlling for class-average achievement. In line with research on the BFLPE, our results suggest that advantageous socioeconomic neighborhood conditions can be harmful to educational outcomes that are susceptible to social comparison processes. We discuss mechanisms that might account for this finding, such as disguised school effects and social comparisons within neighborhoods.

Living in the Big Pond: How Socioeconomic Neighborhood Composition Predicts Students' Academic Self-Concept

Positive self-beliefs are discussed as being one of the socioemotional skills that today's students need in a rapidly changing world (OECD, 2018). They represent "a basic psychological need that has a pervasive impact on daily life, cognition, and behavior, across age and culture" (Elliot & Dweck, 2005, p. 8). One prominent self-belief construct is academic self-concept, which describes students' perceptions of their competence in academic domains (Marsh et al., 2016).

A large body of psychological research shows that students' academic self-concept is negatively predicted by the average achievement of educational environments such as the school or the classroom when controlling for individual achievement (for an overview, see Marsh & Seaton, 2015). In other words, equally able students have a lower academic self-concept in high-achieving learning environments. This finding has been called the big fish little pond effect (BFLPE; Marsh, 1987). The BFLPE is assumed to emerge as a consequence of social comparison processes in which students evaluate their academic capabilities by comparing their achievement with that of other students in their respective educational environment (Huguet et al., 2009; Marsh et al., 2014). Research on the BFLPE suggests that comparisons within proximal student environments—namely those that students are directly exposed to such as the classroom—matter most for academic self-concept formation (Liem et al., 2013; Marsh et al., 2014).

To date, research on the BFLPE focused predominantly on learning environments in formal education settings (e.g., schools, tracks, classes) as potentially relevant frames of reference for academic self-concept formation. However, residential neighborhoods constitute another important non-institutional environment students are directly exposed to (e.g., Boardman & Saint Onge, 2005; Childress, 2016). The association between socioeconomic neighborhood composition and educational outcomes has been investigated predominantly by sociologists and urban geographers (for an overview, see Galster, 2012; Sampson et al., 2002) who were less interested in institutional learning environments such as schools. Whereas research on the BFLPE suggests that selective learning environments harm students' academic self-concept, neighborhood effects research assumes that selective neighborhoods, in terms of advantageous socioeconomic neighborhood conditions, promote academic development in terms of achievement, educational and occupational aspirations and choices, or school behavior (e.g., Bowen & Bowen, 1999; Hartung & Hillmert, 2019; Nieuwenhuis & Hooimeijer, 2016).

To bring together these two supposedly conflicting lines of research, we separately and simultaneously analyzed the effects of selective educational environments (in terms of achievement-related classroom composition) and selective neighborhoods (in terms of socioeconomic neighborhood composition) on students' academic self-concept in several domains. To this end, we used data from Starting Cohort 3 of the German National Educational Panel Study (NEPS; Blossfeld et al., 2011), a survey that integrates measures of psychological constructs with residential information on neighborhoods students live in.

Academic Self-Concept and the BFLPE

Self-concept is broadly defined as a person's self-perceptions that are formed through experiences with his or her environment (Shavelson et al., 1976). More specifically, academic self-concept is students' perception of their academic abilities (Marsh et al., 2016). A positive academic self-concept is seen as a desirable educational outcome because it is assumed to foster academic achievement (Huang, 2011; Valentine et al., 2004). In addition, academic self-concept is also considered as an important predictor of career aspirations and academic choices (Eccles & Wigfield, 2002; Guo et al., 2015; Marsh & Yeung, 1997).

Davis (1966) published a seminal study which showed that compared to students attending low-ability schools, students from high-ability schools reported lower perceptions of their academic abilities and also less often chose a high-performance career—a finding hitherto referred to as the frog pond effect. He interpreted this result as indicating that school selectivity negatively impacts students' academic self-concept through comparison processes which in turn shape career decisions (see also Meyer, 1970). Also, Alwin and Otto (1977) found negative associations between school-average achievement on the one hand, and grades, curriculum choice, college plans, and occupational aspirations on the other hand. To account for these empirical findings theoretically, sociologists referred to the mechanism of relative deprivation (Stouffer et al., 1949), namely the "the judgment that one is worse off compared to some standard accompanied by feelings of anger and resentment" (Smith et al., 2012, p. 203).

Based on the sociological frog pond literature just described as well as psychological social comparison theory (e.g., Festinger, 1957), Marsh (1987) showed that having controlled for individual achievement differences, academic self-concept is negatively impacted by school-average achievement. This so-called big fish little pond effect (BFLPE; for an overview, see Marsh & Seaton, 2015) has been termed a "contrast effect" because it is assumed to emerge due to social comparison processes (e.g., Huguet et al., 2009; Marsh et al., 2014) in which students contrast with their educational environment. Both approaches postulate a negative

effect of the average academic achievement on student outcomes that are susceptible to social comparison processes. Whereas early sociological research on the frog pond effect focused mainly on aspirations, choices, and grades, later psychological research on the BFLPE typically targeted the construct of academic self-concept (Marsh & Seaton, 2015; but see Göllner et al., 2018.)

On the other hand, research on social comparison processes has suggested that the membership in high-status groups might also positively affect self-perceptions (Cialdini & Richardson, 1980; Snyder et al., 1986) by students assimilating with their educational environment. Consequently, average achievement of educational environments would positively affect student academic self-concept, resulting in a "basking in reflected glory effect" (e.g., Felson, 1984; Felson & Reed, 1986; Marsh, 1984). Indeed, Marsh et al. (2000) found perceived school status to positively predict academic self-concept. Additionally, including perceived school status in the BFLPE model amplified the negative frame-of-reference effect. Hence, Marsh et al. (2000) concluded the BFLPE to be the net effect of dominating contrast and less pronounced assimilation processes (see also Chmielewski et al., 2013; Trautwein et al., 2009).

Research within the BFLPE paradigm has concluded that more proximal frames of reference—namely those that students are directly exposed to on a daily basis—are the pivotal ones for academic self-concept formation. Studies that simultaneously tested the effects of school- and class-average achievement on academic self-concept observed that school achievement effects were completely absorbed when simultaneously modeling class achievement in the BFLPE model (e.g., Liem et al., 2013; Marsh et al., 2014). The idea that proximal frames of reference play the primary role in the emergence of self-evaluations is also supported by experimental work by Zell and Alicke (2010) who showed that participants preferred local to global comparison information, a mechanism which they labeled the "local dominance effect".

Neighborhood Effects on Educational Outcomes

Sociological neighborhood effects research investigates the relationship between socioeconomic neighborhood composition and outcomes such as deviant or health-related behavior, but also educational outcomes (for an overview, see Galster, 2012). Measures of socioeconomic neighborhood composition include indices of individual occupations, income, employment, and also ethnic concentration.

Neighborhood effects research that is concerned with educational outcomes predominantly points out the "advantages of advantaged neighbors", also called "Wilson's theory" (Mayer & Jencks, 1989; Wilson, 1987, 1996). For the US context, in line with Wilson's theory, an advantageous socioeconomic neighborhood composition was found to be beneficial for general child development (Brooks-Gunn et al., 1993; Duncan et al., 1994), academic aspirations (Kintrea et al., 2015; Stewart et al., 2016), academic achievement (Aaronson, 1998; Ainsworth, 2002; Catsambis & Beveridge, 2001; Nieuwenhuis & Hooimeijer, 2016), and school dropout (Crane, 1991). The impact of neighborhood conditions on educational outcomes in the European context seems to be weaker and less consistent (Brannstrom, 2008; Garner & Raudenbush, 1991; Helbig, 2010; Sykes & Musterd, 2010; Wicht & Ludwig-Mayerhofer, 2014). One reason for this might be stronger welfare state interventions (e.g., social benefit payments) that prevent extreme residential stratification in terms of race or social status (Friedrichs et al., 2010).

Various theoretical mechanisms that might account for such positive neighborhood effects have been discussed (for an overview see Galster, 2008, 2012). For instance, they can be caused by collective socialization processes in which individuals' behavior is impacted by peer residents who act as role models: Children living in a neighborhood in which adolescents perform well in school and adults have well-paying jobs might be more likely to work hard in school to be as successful as their role models. Besides, social networks that supply residents with assistance in times of need or with institutional resources—for example, the provision of high-quality schooling but also relevant information on schools or jobs—have been proposed to constitute additional potential mechanisms underlying the beneficial effects of advantaged neighborhoods.

Beyond the "advantages of advantaged neighbors", neighborhood effect researchers have also postulated "disadvantages of advantaged neighbors" (Mayer & Jencks, 1989). This idea is strongly related to the concept of relative deprivation (Davis, 1966; Stouffer et al., 1949), meaning that advantageous socioeconomic neighborhood conditions might result in dissatisfaction as a consequence of neighborhood residents' relatively poor evaluation of their own situation compared to their neighbors. Whereas relative deprivation effects of neighborhoods on several non-educational outcomes such as depression (Nieuwenhuis et al., 2017) or rioting (Canache, 1996) have been observed, we are not aware of any study that reports advantageous neighborhood conditions to negatively predict educational outcomes.

In sum, research on the BFLPE suggests that selective student environments, in terms of average academic achievement, negatively affect academic self-concept. However, sociological neighborhood effects research assumes that selective neighborhoods, in terms of socioeconomic neighborhood conditions, positively affect students' academic development. Based on these supposedly conflicting findings the question arises on how socioeconomic neighborhood composition predicts students' academic self-concept. This question comprises two discipline-specific questions. For educational psychology, this question is: Is the neighborhood a frame of reference for academic self-concept formation? For neighborhood effects research this question is: Is academic self-concept an educational outcome that is impacted by "disadvantages of advantaged neighbors"?

The Present Study

The present study brings together two supposedly conflicting lines of argumentation: research on the BFLPE that suggests students' academic self-concept declines when being placed in selective learning environments, and sociological neighborhood effects research that assumes selective neighborhoods positively affect students' academic development. To our knowledge, our study is the first to separately as well as simultaneously analyze the effects of classroom and neighborhood composition on students' academic self-concept. From an educational psychological perspective, the investigation of neighborhood effects on students' academic self-concept is highly relevant because it contributes to the theory of academic self-concept formation by investigating the neighborhood as a potential frame of reference. From a neighborhood effects research perspective, investigating neighborhood effects on students' academic self-concept is also an important contribution because it explicitly investigates neighborhood effects on an educational outcome on which one would expect "disadvantages of advantaged neighbors".

The present investigation is based on data from Starting Cohort 3 of the German National Educational Panel Study (NEPS; Blossfeld et al., 2011), a longitudinal multi-cohort study that includes information on individual students (e.g., academic self-concept, standardized achievement, socioeconomic background), students' learning environments (i.e., class identifiers that enable us to build reliable achievement aggregates), and students' socioeconomic neighborhood conditions (e.g., social status, income, employment). With its interdisciplinary orientation, the NEPS allows for the unique possibility to separately and simultaneously study educational and residential student environments as frames of reference for academic self-concept formation. We chose a rather exploratory approach, using different

indicators of socioeconomic neighborhood conditions and examining their effects on academic self-concept in different domains and grade levels.

More specifically, our study contributed in three ways to the literature. First, we replicated the traditional BFLPE model by analyzing the effects of class-average achievement on students' academic self-concept.

Second, we analyzed the predictive power of socioeconomic neighborhood composition for students' academic self-concept. Based on previous research, two potential patterns of results are plausible. On the one hand, if academic self-concept is impacted by collective socialization processes in neighborhoods, advantageous socioeconomic neighborhood conditions should positively predict students' academic self-concept. This pattern of results has been observed in several sociological neighborhood effect studies; however, the focus of these studies was on other educational outcomes such as performance, transition probabilities, or educational aspirations. On the other hand, if academic self-concept is impacted by social comparison—or, in sociological terms, relative deprivation processes—it should be negatively predicted by advantageous socioeconomic neighborhood conditions. This pattern is supported by research on the BFLPE that showed academic self-concept to be highly susceptible to social comparison processes.

Third, we analyzed the combined effects of both classroom and socioeconomic neighborhood composition. As learning environments are often composed according to residential criteria, students from neighborhoods with advantageous socioeconomic neighborhood conditions might attend educational environments with high average achievement, confounding influences from both sources. Thus, without controlling for both neighborhood and classroom composition, classroom effects might erroneously be attributed to the level of neighborhoods—and vice versa. Consequently, the simultaneous consideration of both students' (a) scholastic learning environment and (b) neighborhood as a non-school based but educationally relevant environment will provide further insight into the mechanisms of respective frame-of-reference effects. Two patterns of results are plausible: First, it may be that the joint consideration of both student environments will result in a disappearance of neighborhood effects. This result might indicate that neighborhood effects in Research Question 2 could be disguised class effects. Second, it may be that the joint consideration of both student environments will result in two independent contextual effects. This result might indicate the existence of social comparison processes within neighborhoods that have not yet been accounted for in research on the BFLPE.

Method

Data

We used data from the Starting Cohort 3 (SC3) of the German National Educational Panel Study (NEPS; Blossfeld et al., 2011). This study sought to establish a representative sample of all children attending fifth grade in Germany in school year 2010-2011. SC3 was drawn using a multistage sampling procedure that sampled schools as a first step and selected all students from two classes of each school in a second step (Skopek et al., 2012). Students in SC3 were followed along their educational career through secondary education. At the time of the study, the majority of the German federal states sorted their students into one of three different school types, namely "Hauptschule" (low track), "Realschule" (intermediate track), and "Gymnasium" (high track), and but also "Gesamtschulen" (comprehensive schools). In comprehensive schools, students were tracked within schools or even within classes into different educational tracks or were not explicitly tracked at all. Some other federal states employed a dyadic system with only comprehensive school and the Gymnasium. Tracks differed in their curriculum, with the high track being the most ambitious school type, preparing students for entering higher education (for a more detailed description of the German educational system, see Hübner, 2017). In NEPS SC3, students' academic self-concept was assessed in wave 1 (students in Grade 5) and wave 5 (students in Grade 9). In Germany, Grades 5 and 9 are important stages of individual educational careers as they are the beginning of secondary education and the end of compulsory education, respectively. The total Grade 5 sample contained 5,778 students. In our multilevel framework, cases could be taken into account only if they could be assigned to a class and a neighborhood. Thus, we had to exclude 1,872 students for whom identifiers for class or neighborhood membership were missing. This resulted in a sample of 3,906 students (48.42% female) that were nested in 234 schools, 466 classes, and 2,617 neighborhoods. The total Grade 9 sample comprised 5,778 students. Following the same procedure used to create the Grade 5 sample, we excluded 2,501 students for whom identifiers for class or neighborhood membership were missing in Grade 9. This resulted in a sample of 3,277 students (50.60% female) nested in 247 schools, 597 classes, and 2,314 neighborhoods.

Instruments

Academic Self-Concept

General self-concept (e.g., *I learn fast in most of the school subjects)*, math self-concept (e.g., *I have always been good at math*), and German self-concept (*I learn fast in German*) were

assessed by three items each (for the exact wording of all academic self-concept items, see Table S1 in the supplementary material; for a detailed description of the self-concept instrument, refer to Wohlkinger et al., 2016). Each academic self-concept item was answered on a 4-point Likert scale ranging from *does not apply at all* to *applies completely*. For subsequent analyses, a mean score comprising these items was constructed (at least two items had to be completed for mean score calculation). Cronbach alphas were α_{g5} = .83 and α_{g9} = .84 for general self-concept, α_{g5} = .87 and α_{g9} = .89 for math self-concept, and α_{g5} = .75 and α_{g9} = .82 for German self-concept. As the three academic self-concept scales each contained one grade-related academic self-concept item ("I do well in written class tests", "I get good grades in math", "I get good grades in German") that might measure school grades rather than academic self-concept, we additionally conducted all analyses with mean scores excluding these items (see Robustness Checks and Additional Analyses section in the Results as well as Tables S2–S3 in the supplementary material).

Academic Achievement

Mathematics academic achievement was assessed with a mathematics competency test that was based on the German Mathematics Education Standard framework as well as the PISA assessment framework (Neumann, 2013). WLE reliability was .778 in Grade 5 and .812 in Grade 9 (for detailed technical information see Duchhardt & Gerdes, 2012; van den Ham et al., 2018). German achievement was computed by averaging achievement estimates from a reading and orthography test. Reading achievement was assessed by a competency test based on the literacy-oriented PISA framework (Gehrer et al., 2013; OECD, 2009). WLE reliability was .767 in Grade 5 and .787 in Grade 9 (for detailed technical information see Pohl et al., 2012; Scharl et al., 2017). The orthography competency test is described in detail by Blatt et al. (2017). EAP/PV reliability was .963 in Grade 5 and .941 in Grade 9. General academic achievement was computed by averaging mathematics and German academic achievement. For more information on the academic achievement measures see also the detailed methods section in the supplemental material.

Socioeconomic Neighborhood Composition

Within the NEPS framework, neighborhood characteristics are provided by the commercial company *microm consumer marketing* (Schönberger & Koberg, 2017). We used neighborhood characteristics on the postal code 8 (PLZ8) level. The PLZ8 system divides geographical space into neighborhoods comprising on average 500 households. As a first measure of socioeconomic neighborhood conditions, we used a composite social status index.

It is computed based on information about the distribution of both academic titles and occupations in PLZ8 neighborhoods and is measured on a scale from 1 to 9 (1: lowest status, 2: far below average, 3: below average, 4: slightly below average, 5: average … 9: highest status). A second indicator of socioeconomic neighborhood conditions is the average income level in the neighborhood which is measured by the purchasing power per household measured in Euros (average net income). Purchasing power for PLZ8 neighborhoods is based on purchasing power on the municipality level and calculated with the help of statistical models accounting for several PLZ8 characteristics (e.g., age, status, etc.). As a third measure of socioeconomic neighborhood conditions, we used the employment rate in the neighborhood (proportion of employed people in relation to the total amount of potentially working people). Unemployment rates for PLZ8 neighborhoods were retrieved from the German Federal Employment Agency. We subtracted the unemployment variable from 1, resulting in the rate of neighborhood residents who are employed. Thus, all neighborhood composition variables were coded in such a way that higher values represented more advantageous socioeconomic neighborhood conditions.

Individual Socioeconomic Background

To control for socioeconomic background on the individual level, we retrieved individual information on social status, income, and employment from the parental questionnaire of SC3. Social status was operationalized as the highest ISEI (level of occupations according to an international standard classification) between both parents (Ganzeboom et al., 1992; Ganzeboom, 2010). In case of missing information for one parent, the information for only the remaining parent was used. Income was measured by the monthly household income after deductions and was surveyed by an open question. Employment was a dichotomous variable (0 for unemployed, 1 for employed). The unemployed group was composed of individuals of whom at least one of the parents received unemployment benefits. The employed group was composed of individuals of whom neither of the parents received unemployment benefits.

Covariates

All analyses were controlled for federal state and school type.

Analyses

The focus of our analyses was to individually as well as simultaneously analyze the effects of classroom and neighborhood composition on students' academic self-concept. In our analysis, we modeled individuals' (*i*) membership in classes (*j*) and neighborhoods (*k*), the latter

two presenting cross-classified factors (for a graphical depiction of the data structure see Figure 1). Thus, we specified cross-classified multilevel models (Hox et al., 2017). In all models, we controlled for federal state, and school type of students.

Generally, all analyses were run in Mplus 8 (Muthén & Muthén, 1998-2018). In Mplus, cross-classified multilevel models are estimates using Bayesian analysis. Thereby Mplus outputs a one-tailed *p*-value based on the posterior distribution. For a positive estimate, the *p*-value is the proportion of the posterior distribution that is below zero. For a negative estimate, the *p*-value is the proportion of the posterior distribution that is above zero (Muthen, 2010). Individual-level and neighborhood-level variables were standardized, and the class-average achievement aggregates were calculated using the standardized individual-level measures. This procedure allows for interpreting higher level effects as contextual effects that are effects of aggregated variables and that are controlled for the same variable on the individual level (Enders & Tofighi, 2007).

To replicate the traditional BFLPE model (Model 1), we regressed academic self-concept on class-average achievement controlling for individual academic achievement:

$$Self\text{-}concept_{i(j,k)} = \gamma_{00} + \gamma_{10} \cdot achievement_{i(j,k)} + \gamma_{01} \cdot class\text{-}average\ achievement_{k} + u_{0j} + v_{0k} + e_{i(j,k)} \quad (1)$$

To analyze the predictive power of socioeconomic neighborhood composition for students' academic self-concept (Model 2), we regressed academic self-concept on socioeconomic neighborhood composition while controlling for both individual academic achievement and individual socioeconomic background:

$$Self\text{-}concept_{i(j,k)} = \gamma_{00} + \gamma_{10} \cdot achievement_{i(j,k)} + \gamma_{01} \cdot socioeconomic\ neighborhood\ composition_{k} + \gamma_{20} \cdot individual\ socioeconomic\ background_{i(j,k)} + u_{0j} + v_{0k} + e_{i(j,k)} \quad (2)$$

For the simultaneous consideration of the class and the neighborhood (Model 3), we regressed academic self-concept on socioeconomic neighborhood composition and class-average achievement, while controlling for individual academic achievement as well as for individual socioeconomic background:

$$Self\text{-}concept_{i(j,k)} = \gamma_{00} + \gamma_{10} \cdot achievement_{i(j,k)} + \gamma_{01} \cdot socioeconomic\ neighborhood\ composition_{k} + \gamma_{20} \cdot individual\ socioeconomic\ background_{i(j,k)} + \gamma_{05} \cdot class\text{-}average\ achievement + v_{0k} + e_{i(j,k)} \quad (3)$$

Missing data rates for academic self-concept and achievement variables were low (between 0% and 3%). Due to parent non-response, missing rates for individual socioeconomic background variables were higher (between 6% and 45%). Missing values were accounted for by using the full-information maximum likelihood procedure (FIML; Enders, 2010; Graham, 2009). In Model 1, we included individual socioeconomic background and socioeconomic neighborhood composition as auxiliary variables. In Model 2, we included class-average achievement as an auxiliary variable. Thus, all Models (1-3) contained the same information (Graham, 2003).

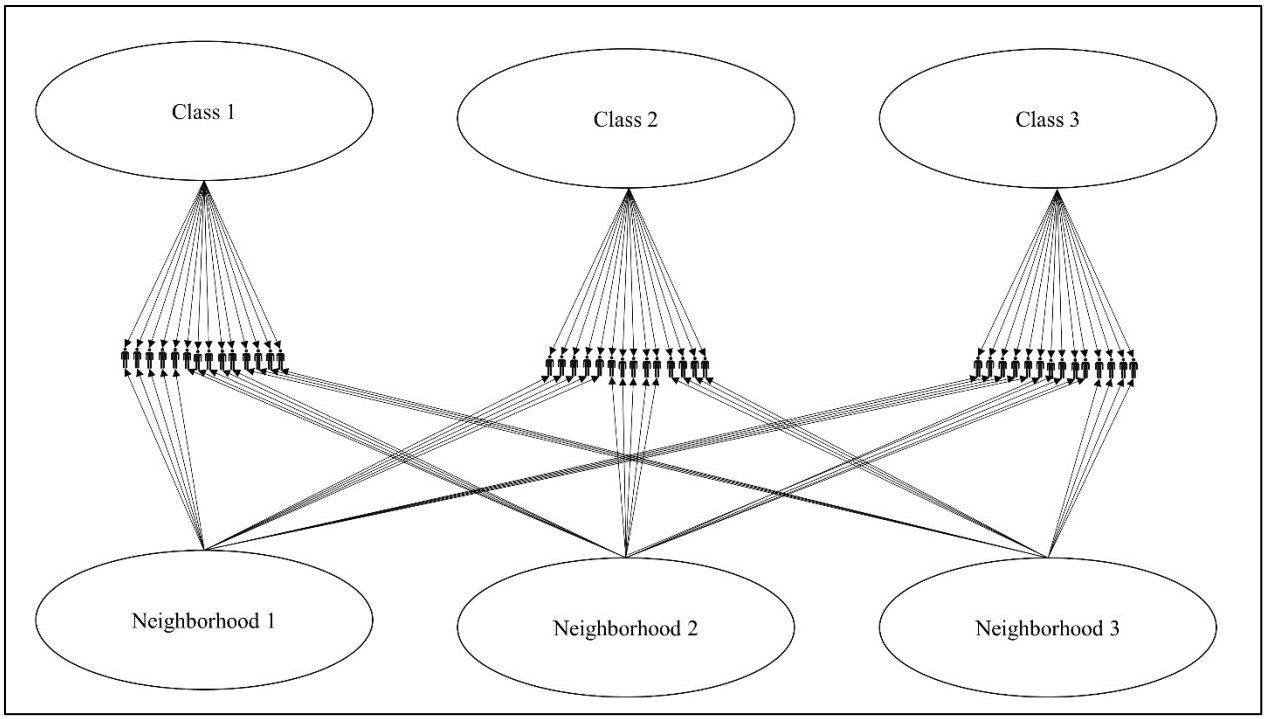

Figure 1: Graphical depiction of data structure.

Results

Descriptive Statistics

We present descriptive statistics for the Grade 5 and Grade 9 samples in Tables 1 and 2, respectively. Generally, the correlation pattern between achievement and self-concept variables was in line with earlier research (for a metaanalysis, see Möller et al., 2009). We found the typically weak correlation between mathematics and German self-concept (r_{G5} = .06/r_{G9} = -.06) as well as moderate correlations between domain-specific achievement and self-concept measures (mathematics: r_{G5} = .28/r_{G9} = .37; German: r_{G5} = .35/r_{G9} = .35). In addition to the reliability measures that we presented earlier, this finding is further evidence for the validity of our self-concept and achievement measures. Variance proportions that resulted from the variance decomposition in an "empty" random intercept model with students nested in classrooms and neighborhoods can also be found in Tables 1 and 2. In Grade 5, self-concept variables mainly varied on the individual level with variance proportions on the other levels being small. This finding suggests that classrooms as well as neighborhoods did not differ a lot regarding mean levels of academic self-concept. Achievement measures varied on the individual as well as the class level with only small variability on the neighborhood level. This finding suggests that classrooms but not neighborhoods differed concerning their academic achievement. The pattern of the variance proportions in Grade 9 was similar to that in Grade 5.

Neighborhood Variables

To gain insight into the neighborhood variables' validity, we took a closer look at their descriptive statistics. Neighborhood social status, which was based on both academic titles and occupations in PLZ8 neighborhoods, was on average M_{G5} = 5.29 and M_{G9} = 5.25 (SD_{G5} = 2.42/SD_{G9} = 2.36), measured on a scale from 1 (lowest) to 9 (highest). Our data indicate that the observed neighborhood status in our sample was slightly above the German average of 5. Neighborhood income was on average M_{G5} = 43,810 € and M_{G9} = 45,070 € (SD_{G5} = 8,930/SD_{G9} = 9,500), and neighborhood employment was on average M_{G5} = 94.01% and M_{G9} = 94.11% (SD_{G5} = 4.82% /SD_{G9} = 4.60%). The neighborhood variables correlated weakly with the academic self-concept measures (*r*s between <.01 and .06), whereas associations with academic achievement were considerably larger (*r*s between .13 and .24). This means that although students from advantageous neighborhoods had higher academic achievement they did not necessarily report a higher academic self-concept.

Expectedly, neighborhood social status was correlated with individual social status by r_{G5} = .30/r_{G9} = .31. Associations between neighborhood income and individual income were

r_{G5} = .12/r_{G9} = .10. Neighborhood employment was correlated with individual employment by r_{G5} = .29/r_{G9} = .20. These results show that the neighborhood measures overlap with respective measures on the individual level and that controlling for them in successive analyses is necessary. Correlations between the three neighborhood variables ranged from *r* = .64 to *r* = .69 suggesting a considerable overlap between the measures.

The BFLPE Model

To replicate the traditional BFLPE model we regressed academic self-concept on class-average academic achievement, controlling for individual achievement (Model 1; results can be found in Table 3 for the fifth-grade sample and in Table 4 for the ninth-grade sample). As expected, a student's academic achievement positively predicted his or her self-concept outcomes. This achievement effect was more pronounced in Grade 9 (coefficients ranging from *b* = .47 to *b* = .56 depending on the domain, all *p*s < .001) as opposed to in Grade 5 (coefficients ranging from *b* = .26 to *b* = .43 depending on the domain, all *p*s < .001). Additionally, class-average achievement negatively predicted respective self-concept outcomes (coefficients ranging from *b* = -.18 to *b* = -.24 depending on the domain as well as the grade level, all *p*s < .001). Thus, an increase of one standard deviation in class-average achievement was associated with a respective decrease of .18 to .24 standard deviations in academic self-concept, when controlling for the covariates. Hence, we replicated the typical findings from BFLPE studies that equally able students have lower academic self-concept in high-achieving classes; this pattern of results is typically interpreted as a consequence of social comparison processes in the classroom.

The Neighborhood Effects Model

To examine how socioeconomic neighborhood composition predicts students' academic self-concept, we regressed academic self-concept on the three neighborhood variables in separate models, controlling for individual achievement and respective social background on the individual level (Models 2a–2c; results can be found in Table 3 for the fifth-grade sample and in Table 4 for the ninth-grade sample).

In Grade 5, general academic self-concept was negatively predicted by neighborhood social status (Model 2a: *b* = -.07, *p* < .001) meaning that an increase of one standard deviation in neighborhood status was associated with a decrease of .07 standard deviations in general academic self-concept, when controlling for the other variables in the model. General academic self-concept was also negatively predicted by neighborhood employment (Model 2c: *b* = -.05, *p* = .004), but not by neighborhood income (Model 2b: *b* = -.02, *p* = .198). For mathematics

self-concept, the results were similar with the addition of a statistically significant neighborhood income effect. Neighborhood social status (Model 2a: b = -.08, p < .001), income (Model 2b: b = -.03, p = .032), and employment (Model 2c: b = -.06, p < .001) were negatively associated with math self-concept. For German self-concept, neighborhood effects were generally negative. However, none of the neighborhood effects were statistically different from zero.

In Grade 9, we did not find any neighborhood effects on general and German self-concept. However, neighborhood status (Model 2a: b = -.04, p = .026), as well as neighborhood employment (Model 2c: b = -.05, p = .002) negatively predicted mathematics self-concept. These results suggest that neighborhoods do not impact academic self-concept via mechanisms of collective socialization but rather by social comparison processes—or, in sociological terms, relative deprivation.

Simultaneous Consideration of Both the Class and the Neighborhood: The Combined Model

To examine how class-average achievement and socioeconomic neighborhood composition simultaneously predict academic self-concept, we regressed academic self-concept on the three neighborhood variables in separate models, controlling for individual and class-average achievement as well as respective social background on the individual level (Models 3a–3c; results can be found in Table 3 for the fifth-grade sample and in Table 4 for the ninth-grade sample). The simultaneous consideration of both the class and the neighborhood is especially important against the background of socioeconomic neighborhood composition being correlated with class-average achievement (as reported in the Results section "Neighborhood Variables"). Thus, a neighborhood effect in the preceding neighborhood effects model might be the result of students from advantageous neighborhoods attending high-achieving classes.

Additionally modeling the neighborhood only slightly impacted the class-level BFLPEs. In Grade 5, general academic self-concept was still negatively predicted by neighborhood status (Model 3a: b = -.06, p < .001; in Model 2a it was b = -.07, p < .001). The effect of neighborhood employment was still negative but not significantly different from zero anymore (Model 3c: b = -.03, p = .088; in Model 2c it was b = -.05, p = .004). The same was true when considering mathematics self-concept. The negative effect of neighborhood status remained (Model 3a: b = -.05, p = .004; in Model 2a it was b = -.08, p < .001), whereas the effect of neighborhood employment was still negative but no longer significantly different from

zero (Model 3c: b = -.05, p = .098; in Model 2c it was b = -.06, p < .001). The fact that neighborhood social status negatively predicted general and math self-concept, even after controlling for class-average achievement might indicate that there are "direct" social comparison processes within neighborhoods. The fact that neighborhood employment did not predict general and math self-concept, after controlling for class-average achievement, might indicate that neighborhood effects found in Models 2a and 2c were disguised class effects. Students living in advantageous neighborhoods attend high-achieving classes which negatively impacts academic self-concept.

In Grade 9, considering mathematics self-concept, both the effects of neighborhood status (Model 3a: b = -.02, p = .290; in Model 2a it was b = -.04, p = .026) and neighborhood employment (Model 3c: b = -.02, p = .242; in Model 2c it was b = -.05, p = .002) did not statistically significantly differ from zero in the model that controlled for class-average achievement. Similar to the Grade 5 results this might imply that the neighborhood effects found in Models 2a and 2c were disguised class effects.

The overall pattern of results can be summed up as follows: We found no positive neighborhood effects that statistically significantly differed from zero. On average, neighborhood effects were negative and small. Neighborhood variables were more predictive for students' general and math self-concept as opposed to German self-concept. In Grade 5, neighborhood variables more negatively predicted students' academic self-concept as opposed to in Grade 9. And neighborhood effects were stronger for social status and employment than for income.

Robustness Checks and Additional Analyses

As described in the Instruments section, the NEPS academic self-concept scales each contain one item that explicitly refers to school grades (e.g., "*I get good grades in math*"; for the exact wording of all items, see Table S1 in the supplementary material). Thus, we additionally conducted all analyses with academic self-concept mean scores in which these grade-related items were excluded (see Tables S2 and S3 in the supplementary material). This did not change the pattern of results.

Additionally, we also conducted all analyses using a socioeconomic neighborhood conditions composite score that was the mean of the three neighborhood characteristics. In the respective analyses, we regressed academic self-concept on this composite score and controlled for individual and class-average achievement as well as respective social background, income, and employment on the individual level. In the Grade 5 sample (Table S4), we found that the

socioeconomic neighborhood conditions composite score negatively predicted general and math self-concept, whether class-average achievement was controlled (Model 3; general: b = -.05, p = .004; math: b = -.04, p < .001) or not (Model 2; general: b = -.06, p = .002; math: b = -.05, p < .001). The socioeconomic neighborhood conditions composite score negatively predicted German self-concept only when not controlling for class-average achievement (Model 2; b = -.04, p = .028); the composite score was not statistically significant when controlling for class achievement (Model 3; b = -.03, p = .102). In the Grade 9 sample (Table S5), the socioeconomic neighborhood conditions composite score negatively predicted math self-concept only when not controlling for class-average achievement (Model 2; b = -.05, p = .020) and not when controlling for class-average achievement (Model 3; b = -.02, p = .206). General and German self-concept were not predicted by the socioeconomic neighborhood conditions composite score in Grade 9.

Table 1

Descriptive Statistics of Model Variables in Grade 5 Sample

	Mis	M	SD	VP_i	VP_c	VP_n	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Self-concept general	.03	3.16	0.60	.94	.05	.01														
2. Self-concept math	.03	2.93	0.85	.93	.06	.01	.36													
3. Self-concept German	.03	3.01	0.66	.93	.05	.02	.52	.06												
4. Achievement general	.00	0.04	1.04	.47	.52	.01	.22	.18	.28											
5. Achievement math	.00	0.06	1.16	.55	.42	.03	.16	.28	.16	.91										
6. Achievement German	.00	0.02	1.12	.54	.45	.01	.24	.05	.35	.91	.66									
7. Class achievement general	.00	0.04	0.77				.14	.11	.18	.74	.67	.68								
8. Class achievement math	.00	0.06	0.79				.13	.12	.17	.73	.69	.64	.98							
9. Class achievement German	.00	0.02	0.79				.15	.09	.19	.73	.62	.70	.98	.91						
10. Status	.12	53.06	16.62	.70	.21	.09	.12	.04	.14	.35	.31	.32	.40	.39	.39					
11. Income	.22	3.62	3.35	.94	.04	.02	.06	.03	.04	.16	.14	.16	.19	.19	.19	.27				
12. Employment	.09	0.92	0.28	.64	.30	.06	.04	.02	.08	.26	.23	.24	.30	.29	.30	.24	.18			
13. Neighborhood status	.00	5.29	2.42				.01	.01	.05	.24	.22	.22	.28	.26	.28	.30	.17	.23		
14. Neighborhood income	.00	43.81	8.93				.02	.04	.02	.16	.14	.16	.18	.16	.19	.18	.12	.19	.69	
15. Neighborhood employment	.00	94.09	4.82				.01	.02	.03	.24	.22	.23	.28	.26	.28	.21	.13	.29	.68	.65

Note. Mis is the percentage of missing data. VP_i, VP_c, and VP_n are variance proportions on the individual, the class, and the neighborhood level, respectively.

Table 2

Descriptive Statistics of Model Variables in Grade 9 Sample

	Mis	M	SD	VP_i	VP_c	VP_n	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Self-concept general	.01	2.91	0.59	.95	.03	.01														
2. Self-concept math	.01	2.52	0.90	.94	.04	.02	.35													
3. Self-concept German	.01	2.97	0.61	.88	.10	.02	.42	-.06												
4. Achievement general	.00	0.10	0.96	.40	.55	.05	.30	.25	.24											
5. Achievement math	.00	0.11	1.20	.52	.43	.05	.25	.37	.12	.89										
6. Achievement German	.00	0.11	1.22	.45	.48	.06	.27	.05	.35	.85	.59									
7. Class achievement general	.00	0.10	0.76				.15	.10	.17	.79	.69	.71								
8. Class achievement math	.00	0.10	0.87				.14	.15	.12	.75	.73	.62	.95							
9. Class achievement German	.00	0.11	0.92				.14	.04	.20	.74	.60	.76	.94	.82						
10. Status	.36	54.27	16.53	.67	.24	.09	.13	.06	.13	.37	.33	.32	.42	.39	.39					
11. Income	.45	3.74	3.75	.86	.08	.06	.05	.04	.05	.13	.12	.12	.17	.16	.18	.25				
12. Employment	.19	0.96	0.21	.86	.08	.06	.03	.03	.03	.14	.12	.12	.19	.18	.18	.19	.12			
13. Neighborhood status	.00	5.25	2.36				.05	.02	.03	.24	.22	.21	.29	.29	.26	.31	.13	.18		
14. Neighborhood income	.00	45.07	9.50				.06	.04	.01	.17	.16	.13	.19	.20	.17	.19	.10	.14	.68	
15. Neighborhood employment	.00	94.11	4.60				.05	.02	.01	.21	.20	.17	.24	.26	.21	.20	.10	.20	.68	.64

Note. Mis is the percentage of missing data. VP_i, VP_c, and VP_n are variance proportions on the individual, the class, and the neighborhood level, respectively.

Table 3

Results From Cross-Classified Multilevel Models in Grade 5

	Model 1		Model 2a		Model 2b		Model 2c		Model 3a		Model 3b		Model 3c	
General	b	p	b	p	b	p	b	p	b	p	b	p	b	p
Achievement	.26	<.001	.23	<.001	.23	<.001	.24	<.001	.26	<.001	.26	<.001	.26	<.001
Class achievement	-.19	<.001							-.17	.002	-.18	<.001	-.16	<.001
Status			.06	.004					.06	.002				
Neighborhood status			-.07	<.001					-.06	<.001				
Income					.02	.268					.02	.210		
Neighborhood income					-.02	.198					-.02	.244		
Employment							.02	.802					.08	.288
Neighborhood employment							-.05	.004					-.03	.088
Math														
Achievement	.37	<.001	.34	<.001	.34	<.001	.35	<.001	.37	<.001	.37	<.001	.37	<.001
Class achievement	-.23	<.001							-.19	<.001	-.22	<.001	-.18	<.001
Status			-.03	.130					-.03	.190				
Neighborhood status			-.08	<.001					-.05	.004				
Income					-.01	.800					.00	.914		
Neighborhood income					-.03	.032					-.01	.448		
Employment							-.08	.160					-.07	.318
Neighborhood employment							-.06	<.001					-.05	.098
German														
Achievement	.43	<.001	.39	<.001	.40	<.001	.40	<.001	.42	<.001	.43	<.001	.43	<.001
Class achievement	-.18	<.001							-.17	<.001	-.16	.004	-.17	.002
Status			.04	.020					.04	.016				
Neighborhood status			-.03	.090					-.03	.152				
Income					-.01	.474					-.01	.498		
Neighborhood income					-.03	.136					-.01	.378		
Employment							.05	.432					.09	.080
Neighborhood employment							-.03	.080					-.02	.254

Note. The dependent variable is academic self-concept. All analyses were controlled for federal state and school type.

Table 4

Results From Cross-Classified Multilevel Models in Grade 9

	Model 1		Model 2a		Model 2b		Model 2c		Model 3a		Model 3b		Model 3c	
General	b	p	b	p	b	p	b	p	b	p	b	p	b	p
Achievement	.47	<.001	.40	<.001	.41	<.001	.41	<.001	.47	<.001	.47	<.001	.48	<.001
Class achievement	-.21	<.001							-.22	<.001	-.22	<.001	-.23	<.001
Status			.05	.016					.05	.012				
Neighborhood status			-.03	.074					-.02	.306				
Income					.02	.412					.02	.378		
Neighborhood income					.01	.652					.02	.396		
Employment							-.07	.340					.01	.892
Neighborhood employment							-.01	.632					.01	.492
Math														
Achievement	.56	<.001	.51	<.001	.51	<.001	.51	<.001	.58	<.001	.58	<.001	.58	<.001
Class achievement	-.24	<.001							-.32	<.001	-.32	<.001	-.32	<.001
Status			-.01	.670					.00	.994				
Neighborhood status			-.04	.026					-.02	.290				
Income					.02	.300					.02	.224		
Neighborhood income					-.03	.150					.00	.814		
Employment							-.03	.696					.05	.508
Neighborhood employment							-.05	.002					-.02	.242
German														
Achievement	.48	<.001	.43	<.001	.43	<.001	.43	<.001	.48	<.001	.48	<.001	.48	<.001
Class achievement	-.20	<.001							-.29	<.001	-.29	<.001	-.29	<.001
Status			.02	.250					.03	.102				
Neighborhood status			-.03	.112					-.02	.484				
Income					.01	.638					.02	.500		
Neighborhood income					-.01	.680					.01	.600		
Employment							.00	.984					.05	.644
Neighborhood employment							-.01	.494					.01	.780

Note. The dependent variable is academic self-concept. All analyses were controlled for federal state and school type.

Discussion

In the present study, we separately and simultaneously analyzed the effects of classroom and neighborhood composition on students' academic self-concept. Our results can be summarized as follows: First, in line with research on the BFLPE, we found classroom selectivity, operationalized by class-average achievement, to negatively predict academic self-concept. In other words, equally able students had lower self-concept in high-achieving classrooms. This frame-of-reference effect is typically interpreted as a consequence of social comparison processes in informal educational settings. Second, in supposed contrast to the bulk of the neighborhood effects literature, we found neighborhood selectivity, operationalized by socioeconomic neighborhood composition, not to positively predict academic self-concept. Quite the contrary, neighborhood status and employment rate negatively predicted math self-concept in Grade 5 and Grade 9; in the former grade, these neighborhood measures negatively predicted general self-concept as well. Third, when simultaneously analyzing the effects of classroom and neighborhood composition, math and general self-concept in Grade 5 were negatively predicted by neighborhood status, whereas the other neighborhood effects were no longer statistically significant. Class-average achievement continued to be a strong negative predictor of academic self-concept (the BFLPE).

Our study's unique contribution to the literature on academic self-concept formation is that it investigates the neighborhood as a non-institutional student environment for academic self-concept formation. On a theoretical level, our findings suggest that students' academic self-concept may result from social comparison processes within not only classrooms but also neighborhoods. At the same time, our study contributes to the literature on neighborhood effects research by focusing on an educational outcome that is highly susceptible to social comparison processes. On a theoretical level, our findings suggest that academic self-concept is an educational outcome that might be negatively impacted by neighborhood selectivity.

Heterogeneity of Neighborhood Effects

As our study—to our knowledge—was the very first one to examine how socioeconomic neighborhood composition predicts students' academic self-concept, we chose an exploratory approach and investigated neighborhood effects on different self-concept domains (general, math, German), using different indicators for socioeconomic neighborhood composition (social status, income, employment) within different grade levels (Grade 5, Grade 9). Mathematics self-concept was the domain that turned out to be most susceptible to neighborhood effects. This may have been because mathematics represents a domain that

students perceive to be of crucial importance for intellectual ability. Additionally, mathematics might be the domain in which student achievement is the most salient. We did not find neighborhood effects on German self-concept at all. This may have been caused by a lower reliability of the respective scale, which included one item ("In the subject German, I am a hopeless case") that caused a drop in construct reliability. Among the three indicators of socioeconomic neighborhood composition, neighborhood status was most strongly associated with students' academic self-concept. This may be due to this measure depending on the distribution of academic titles and occupations in the neighborhood, thus presenting the best approximation of the intellectual capacity of a neighborhood. Additionally, neighborhood effects were more prevalent in Grade 5 as opposed to Grade 9. Such a finding is not uncommon in the neighborhood effects literature as the effect of different neighborhood features might vary with age (e.g., Ellen & Turner, 1997; Sharkey & Faber, 2014; van Ham & Tammaru, 2016; Wheaton & Clarke, 2003).

Potential Mechanisms for Negative Neighborhood Effects on Academic Self-Concept

In general, our findings of negative neighborhood effects on students' academic self-concept call for a more elaborate discussion of the hypothesized underlying mechanisms which, of course, can only be theorized within the obvious limitations of a study that is correlational by design. First, some of the neighborhood effects we found vanished when additional controls for class achievement were included in our analytical model. This finding suggests that at least these neighborhood effects might have been hidden classroom effects. Because school classes are often composed according to local criteria, students who live in neighborhoods with advantageous socioeconomic conditions have a higher likelihood to end up in high-achieving classes and consequently experience a decline in their academic self-concept in terms of BFLPEs. Second, as some of the neighborhood effects remained even when controlling for class achievement, these effects might indeed reflect social comparison processes within the neighborhood. As stated above, children living in a neighborhood in which the majority of children commute to a high track school might have a lower academic self-concept compared to equally able children living in neighborhoods in which the majority of children commute to a low track school as between-school tracking conveys to them a notion of the academic capabilities of their neighborhood. Beyond that, other, potentially less apparent mechanisms might be driving the neighborhood effects we found. For example, the effects we found in the fifth-grade sample might have been a residual effect of primary education. Academic self-concept was measured 2 to 5 months after students entered secondary education and might have been impacted by elementary school class composition, which usually represents students'

neighborhood composition to a much stronger degree than secondary education does. In other words, equally able students might have reported lower academic self-concept in high-SES neighborhoods because they attended a high-achieving class in elementary school. In technical terms, this means that we might not have found negative neighborhood effects in Grade 5 if we had also controlled for class-average achievement in elementary school. On the other hand, this potential objection is weakened by a recent study by Becker and Neumann (2018) which showed that BFLPEs on domain-specific academic self-concept fade away in the transition from primary to secondary education. Yet, given our limited observation window, it remains an open question and a corresponding direction for future research which mechanism(s) are actually driving our particular result.

Local Dominance Theory

In previous research on the BFLPE, the local class environment was observed to be the pivotal frame of reference for academic self-concept formation (in contrast to the more global school environment; Marsh et al., 2014). This finding was explained as the local dominance effect (Zell & Alicke, 2010), that is, the tendency of individuals to use proximal comparison information for ability self-evaluations. The neighborhood presents another, non-scholastic environment to which children and adolescents are directly exposed in everyday life, but which has so far never been tested as a potential frame of reference for academic self-concept formation. Depending on both the particular domain of student academic self-concept and their grade level, our empirical analyses support our main argument that students' neighborhood can constitute an additional frame of reference for academic self-concept formation. Thus, our results suggest that students are capable of making use of several comparison standards at the same time, which once more underlines the fascinating complexity of academic self-concept formation.

Practical Implications

The neighborhood effects we observed were generally small in size (between b = -.04 and b = -.08), which mirrors findings of previous studies on neighborhood effects on other outcomes. Thus, one may argue that socioeconomic neighborhood conditions are not practically relevant for academic self-concept formation. On the other hand, the neighborhood effects we observed were still up to 50% the size of respective BFLPEs (which ranged between b = -.16 and b = -.32). Besides, our neighborhood-level indicators can be assumed to be only an approximation to the underlying constructs of interest. More accurately measuring socioeconomic neighborhood conditions (e.g., by averaging ISEIs of all neighborhood

inhabitants) might have led to even stronger negative neighborhood effects. Moreover, as neighborhood social polarization is less pronounced in European countries compared to, for example, the U.S., contrastive neighborhood effects on academic self-concept might be stronger in the latter context. Generally, our study does not propose a social stratification of neighborhoods to establish equality in students' academic self-concept. However, it offers an alternative perspective in that there might exist educational outcomes that are not or are even negatively impacted by socioeconomic neighborhood conditions. Thus social destratification of neighborhoods will not necessarily contribute to closing the gaps with regards to all educational outcomes.

Interdisciplinary Value of the Study

By predicting academic self-concept—an educational outcome that is typically considered in educational psychology—by indicators of socioeconomic neighborhood composition, our study integrated psychological social comparison theory and sociological neighborhood effects research. Thereby it calls attention to the considerable conceptual similarity of the social-psychological mechanisms described by different terminologies between the two disciplines. Contrastive frame-of-reference effects are the psychological counterpart to the sociological concept of relative deprivation. And assimilation effects have much in common with the sociological concept of collective socialization. We contributed to sociological neighborhood effects research by showing that advantageous socioeconomic neighborhood conditions do not positively impact all educational outcomes. In fact, advantageous socioeconomic neighborhood conditions might indeed negatively impact educational outcomes, especially those that are highly susceptible to social comparison processes. Although "relative deprivation" is discussed as a potential mechanism of neighborhood effects in the literature (see Galster, 2012), surprisingly few studies took a closer look at educational outcomes that might be negatively impacted by advantageous socioeconomic neighborhood conditions (for an exception see Turley, 2002). Additionally, in a number of our statistical models, we found neighborhood effects to be eradicated after controlling for class achievement. Thus, our study cautions researchers to carefully translate the theoretical neighborhood mechanism of interest into an adequate statistical multilevel model. An identification of neighborhood effects as "true" contextual effects, that is, effects as a consequence of direct neighborhood interaction or other forms of exposure, is possible only if compositional effects of all lower levels, for example, institutional effects operating within the school environment, are rigorously controlled for.

Limitations and Future Research

To the best of our knowledge, the present study was the first to investigate how socioeconomic neighborhood composition predicts students' academic self-concept. We drew on the German National Educational Panel Study (NEPS; Blossfeld et al., 2011), which was well suited for such an endeavor as it comprises information about students' academic self-concept and their socioeconomic neighborhood composition. Despite these advantages, some potential limitations should be addressed in the future.

First, the present study is based on non-experimental cross-sectional data. Consequently, causal interpretations of our results require caution. However, we explicitly modeled several possible confounders and have good reason to conclude that depending on the domain under evaluation as well as students' grade level, equally able students in equally able classes have lower academic self-concept in advantageous neighborhoods. Generally, field-experimental approaches in the research of neighborhood effects are not easily feasible and have been criticized for ethical reasons (Geronimus & Thompson, 2004). Also, laboratory experiments will be hardly able to model the complexity of simultaneous operating influences of student environments on academic self-concept. Nonetheless, future studies of the neighborhood as a potential frame of reference for academic self-concept formation should make use of natural experiments (e.g., analyze individuals' between-neighborhood mobility) or elaborated statistical methodologies that facilitate causal inference (e.g., instrumental variable approaches).

Second, we did not model schools as a distinct level of analysis. This was due to NEPS drawing only two classes from each school, making it hard to disentangle class and school effects. In particular, we were not able to additionally control for school achievement. Therefore, critics might argue that the neighborhood effects in our models are caused by school effects. On the other hand, experimental social comparison research assumes that proximal environments matter the most (Zell & Alicke, 2010) for academic self-concept formation. Marsh et al. (2014), as well as Liem et al. (2013), showed that the class environment represents the pivotal frame of reference for self-concept formation. As we also controlled for school type in our analyses, there are few reasons to believe that additional controls for school achievement would have substantially impacted our results.

Third, there are limitations in terms of the generalizability of our results to other countries and educational systems. Future research is needed to investigate neighborhood

effects on educational outcomes that might be susceptible to social comparison processes in non-European countries.

Conclusion

Our study showed that advantageous socioeconomic neighborhood conditions do not positively predict academic self-concept. Quite the contrary, depending on both the domain under evaluation and students' grade level, advantageous socioeconomic neighborhood conditions indeed negatively predicted academic self-concept when controlling for possible confounding variables. Our study advances educational psychological research by introducing an additional determinant of academic self-concept formation. By doing so, it suggests the neighborhood as being an additional frame of reference for academic self-concept formation. Complementarily, our study advances neighborhood effects research by explicitly investigating relative deprivation processes on students' academic self-concept as an educational outcome that has been neglected hitherto. Our results are especially interesting in the light of neighborhood effects research that generally reports advantageous socioeconomic neighborhood conditions to positively predict educational outcomes, but has not yet focused on educational outcomes that are highly susceptible to social comparison processes. Consequently, our study is important for educational planners and practitioners as it suggests that, depending on the outcome of interest, social destratification of neighborhoods will activate different mechanisms (e.g., collective socialization or social comparison processes/relative deprivation) that might not necessarily contribute to closing the gaps with regards to all educational outcomes.

References

Aaronson, D. (1998). Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. *The Journal of Human Resources*, *33*(4), 915–946. https://doi.org/10.2307/146403

Ainsworth, J. W. (2002). Why does it take a village? The mediation of neighborhood effects on educational achievement. *Social Forces*, *81*(1), 117–152. https://doi.org/10.1353/sof.2002.0038

Alwin, D. F., & Otto, L. B. (1977). High school context effects on aspirations. *Sociology of Education*, *50*(4), 259. https://doi.org/10.2307/2112499

Becker, M., & Neumann, M. (2018). Longitudinal big-fish-little-pond effects on academic self-concept development during the transition from elementary to secondary schooling. *Journal of Educational Psychology*, *110*(6), 882–897. https://doi.org/10.1037/edu0000233

Blatt, I., Jarsinski, S., & Prosch, A. (2017). *Technical Report for Orthography: Scaling results of Starting Cohort 3 in Grades 5, 7, and 9 (NEPS Survey Paper No. 15)*. Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SC3:5.0.0

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

Boardman, J. D., & Saint Onge, J. M. (2005). Neighborhoods and adolescent development. *Children, Youth and Environments*, *15*, 138–164.

Bowen, N. K., & Bowen, G. L. (1999). Effects of crime and violence in neighborhoods and schools on the school behavior and performance of adolescents. *Journal of Adolescent Research*, *14*(3), 319–342. https://doi.org/10.1177/0743558499143003

Brannstrom, L. (2008). Making their mark: The effects of neighbourhood and upper secondary school on educational achievement. *European Sociological Review*, *24*(4), 463–478. https://doi.org/10.1093/esr/jcn013

Brooks-Gunn, J., Duncan, G. J., Klebanov, P. K., & Sealand, N. (1993). Neighborhoods influence child and adolescent development? *American Journal of Sociology*, *99*(2), 353–395. https://doi.org/10.1086/230268

Canache, D. (1996). Looking out my back door: The neighborhood context and perceptions of relative deprivation. *Political Research Quarterly*, *49*(3), 547–571. https://doi.org/10.1177/106591299604900305

Catsambis, S., & Beveridge, A. A. (2001). Does neighborhood matter? Family, neighborhood, and school influences on eighth-grade mathematics achievement. *Sociological Focus*, *34*(4), 435–457. https://doi.org/10.1080/00380237.2001.10571212

Childress, H. (2016). Teenagers, territory and the appropriation of space. *Childhood*, *11*(2), 195–205. https://doi.org/10.1177/0907568204043056

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type. *American Educational Research Journal*, *50*(5), 925–957. https://doi.org/10.3102/0002831213489843

Cialdini, R. B., & Richardson, K. D. (1980). Two indirect tactics of image management: Basking and blasting. *Journal of Personality and Social Psychology*, *39*(3), 406–415. https://doi.org/10.1037/0022-3514.39.3.406

Crane, J. (1991). Effects of neighorhoods on dropping out of school and teenage childbearing. In C. Jencks (Ed.), *The urban underclass* (pp. 299–320). Brookings Institution.

Davis, J. A. (1966). The campus as a fog pod: An application of the theory of relative deprivation to career decisions of college men. *American Journal of Sociology*, *72*(1), 17–31. https://doi.org/10.1086/224257

Duchhardt, C., & Gerdes, A. (2012). *NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 3 in Fifth Grade (NEPS Working Paper No. 19)*. Otto-Friedrich-Universität, Nationales Bildungspanel.

Duncan, G. J., Brooks-Gunn, J., & Klebanov, P. K. (1994). Economic deprivation and early childhood development. *Child Development*, *65*(2), 296–318. https://doi.org/10.1111/j.1467-8624.1994.tb00752.x

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

Ellen, I. G., & Turner, M. A. (1997). Does neighborhood matter? Assessing recent evidence. *Housing Policy Debate*, *8*(4), 833–866. https://doi.org/10.1080/10511482.1997.9521280

Elliot, A. J., & Dweck, C. (2005). Competence and motivation. Competence as the core of achievement motivation. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of Competence and Motivation* (pp. 3–12). Guilford Publications.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Felson, R. B. (1984). The effect of self-appraisals of ability on academic performance. *Journal of Personality and Social Psychology*, *47*(5), 944–952. https://doi.org/10.1037/0022-3514.47.5.944

Felson, R. B., & Reed, M. D. (1986). Reference groups and self-appraisals of academic ability and performance. *Social Psychology Quarterly*, *49*(2), 103–109. https://doi.org/10.2307/2786722

Festinger, L. (1957). A theory of social comparison processes. *Human Relations*, *7*, 117–140. https://doi.org/10.1177/001872675400700202

Friedrichs, J., Galster, G., & Musterd, S. (2010). Neighbourhood effects on social opportunities: The European and American research and policy context. *Housing Studies*, *18*(6), 797–806. https://doi.org/10.1080/0267303032000156291

Galster, G. C. (2008). Quantifying the effect of neighbourhood on individuals: Challenges, alternative approaches, and promising directions. *Schmollers Jahrbuch*, *128*(1), 1–42. https://doi.org/10.3790/schm.128.1.7

Galster, G. C. (2012). The mechanism(s) of neighbourhood effects: Theory, evidence, and policy implications. In M. van Ham, N. Manley, L. Bailey, D. Simpson, & D. MacLennan (Eds.), *Neighbourhood effects research: New perspectives* (pp. 23–56). Springer.

Ganzeboom, H. B.G. (2010). *A new international socio-economic index (ISEI) of occupational status for the international standard classification of occupation 2008 (ISCO-08) constructed with data from 11 the ISSP 2002-2007*. Paper presented at the Annual Conference of International Social Survey Programme, Lisbon.

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

Garner, C. L., & Raudenbush, S. W. (1991). Neighborhood effects on educational attainment: A multilevel analysis. *Sociology of Education*, *64*(4), 251–262. https://doi.org/10.2307/2112706

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online*, 50–79.

Geronimus, A. T., & Thompson, J. P. (2004). To denigrate, ignore, or disrupt: Racial inequality in health and the impact of a policy-induced breakdown of african american communities. *Du Bois Review*, *1*(2), 612. https://doi.org/10.1017/S1742058X04042031

Göllner, R., Damian, R. I., Nagengast, B., Roberts, B. W., & Trautwein, U. (2018). It's not only who you are but who you are with: High school composition and individuals' attainment over the life course. *Psychological Science*, 1-12. https://doi.org/10.1177/0956797618794454

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

Guo, J., Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2015). Directionality of the associations of high school expectancy-value, aspirations, and attainment. *American Educational Research Journal*, *52*(2), 371–402. https://doi.org/10.3102/0002831214565786

Hartung, A., & Hillmert, S. (2019). Assessing the spatial scale of context effects: The example of neighbourhoods' educational composition and its relevance for individual aspirations. *Social Science Research*, *83*, 1–13. https://doi.org/10.1016/j.ssresearch.2019.05.001

Helbig, M. (2010). Neighborhood does matter! *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie*, *62*(4), 655–679. https://doi.org/10.1007/s11577-010-0117-y

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge Taylor & Francis Group.

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, *49*(5), 505–528. https://doi.org/10.1016/j.jsp.2011.07.001

Hübner, N. (2017). *Educational effectiveness at the end of upper secondary school: Further insights into the effects of statewide policy reforms*. Dissertation.

Huguet, P., Dumas, F., Marsh, H. W., Wheeler, L., Seaton, M., Nezlek, J., Suls, J., & Régner, I. (2009). Clarifying the role of social comparison in the big-fish-little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology*, *97*(1), 156–170. https://doi.org/10.1037/a0015558

Kintrea, K., St Clair, R., & Houston, M. (2015). Shaped by place? Young people's aspirations in disadvantaged neighbourhoods. *Journal of Youth Studies*, *18*(5), 666–684. https://doi.org/10.1080/13676261.2014.992315

Liem, G. A. D., Marsh, H. W., Martin, A. J., McInerney, D. M., & Yeung, A. S. (2013). The big-fish-little-pond effect and a national policy of within-school ability streaming. *American Educational Research Journal*, *50*(2), 326–370. https://doi.org/10.3102/0002831212464511

Marsh, H. W. (1984). Self-concept: The application of a frame of reference model to explain paradoxical results. *Australian Journal of Education*, *28*(2), 165–181. https://doi.org/10.1177/000494418402800207

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., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology*, *78*(2), 337–349. https://doi.org/10.1037//0022-3514.78.2.337

Marsh, H. W., Kuyper, H., Morin, A. J., Parker, P. D., & Seaton, M. (2014). Big-fish-little-pond social comparison and local dominance effects: Integrating new statistical models, methodology, design, theory and substantive implications. *Learning and Instruction*, *33*, 50–66. https://doi.org/10.1016/j.learninstruc.2014.04.002

Marsh, H. W., Martin, A. J., Yeung, A. S., & Craven, R. (2016). Competence self-perceptions. In C. Dweck & D. Yaeger (Eds.), *Handbook of competence and motivation.* Guilford Press.

Marsh, H. W., & Seaton, M. (2015). The big-fish–little-pond effect, competence self-perceptions, and relativity: Substantive advances and methodological innovation. *Advances in Motivation Science*, *2*, 127–184.

Marsh, H. W., & Yeung, A. S. (1997). Coursework selection: Relations to academic self-concept and achievement. *American Educational Research Journal*, *34*(4), 691–720. https://doi.org/10.3102/00028312034004691

Mayer, S. E., & Jencks, C. (1989). Growing up in poor neighborhoods: How much does it matter? *Science*, *243*(4897), 1441–1445. https://doi.org/10.1126/science.243.4897.1441

Meyer, J. W. (1970). High school effects on college intentions. *American Journal of Sociology*, *76*(1), 59–70. https://doi.org/10.1086/224906

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research*, *79*, 1129–1167. https://doi.org/10.3102/0034654309337522

Muthen, B. (2010). *Bayesian analysis in Mplus: A brief introduction*. (Technical Report). Version 3.

Muthén, L. K., & Muthén, B.O. (1998-2018). *Mplus user's guide*. Muthén & Muthén.

Neumann, M. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online, 5*, 80–109.

Nieuwenhuis, J., & Hooimeijer, P. (2016). The association between neighbourhoods and educational achievement, a systematic review and meta-analysis. *Journal of Housing and the Built Environment, 31*, 321–347. https://doi.org/10.1007/s10901-015-9460-7

Nieuwenhuis, J., van Ham, M., Yu, R., Branje, S., Meeus, W., & Hooimeijer, P. (2017). Being poorer than the rest of the neighborhood: Relative deprivation and problem cehavior of youth. *Journal of Youth and Adolescence*, *46*(9), 1891–1904. https://doi.org/10.1007/s10964-017-0668-6

OECD. (2009). *PISA 2009 assessment framework. Key competencies in reading, mathematics, and science*. OECD.

OECD. (2018). *The future of education and skills. Education 2030*. OECD.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grade (NEPS Working Paper No. 15)*. Otto-Friedrich-Universität, Nationales Bildungspanel.

Sampson, R. J., Morenoff, J. D., & Gannon-Rowley, T. (2002). Assessing "neighborhood effects": Social processes and new directions in research. *Annual Review of Sociology, 28*(1), 443–478. https://doi.org/10.1146/annurev.soc.28.110601.141114

Scharl, A., Fischer, L., Gnambs, T., & Rohm, T. (2017). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 3 for Grade 9 (NEPS Survey Paper No. 20)*. Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Schönberger, K., & Koberg, T. (2017). *Regional Data: Microm*. Leibniz Institute for Educational Trajectories.

Sharkey, P., & Faber, J. W. (2014). Where, when, why, and for whom do residential contexts matter? Moving away from the dichotomous understanding of neighborhood effects. *Annual Review of Sociology, 40*(1), 559–579. https://doi.org/10.1146/annurev-soc-071913-043350

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, *46*(3), 407–441. https://doi.org/10.2307/1170010

Skopek, J., Pink, S., & Bela, D. (2012). *Data manual. Starting cohort 3 - From lower to upper secondary school: NEPS SC3 1.0.0. NEPS Research data paper*. University of Bamberg.

Smith, H. J., Pettigrew, T. F., Pippin, G. M., & Bialosiewicz, S. (2012). Relative deprivation: A theoretical and meta-analytic review. *Personality and Social Psychology Review*, *16*(3), 203–232. https://doi.org/10.1177/1088868311430825

Snyder, C. R., Lassegard, M., & Ford, C. E. (1986). Distancing after group success and failure: Basking in reflected glory and cutting off reflected failure. *Journal of Personality and Social Psychology*, *51*(2), 382–388. https://doi.org/10.1037/0022-3514.51.2.382

Stewart, E. B., Stewart, E. A., & Simons, R. L. (2016). The effect of neighborhood context on the college aspirations of African American adolescents. *American Educational Research Journal*, *44*(4), 896–919. https://doi.org/10.3102/0002831207308637

Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A., & Williams, R. M. (1949). *The American soldier: Adjustment during army life.* Princeton University Press.

Sykes, B., & Musterd, S. (2010). Examining neighbourhood and school effects simultaneously. *Urban Studies*, *48*(7), 1307–1331. https://doi.org/10.1177/0042098010371393

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

Turley, R. N. L. (2002). Is relative deprivation beneficial? The effects of richer and poorer neighbors on children's outcomes. *Journal of Community Psychology*, *30*(6), 671–686. https://doi.org/10.1002/jcop.10033

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist*, *39*(2), 111–133. https://doi.org/10.1207/s15326985ep3902_3

van den Ham, A.-K., Schnittjer, I., & Gerken, A.-L. (2018). *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 3 for Grade 9 (NEPS Survey Paper No. 38).* Leibniz Institute for Educational Trajectories, National Educational Panel Study.

van Ham, M., & Tammaru, T. (2016). New perspectives on ethnic segregation over time and space. A domains approach. *Urban Geography*, *37*(7), 953–962. https://doi.org/10.1080/02723638.2016.1142152

Wheaton, B., & Clarke, P. (2003). Space meets time: Integrating temporal and contextual influences on mental health in early adulthood. *American Sociological Review*, *68*(5), 680. https://doi.org/10.2307/1519758

Wicht, A., & Ludwig-Mayerhofer, W. (2014). The impact of neighborhoods and schools on young people's occupational aspirations. *Journal of Vocational Behavior*, *85*(3), 298–308. https://doi.org/10.1016/j.jvb.2014.08.006

Wilson, W. J. (1987). *The truly disadvantaged*. University of Chicago Press.

Wilson, W. J. (1996). *When work disappears: The world of the new urban poor*. Vintage.

Wohlkinger, F., Bayer, M., & Ditton, H. (2016). Measuring self-concept in the NEPS. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological Issues of Longitudinal Surveys* (pp. 181–194). Springer Fachmedien Wiesbaden.

Zell, E., & Alicke, M. D. (2010). The local dominance effect in self-evaluation: Evidence and explanations. *Personality and Social Psychology Review*, *14*(4), 368–384. https://doi.org/10.1177/1088868310366144

Supplemental Material

Table S1

Academic Self-Concept Items

	General self-concept	Math self-concept	German self-concept
1	I learn fast in most of the school subjects	Math is one of my best subjects	In the subject German, I am a hopeless case (reverse coded)
2	I do well in most school subjects	I have always been good at math	I learn fast in German
3	I do well in written class tests	I get good grades in math	I get good grades in German

Detailed methods section

Additional information on the standardized math achievement test

In Grade 5, all students received the same test, which comprised 24 items with different response formats (simple multiple choice: 12; complex multiple-choice: 1; short constructed response: 11). In Grade 9, three different test booklets existed that differed in their difficulty. Each of the booklets comprised 23 items and there were 7 common items in all three tests.

Additional information on the standardized reading achievement test

In Grade 5, all students received the same test, which comprised 32 items with different response formats (simple multiple choice: 26; complex multiple-choice: 3; matching: 3). In Grade 9, two different test booklets existed that differed in their difficulty. The easy test comprised 30 items, the difficult test comprised 32 items. There were 18 common items between the two tests.

Additional information on the standardized orthography achievement test

In Grade 5, the test comprised spelling 30 words in a cloze test and 44 words in full sentences. In Grade 9, it comprised spelling 11 words in a cloze test and 126 words in full sentences.

Table S2

Robustness Checks With an Academic Self-Concept Measure Excluding the Grade-Related Item for Grade 5

	Model 1		Model 2a		Model 2b		Model 2c		Model 3a		Model 3b		Model 3c	
General	b	p	b	p	b	p	b	p	b	p	b	p	b	p
Achievement	.22	<.001	.19	<.001	.19	<.001	.20	<.001	.22	<.001	.22	<.001	.22	<.001
Class achievement	-.19	<.001							-.17	<.001	-.18	<.001	-.17	<.001
Status			.05	.006					.05	.004				
Neighborhood status			-.07	.002					-.06	<.001				
Income					.02	.326					.02	.304		
Neighborhood income					-.03	.020					-.02	.186		
Employment							.02	.796					.03	.664
Neighborhood employment							-.06	<.001					-.04	.042
Math														
Achievement	.35	<.001	.33	<.001	.32	<.001	.33	<.001	.36	<.001	.35	<.001	.35	<.001
Class achievement	-.23	<.001							-.20	<.001	-.23	<.001	-.21	<.001
Status			-.03	.106					-.03	.172				
Neighborhood status			-.07	<.001					-.07	<.001				
Income					-.01	.726					.00	.828		
Neighborhood income					-.03	.152					-.01	.442		
Employment							-.10	.180					-.07	.244
Neighborhood employment							-.07	<.001					-.04	.102
German														
Achievement	.40	<.001	.37	<.001	.37	<.001	.37	<.001	.39	<.001	.40	<.001	.39	<.001
Class achievement	-.13	<.001							-.12	.028	-.13	.022	-.11	.034
Status			.03	.070					.03	.066				
Neighborhood status			-.02	.156					-.02	.278				
Income					-.02	.382					-.01	.442		
Neighborhood income					-.03	.082					-.01	.606		
Employment							-.01	.838					-.01	.814
Neighborhood employment							-.01	.458					-.02	.434

Note. The dependent variable is academic self-concept. All analyses were controlled for federal state and school type.

Table S3

Robustness Checks With an Academic Self-Concept Measure Excluding the Grade-Related Item for Grade 9

	Model 1		Model 2a		Model 2b		Model 2c		Model 3a		Model 3b		Model 3c	
General	b	p	b	p	b	p	b	p	b	p	b	p	b	p
Achievement	.44	<.001	.37	<.001	.37	<.001	.37	<.001	.43	<.001	.43	<.001	.44	<.001
Class achievement	-.23	<.001							-.22	<.001	-.22	<.001	-.22	<.001
Status			.04	.088					.04	.052				
Neighborhood status			-.03	.226					-.02	.320				
Income					.02	.550					.02	.512		
Neighborhood income					-.01	.724					.02	.442		
Employment							-.11	.366					-.06	.508
Neighborhood employment							-.02	.450					.00	.828
Math														
Achievement	.55	<.001	.50	<.001	.50	<.001	.50	<.001	.57	<.001	.57	<.001	.57	<.001
Class achievement	-.23	<.001							-.33	<.001	-.35	<.001	-.32	<.001
Status			-.02	.370					-.01	.784				
Neighborhood status			-.04	.120					-.03	.204				
Income					.02	.308					.03	.224		
Neighborhood income					-.03	.128					-.01	.606		
Employment							.09	.304					.09	.344
Neighborhood employment							-.05	<.001					-.03	.200
German														
Achievement	.46	<.001	.41	<.001	.41	<.001	.41	<.001	.46	<.001	.46	<.001	.46	<.001
Class achievement	-.22	<.001							-.29	<.001	-.30	<.001	-.30	<.001
Status			.01	.628					.02	.310				
Neighborhood status			-.03	.214					-.01	.740				
Income					.01	.616					.02	.502		
Neighborhood income					-.02	.352					.00	.880		
Employment							-.02	.850					.00	.974
Neighborhood employment							-.02	.438					.01	.616

Note. The dependent variable is academic self-concept. All analyses were controlled for federal state and school type.

Table S4

Robustness Checks With Socioeconomic Neighborhood Composition Operationalized by a Composite Score of Neighborhood Status, Income, and Employment for Grade 5

	Model 1		Model 2		Model 3	
General	b	p	b	p	b	p
Achievement	.26	<.001	.23	<.001	.26	0
Class achievement	-.19	<.001			-.17	.004
Status			.05	.010	.05	.006
Income			.02	.430	.02	.432
Employment			-.02	.818	.00	.972
Neighborhood conditions			-.06	.002	-.05	.004
Math						
Achievement	.37	<.001	.35	<.001	.38	<.001
Class achievement	-.23	<.001			-.19	<.001
Status			-.03	.092	-.03	.152
Income			.01	.702	.01	.658
Employment			-.08	.200	-.09	.086
Neighborhood conditions			-.05	<.001	-.04	<.001
German						
Achievement	.43	<.001	.39	<.001	.42	<.001
Class achievement	-.18	<.001			-.17	.002
Status			.04	.020	.05	.016
Income			-.02	.246	-.02	.238
Employment			.06	.358	.06	.298
Neighborhood conditions			-.04	.028	-.03	.102

Note. The dependent variable is academic self-concept. All analyses were controlled for federal state and school type.

Table S5

Robustness Checks With Socioeconomic Neighborhood Composition Operationalized by a Composite Score of Neighborhood Status, Income, and Employment for Grade 9

	Model 1		Model 2		Model 3	
General	b	p	b	p	b	p
Achievement	.47	<.001	.40	<.001	.47	<.001
Class achievement	-.21	<.001			-.23	<.001
Status			.00	.048	.00	.058
Income			.00	.558	.00	.534
Employment			.03	.800	.03	.764
Neighborhood conditions			-.02	.276	.00	.946
Math						
Achievement	.56	<.001	.51	<.001	.58	<.001
Class achievement	-.24	<.001			-.32	<.001
Status			-.01	.512	.00	.852
Income			.02	.256	.03	.198
Employment			.04	.662	.01	.848
Neighborhood conditions			-.05	.020	-.02	.206
German						
Achievement	.48	<.001	.43	<.001	.48	<.001
Class achievement	-.20	<.001			-.29	<.001
Status			.02	.288	.03	.142
Income			.01	.690	.01	.634
Employment			.01	.962	.05	.514
Neighborhood conditions			-.03	.182	-.01	.742

Note. The dependent variable is academic self-concept. All analyses were controlled for federal state and school t

5 Study 3: Can Grades Move the Big Fish? The Consequences of Receiving Report Cards for Frame-of-Reference Effects on Academic Self-Concept

Fleischmann, M., Hübner, N., Marsh, H. W., Trautwein, U., Nagengast, B. (2020). Can Grades Move the Big Fish? The Consequences of Receiving Report Cards for Frame-of-Reference Effects on Academic Self-Concept. Manuscript ready for submission.

Abstract

Equally able students have lower academic self-concepts in high-achieving classes—a phenomenon known as the big fish little pond effect (BFLPE). School grades have been speculated to contribute to the BFLPE as they provide relative class ranking information and increase competition. However, empirical evidence for this assumption is not conclusive as it stems from correlational studies. Our sample comprised 9,104 Swedish elementary school students from the 1970s, a time period in which Swedish municipalities were free to decide to abolish grading. We found the frame-of-reference effect not to differ between nongraded and graded students. In line with the evolutionary basis of the BFLPE, these results suggest that students engage in social comparisons independent of whether or not they are graded.

Can Grades Move the Big Fish? The Consequences of Receiving Report Cards for Frame-of-Reference Effects on Academic Self-Concept

Academic self-concepts are students' self-perceptions of their competence in academic domains (Marsh, Martin, Yeung, & Craven, 2016). They have been found to have high power for predicting subsequent academic achievement (see Huang, 2011; Valentine, DuBois, & Cooper, 2004, for meta-analyses) as well as academic aspirations and choices (e.g., Guo, Marsh, Morin, Parker, & Kaur, 2015). A long tradition of research has suggested that academic self-concept is impacted by social comparison processes as is evident from the negative effect of the average level of achievement in educational contexts (e.g., school or classroom) on individuals' self-concept after individual achievement is controlled for. This finding is referred to as the big fish little pond effect (BFLPE; Marsh, 1987). The BFLPE predicts that equally able students will have lower self-concepts when placed in high-achieving schools and classes as opposed to low-achieving ones.

An important open research question regarding the BFLPE is whether the social comparison processes that underlie the frame-of-reference effect are reinforced by class-referenced grading (also known as "grading on a curve"). In other words: Does class-referenced grading contribute to the BFLPE because equally able students receive worse grades in higher achieving classrooms, which in turn negatively impact their self-concept? This assumption is strengthened by empirical findings that suggest that the BFLPE is smaller when differences in teacher-assigned grades are statistically controlled for. However, moderation studies have found that the BFLPE does not vary substantially with regard to individual or classroom characteristics variables, underscoring its immutable nature (for an overview, see Marsh & Seaton, 2015).

To address this research gap, in the present study, we evaluated a unique natural quasi-experiment in Sweden. Study participants attended elementary school during a time period in which municipalities were free to decide whether to keep or abolish grading. To our knowledge, the present investigation is the first to use a natural experiment to examine the effects of grading practices on academic self-concept formation and the BFLPE. By comparing nongraded and graded students, our study provides a much stronger test than any previous research of the widely accepted but untested assumption that class-referenced grading reinforces the BFLPE, making the results especially valuable for both research and educational practice.

The Big Fish Little Pond Effect and Potential Moderators

The BFLPE, specifically the negative effect of class-average achievement on academic self-concept, has been suggested to be induced by social comparison processes by which students compare their academic achievement with that of their school- and classmates by using them as a frame-of-reference for the evaluation of their own achievement (Huguet et al., 2009). Over the years, the BFLPE has received a great deal of empirical support (for an overview, see Marsh & Seaton, 2015). The frame-of-reference effect has been shown to be generalizable across different cultures (e.g., Guo, Marsh, Parker, & Dicke, 2018; Marsh, Parker, & Pekrun, 2018; Marsh, Pekrun, et al., 2018). Additionally, negative frame-of-reference effects have also been found to affect academic interest (Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006), task values (Cambria, Brandt, Nagengast, & Trautwein, 2017), academic aspirations (Nagengast & Marsh, 2012), academic emotions (Pekrun, Murayama, Marsh, Goetz, & Frenzel, 2019), and a large number of other desirable academic outcomes (Marsh, 1991). A question regarding the mechanisms behind the BFLPE is whether social comparisons underlying the frame-of-reference effect are reinforced by individual or contextual factors. This question was tackled in studies that probed for potential moderator effects. Motivated by the urge for a better understanding of the psychological mechanisms underlying the BFLPE and by its potentially negative consequences, there have been numerous attempts to detect possible moderators. Because the BFLPE is assumed to be a consequence of social comparisons within the classroom, characteristics that stimulate or attenuate these comparisons are candidates for potential moderator analyses.

Research on Individual-Level Moderators

The most frequently tested individual-level moderator is individual achievement (i.e., a test of whether the BFLPE is present to the same extent for low and high achievers). Generally, interaction terms between individual and aggregated achievement measures have been found to be small, nonsignificant, and not even consistent in direction (Marsh & Seaton, 2015). Conflicting results were also found regarding gender as an individual-level BFLPE moderator (e.g., Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007; Plieninger & Dickhäuser, 2013). Jonkmann, Becker, Marsh, Lüdtke, and Trautwein (2012) investigated personality traits as potential BFLPE moderators, and in addition to several null results, they found that students high on neuroticism tended to experience a stronger BFLPE, whereas the opposite was true for children high on narcissism. Seaton, Marsh, and Craven (2010) as well as Seaton, Marsh, Yeung, and Craven (2011) conducted an exploratory examination of BFLPE robustness. They

tested numerous individual difference variables as potential BFLPE moderators. For some of them, they found small interaction effects (e.g., anxious students experienced a more negative BFLPE). In sum, interaction effects between individual and aggregated achievement measures are rare findings. When they are detected, they are small in size and not able to eliminate the BFLPE or to even change its direction.

Research on Classroom-Level Moderators

Research on classroom-level moderators of the BFLPE is rather scarce. Investigations have focused on either aggregated student characteristics or instructional variables. In view of the former, Wouters, Colpin, van Damme, and Verschueren (2013) investigated the moderating effect of several goal orientations from Achievement Goal Theory. They also tested interactions between the class means of these variables and the BFLPE. They assumed that not only individual but also class-average motivation fosters social comparison processes. However, these aggregated motivational variables did not moderate the BFLPE. With regard to teacher variables, Lüdtke, Köller, Marsh, and Trautwein (2005) examined the moderating role of a social versus an individual teacher frame-of-reference for the BFLPE. They hypothesized that an individual reference standard—based on intraindividual improvement—would counteract social comparisons in the classroom and would reduce the BFLPE. Lüdtke et al. (2005) found that an individualized teacher frame-of-reference was positively associated with academic self-concept but did not change the BFLPE.

In sum, neither individual- nor classroom-level variables have been found to substantially moderate the BFLPE. Accordingly, it has been argued that students inevitably rank order themselves within educational environments. Such a conception has been supported by classical social comparison theory, which considers social comparison to be a universal human drive (e.g., Festinger, 1957). Also, Frank (2011)—in his evolutionary approach to social comparison—described the tendency to compare oneself to others as an immutable aspect of human nature. He argued that social comparison processes are the result of human evolution as individuals compete with others in their immediate social surroundings for all kinds of resources. However, to date, one promising classroom-level moderator—namely, teacher-assigned grades—has not yet been investigated.

Class-Referenced Grading and the BFLPE

A number of articles have shown that frame-of-reference effects not only affect academic self-concept but also teacher-assigned grades (e.g., Hochweber, Hosenfeld, & Klieme, 2014; Westphal et al., 2016). Thus, grades are negatively predicted by the average level

of academic achievement of the learning group when individual achievement is controlled. This finding has been interpreted as evidence of class-referenced grading, namely, the practice by which teachers assign very good grades to the best students in the class, assign the worst grades to the worst students in the class, and place the others somewhere in between (Neumann, Trautwein, & Nagy, 2011).

Because teacher-assigned grades have been found to be of great importance for domain-specific academic self-concept formation (e.g., Marsh & Craven, 1997; Skaalvik & Skaalvik, 2002), very early research had already theorized that class-referenced grading contributes to the BFLPE (e.g., Marsh, 1987). In other words, the BFLPE might be reinforced because equally able students receive lower grades in high-achieving learning environments, and this in turn results in a lower academic self-concept. According to this assumption, the BFLPE is not only due to an active social comparison process in which students engage in comparisons with classmates but also a passive comparison process by which students are compared with each other by their teacher. This idea was supported by a study by Trautwein, Gerlach, and Lüdtke (2008) who investigated frame-of-reference effects on physical activity self-concept at two measurement points. At T1, when students had not received grades, the BFLPE was smaller than at T2 when grading was introduced.

Researchers have tried to tackle the question of whether the BFLPE is reinforced by class-referenced grading by controlling the frame-of-reference effect for teacher-assigned grades, thus investigating whether equally able students who are provided with equal grades still have lower academic self-concepts in high-achieving classes. Trautwein et al. (2006) found that such an approach reduced the negative direct effect of class achievement on self-concept by about 50%. On a theoretical level, these results suggest that class-referenced grading may contribute to the BFLPE by explicitly providing students with information regarding their relative class ranking. Their finding even led Trautwein et al. (2006) to raise the critical question: “Would we still find a BFLPE if no school grades were assigned?” (p. 802).

However, the BFLPE could potentially be reinforced not only by the provision of grades but also by the expectation of receiving class-referenced grades. The expectation of receiving written grades has also been linked to enhanced competition in educational contexts in qualitative research (Covington, 2000; Elliot & Moller, 2003; Kohn, 1999; Pulfrey, Buchs, & Butera, 2011; Romanowski, 2004). In particular, the expectation of receiving class-referenced grades that strongly reflect the relative position of an individual student’s level of achievement in the classroom—as opposed to criterion- or self-referenced grades—have been theorized to

foster competition (Schinske & Tanner, 2014; Seymour & Hewitt, 1997). In turn, an increase in competition is theoretically expected to promote interest in social comparison (Ruble & Frey, 1991).

The assumption that class-referenced grading reinforces the BFLPE is largely based on correlational work and there is only weak evidence for causal relations. For instance, as grades are strongly correlated with both academic self-concept and standardized achievement, it is not surprising that controlling for grades typically reduces the BFLPE. It therefore seems more promising to compare BFLPEs of nongraded and graded students to examine if grading reinforces the BFLPE. However, to investigate grading as a moderator of the BFLPE depends on the identification of exogenous variation of grading practices in the field.

The Present Study

The present study makes use of a unique Swedish data set from 1980, a time in which grading practices in elementary school varied between students due to a school reform. This reform gave municipalities the option to either abolish or keep providing written grades and report cards. To our knowledge, this is the only available data set in which grading was quasi-experimentally manipulated. Therefore, these data offer the unprecedented opportunity to evaluate the mechanisms behind the BFLPE. Thereby, our study provides a much stronger test than any previous research of the widely accepted but untested assumption of class-referenced reinforcing the BFLPE. The study addresses three research questions:

Research Question 1: Did teachers in municipalities that continued to provide written grades and report cards assign class-referenced grades? In other words, did they “grade on a curve”? Research Question 1 is an important preliminary analysis because the assumption that grading reinforces the BFLPE depends on the provision and expectation of class-referenced grades.

Research Question 2: Is there support for the BFLPE in the present sample? This research question is aimed at replicating the well-known BFLPE finding. Moreover, it serves as a validation that the measures that were used for this study (see Method section) were appropriate for calculating frame-of-reference effects on academic self-concept.

Research Question 3: Did the size of the BFLPE differ between students who attended schools in municipalities that provided school grades and those that had abolished grading? The results for this third research question are at the core of the present article because they will provide evidence for whether grading reinforces the BFLPE.

Method

Study Background and Design: The Swedish Grading Reform in the 1970s

In the 1970s, Swedish children entered elementary education at the age of 7. They were assigned to schools on the basis of predefined catchment areas determined by their residence and were not allowed to choose a different learning institution. Elementary education, in which class composition did not change, included Grades 1 to 6. Every class was typically taught by the same teacher from Grade 1 to the middle of Grade 4 when another teacher took over for the rest of elementary education (Klapp, 2015; Sjögren, 2010).

Until the 1968/1969 school year, students were provided with written grades and report cards in the core subjects of mathematics, Swedish, and English at the ends of Grades 3 and 6. Beginning with the 1969/1970 school year, municipalities were free to decide to abolish grading. The reform made schools gradually abandon the practice of providing written grades in the 1970s before grading was finally abolished in the 1982/1983 school year throughout Sweden.

Generally, arguments for the shift in the grading policy were strongly influenced by the idea that providing grades promotes unhealthy competition between students and fosters inequalities in educational outcomes by encouraging high-performers and discouraging low-performers (Sjögren, 2010).

Data

The analyses were based on data coming from the Swedish "Evaluation through follow-up study" (ETF Study; Härnquist, 2000). For the present investigation, we used data from the first measurement occasion of the third ETF cohort (born in 1967) in spring 1980 when students were in Grade 6 of elementary education. This cohort is of special interest because these children attended elementary school during the reform window described above (from the 1974/1975 school year to the 1979/1980 school year) in which municipalities were free to decide to abolish grading. Generally, sampling from the third ETF cohort was conducted by means of a multistage sampling procedure in which a stratified sample of 29 municipalities was drawn in a first step, and school classes from these municipalities were drawn in a second step. The total sample consisted of N = 9,104 students who were nested in 421 classes from 138 schools. In the data, each school contained an average of M = 3.05 (SD = 2.73) classes and each class an average of M = 21.62 (SD = 6.55) students. A total of 49.14% of the sample was female, and students were on average M = 12.85 (SD = 0.33) years old. A total of 4,656 students were not graded, whereas the other 4,448 students received grades (for more information on the

grading variable see Appendix A). It is important to note that in spring 1980 when participants were measured, students in grading municipalities had not yet received their Grade 6 report cards. As municipalities were free in their decision to abolish grading, we compared nongraded and graded students with regard to the independent variables and covariates. We did not find differences between subgroups in any of these variables (table B in appendix B). In sum, this quasi-experimental design allows for the strongest test of the untested assumption of class-referenced reinforcing the BFLPE.

Instruments

Domain-specific academic self-concept. Domain-specific academic self-concept was measured with items that were presented along with pictures and had to be answered with no or yes. For mathematics self-concept, the item was: "The girl in the picture thinks she is good at sums. Do you think you are good at sums?" Reading self-concept was the only reverse-scored item, which asked: "The boy in the picture thinks he is bad at reading. Do you think you are bad at reading?" Spelling self-concept was measured with: "The boy in the picture thinks he is good at spelling. Do you think you are good at spelling?" Moreover, general academic self-concept was assessed with: "The boy in the picture thinks he does well in school. Do you think you do well in school?" Research has shown reliability and validity of single-item measures to be acceptable when the measure is homogenous and clearly defined (Gardner, Cummings, Dunham, & Pierce, 1998). As a consequence, single-item measures have been successfully used for measuring a variety of psychological constructs (Postmes, Haslam, & Jans, 2013; Wanous, Reichers, & Hudy, 1997).

Domain-specific academic achievement. Domain-specific academic achievement was measured with standardized national tests. The standardized tests consisted of items from different subcategories (see Appendix C for a detailed description). As ETF data does provide total points within each of the subcategories, we calculated a sum score comprising total points from all subcategories. Reliability between the subcategories in, as measured by Cronbach's Alpha, was α = .89 in math, α = .85 in Swedish, and α = .93 in English.

To further assess the measurement quality of self-concept and achievement scales we closely inspected their interrelations (table D in appendix D). As expected, domain-specific self-concept measures were strongly correlated with their respective achievement variables (math: r = .40, reading: r = .31, spelling: r = .34, general: r = .35). These correlations are nearly identical to those reported in a meta-analysis by Möller, Pohlmann, Köller, and Marsh (2009) who reanalyzed 69 datasets and found average correlations between math self-concept and math

achievement of r = .37 and verbal self-concept and verbal achievement of r = .34. These results empirically support findings by Gogol et al. (2014) who found nearly identical relations within a nomological network for single-item measures as compared to multi-item scales, thereby further supporting the reliability and validity of our single-item self-concept measures.

Domain-specific teacher-assigned grades in grade 6. Domain-specific teacher-assigned grades in Grade 6 were retrieved from school administrative data. Grades were delivered on a scale from 1 to 5 with 5 representing the highest grade.

Covariates. As covariates, we used students' age, sex, SES (based on parents' occupations), and cognitive abilities (the mean of the total number of points scored on the verbal opposite ability test, the spatial ability test, and the inductive ability test).

Analyses

Analyses were run in Mplus 8 (Muthén & Muthén, 1998-2018). We took a multilevel structural equation modeling approach in which we explicitly modeled the individual as well as the class level. We did not explicitly model the school level as research has shown the class to be the pivotal frame-of-reference for academic self-concept formation (Marsh, Kuyper, Morin, Parker, & Seaton, 2014). But we controlled for the dependency of observations at the school level using a design-based correction of standard errors and fit statistics (implemented with the Mplus command TYPE = TWOLEVEL COMPLEX). Because domain-specific academic self-concepts were assessed with a binary variable (e.g., Do you think you are good at sums? No/Yes), we used multilevel linear probability models (Breen, Karlson, & Holm, 2018). In contrast to logistic regression, linear probability models directly model the probability of choosing a binary category, thus facilitating parameter interpretation. Further, linear probability models allow the comparison of parameters across different models in contrast to logistic regression, where the error variance is fixed (Mood, 2010). As robustness checks, we additionally analyzed all models with multilevel logistic regression models.

The proportions of missing values for model variables are presented in can be found in Table D in Appendix D. In all statistical models, full maximum likelihood estimation (FIML) was used to account for missing values (Enders, 2010; Graham, 2009). In the contextual effect models, all continuous predictor variables were standardized, and class-average achievement was calculated on the basis of standardized individual-level measures.

Results

Descriptive Statistics

Descriptive statistics for the total student sample are reported in Table D in Appendix D. Class- and school-level proportions of variance for self-concept were low. By contrast, variation in achievement on the class level was larger (between VP_{cla} = .08 for Swedish and VP_{cla} = .12 for general achievement), whereas variation at the school level was low (between VP_{sch} = .01 for Swedish and VP_{sch} = .02 for math and general). These low school-level proportions of variance show that next to the theoretical reasons presented above, there were no empirical reasons for explicitly modeling the school level. Descriptive statistics presented separately for the nongraded and graded student samples can be found in the supplementary material (Tables S1 and S2). Correlations between self-concept and achievement measures were similar across nongraded and graded students. As expected, the proportions of variance for grades in the sample of graded students were relatively low for math (VP_{cla} = .04; VP_{sch} = .02) and Swedish (VP_{cla} = .02; VP_{sch} = .01). Additionally, grades were strongly correlated with the respective achievement measures (*r* = .85 for math and *r* =.86 for Swedish).

Research Question 1: Did Teachers use Class-Referenced Grading?

As mentioned above, answering Research Question 1 served two aims. First—as grades were mostly based on the results from national standardized tests but teachers were allowed to go beyond these tests in their grading—this research question tested for whether class-referenced grading existed, even when teachers knew about students' absolute academic achievement. Second, it provided an important preliminary analysis because the theory that grading reinforces the BFLPE depends on the provision and expectations of class-referenced grades. To answer Research Question 1, we took the complete set of graded students and regressed their grades on the covariates and achievement as well as class achievement. The results can be found in Table 1. In math, class achievement negatively predicted grades when the other variables were controlled for (b = -0.31, p < .001). In other words, an increase in class achievement by one standard deviation was associated with a decrease in grades by 0.31 standard deviations. Equally able students had worse grades in high-achieving classes and vice versa. Frame-of-reference effects were also found for Swedish (b = -0.26, p < .001) and English grades (b = -0.34, p < .001). Generally, the results suggest that students were graded on a class-referenced basis, even though grades were mostly based on the results from national standardized tests.

Research Question 2: Was there a BFLPE?

Research Question 2 asked whether the BFLPE could be found in the total sample. To answer Research Question 2, we took the total student sample and regressed self-concept on the covariates, achievement, class achievement, and grading. Results from these multilevel linear probability models are presented in Tables 2 and 3. In all four domains, individual achievement positively predicted self-concept (math: b = 0.20; reading: b = 0.16; spelling: b = 0.25; general: b = 0.17; all ps < .001). This means that an increase of one standard deviation in academic achievement was associated with a 20, 16, 25, and 17 percentage point increase in the probability of stating that one was good at the respective domain. Grading negatively predicted general self-concept (b = -0.05, p = .001). This means that graded students had a 5 percentage point lower probability of stating they were good at school. In all four domains, class achievement negatively predicted self-concept (math: b = -0.10, p < .001; reading: b = -0.05, p = .003; spelling: -0.08, p < .001; general: b = -0.09, p < .001). This means that an increase of one standard deviation in class achievement was associated with a 10, 5, 8, and 9 percentage point decrease in stating that one is good at the respective domain. The fact, that the BFLPE could be found in all these domains gives further evidence for the reliability and validity of our single-item academic self-concept measures. These BFLPEs were also found in the respective logistic regression analyses (see Tables S3 and S4 in the supplemental online materials).

Research Question 3: Did the BFLPE Differ between Nongraded and Graded Students?

Research Question 3 investigated whether the BFLPE differed between nongraded and graded students (i.e., Was the frame-of-reference effect reinforced by providing class-referenced grades?). Research Question 3 represents the main research question of the present paper. It builds on previous correlational work that suggested class-referenced grading to contribute to the BFLPE by providing relative class-ranking information. As grading in our study was quasi-experimentally manipulated, our study provides a much stronger test than any previous research on the assumption of class-referenced grading reinforcing the BFLPE. To answer Research Question 3, we extended the statistical model from Research Question 2 and additionally modeled the interaction between grading and class achievement. The results from these multilevel linear probability models are presented in Tables 2 and 3. None of the interactions between the grading dummy and class achievement were significantly different from zero (math: b = -0.07, p = .150; reading: b = 0.01, p = .676; spelling: b = 0.00, p = .962; general: b = -0.03, p = .467). Thus, the BFLPEs did not differ between nongraded and graded

students. These results were the same in the respective logistic regression analyses (see Tables S3 and S4 in the supplemental online materials).

Table 1

Results from Contextual Effects Models with Teacher-Assigned Grades as the Outcome

	Math				Swedish				English			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	.79	.02	[.75, .82]	< .001	.80	.02	[.76, .83]	< .001	.77	.02	[.74, .81]	< .001
Class achievement	-.31	.04	[-.39, -.23]	< .001	-.26	.06	[-.39, -.14]	< .001	-.34	.04	[-.42, -.26]	< .001

Note. Analyses were conducted with the graded student sample (N = 4,448). Outcomes are grades in the respective domain. Achievement and class achievement resemble standardized achievement scores in the respective domains. All analyses are controlled for age, sex, SES, and cognitive abilities.

Table 2

Results from Contextual Effects Models with Math and Reading Self-Concept as the Outcome

	Math								Reading							
	Model 1				Model 2				Model 1				Model 2			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	.20	.01	[.18, .22]	< .001	.20	.01	[.18, .22]	< .001	.16	.01	[.14, .17]	< .001	.16	.01	[.14, .17]	< .001
Grading	-.03	.01	-[.05, .00]	.051	-.02	.01	-[.05, .00]	.074	-.01	.01	-[.03, .01]	.348	-.01	.01	-[.03, .01]	.348
Class achievement	-.10	.02	-[.14, -.05]	< .001	-.05	.04	-[.12, .02]	.172	-.05	.02	-[.08, -.02]	.003	-.06	.02	-[.10, -.01]	.014
Grading x Class achievement					-.07	.05	-[.15, .02]	.150					.01	.03	-[.05, .08]	.676

Note. The table contains results from multilevel linear probability analyses. Outcomes are dichotomous self-concept items in the respective domain (e.g., Do you think you are good at sums? 0 for No and 1 for Yes). Achievement and class achievement resemble standardized achievement scores in the respective domains. Grading is a dichotomous variable (0 for nongraded and 1 for graded). All analyses are controlled for age, sex, SES, and cognitive abilities.

Table 3

Results from Contextual Effects Models with Spelling and General Self-Concept as the Outcome

	Spelling								General							
	Model 1				Model 2				Model 1				Model 2			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	.25	.01	[.23, .26]	< .001	.25	.01	[.23, .26]	< .001	.17	.01	[.15, .20]	< .001	.17	.01	[.15, .20]	< .001
Grading	-.02	.01	-[.05, .00]	.055	-.02	.01	-[.05, .00]	.055	-.05	.02	-[.08, -.02]	.001	-.05	.02	-[.08, -.02]	.001
Class achievement	-.08	.02	-[.12, -.05]	< .001	-.08	.03	-[.13, -.03]	.001	-.09	.02	-[.14, -.05]	< .001	-.08	.03	-[.13, -.03]	.003
Grading x Class achievement					.00	.04	-[.07, .07]	.962					-.03	.04	-[.11, .05]	.467

Note. The table contains results from multilevel linear probability analyses. Outcomes are dichotomous self-concept items in the respective domain (e.g., Do you think you are good at sums? 0 for No and 1 for Yes). Achievement and class achievement resemble standardized achievement scores in the respective domains. Grading is a dichotomous variable (0 for nongraded and 1 for graded). All analyses are controlled for age, sex, SES, and cognitive abilities.

Discussion

Previous studies found the BFLPE to be mediated by teacher-assigned grades and followingly argued that the BFLPE is driven by class-referenced grades that provide relative class ranking information. However, as these studies did not experimentally manipulate grading practices in the field, they were limited regarding their internal validity concerning the assumption the class-referenced grades drive the BFLPE. In the current study, we built on this research and evaluated a unique natural experiment in Sweden. Study participants attended elementary school during a period of time in which municipalities were free to decide to either keep or abolish the provision of written grades and report cards in elementary education. We found no differences in the size of the BFLPEs between nongraded and graded students. Our results support the contention that the provision of class-referenced grades does not reinforce the BFLPE. By comparison of nongraded and graded students, our study provides a much stronger test than any previous research of the untested assumption of class-referenced grades reinforcing the BFLPE.

Contributions to the State of Research

Numerous studies have found that the BFLPE declines when teacher-assigned grades were controlled for (e.g., Marsh, 1987; Trautwein et al., 2006). This finding was typically interpreted as evidence that the provision of class-referenced grades works as an amplifier of the frame-of-reference effect. Equally able students receive worse grades in high-achieving classrooms, thus leading to a decline in academic self-concept. Because of the low internal validity of correlational mediation analysis, the question if grades reinforce the BFLPE can much better be answered by means of a moderation analysis that compares BFLPEs of nongraded and graded students. Using a quasi-experimental field study, we found no differences in the BFLPE between nongraded and graded students and therefore caution researchers not to make causal attributions to teacher-assigned grades in the BFLPE mediation model. The findings of our study are in line with previous empirical results from BFLPE moderation studies, which argue for the broad generalizability of the BFLPE (e.g., Seaton et al., 2010, 2009).

Additionally, we found that grading negatively affected general academic self-concept. This negative main effect was rather unexpected and may have resulted from more of a competitive atmosphere in graded classes, causing all students to more critically evaluate their academic achievements. Generally, one would rather expect differential grading effects for low and high achievers because grading accentuates self-concept differences between the two

groups of students. However, in additional analyses in which we modeled the interaction between grading and individual achievement (see table S5 in the supplemental online materials), we found no differential effects of grading on self-concept for low and high achievers.

Along with sobering results from BFLPE moderation studies, our investigation suggests that social comparisons underlying the BFLPE happen spontaneously because students tend to inevitably rank order themselves in educational environments (see also Marsh, Parker, Guo, Pekrun, & Basarkod, 2020). For example, students may make these comparisons when talking about homework with peers, or on the basis of their classmates' classroom participation. Such a conception is supported by classical social comparison theory, which views social comparison as a universal human drive (cf. Festinger, 1957). Also, more recent evolutionary approaches to social comparison view the tendency to compare oneself with others as a largely immutable aspect of human behavior (e.g., Frank, 2011).

The evolutionary approach to social comparison has implications for educational practice. It has repeatedly been argued that class-referenced grading encourages social comparisons in the classroom, thus negatively affecting student outcomes (e.g., Covington, 2000; Elliot & Moller, 2003; Kohn, 1999; Pulfrey et al., 2011; Romanowski, 2004). The evolutionary approach to social comparison suggests that the grading controversy might be less important than believed because students compare themselves with one another anyway, independent of grade provision. Grading opponents might also argue that grading increases the self-concept of high achievers by providing them with positive performance feedback. Such a practice would decrease the self-concept of low achievers because this group of students receives negative performance feedback, thus amplifying inequalities in educational outcomes. As reported above, we found no differential grading effects for low and high achievers, supporting the idea that students rank order themselves in educational environments independent of whether they receive written grades.

Limitations

Our study is unique in that we made use of a natural quasi-experiment to gain a deeper understanding of the BFLPE. Typically, such field experiment studies are based on data that were not collected with the primary aim of answering the research question under investigation. Such a practice usually leads to some limitations, which was also the case for our study.

First, it is possible that teachers in both nongrading and grading municipalities conducted continuous classroom assessments. No information exists about whether these tests

resulted in qualitative or quantitative (e.g., grade-like) performance feedback. On the other hand, our study showed that the abolishment of highly salient social comparison information such as class-referenced written grades and report cards will probably not be able to alter the BFLPE. Additionally, whereas grades that were given in Grade 3 provided relative performance feedback, grades from Grade 6 only were able to contribute to increased classroom competition because students had not received their report cards when they completed the academic self-concept instrument. These issues do not have any consequences for the processing of our primary research question, which asked about the effects on the BFLPE from a school reform that abolished written grades and report cards because they were assumed to induce unhealthy competition. Concerning this question, we can indeed say that the reform did not affect the BFLPE. On a theoretical level, we raised the question of whether grading reinforces the BFLPE. This question indeed could not be answered conclusively with the present study design for the abovementioned reasons. Because of the complexity of educational field research (e.g., the experimental manipulation of grading practices is virtually impossible), we argue that our study is one very important puzzle piece in testing the nature of social comparisons that underlie the BFLPE.

Another limitation of the present study is related to measurement issues. Academic self-concept was measured with items that referred to a sex-specific comparison target (e.g., "The girl in the picture thinks she is good at sums. Do you think you are good at sums? Yes/No"). In the other self-concept items, the target of comparison was a boy. On the one hand, one can argue that on the basis of prevailing stereotypes (e.g., boys are better at math), participants may have reacted differently to the items, thus resulting in an unreliable measure of our outcome. On the other hand, the item-specific target of comparison was the same for boys and girls, and sex was included as a covariate in our analyses. In supplementary analyses, we tested for sex differences in domain-specific academic self-concept. In line with the literature (e.g., Marsh & Hattie, 1996; Watt & Eccles, 2008), boys had higher self-concept in math, whereas girls showed higher spelling and reading self-concepts. We interpret these results as indicating that the sex-related item format did not limit the validity of our self-concept items. Further, academic self-concept was measured with the help of binary single-item scales that asked whether students "are good" at the respective domains. As argued in the method section, we assumed that the single-item measures would be sufficient for measuring schematic, unidimensional, and subjective constructs such as academic self-concept. In an additional robustness-check, we also constructed a multi-item general self-concept variable by averaging the four domain-specific self-concept indicators (Table S6). Again, we found no BFLPE differences between nongraded

and graded students. Third, in our study, grading practices were quasi-experimentally manipulated in that municipalities could abolish the provision of written grades and report cards. Although we did not find differences between nongraded and graded students in regard to the independent variables or covariates, we were not able to control for municipality-level characteristics. However, this quasi-experimental design allows for the strongest test of the untested assumption of class-referenced reinforcing the BFLPE.

Future Prospects

The present study took advantage of a natural experiment within the context of a unique educational reform, an opportunity unlikely to be available again in the near future. Indeed, investigating whether grading reinforces the BFLPE would ideally be tested by conducting a randomized controlled field trial. Given that it seems nearly impossible to randomly vary grading practices in the field, this issue cannot be resolved in a single study but has to be approached from different angles. Our study provides very good conditions from an internal validity perspective, with limitations concerning the treatment and measurement issues as described above. We argue that the unique strengths of this study far outweigh potential limitations and provide a good basis for seriously questioning the "accepted" results that have not been tackled with a suitable research design until now. In the future, deeper insights into the association between the BFLPE and grading practices can be investigated by analyzing grading reforms with the help of cohort-control designs. In the present study, cohort comparisons were not possible due to missing information about class membership in the cohorts from before and after the time period under study. These cohort comparisons may overcome some of the present limitations but will yield other drawbacks such as the confounding of grading and cohort effects. Because the assumption that grading reinforces the BFLPE is mainly based on the idea that grades to a certain extent are class-norm referenced, the issue can also be approached by comparing BFLPEs in class- and population-referenced grading systems.

References

Breen, R., Karlson, K. B., & Holm, A. (2018). Interpreting and understanding logits, probits, and other nonlinear probability models. Annual Review of Sociology, 44, 39–54. https://doi.org/10.1146/annurev-soc-073117-041429

Cambria, J., Brandt, H., Nagengast, B., & Trautwein, U. (2017). Frame of reference effects on values in mathematics: Evidence from German secondary school students. *ZDM*, *49*, 435–447. https://doi.org/10.1007/s11858-017-0841-0

Covington, M. V. (2000). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology*, *51*, 171–200. https://doi.org/10.1146/annurev.psych.51.1.171

Elliot, A. J., & Moller, A. C. (2003). Performance-approach goals: Good or bad forms of regulation? *International Journal of Educational Research*, *39*, 339–356. https://doi.org/10.1016/j.ijer.2004.06.003

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Festinger, L. (1957). A theory of social comparison processes. *Human Relations*, *7*, 117–140. https://doi.org/10.1177/001872675400700202

Frank, R. H. (2011). *The Darwin economy: Liberty, competition, and the common good*. Princeton: Princeton University Press.

Gardner, D. G., Cummings, L. L., Dunham, R. B., & Pierce, J. L. (1998). Single-item versus multiple-item measurement scales: An empirical comparison. *Educational and Psychological Measurement*, *58*, 898–915. https://doi.org/10.1177/0013164498058006003

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., . . . Preckel, F. (2014). "My questionnaire is too long!": The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology*, *39*, 188–205. https://doi.org/10.1016/j.cedpsych.2014.04.002

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

Guo, J., Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2015). Directionality of the associations of high school expectancy-value, aspirations, and attainment. *American Educational Research Journal*, *52*, 371–402. https://doi.org/10.3102/0002831214565786

Guo, J., Marsh, H. W., Parker, P. D., & Dicke, T. (2018). Cross-cultural generalizability of social and dimensional comparison effects on reading, math, and science self-concepts for

primary school students using the combined PIRLS and TIMSS data. *Learning and Instruction, 58,* 210–219.

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

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology, 49,* 505–528. https://doi.org/10.1016/j.jsp.2011.07.001

Huguet, P., Dumas, F., Marsh, H. W., Wheeler, L., Seaton, M., Nezlek, J., . . . Régner, I. (2009). Clarifying the role of social comparison in the big-fish-little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology, 97,* 156–170. https://doi.org/10.1037/a0015558

Jonkmann, K., Becker, M., Marsh, H. W., Lüdtke, O., & Trautwein, U. (2012). Personality traits moderate the big-fish–little-pond effect of academic self-concept. *Learning and Individual Differences, 22,* 736–746. https://doi.org/10.1016/j.lindif.2012.07.020

Klapp, A. (2015). Does grading affect educational attainment? A longitudinal study. *Assessment in Education: Principles, Policy & Practice, 22,* 302–323. https://doi.org/10.1080/0969594X.2014.988121

Kohn, A. (1999). *Punished by rewards*. Boston: Houghton Mifflin.

Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish-little-pond effect. *Contemporary Educational Psychology, 30,* 263–285. https://doi.org/10.1016/j.cedpsych.2004.10.002

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

Marsh, H. W. (1991). Failure of high-ability high schools to deliver academic benefits commensurate with their students' ability levels. *American Educational Research Journal, 28,* 445–480. https://doi.org/10.2307/1162948

Marsh, H. W., & Craven, R. (1997). Academic self-concept: Beyond the dustbowl. In G. D. Phye (Ed.), *Handbook of classroom assessment* (pp. 131–198). San Diego: Academic Press.

Marsh, H. W., & Hattie, J. (1996). Theoretical perspectives on the structure of self-concept. In B. A. Bracken (Ed.), *A Wiley-Interscience publication. Handbook of self-concept: Developmental, social, and clinical considerations* (pp. 38–90). New York: Wiley.

Marsh, H. W., Kuyper, H., Morin, A. J., Parker, P. D., & Seaton, M. (2014). Big-fish-little-pond social comparison and local dominance effects: Integrating new statistical models, methodology, design, theory and substantive implications. *Learning and Instruction, 33,* 50–66. https://doi.org/10.1016/j.learninstruc.2014.04.002

Marsh, H. W., Martin, A. J., Yeung, A. S., & Craven, R. (2016). Competence self-perceptions. In C. Dweck & D. Yaeger (Eds.), *Handbook of competence and motivation.* New York: Guilford Press.

Marsh, H. W., Parker, P. D., Guo, J., Pekrun, R., & Basarkod, G. (2020). Psychological comparison processes and self-concept in relation to five distinct frame-of-reference effects: Pan-human cross-cultural generalizability over 68 countries. *European Journal of Personality, 3,* 180–202. https://doi.org/10.1002/per.2232

Marsh, H. W., Parker, P. D., & Pekrun, R. (2018). Three paradoxical effects on academic self-concept across countries, schools, and students. *European Psychologist*, 1–12. https://doi.org/10.1027/1016-9040/a000332

Marsh, H. W., Pekrun, R., Murayama, K., Arens, A. K., Parker, P. D., Guo, J., & Dicke, T. (2018). An integrated model of academic self-concept development: Academic self-concept, grades, test scores, and tracking over 6 years. *Developmental Psychology, 54,* 263–280. https://doi.org/10.1037/dev0000393

Marsh, H. W., & Seaton, M. (2015). The big-fish–little-pond effect, competence self-perceptions, and relativity: Substantive advances and methodological innovation. *Advances in Motivation Science, 2,* 127–184.

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,* 631–669. https://doi.org/10.3102/0002831207306728

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79,* 1129–1167. https://doi.org/10.3102/0034654309337522

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26,* 67–82. https://doi.org/10.1093/esr/jcp006

Muthén, L. K., & Muthén, B.O. (1998-2018). *Mplus user's guide*. Los Angeles: Muthén & Muthén.

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

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*, 206–217. https://doi.org/10.1016/j.stueduc.2012.02.002

Pekrun, R., Murayama, K., Marsh, H. W., Goetz, T., & Frenzel, A. C. (2019). Happy fish in little ponds: Testing a reference group model of achievement and emotion. *Journal of Personality and Social Psychology*, *117*, 166–185. https://doi.org/10.1037/pspp0000230

Plieninger, H., & Dickhäuser, O. (2013). The female fish is more responsive: Gender moderates the BFLPE in the domain of science. *Educational Psychology*, *35*, 213–227. https://doi.org/10.1080/01443410.2013.814197

Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure of social identification: Reliability, validity, and utility. *The British Journal of Social Psychology*, *52*, 597–617. https://doi.org/10.1111/bjso.12006

Pulfrey, C., Buchs, C., & Butera, F. (2011). Why grades engender performance-avoidance goals: The mediating role of autonomous motivation. *Journal of Educational Psychology*, *103*, 683–700. https://doi.org/10.1037/a0023911

Romanowski, M. H. (2004). Student obsession with grades and achievement. *Kappa Delta Pi Record*, *40*, 149–151. https://doi.org/10.1080/00228958.2004.10516425

Ruble, D., & Frey, K. (1991). Changing patterns of comparative behavior as skills are acquired: A functional model of self-evaluation. In J. Suls & T. A. Wills (Ed.), *Social comparison: Contemporary theory and research* (pp. 79–113). New York: Hillsdale.

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

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

Seaton, M., Marsh, H. W., & Craven, R. G. (2010). Big-fish-little-pond effect: Generalizability and moderation - Two sides of the same coin. *American Educational Research Journal*, *47*, 390–433. https://doi.org/10.3102/0002831209350493

Seaton, M., Marsh, H. W., Yeung, A. S., & Craven, R. (2011). The big fish down under: Examining moderators of the 'big-fish little-pond' effect for Australia's high achievers. *Australian Journal of Education*, *55*, 93–114. https://doi.org/10.1177/000494411105500202

Seymour, E., & Hewitt, N. M. (1997). Talking about leaving: Why undergraduates leave the sciences. *Choice Reviews Online*, *34*, 34-5652-34-5652. https://doi.org/10.5860/CHOICE.34-5652

Sjögren, A. (2010). *Graded children - Evidence of long-run consequences of school grades from a nationwide reform*. Uppsala: IFAU – Institute for Labour Market Policy.

Skaalvik, E. M., & Skaalvik, S. (2002). Internal and external frames of reference for academic self-concept. *Educational Psychologist*, *37*, 233–244. https://doi.org/10.1207/S15326985EP3704_3

Trautwein, U., Gerlach, E., & Lüdtke, O. (2008). Athletic classmates, physical self-concept, and free-time physical activity: A longitudinal study of frame of reference effects. *Journal of Educational Psychology*, *100*, 988–1001. https://doi.org/10.1037/0022-0663.100.4.988

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

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist*, *39*, 111–133. https://doi.org/10.1207/s15326985ep3902_3

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, *82*, 247–252. https://doi.org/10.1037/0021-9010.82.2.247

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.

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

Wouters, S., Colpin, H., van Damme, J., & Verschueren, K. (2013). Endorsing achievement goals exacerbates the big-fish-little-pond effect on academic self-concept. *Educational Psychology*, *35*, 252–270. https://doi.org/10.1080/01443410.2013.822963

Appendix A

The data contains information about whether students were graded in Grade 6. This information was derived from the grade variables that were based on school administrative data. When every student in a municipality had missing data on the grade variables, the students from the respective municipality were identified as nongraded students. When a majority of students in a municipality had nonmissing values on the grade variables (note that in the graded municipalities some students had "real" missing values on grade variables), students from the respective municipalities were identified as graded students. When students in our sample were not graded in Grade 6, the probability was very high (about 77%) that they were not graded in Grade 3 (see Table A).

Table A

Introduction of the Grading Reform

School year	Cohort in year	Percentage of municipalities that abolished grading in Grade 3	Percentage of municipalities that abolished grading in Grade 6
1974/1975	1	9.09	5.35
1975/1976	2	18.18	9.09
1976/1977	**3**	**34.22**	14.44
1977/1978	4	57.75	25.67
1978/1979	5	67.91	35.29
1979/1980	**6**	75.94	**44.39**

Note. The information in this table was retrieved from Sjögren (2010). Sjögren (2010) showed that in the 1976/1977 school year, (when the present cohort was in Grade 3), approximately 34% of the municipalities had abolished grading in Grade 3. In the 1979/1980 school year (when our cohort was in Grade 6), approximately 44% of the municipalities had abolished grading in Grade 6. This means that 77% (34.22/44.39) of the municipalities that had abolished grading in the 1979/1980 school year (when our cohort was in Grade 6) had already abolished grading in Grade 3 (when our cohort was in Grade 3). Thus, when the students in our sample were not graded in Grade 6, the probability was high that they had not been graded in Grade 3.

Appendix B

Table B

Mean Differences in Model Variables between Nongraded and Graded Students

	b	*p*
Math self-concept	-0.01	.414
Spelling self-concept	-0.01	.578
Reading self-concept	-0.02	.081
General self-concept	-0.04	.003
Math achievement	0.00	.999
Swedish achievement	0.03	.449
General achievement	-0.04	.403
Age	-0.03	.304
Sex	-0.01	.152
Ses	0.07	.073
Cognitive ability	0.07	.062

Note. Mean differences were calculated by regressing the respective outcomes on the grading dummy (0 for nongraded and 1 for graded). Continuous outcomes were standardized.

Appendix C

The standardized mathematics test consisted of items from different subcategories (e.g., percentage ability or geometry ability). The standardized Swedish language test contained items from six subcategories (e.g., reading or spelling). The standardized English language test contained items from four subcategories (e.g., vocabulary or listening). Additionally, we constructed a measure of general academic achievement by averaging the math, Swedish, and English achievement scores.

Appendix D

Table D

Descriptive Statistics for the Total Sample

	Mis	*M*	*SD*	VP_{cla}	VP_{sch}	1	2	3	4	5	6	7	8	9	10
1. Math self-concept	0.11	0.69	0.46	.03	.01										
2. Reading self-concept	0.10	0.80	0.40	.01	.01	.16									
3. Spelling self-concept	0.10	0.67	0.47	.02	.00	.10	.31								
4. General self-concept	0.13	0.67	0.47	.03	.02	.48	.30	.28							
5. Math achievement	0.44	50.05	15.55	.10	.02	.40	.16	.12	.34						
6. Swedish achievement	0.39	66.44	18.26	.08	.01	.22	.31	.34	.33	.70					
7. General achievement	0.38	70.32	18.66	.12	.02	.27	.28	.30	.35	.82	.91				
8. Age	0.00	12.85	0.33	.01	.00	.00	-.02	-.03	-.02	-.05	-.05	-.06			
9. Sex	0.00	0.49	0.50	.00	.00	-.09	.03	.13	-.03	-.02	.17	.14	-.03		
10. SES	0.05	2.28	0.67	.08	.03	-.10	-.09	-.05	-.12	-.26	-.28	-.29	.04	-.02	
11. Cognitive abilities	0.10	22.84	5.82	.09	.01	.32	.17	.12	.28	.75	.71	.73	-.07	.02	-.25

Note. Descriptive statistics were based on the total sample (N = 9,104). Descriptive statistics were estimated using full information maximum likelihood estimation (FIML). The column *Mis* contains proportions of missing values. Note that achievement variables were often missing for whole classes (170 classes in math, 150 classes in Swedish, and 155 classes in general). The self-concept variables are binary such that 0 indicates that the student stated that he/she was not good and 1 that he/she was good at the respective domain. The sex variable is binary with 0 for male and 1 for female. VP_{cla} is the proportion of class-level variation out of the total variation of a variable derived from a three-level (individual – class – school) random intercept model. VP_{sch} is the proportion of school-level variation out of the total variation of a variable derived from a three-level (individual – class – school) random intercept model.

Supplemental Material

Table S1

Descriptive Statistics for the Nongraded Student Sample

	M	*SD*	VP_{cla}	VP_{sch}	1	2	3	4	5	6	7	8	9	10
1. Math self-concept	0.69	0.46	.02	.02										
2. Reading self-concept	0.80	0.40	.01	.02	.17									
3. Spelling self-concept	0.68	0.47	.01	.01	.08	.30								
4. General self-concept	0.69	0.46	.03	.02	.50	.30	.24							
5. Math achievement	50.09	15.16	.09	.03	.39	.15	.11	.33						
6. Swedish achievement	66.11	18.18	.10	.02	.20	.28	.33	.32	.69					
7. General achievement	70.76	18.70	.13	.05	.24	.25	.28	.34	.81	.90				
8. Age	12.86	0.34	.01	.00	-.02	-.02	-.04	-.04	-.06	-.06	-.06			
9. Sex	0.50	0.50	.00	.00	-.11	.00	.12	-.05	-.03	.16	.14	-.04		
10. Ses	2.26	0.68	.12	.05	-.12	-.08	-.04	-.14	-.28	-.30	-.30	.03	-.03	
11. Cognitive ability	22.64	5.80	.10	.02	.30	.16	.12	.28	.73	.70	.71	-.08	.01	-.27

Note. Descriptive statistics were based on the nongraded student sample (N = 4,656). Descriptive statistics were estimated using full information maximum likelihood estimation (FIML). The self-concept variables are binary with 0 indicating that the student stated that he/she was not good and 1 that he/she was good at the respective domain. The sex variable is binary with 0 for male and 1 for female. VP_{cla} is the proportion of class-level variation out of the total variation of a variable derived from a three-level (individual – class – school) random intercept model. VP_{sch} is the proportion of school-level variation out of the total variation of a variable derived from a three-level (individual – class – school) random intercept model.

Table S2

Descriptive Statistics for the Graded Student Sample

	M	*SD*	VP_{cla}	VP_{sch}	1	2	3	4	5	6	7	8	9	10	11	12
1. Math self-concept	0.68	0.47	.03	.00												
2. Reading self-concept	0.79	0.40	.01	.00	.15											
3. Spelling self-concept	0.66	0.47	.02	.00	.12	.32										
4. General self-concept	0.65	0.48	.03	.01	.46	.30	.31									
5. Math achievement	50.17	15.66	.10	.02	.42	.17	.12	.35								
6. Swedish achievement	66.96	18.29	.06	.01	.25	.32	.35	.35	.71							
7. General achievement	70.29	18.61	.08	.01	.30	.30	.30	.37	.82	.92						
8. Age	12.85	0.33	.01	.00	.02	-.03	-.03	.01	-.04	-.04	-.05					
9. Sex	0.48	0.50	.00	.00	-.08	.05	.13	-.02	-.01	.19	.15	-.02				
10. Ses	2.31	0.65	.04	.02	-.08	-.10	-.06	-.10	-.24	-.27	-.27	.05	-.01			
11. Cognitive ability	23.04	5.85	.08	.01	.35	.18	.13	.30	.76	.71	.74	-.07	.03	-.24		
12. Grade math	3.19	1.00	.04	.02	.43	.17	.13	.36	.85	.68	.74	-.02	.05	-.23	.70	
13. Grade Swedish	3.14	0.94	.02	.01	.26	.30	.35	.36	.65	.86	.82	-.04	.27	-.25	.63	.68

Note. Descriptive statistics were based on the graded student sample (N = 4,448). Descriptive statistics were estimated using full information maximum likelihood estimation (FIML). The self-concept variables are binary with 0 indicating that the student stated that he/she was not good and 1 that he/she was good at the respective domain. The sex variable is binary with 0 for male and 1 for female. VP_{cla} is the proportion of class-level variation out of the total variation of a variable derived from a three-level (individual – class – school) random intercept model. VP_{sch} is the proportion of school-level variation out of the total variation of a variable derived from a three-level (individual – class – school) random intercept model.

Table S3

Results from Logistic Regression Contextual Effects Models with Math and Reading Self-Concept as the Outcome

	Math								Reading							
	Model 1				Model 2				Model 1				Model 2			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	1.09	.07	[.95, 1.23]	< .001	1.09	.07	[.95, 1.24]	< .001	1.09	.06	[.97, 1.20]	< .001	1.09	.06	[.97, 1.20]	< .001
Age	.05	.03	[-.02, .11]	.167	.05	.03	[-.02, .11]	.162	-.03	.03	[-.09, .03]	.323	-.03	.03	[-.09, .03]	.323
Sex	-.49	.07	[-.63, -.35]	< .001	-.49	.07	[-.63, -.35]	< .001	-.24	.07	[-.38, -.10]	.001	-.24	.07	[-.38, -.10]	.001
SES	.00	.03	[-.05, .06]	.922	.01	.03	[-.05, .06]	.841	-.06	.03	[-.12, .01]	.076	-.06	.03	[-.12, .01]	.076
Cognitive abilities	.08	.06	[-.03, .20]	.152	.08	.06	[-.03, .19]	.169	-.29	.05	[-.38, -.19]	< .001	-.29	.05	[-.38, -.19]	< .001
Grading	-.10	.08	[-.26, .07]	.251	-.09	.08	[-.25, .08]	.304	-.03	.07	[-.17, .11]	.683	-.03	.07	[-.17, .11]	.687
Class achievement	-.56	.16	[-.87, -.26]	< .001	-.23	.24	[-.71, .25]	.343	-.35	.09	[-.53, -.17]	< .001	-.35	.11	[-.57, -.13]	.002
Grading x Class achievement					-.51	.28	[-1.06, .03]	.065					.01	.17	[-.31, .34]	.940

Note. The table contains results from logistic regression analyses. Outcomes are dichotomous self-concept items in the respective domain (e.g., Do you think you are good at sums? 0 for No and 1 for Yes). Achievement and class achievement resemble standardized achievement scores in the respective domains. Grading is a dichotomous variable (0 for nongraded and 1 for graded).

Table S4

Results from Logistic Regression Contextual Effects Models with Spelling and General Self-Concept as the Outcome

	Spelling								General							
	Model 1				Model 2				Model 1				Model 2			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	1.31	.05	[1.21, 1.42]	< .001	1.31	.05	[1.21, 1.42]	< .001	.93	.07	[.79, 1.06]	< .001	.92	.07	[.79, 1.06]	< .001
Age	-.05	.03	[-.11, .00]	.057	-.05	.03	[-.11, .00]	.057	.02	.03	[-.04, .07]	.528	.02	.03	[-.04, .07]	.528
Sex	.22	.06	[.11, .33]	< .001	.22	.06	[.11, .33]	< .001	-.46	.06	[-.58, -.34]	< .001	-.46	.06	[-.58, -.34]	< .001
SES	.07	.03	[.01, .13]	.021	.07	.03	[.01, .13]	.021	-.05	.03	[-.11, .01]	.084	-.05	.03	-[.11, .01]	.088
Cognitive abilities	-.58	.05	[-.67, -.49]	< .001	-.58	.05	[-.67, -.49]	< .001	.10	.05	[.00, .20]	.058	.10	.05	[.00, .20]	.056
Grading	-.09	.06	[-.21, .03]	.129	-.09	.06	[-.21, .03]	.134	-.31	.09	[-.49, -.13]	.001	-.31	.09	[-.49, -.13]	.001
Class achievement	-.38	.07	[-.53, -.24]	< .001	-.40	.09	[-.56, -.23]	< .001	-.48	.11	[-.69, -.27]	< .001	-.41	.14	[-.68, -.13]	.003
Grading x Class achievement					.02	.12	[-.22, .26]	.867					-.15	.19	[-.52, .22]	.422

Note. The table contains results from logistic regression analyses. Outcomes are dichotomous self-concept items in the respective domain (e.g., Do you think you are good at sums? 0 for No and 1 for Yes). Achievement and class achievement resemble standardized achievement scores in the respective domains. Grading is a dichotomous variable (0 for nongraded and 1 for graded).

Table S5

Results from Linear Probability Models Investigating Differential Grading Effects for Low and High Achievers on Math and Reading Self-Concept

	Math				Reading				Spelling				General			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	.20	.01	[.17, .23]	< .001	.15	.01	[.13, .17]	< .001	.24	.01	[.22, .26]	< .001	.17	.01	[.14, .19]	< .001
Age	.01	.01	-[.01, .02]	.270	-.01	.01	-[.02, .00]	.197	-.01	.01	-[.02, .00]	.041	.00	.01	-[.01, .01]	.665
Sex	-.08	.01	-[.10, -.05]	< .001	-.03	.01	-[.05, -.01]	.002	.04	.01	[.02, .06]	< .001	-.08	.01	-[.10, -.06]	< .001
SES	.00	.01	-[.01, .01]	.549	-.01	.00	-[.01, .00]	.174	.01	.01	[.00, .02]	.017	-.01	.01	-[.02, .00]	.162
Cognitive abilities	.01	.01	-[.01, .03]	.303	-.04	.01	-[.05, -.03]	< .001	-.11	.01	-[.12, -.09]	< .001	.02	.01	[.00, .04]	.067
Grading	-.03	.01	-[.05, .00]	.057	-.01	.01	-[.03, .01]	.337	-.03	.01	-[.05, .00]	.054	-.05	.02	-[.08, -.02]	.001
Class achievement	-.09	.02	-[.14, -.05]	< .001	-.05	.02	-[.08, -.02]	.003	-.08	.02	-[.12, -.05]	< .001	-.09	.02	-[.13, -.05]	< .001
Grading x Achievement	-.01	.01	-[.03, .02]	.687	.02	.01	[.00, .04]	.069	.01	.01	-[.01, .03]	.195	.01	.01	-[.02, .03]	.514

Note. The table contains results from multilevel linear probability analyses. Outcomes are dichotomous self-concept items in the respective domain (e.g., Do you think you are good at sums? 0 for No and 1 for Yes). Achievement and class achievement resemble standardized achievement scores in the respective domains. Grading is a dichotomous variable (0 for nongraded and 1 for graded). As grading might accentuate self-concept differences between low and high achievers, we conducted additional analyses in which we modeled the interaction between grading and individual achievement. The interaction was not significantly different from zero in any of the domains. Hence, low and high achievers did not differ in effects of grading on academic self-concept.

Table S6

Results from Contextual Effects Models with Multi-Item General Self-Concept as the Outcome

	General							
	Model 1				Model 2			
	b	*SE*	95% CI	*p*	*b*	*SE*	95% CI	*p*
Achievement	.15	.01	[.13, .16]	<.001	.15	.01	[.13, .16]	<.001
Grading	-.03	.01	[-.05, -.01]	.009	-.03	.01	[-.05, -.01]	.008
Class achievement	-.07	.01	[-.10, -.04]	<.001	-.06	.02	[-.11, -.02]	.003
Grading x Class achievement					-.01	.03	[-.07, .05]	.739

Note. The table contains results from multilevel linear probability analyses. Grading is a dichotomous variable (0 for nongraded and 1 for graded). All analyses are controlled for age, sex, SES, and cognitive abilities

6 Study 4: The Dark Side of Detracking: Mixed-Ability Classrooms Hurt Low-Achievers' Math Motivation

Fleischmann, M., Hübner, N., Nagengast, B., Trautwein, U. (2020). The Dark Side of Detracking: Mixed-Ability Classrooms Hurt Low-Achievers' Math Motivation. Manuscript ready for submission.

Abstract

We tested how detracking school reforms, which abolish ability grouping and introduce mixed-ability classrooms, affect students' math motivation. To do so, we made use of data from two unique natural experiments ($N_{Study\ 1}$ = 78,376, $N_{Study\ 2}$ = 2,257) and compared student cohorts before and after detracking. In both studies, we found low achievers' math motivation to be substantially lower after the reform, whereas this was not the case for high achievers. Our study reminds researchers and policymakers that detracking school reforms can come with unintended side effects on student motivation. Only when such negative side effects are reduced or eliminated, detracking school reforms can unfold their full potential in establishing educational equality.

The Dark Side of Detracking: Mixed-Ability Classrooms Hurt Low-Achievers' Math Motivation

Tracking—that is, grouping students with similar ability levels into different schools, study programs, or courses—has been a hotly discussed issue in educational research and policy-making (Ansalone, 2003; Loveless, 1999). Generally, tracking is intended to create homogenous learning environments that, in theory, optimally correspond to the specific needs of students with different abilities (Hallinan, 1994). A form of tracking that is widespread in the United States as well as many other Anglo-Saxon countries is course-by-course tracking, the grouping of students with similar achievement levels into different courses which are then taught according to curricula that differ in their performance requirements (Chmielewski, 2014). Tracking opponents have argued that track placement is class- and race-biased (Kershaw, 1992), providing unequal learning opportunities (Riordan, 1997), consequently promoting academic achievement of high achievers and holding back low achievers (Rui, 2009). Thus, practices of detracking have gained popularity and have been implemented in school systems all around the world (Burris & Welner, 2005; Domina & Saldana, 2012; Hallinan, 2004). However, there might also be paradoxical side effects of such reform efforts.

In contrast to the question of how tracking affects student achievement, less attention has been paid to the question of how it affects student motivation. One important motivational variable is students' self-perception of competence, referred to as academic self-concept (Marsh et al., 2016). Academic self-concept is regarded as important because it is a significant predictor of academic effort and achievement (Huang, 2011; Valentine et al., 2004), as well as academic aspirations and choices (Betz & Hackett, 1981; Eccles & Wigfield, 2002; Hackett & Betz, 1981). Because of the strong domain-specific nature of academic self-concept (Marsh et al., 2016), it comes as no surprise that fostering academic self-concept in STEM domains is seen as a promising measure to increase enrollment rates in STEM-related study programs (Keyserlingk et al., 2019).

Proponents of detracking school reforms have argued that "track membership provides a single, highly visible, unambiguous label that instantaneously communicates stigma" (Rosenbaum, 1976, p. 169), thus negatively affecting low-achieving students' academic self-concept. Consequently, detracking has been assumed to raise the motivation of low achievers by removing the negative branding of the track level for this group of students (Esposito, 1973; Oakes, 2005). Empirical evidence for this labeling hypothesis stems mainly from studies that showed track level to be positively associated with academic self-beliefs (e.g., Byrne, 1988,

1990; Gamoran & Berends, 1987; Hargreaves, 1967; Kelly, 1975; Lacey, 1974; Mann, 1960; Nachmias, 1977). However, these correlational studies are prone to selection bias as students from high tracks typically differ drastically from those in low tracks.

In contrast, there is also evidence for the opposite effect: In non-experimental studies, students' academic self-concept is negatively predicted by the average achievement of the other students in their educational environment, such as the school or the classroom (Marsh & Seaton, 2015). In other words, given similar individual achievement, membership in low-achieving educational environments boosts academic self-concept—a phenomenon known as the big fish little pond effect (BFLPE; Marsh, 1987). The BFLPE is assumed to be the consequence of social comparison processes (Huguet et al., 2009; Marsh et al., 2014). Research has shown that the BFLPE is a cross-cultural phenomenon (Marsh & Hau, 2003; Nagengast & Marsh, 2012), which persists even after the completion of high school (Marsh et al., 2007). Negative frame-of-reference effects have been found also to affect academic interest (Trautwein et al., 2006), academic aspirations (Nagengast & Marsh, 2012), and a large number of other academic outcomes (Marsh, 1991).

Based on research on the BFLPE, one can make clear predictions on how detracking affects student motivation. After detracking, low achievers are taught in mixed-ability classrooms with higher average ability compared to tracked classrooms; high achievers, on the other hand, have classmates with lower average ability after detracking. Consequently, based on BFLPE theory, one would expect that detracking negatively affects low achievers' motivation, whereas it has no such or even a positive effect for high-achieving students. To investigate how detracking impacts student motivation in terms of academic self-concept, we evaluated one Austrian and one German detracking school reform. Both reforms provide us with the unprecedented opportunity to test via a natural experiment the differential effects of detracking on student motivation for high and low achievers—an issue that is of high theoretical and practical relevance. The two reforms are ideally suited for an investigation of detracking effects on students because they focus on two stages of secondary education and focus on two types of tracking, namely, achievement grouping and opt-in tracking. Educational reforms that are implemented from one school year to another often can be regarded as natural experiments (Murnane & Willett, 2010). We evaluated these experiments by using a cohort-control design, that is, comparing student cohorts before and after respective school reforms (Shadish et al., 2002).

Until the end of school year 2011/2012, Austrian general secondary school students were divided into three tracks according to their ability level in the core subjects mathematics,

German, and English. Track assignment was based on prior achievement. This type of tracking is called achievement tracking (Trautwein et al., 2005). Starting in the school year 2012/2013, these schools were successively transformed into "new general secondary schools". The most drastic reform element was detracking: Ability grouping in the core subjects was abolished and replaced by mixed-ability grouping. After the reform, the curriculum for all students resembled the high-track curricula from before the reform (Eder et al., 2015). Similarly, until the school year 2008/2009, German upper secondary school students in the German state Thuringia could choose between two tracks (low track, high track) in mathematics as well as German. Unlike the tracking in Austria, track assignment was based on student choice. This type of tracking is called opt-in tracking (Trautwein et al., 2005). After the reform (starting in the school year 2009/2010), ability grouping was abolished, and the curriculum for all students resembled the high-track curricula from before the reform (for a more detailed description of the detracking school reforms in Austria and Germany, see Supplementary Material).

For evaluating the Austrian detracking reform (Study 1), we use data from the national educational standard assessment in 2012 and 2017 (BIFIE, 2016, 2018). In our analysis, we included only students from those schools that were lower secondary schools in 2012 and new lower secondary schools in 2017. This procedure resulted in a sample of 78,376 students (40,931 before and 37,445 after the reform). The sample was 47.49% female, and the average age was M = 14.46 (SD = 0.55) years. For evaluating the detracking reform in the German state Thuringia (Study 2), we used data from the Additional Study Thuringia from the German National Educational Panel Study (NEPS; Blossfeld et al., 2011). This study measured a random student sample from 32 schools before (2010) and after (2011) the detracking reform. This resulted in a sample of 2,257 students (1,372 before and 885 after the reform). The sample was 54.17% female, and the average age was M = 17.41 years (SD = 0.78). The Additional Study Thuringia has already been used to investigate differential detracking effects for boys and girls (see Hübner et al., 2019); however, low and high achievers have so far not been considered.

In both studies, mathematics self-concept (in the following: self-concept) was assessed with self-report items (e.g., *Usually, I am good at mathematics;* see Tables S3 and S4 for complete item batteries), which were answered on a 4-point response scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). Additionally, mathematics achievement (in the following: achievement) was measured using standardized tests focusing on mathematical literacy, that is, the mastery of processes, understanding of concepts, and the ability to deal with

different everyday situations and problems within a competence area (cf. OECD, 2017a). The data were analyzed using three-level regression models with students nested in classes and schools. We regressed self-concept on standardized achievement, a cohort dummy (with values 0 for "before detracking" and 1 for "after detracking"), and the interaction between both variables. To control for potential differences between cohorts, we additionally controlled for a set of covariates, namely gender, age, socioeconomic background, and migration background (for a more detailed description of instruments and analyses, see Supplementary Material).

To address the differential effects of detracking reforms on high- and low-achieving students' academic self-concepts, we focus on the interaction between the cohort dummy variable and student achievement. We predicted this interaction effect to be positive, implying that detracking reforms have more positive effects on the academic self-concept of high achievers as opposed to low achievers and vice versa. In line with our hypothesis, we found the interaction between the cohort dummy and student achievement to be significantly positive in both studies (b = .18, p < .001 in Austria and b = .13, p = .002 in Thuringia; Table 1). Thus, the cohort effect was more positive for high-achieving students compared to low-achieving students and thus more negative for low-achieving compared to high-achieving students. To examine the specific detracking effects for students with different abilities, we calculated predicted values for high-, average-, and low-achieving students (see Figure 1). In both studies, high-achieving students (one standard deviation above the mean) had similar self-concept before and after the detracking reforms (b < .01, p = .938 in Austria and b = .05, p = .456 in Thuringia). Average-achieving students had lower self-concept after detracking in Austria (b = -.18, p = <.001) but not in Thuringia (b = -.08, p = .118). In both studies, low-achieving students (one standard deviation below the mean) had significantly lower academic self-concept after detracking (b = -.36, p < .001 in Austria and b = -.12, p = .003 in Thuringia).

Table 1

Results from Multilevel Regression Models

	Austria				Thuringia (Germany)			
	b	SE	95% CI	*p*	*b*	SE	95% CI	*p*
Sex	-.30	.01	[.28, .31]	<.001	-.04	.04	[-.12, .05]	.382
Age	-.01	.00	[-.02, -.01]	<.001	-.02	.02	[-.05, .03]	.463
SES	.01	.00	[.00, .02]	<.001	.03	.02	[-.01, .08]	.147
Migration	.24	.01	[.22, .25]	<.001	.15	.07	[.01, .29]	.038
Cohort	-.18	.01	[-.20, -.15]	<.001	-.08	.05	[-.19, .02]	.118
Achievement	.49	.01	[.48, .50]	<.001	.33	.03	[.27, .38]	<.001
Cohort x achievement	.18	.01	[.16, .19]	<.001	.13	.04	[.05, .21]	.002

Note. The dependent variable is academic self-concept (standardized). Sex is a dichotomous variable with 0 for males and 1 for females. Migration is a dichotomous variable with 0 for no migration background and 1 for migration background. Cohort is a dichotomous variable with 0 for cohort 1 (before detracking) and 1 for cohort 2 (after detracking). All continuous predictors are standardized.

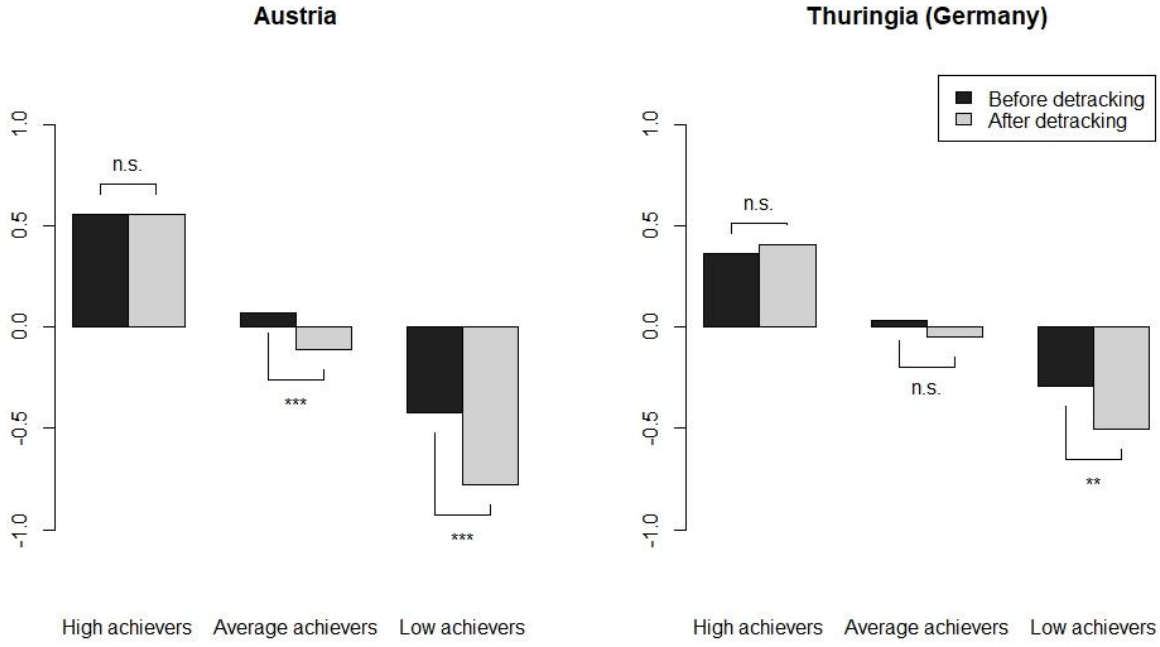

Figure 1. Predicted values of self-concept for high, average, and low achievers.

Using data from two natural experiments, our study found low achievers to have substantially lower academic self-concept after detracking reforms, whereas this was not the case for high achievers. Thus our results suggest that detracking—that is, the abolishment of ability grouping and introduction of mixed-ability classrooms—negatively affects low achievers' academic self-concept. Our findings clearly speak against the labeling hypothesis, namely the assumption that detracking relieves low achievers from being labeled as low-track students, thus increasing their academic self-concept. Instead, our results are in line with research on the BFLPE that predicts low achievers' self-concept to suffer as a consequence of being exposed to unfavorable social comparisons in mixed-ability classrooms.

Our study comes with several implications for educational practice. Detracking school reforms are often conducted in STEM subjects (science, technology, engineering, and mathematics; Domina & Saldana, 2012; Hübner et al., 2019), which are believed to be of specific importance to individuals as well as society (OECD, 2010). Detracking in STEM areas is often conducted with the aim of homogenizing student achievement in those specific subjects and consequently providing equal career chances for all students. Our study shows that this calculation might not necessarily pay off. As academic self-concept is a decisive determinant of occupational aspirations and choices (Eccles, 2009), detracking may hamper lower achieving students' pursuit of a STEM career.

Our study evaluated two natural experiments that perfectly complemented each other by focusing on two stages of secondary education as well as by examining two types of course-by-course tracking, namely achievement grouping and opt-in tracking. Although cohort-control designs also come with some limitations, they are among the strongest designs to evaluate educational reform efforts as randomized controlled trials are virtually impossible to implement because of political and ethical reasons (Murnane & Willett, 2010; Rochon et al., 2005; Shadish et al., 2002).

Our study highlights the potential side effects of detracking school reforms that aim at establishing equality in educational outcomes but might have the opposite effect, at least in some respects. It thereby encourages educational policymakers to think of how educational reforms might affect non-cognitive factors such as student motivation. More specifically, researchers and practitioners have to think about how to cushion negative side effects. Future research is needed in this area. To get an overall picture, more research is also required on how detracking affects student achievement. Whereas findings are mixed here (e.g., Duflo et al., 2011; Hanushek & Wößmann, 2006), reducing negative side effects on academic self-concept

will likely support the desired homogenizing effect on achievement. Only when negative side effects of detracking school reforms are reduced or eliminated, such efforts can achieve their full potential in establishing equality in educational outcomes.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*(2-3), 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008

Ansalone, G. (2003). Poverty, tracking, and the social construction of failure: International perspectives on tracking. *Journal of Children and Poverty, 9*(1), 3–20. https://doi.org/10.1080/1079612022000052698

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Betz, N. E., & Hackett, G. (1981). The relationship of career-related self-efficacy expectations to perceived career options in college women and men. *Journal of Counseling Psychology, 28*(5), 399–410. https://doi.org/10.1037/0022-0167.28.5.399

BIFIE. (2016). *Datensatz zur Standardüberprüfung 2012 Mathematik, 8. Schulstufe. Schülerebene, M812I. Forschungsdatenbibliothek (FDB). Nicht-imputierter Datensatz, v2.0.* BIFIE.

BIFIE. (2018). *Datensatz zur Standardüberprüfung 2017 Mathematik, 8. Schulstufe. Schülerebene, M817I. Forschungsdatenbibliothek (FDB). Nicht-imputierter Datensatz, v1.0-E.*

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

Burris, C. C., & Welner, K. G. (2005). Closing the achievement gap by detracking. *Phi Delta Kappan, 86*(8), 594–598. https://doi.org/10.1177/003172170508600808

Byrne, B. M. (1988). Adolescent self-concept, ability grouping, and social comparison. *Youth & Society, 20*(1), 46–67. https://doi.org/10.1177/0044118X88020001003

Byrne, B. M. (1990). Self-concept and academic achievement: Investigating their importance as discriminators of academic track membership in high school. *Canadian Journal of Education, 15*(2), 173. https://doi.org/10.2307/1495374

Chmielewski, A. K. (2014). An international comparison of achievement inequality in within- and between-school tracking systems. *American Journal of Education, 120*(3), 293–324. https://doi.org/10.1086/675529

Domina, T., & Saldana, J. (2012). Does raising the bar level the playing field? *American Educational Research Journal, 49*(4), 685–708. https://doi.org/10.3102/0002831211426347

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, *101*(5), 1739–1774. https://doi.org/10.1257/aer.101.5.1739

Eccles, J. S. (2009). Who am i and what am i going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist*, *44*(2), 78–89. https://doi.org/10.1080/00461520902832368

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

Eder, F., Altrichter, H., Hofmann, F., & Weber, C. (Eds.). (2015). *Evaluation der Neuen Mittelschule (NMS). Befunde aus den Anfangskohorten.* Leykam.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Esposito, D. (1973). Homogeneous and heterogeneous ability grouping: Principal findings and implications for evaluating and designing more effective educational environments. *Review of Educational Research*, *43*(2), 163–179. https://doi.org/10.3102/00346543043002163

Gamoran, A., & Berends, M. (1987). The effects of stratification in secondary schools: Synthesis of survey and ethnographic research. *Review of Educational Research*, *57*(4), 415–435. https://doi.org/10.3102/00346543057004415

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

Hackett, G., & Betz, N. E. (1981). A self-efficacy approach to the career development of women. *Journal of Vocational Behavior*, *18*(3), 326–339. https://doi.org/10.1016/0001-8791(81)90019-1

Hallinan, M. T. (1994). Tracking: From theory to practice. *Sociology of Education*, *67*(2), 79. https://doi.org/10.2307/2112697

Hallinan, M. T. (2004). The detracking movement. *Education Next*, *4*(4), 72–76.

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

Hargreaves, D. H. (1967). *Social relations in a secondary school*. Routledge.

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, *49*(5), 505–528. https://doi.org/10.1016/j.jsp.2011.07.001

Hübner, N. (2017). *Educational effectiveness at the end of upper secondary school: Further insights into the effects of statewide policy reforms*. Dissertation.

Hübner, N., Wagner, W., Nagengast, B., & Trautwein, U. (2019). Putting all students in one basket does not produce equality: Gender-specific effects of curricular intensification in upper secondary school. *School Effectiveness and School Improvement*, *14*(2), 1–25. https://doi.org/10.1080/09243453.2018.1504801

Huguet, P., Dumas, F., Marsh, H. W., Wheeler, L., Seaton, M., Nezlek, J., Suls, J., & Régner, I. (2009). Clarifying the role of social comparison in the big-fish-little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology*, *97*(1), 156–170. https://doi.org/10.1037/a0015558

Kelly, D. H. (1975). Tracking and its impact upon self-esteem: A neglected dimension. *Education*, *96*, 2–9.

Kershaw, T. (1992). The effects of educational tracking on the social mobility of african americans. *Journal of Black Studies*, *23*(1), 152–169. https://doi.org/10.1177/002193479202300111

Keyserlingk, L. von, Becker, M., Jansen, M., & Maaz, K. (2019). Leaving the pond—Choosing an ocean: Effects of student composition on STEM major choices at university. *Journal of Educational Psychology.* Advance online publication. https://doi.org/10.1037/edu0000378

Lacey, C. (1974). Destreaming in a pressured academic environment. In J. Eggleston (Ed.), *Contemporary Research in the Sociology of Education* (pp. 148–166). Methuen.

Loveless, T. (1999). *The tracking wars*. Brookings Institution Press.

Mann, M. (1960). What does ability grouping do to the self-concept? *Childhood Education*, *36*(8), 357–360. https://doi.org/10.1080/00094056.1960.10727801

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. (1991). Failure of high-ability high schools to deliver academic benefits commensurate with their students' ability levels. *American Educational Research Journal*, *28*(2), 445–480. https://doi.org/10.2307/1162948

Marsh, H. W., & Hau, K.-T. (2003). Big-fish-little-pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist*, *58*(5), 364–376. https://doi.org/10.1037/0003-066X.58.5.364

Marsh, H. W., Kuyper, H., Morin, A. J., Parker, P. D., & Seaton, M. (2014). Big-fish-little-pond social comparison and local dominance effects: Integrating new statistical models, methodology, design, theory and substantive implications. *Learning and Instruction, 33,* 50–66. https://doi.org/10.1016/j.learninstruc.2014.04.002

Marsh, H. W., Martin, A. J., Yeung, A. S., & Craven, R. (2016). Competence self-perceptions. In C. Dweck & D. Yaeger (Eds.), *Handbook of competence and motivation.* Guilford Press.

Marsh, H. W., & Seaton, M. (2015). The big-fish–little-pond effect, competence self-perceptions, and relativity: Substantive advances and methodological innovation. *Advances in Motivation Science, 2,* 127–184.

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

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Muthén, L. K., & Muthén, B.O. (1998-2018). *Mplus user's guide*. Muthén & Muthén.

Nachmias, C. (1977). The issue of saliency and the effect of tracking on self-esteem. *Urban Education, 12,* 327–344. https://doi.org/10.1177/0042085977123007

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

Oakes, J. (2005). *Keeping track: How schools structure inequality*. Yale University Press.

OECD. (2010). *OECD Information Technology Outlook*. OECD.

OECD. (2017a). *Pisa 2015 assessment and analytical framework. Science, reading, mathematic, financial literacy and collaborative problem solving*. OECD.

OECD. (2017b). *PISA 2015 Technical Report*. OECD.

R Core Team. (2008). *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage.

Riordan, C. (1997). *Equality and achievement*. Longman Press.

Robitzsch, A., Grund, S., & Henke, T. (2018). *Miceadds: Some additional multiple imputation functions, especially for mice.*

Rochon, P. A., Gurwitz, J. H., Sykora, K., Mamdani, M., Streiner, D. L., Garfinkel, S., Normand, S.-L. T., & Anderson, G. M. (2005). Reader's guide to critical appraisal of cohort studies: 1. Role and design. *BMJ, 330*, 895–897. https://doi.org/10.1136/bmj.330.7496.895

Rosenbaum, J. E. (1976). *Making inequality: The hidden curriculum of high school tracking.* Wiley.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Rui, N. (2009). Four decades of research on the effects of detracking reform: Where do we stand? - A systematic review of the evidence. *Journal of Evidence-Based Medicine*, *2*(3), 164–183. https://doi.org/10.1111/j.1756-5391.2009.01032.x

Schreiner, C., & Breit, S. (2012). *Standardüberprüfung 2012 Mathematik, 8. Schulstufe. Bundesergebnisbericht*. BIFIE.

Schreiner, C., Breit, S., Pointinger, M., Pacher, K., Neubacher, M., & Wiesner Christian. (2017). *Standardüberprüfung 2017 Mathematik, 8. Schulstufe. Bundesergebnisbericht.* BIFIE.

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

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

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 so exist after all. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *New frontiers for self research* (pp. 307–327). CT: IAP.

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

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist*, *39*(2), 111–133. https://doi.org/10.1207/s15326985ep3902_3

von Davier, M. von, Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Vol. 2. Issues and methodologies in large scale assessments* (pp. 9–36). IEA-ETS Research Institute.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2-3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Supplemental Material

Detracking School Reforms in Austria and Germany

The Austrian Educational System and the Introduction of the New Secondary School (Study 1)

In Austria, children enter four-year elementary school at the age of six. After Grade 4, they are tracked in either a vocational or academic track school. The low track prepares students for vocational training or the transition to upper secondary schools.

Until the school year 2011/2012, the low-track schools were called "Hauptschulen" (in the following: secondary schools). Students in secondary schools were divided into three performance groups according to their ability level in the subjects mathematics, German, and English. In the school year 2011/2012, about 44%, 34%, and 17% of secondary school students attended performance group 1 (high-performance group), group 2 (average-performance group), and group 3 (low-performance group) in math, respectively. From the school year 2012/2013, these secondary schools were successively transformed into "Neue Mittelschulen" (in the following: new secondary schools). New secondary schools came with far-reaching changes concerning teaching practices (Eder et al., 2015), which are summarized in Table S1. The most drastic reform element was detracking. With regard to class composition, ability grouping in the core subjects was abolished and replaced by mixed-ability grouping. With regard to the curricula, all students in new secondary schools were taught according to the curricula that in secondary schools were reserved for students in the highest performance group.

Next to detracking, other reform elements were implemented. Another reform feature concerned instruction. Students should receive instruction that fits their individual needs, for example, by individualized assignments adjusted to their performance level. Team teaching—namely, cooperative teaching of two instructors—was established in the core subjects to cope with the increased heterogeneity of the student body and to facilitate individualized instruction. Another reform feature concerned performance feedback. In years 7 to 8, students were evaluated concerning two different grading standards, namely "basic education" and "deepened education". In deepened education, students could earn the grades 1 (very good) to 4 (sufficient). In basic education, students could earn the grades 3 (satisfying) to 5 (insufficient). Thus, the grading system of the new secondary school resulted in a 7-stage grading scale. Additional to regular report cards, a complimentary evaluation in the form of differentiated performance feedback, that in written form emphasized individual progress, was delivered to students. Also semi-annual teacher-parent-child conversations were introduced in which learning aims were formulated.

The German Educational System and Reform of the Upper Secondary School System (Study 2)

In Germany, children enter four-year elementary school at the age of six. After Grade 4, they attend either Hauptschule, Realschule, or Gymnasium. All three school types differ in their curriculum and respective performance requirements. Hauptschule is the less ambitious school type that prepares students for vocational training. Realschule is a more ambitious school type preparing students for vocational training or the transition to the Gymnasium. The Gymnasium is the most ambitious school type, preparing students for university (for an overview of the German school system, see also Hübner, 2017).

Until the school year 2008/2009, Gymnasium students in the German state Thuringia could, by choice, enroll in a basic or advanced course in either math or German. After the reform , they had to attend an advanced course in both subjects. Additionally, after the reform, all students were taught according to the curricula that before the reform were reserved for students in advanced courses.

Next to detracking in the subjects mathematics and German, two other reform elements were implemented. Whereas before the reform, students had to choose two advanced courses (one of them had to be German or math) and two basic courses, after the reform, three other advanced courses next to compulsory ones in German and math had to be chosen. Additionally, whereas before the reform, different courses weighted differentially for the final examination grades, after the reform, the weighting was identical for all students. For a summary of the reform elements, see Table S2.

Data

Data for Study 1

Data for Study 1 was drawn from the Austrian national educational standard assessments from 2012 (BIFIE, 2016; Schreiner & Breit, 2012) and 2017 (BIFIE, 2018; Schreiner et al., 2017). The Austrian national educational standard assessments are conducted by the Federal Institute for Educational Research, Innovation, and Development of the Austrian School System (BIFIE; now Federal Institute for Quality Assurance of the Austrian School System, IQS), which is responsible for educational monitoring.

In both assessment years, all Austrian eighth-grade students without special educational needs were tested in the domain of mathematics. As the detracking school reform was implemented in the school year 2012/2013, the majority of secondary school students in the Austrian national educational standard assessment from 2012 were tracked in Math, German, and English. By contrast, the majority of secondary school students in the Austrian national educational standard assessment from 2017 were not tracked in Math, German, and English.

In our analysis, we included only students from those schools that were lower secondary schools in 2012 and new lower secondary schools in 2017. This procedure resulted in a sample of 78,376 students that were nested in 1727 schools and 4898 math classes. The average age was M = 14.46 (SD = 0.55) years, and 47.49% of students were female. Cohort 1 contained 40,931 students that were nested in 865 schools and 2786 math classes. There were M = 3.22 (SD = 1.11) classes per school and M = 14.96 (SD = 5.39) students per class. The average age was M = 14.45 (SD = 0.54) years, and 47.42% of students were female. Cohort 2 contained 37,445 students that were nested in 862 schools and 2112 math classes. There were M = 2.45 (SD = 0.92) classes per school and M = 17.73 (SD = 4.13) students per class. The average age was M = 14.47 (SD = 0.55) years, and 47.57% of students were female.

Data for Study 2

Data for Study 2 were drawn from the Additional Study Thuringia from the German National Educational Panel Study (NEPS; Blossfeld et al., 2011). NEPS is a longitudinal multi-cohort study carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg. The Additional Study Thuringia was explicitly designed for investigating the effects of the upper secondary school reform in Thuringia.

Study participants were selected with the help of a two-stage procedure. First, 32 Gymnasiums in the German state Thuringia were randomly sampled. Second, all twelfth-grade

students in the school year 2009/2010 (measured in January and February 2010) and in the school year 2010/2011 (measured in January 2011) were invited to participate. As the detracking school reform in Thuringia was implemented in the school year 2009/2010, students that were measured in 2010 were tracked in German and Math. By contrast, students that were measured in 2011 were not tracked in German and Math.

This procedure resulted in a sample of 2,257 students. Cohort 1 contained 1,372 students that were nested in 32 schools and 113 math classes. There were M = 3.53 (SD = 1.27) math classes per school and M = 12.02 (SD = 4.77) students per math class. Fifty-three percent of students were female, and the average age was M = 17.91 (SD = 0.69) years. Cohort 2 contained 885 students that were nested in 32 schools and in 74 math classes. There were M = 2.5 (SD = 0.73) classes per school and M = 11.68 (SD = 5.02) students per math class. Fifty-six percent of students were female, and the average age was M = 17.79 (SD = 0.64) years.

Instruments and Analyses

Instruments and Analyses for Study 1

Mathematics self-concept (in the following: self-concept) was assessed using four items (*Usually, I am good at mathematics; Mathematics is harder for me than for many of my classmates; I am just not good at mathematics; I learn quickly in mathematics;* see also Table S3), which were answered on a 4-point Likert scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). For subsequent analyses, a mean score comprising these items was built (at least two items had to be completed for mean score calculation, α = .86).

Math achievement was measured with a math competency test that lasted about 90 minutes. Students completed approximately 48 multiple-choice items based on a multi-matrix booklet design. Additionally, there was also a limited amount of half-open and open items. The BIFIE provides ten plausible values sampled from the likely distribution of a person's ability (von Davier et al., 2009; Wu, 2005). Large-scale assessment studies typically use plausible values because it allows taking into account the uncertainty of person parameter estimation. We calculated a reliability coefficient by deducting the within-person PV variance proportion from 1 (Adams, 2005; OECD, 2017b). A reliability coefficient close to one indicates that PVs vary within individuals only to a small extent, thus pointing to high measurement accuracy. The reliability coefficient in our sample was 0.91 and thus exceeded the one from the PISA 2015 assessment (0.85). We take this as evidence for the high psychometric quality of the mathematics achievement test**.** To inspect general cohort differences and to further strengthen the internal validity of our statistical models, we made use of several covariates, namely gender (dichotomous: 0 for male and 1 for female), age, socioeconomic background (indicated by highest occupational status of parents; HISEI), and migration background (dichotomous: 0 for no migration background, 1 for migration background).

For addressing our research questions, we applied hierarchical linear regression analysis (e.g., Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). As students were hierarchically nested in classes and schools, we applied a three-level approach, modeling the individual, the class, and the school levels. For all analyses, we used the statistical computing software R (R Core Team, 2008) and the package lme4 (Bates et al., 2015). All analyses were run ten times—once for each plausible value as the outcome variable. Results were then pooled according to Rubin's (1987) rule using the lmer_pool function from the R-package miceadds (Robitzsch et al., 2018).

We ran all our statistical models using a complete case analysis approach (also known as "listwise deletion"). Thus, cases that had missing values on at least one model variable were excluded. The exclusion rates for all statistical models were below the 5% boundary that does not lead to substantial parameter bias or loss in statistical power (Graham, 2009). All level-1 variables were standardized, and all achievement aggregates (class- and school-average achievement) were calculated based on the standardized level-1 measure and not re-standardized. As a result, all achievement variables are in the same metric (standard deviations of individual achievement), making coefficients comparable across levels.

Instruments and Analyses for Study 2

Self-concept was assessed using three items (*I simply have no talent for mathematics, I'm not very good at mathematics, I am good at mathematics*; see also Table S4), which were answered on a 4-point Likert scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). For subsequent analyses, a mean score comprising these items was built (at least two items had to be completed for mean score calculation, α = .94).

Mathematics achievement was measured with a math competency test that lasted about 30 minutes. Similar to the PISA assessment, mathematics achievement was conceptualized as mathematical literacy. NEPS provides a WLE estimate for mathematics achievement. The reliability of the test was acceptable (reliability of the weighted likelihood estimator: WLE = .67). To inspect general cohort differences and to further strengthen the internal validity of our statistical models, we made use of several covariates, namely gender (dichotomous: 0 for male and 1 for female), age, socioeconomic background (indicated by highest occupational status of parents; HISEI), and migration background (dichotomous: 0 for no migration background, 1 for migration background).

We applied a two-level hierarchical linear regression analysis using the statistical computing software Mplus (Muthén & Muthén, 1998-2018). Identical to Study 1, we regressed self-concept on the cohort dummy, achievement, the interaction between the cohort dummy and achievement, and the covariates. In the statistical models, full information maximum likelihood estimation (FIML) was used to account for missing values (Enders, 2010; Graham, 2009).

Table S1

Reform Overview Study 1

Before the reform	After the reform
Tracking	
In Math, German, and English, students were divided into three tracks	In Math, German, and English, students were taught in mixed-ability classrooms
In Math, German, and English, different tracks were taught according to curricula that differed in their performance requirements	In Math, German, and English, all students were taught with the same curriculum that before the reform was reserved for students in the highest track
Additional reform elements	
No special focus on individualized instruction	Special focus on individualized instruction
One teacher per classroom	Team teaching
Students were graded on track-specific grading scales	Single grading scale in years 5 and 6; Two grading standards (basic education vs. deepened education) in years 7 and 8
	Additional written performance feedback
	Additional semi-annual teacher-parent-child conversations

Table S2

Reform Overview Study 2

Before the reform	After the reform
Tracking	
Students could choose to attend a high-track class in German or Math while attending the other subject in a low-track class	In both German and Math, students were taught in mixed-ability classrooms
Different tracks were taught according to curricula that differed in their performance requirements	In German and Math, all students were taught with the same curriculum that before the reform was reserved for students in the high track
Additional reform elements	
Students had to choose another advanced course (besides the one in German or Math) and two other basic courses	Students had to choose three other advanced courses (in addition to the compulsory ones in German and Math)
Courses weighted differentially for the final examination grades	Courses weighted identically for the final examination grades

Table S3

Mathematics Self-Concept Items in Study 1

Item 1	Usually, I am good at mathematics
Item 2	Mathematics is harder for me than for many of my classmates (recoded)
Item 3	I am just not good at mathematics (recoded)
Item 4	I learn quickly in mathematics

Table S4

Mathematics Self-Concept Items in Study 2

Item 1	I simply have no talent for mathematics (recoded)
Item 2	I'm not very good at mathematics (recoded)
Item 3	I am good at mathematics

Table S5

Descriptives of Model Variables in Study 1 in Cohort 1

	M	*SD*	$\hat{\rho}_{class}$	$\hat{\rho}_{school}$	1	2	3	4	5
1. Self-concept	2.96	0.76	.05	.01					
2. Achievement	499.23	86.63	.50	.07	.39				
3. Age	14.45	0.54	.10	.06	-.07	-.26			
4. Sex	0.47	0.50	.02	.02	-.20	-.05	-.08		
5. HISEI	43.03	19.01	.06	.05	.07	.23	-.10	-.03	
6. Migration	0.21	0.41	.04	.28	-.01	-.31	.22	-.01	-.23

Note. Sex is a dummy variable with 1 for male and 2 for female. Migration is a dummy variable with 0 for no migration background and 1 for migration background. HISEI = highest occupational status of parents.

Table S6

Descriptives of Model Variables in Study 1 in Cohort 2

	M	*SD*	$\hat{\rho}_{class}$	$\hat{\rho}_{school}$	1	2	3	4	5
1. Self-concept	2.89	0.80	.00	.01					
2. Achievement	508.27	85.30	.05	.24	.53				
3. Age	14.47	0.55	.04	.08	-.11	-.25			
4. Sex	0.48	0.50	.02	.01	-.16	-.05	-.08		
5. HISEI	43.93	19.06	.03	.07	.13	.23	-.12	-.03	
6. Migration	0.24	0.43	.02	.27	-.04	-.31	.25	.00	-.25

Note. Sex is a dummy variable with 1 for male and 2 for female. Migration is a dummy variable with 0 for no migration background and 1 for migration background. HISEI = highest occupational status of parents.

Table S7

Descriptives of Model Variables in Study 2 in Cohort 1

	M	*SD*	$\hat{\rho}_{class}$	1	2	3	4	5
1. Self-concept	2.78	0.91	.19					
2. Achievement	0.01	1.12	.34	.42				
3. Age	17.91	0.69	.01	-.01	-.06			
4. Sex	0.53	0.5	.08	-.09	-.29	-.08		
5. HISEI	49.55	14.31	.10	.09	.16	-.01	-.09	
6 Migration	0.07	0.26	.01	0	0.03	0	.01	.08

Note. Sex is a dummy variable with 0 for male and 1 for female. Migration is a dummy variable with 0 for no migration background and 1 for migration background. HISEI = highest occupational status of parents.

Table S8

Descriptives of Model Variables in Study 2 in Cohort 2

	M	*SD*	$\hat{\rho}_{class}$	1	2	3	4	5
1. Self-concept	2.7	1	.03					
2. Achievement	0.02	1.08	.19	.42				
3. Age	17.79	0.64	.07	-.07	-.12			
4. Sex	0.56	0.5	.09	-.2	-.28	-.12		
5. HISEI	50.31	14.64	.11	.12	.2	-.07	-.03	
6 Migration	0.05	0.22	0	.1	.09	.02	-.07	.05

Note. Sex is a dummy variable with 0 for male and 1 for female. Migration is a dummy variable with 0 for no migration background and 1 for migration background. HISEI = highest occupational status of parents.

Table S9

Cohort Differences in Study 1

	b	SE	*p*
1. Self-concept	-.09	.01	< .001
2. Achievement	.19	.02	< .001
3. Age	.00	.01	.858
4. Sex	.02	.02	.316
5. HISEI	.06	.02	< .001
6. Migration	.02	.01	.052

Note. Sex is a dummy variable with 0 for male and 1 for female. Migration is a dummy variable with 0 for no migration background and 1 for migration background. Outcomes were regressed on the cohort dummy (0 for cohort 1 and 1 for cohort 2). HISEI = highest occupational status of parents.

Table S10

Cohort Differences in Study 2

	b	SE	*p*
1. Self-concept	-0.08	0.07	0.21
2. Achievement	0	0.09	0.96
3. Age	-0.18	0.05	<.001
4. Sex	0.02	0.03	0.47
5. HISEI	0.06	0.07	0.42
6. Migration	-0.02	0.01	0.05

Note. Sex is a dummy variable with 0 for male and 1 for female. Migration is a dummy variable with 0 for no migration background and 1 for migration background. Outcomes were regressed on the cohort dummy (0 for cohort 1 and 1 for cohort 2). HISEI = highest occupational status of parents.

7 General Discussion

The big fish little pond effect (BFLPE), or the negative effect of school- or class-average achievement on academic self-concept when controlling for individual achievement differences, is a well-researched phenomenon in educational psychology. However, there are still unresolved issues concerning the mechanisms, implications, and interdisciplinary integration of the BFLPE. These exist partly because of the homogeneity of research designs that have been used to date. In the present dissertation, I identified four of these unresolved issues and described the underlying design-based challenges on a conceptual level. The present dissertation aimed to address these unresolved issues in research on the BFLPE by using new designs. To achieve this aim, Studies 1 and 2 used extensive large-scale data in the form of comprehensive educational monitoring data and interdisciplinary data, and Studies 3 and 4 focused on natural experiments in the form of educational policy reforms.

In this chapter, I will summarize the findings of the four empirical studies and discuss them in a broader research context (Section 7.1). This discussion is structured according to the three research areas identified in the theoretical background section (mechanisms, implications, and interdisciplinary integration). Next, I will reflect on the extent to which the design-based challenges of previous research (raised in the theoretical background section) could be successfully addressed and give a final evaluation of the present dissertation's subordinate aims (Section 7.2). After elaborating on the dissertation's strengths and limitations (Section 7.3), I will discuss implications for educational practice (Section 7.4). The general discussion closes with directions for future research (Section 7.5) and a conclusion (Section 7.6).

7.1 Contribution to the BFLPE Literature

The present dissertation investigated four unresolved issues regarding a) multiple class environments as frames of reference, b) the association between grading on a curve and the BFLPE, c) tracking effects on academic self-concept, and d) neighborhood effects on academic self-concept. By applying new designs, the present work offers new insights into frame-of-reference effects on academic self-concept. In this section, I clarify the present dissertation's contribution to the three research areas reviewed in the introduction, namely mechanisms, implications, and interdisciplinary integration.

7.1.1 Mechanisms

Traditionally, research on the BFLPE has assumed that students build their academic self-concept in relation to the school as a frame of reference. The empirical foundation for such thinking was the negative effect of school-average achievement on students' academic self-concept when controlling for individual achievement differences (e.g., Marsh, 1987; Marsh & Parker, 1984). However, this traditional two-level approach is only weak evidence for the school as the pivotal frame of reference for academic self-concept formation because aggregate achievement measures from multiple educational environments (e.g., schools, classrooms, or tracks) can be assumed to be highly correlated. That means that traditional school-level BFLPE might just be a noisy reflection of a frame-of-reference effect at another level. Indeed, Marsh, Kuyper, et al. (2014) found school-average achievement to not affect academic self-concept when simultaneously controlling for class-average achievement. However, in school systems with course-by-course tracking in which students are grouped according to ability separately in one or more subjects, students are members of several class environments. Thus, to date, it is not clear how multiple class environments act as frames of reference for academic self-concept formation. The present dissertation filled the research gap concerning multiple class environments as frames of reference for academic self-concept formation by using comprehensive national educational monitoring data (Study 1). In line with previous research, we found a strong school-level BFLPE that substantially declined when additionally considering average achievement on the classroom level. Extending previous research, we found the domain-specific class BFLPE (negative effect of math class math achievement on math self-concept) to be more pronounced than the domain-unspecific class BFLPE (negative effect of regular class math achievement on math self-concept).

The present dissertation advanced research on mechanisms (research issues: pivotal frames of reference) by suggesting that students in school systems with course-by-course

tracking use multiple class environments as frames of reference for academic self-concept formation. In line with local dominance theory (Zell & Alicke, 2010), comparisons with proximal comparison targets influence self-evaluations much more than comparisons with distal ones. More specifically, Study 1 shows that with respect to domain-specific academic self-concept formation, *proximity* in the local dominance theory can be understood not only as *spatial proximity* but also as *domain-specific* proximity. In Study 1, we also found that the math class BFLPE becomes more negative when additionally controlling for track level and becomes less negative when additionally controlling for teacher-assigned grades whereas this was not the case for the regular class BFLPE. This result contributes to BFLPE theory, as it confirms the assumption that the frame-of-reference effects of multiple class environments might differ in their mechanisms. For example, in school systems with course-by-course tracking, domain-specific BFLPEs (e.g., the effect of domain-specific math class achievement on math self-concept) might be counterbalanced by assimilation effects, while this is not the case for a domain-unrelated BFLPE (e.g., the effects of domain-specific regular class achievement on math self-concept). Taken together, the present dissertation provides further evidence for the complex pattern of influences on academic self-concept formation.

From the very beginning of research on the BFLPE, it has been assumed that grading on a curve—namely teachers' tendency to provide the best grades to the best students, the worst grades to the worst students and place the others somewhere in between (e.g., Hübner et al., 2020)—contributes to the BFLPE (Marsh, 1987). The underlying idea behind this assumption was that equally able students have lower academic self-concept in high-achieving learning environments because they receive worse grades in high-ability classrooms. Empirical evidence for the hypothesis that grading on a curve contributes to the BFLPE came from analyses based on traditional mediation models (Baron & Kenny, 1986). These studies found that controlling for class-referenced school grades led to a substantial decline in the BFLPE (e.g., Marsh, Kuyper, et al., 2014; Trautwein, Lüdtke, Marsh, et al., 2006). Thus, students with equal school grades experience a less severe BFLPE. However, the traditional mediation approach only provides weak evidence for the notion that grading on a curve reinforces the BFLPE. It might just as well be that the two processes, grading on a curve and the BFLPE, coexist, with students evaluating themselves in comparison with their classmates and teachers assessing their students in comparison to each other. The present dissertation filled the research gap concerning the association between grading on a curve and the BFLPE by exploiting a natural experiment, namely a Swedish school reform in which municipalities were free to decide to abolish grading if they wished to do so. This natural experiment allowed for investigating the association

between grading on a curve and the BFLPE by comparing the BFLPEs of students who were not graded with students who received class-referenced grades (Study 3). If grading on curve contributes to the BFLPE, graded students should experience a stronger frame-of-reference effect than non-graded students. Against the expectations of previous research and in line with an evolutionary approach to social comparisons, we found no differences in the BFLPE between non-graded and graded students.

The present dissertation advanced research on mechanisms (research issue: moderators) by testing one of the most theoretically and empirically promising moderators of the BFLPE, namely the provision of class-referenced grades. Therefore, it adds to the considerable number of studies investigating potential BFLPE moderators that have found the frame-of-reference effects to be remarkable robust (e.g., Lüdtke et al., 2005; Seaton et al., 2010). On a more general level, Study 3 is in line with the notion that social comparison processes in the classroom might not necessarily be driven by teaching practices but rather result from students' innate drive to evaluate themselves against their peers. More specifically, Study 3 is in line with the Darwinian-economic perspective on social comparison processes laid down by Frank (2011), who considers social comparison processes a fundamental endowment of human evolution.

7.1.2 Implications

One of the major implications of the BFLPE concerns the achievement-related composition of educational environments, also referred to as tracking (Chmielewski, 2014). Very generally, the BFLPE predicts that low achievers will have higher academic self-concept in segregated educational systems as opposed to comprehensive ones because they will be exposed to classmates with lower average achievement (Trautwein & Möller, 2016). To date, empirical evidence for the BFLPE's prediction concerning tracking comes from two types of research designs. The first compares high-track students' academic self-concept with that of low-track students, finding equally able students to have a lower academic self-concept in high tracks than low tracks (e.g., Liem et al., 2013; Marsh et al., 2018; Trautwein, Lüdtke, Marsh, et al., 2006). However, this is only indirect evidence for the BFLPE's predictions concerning tracking because this approach does not compare non-tracked with tracked students. The second empirical approach does compare students from non-tracked and tracked school systems, finding low achievers to have a higher self-concept when tracked as opposed to not tracked (e.g., Dupriez et al., 2008; Kulik & Kulik, 1992; Marsh, Köller, et al., 2001). However, this is also not necessarily compelling evidence for the BFLPE's predictions concerning tracking because tracking practices in these studies are typically confounded with a range of third

variables (e.g., culture, teaching style). The present dissertation filled the research gap concerning tracking effects on academic self-concept by exploiting two natural experiments, namely an Austrian and a German detracking school reform, and comparing student cohorts before and after detracking (Study 3). In this cohort-control design (Shadish et al., 2002), variation in tracking practices was as-if random. In line with the BFLPE's predictions, we found that low achievers experienced a substantial academic self-concept decline after detracking, whereas this was not the case for high achievers.

The present dissertation advanced research on implications (research issue: educational systems) by testing and confirming the BFLPE's predictions concerning tracking based on as-if random variation in tracking practices. To my knowledge, this is the strongest evidence for differential detracking effects on academic self-concept for low and high achievers—as predicted by the BFLPE—so far. While this is not the paper's major focus, Study 4 also provides additional insights into a related issue that has caused controversy. Research typically assumes that high-achieving peers positively affect academic achievement but negatively affect academic self-concept via the BFLPE (e.g., Stäbler et al., 2017). At first glance, these findings are incompatible with the reciprocal effects model, which suggests that academic self-concept causally predicts academic achievement (Dicke et al., 2018). Study 4 shows on a descriptive level that, on average, students had a lower academic self-concept but higher academic achievement after detracking. Additionally, students had slightly higher variability in academic self-concept after detracking (as predicted by the BFLPE) but somewhat lower variability in academic achievement. One interpretation of this result is that detracking decreases low achievers' academic self-concept while simultaneously increasing their academic achievement. In the same vein, Hübner et al. (2017) found detracking to lower girls' academic self-concept while simultaneously raising their academic achievement. These results are remarkable because the reciprocal effects model, which postulates a causal effect of academic self-concept on academic achievement (e.g., Marsh et al., 2005), would predict increases in academic self-concept to go along with gains in academic achievement. On a theoretical level, this finding suggests that academic self-concept might not fully determine academic achievement and that there might be no paradox in self-concept declines accompanied by achievement gains (Dicke et al., 2018). Finally, it has to be noted that Study 4, which confirmed the BFLPE's predictions concerning tracking using as-if random variation in tracking practices, also strengthens BFLPE theory by providing evidence from a research design that is more internally valid than the typical cross-sectional BFLPE model.

7.1.3 Interdisciplinary Integration

The neighborhood effects literature in the field of sociology assumes that advantageous neighborhood socioeconomic conditions positively impact students' academic development via collective socialization mechanisms (e.g., Nieuwenhuis & Hooimeijer, 2016). This means that adolescents growing up in *good* neighborhoods emulate their neighborhood peers, who act as role models. Empirical evidence for this assumption comes from the predictive power of neighborhood socioeconomic composition indicators (e.g., neighborhood social status, income, or employment) on a broad range of educational outcomes (e.g., Brooks-Gunn et al., 1993; Hartung & Hillmert, 2019). However, many of those studies did not consider school characteristics, which might overlap with neighborhood composition because educational environments are typically based on geographic criteria (Jargowsky & Komi, 2011). Additionally, the neighborhood effects studies did not examine academic self-concept, which educational psychology research has shown to be highly susceptible to social comparison processes and might be negatively affected by advantageous neighborhood socioeconomic conditions via relative deprivation mechanisms. The present dissertation addressed this unresolved issue by simultaneously investigating neighborhood and classroom effects on academic self-concept using interdisciplinary large-scale data (Study 2). In seeming contrast to the neighborhood effects literature, we found advantageous neighborhood socioeconomic conditions to not or even negatively predict academic self-concept. Some of these neighborhood effects still existed when simultaneously considering class-average achievement, although they were then smaller in size.

The present dissertation advanced interdisciplinary integration by simultaneously considering school and neighborhood environments as frames of reference for academic self-concept formation. Study 2 contributes to BFLPE theory by suggesting that institutional learning environments such as schools or classrooms might overlap with non-institutional ones such as neighborhoods because these learning environments are often constructed based on geographic criteria (e.g., catchment areas). On a theoretical level, this raises the question of whether neighborhood and school effects "belong together" or can be seen as independent from each other. Furthermore, the present dissertation contributed to interdisciplinary integration by bringing together educational psychology research on the BFLPE, which emphasizes learning environments, and sociological neighborhood effects research, which additionally focuses on non-institutional student environments. This interdisciplinary endeavor led to new discipline-specific insights. Educational psychology research on the BFLPE was extended for a non-institutional student environment, the neighborhood, which might contribute to academic self-

concept formation in addition to institutional learning environments. Meanwhile, sociological neighborhood effects research was reminded of the importance of simultaneously considering neighborhood and school composition and the fact that there might be educational outcomes where the relative deprivation mechanism comes into play. In Study 2, the present dissertation also highlights the linkages between the sociological relative deprivation literature and psychological social comparison research, which share a close historical connection. For instance, the original BFLPE publication within educational psychology by Marsh (1987) was grounded in Davis's (1966) seminal sociological study, which was in turn theoretically rooted in work on relative deprivation by Stouffer et al. (1949).

In sum, the results of all four studies provide renewed evidence for the ubiquity of the BFLPE across different countries (Austria, Germany, Sweden), different age groups (elementary education, early and late secondary education), and academic self-concept in different domains (global, math, language). On a general level, all four studies contributed to refining BFLPE theory by adding multiple class environments to the BFLPE model (Study 1), including the neighborhood as an additional frame of reference for academic self-concept formation (Study 2), elaborating the association between grading and the BFLPE (Study 3), and investigating the BFLPE's predictions concerning tracking (Study 4).

7.2 Opportunities and Challenges of Integrating New Designs

The present dissertation addressed four unresolved issues in research on the BFLPE by tackling four design-based challenges of previous research. These design-based challenges were the (a) high correlation between multiple student environments, (b) weak internal validity of traditional mediation models, (c) non-random variation of tracking practices, and (d) confounding of neighborhood and school characteristics. The present dissertation's overarching aim was to address unresolved issues by tackling design-based challenges using new designs, specifically extensive large-scale data (comprehensive educational monitoring data and interdisciplinary data) and natural experiments (a school reform abolishing grades and two de-tracking school reforms). In this section, I will reflect on the extent to which these design-based challenges were successfully tackled.

7.2.1 Extensive Large-scale Data

This dissertation's first subordinate aim was to use extensive large-scale data in the form of comprehensive educational monitoring data as well as interdisciplinary data to tackle design-based challenges of previous research on the BFLPE, thus addressing unresolved issues.

Study 1 tackled the design-based challenge posed by the high correlation between multiple student environments using extensive large-scale data in the form of the 2012 Austrian National Educational Standard Assessments (BIFIE, 2016; Schreiner & Breit, 2012), which is unique in its comprehensive assessment of the total student population and the availability of information on students' multiple class environments. More specifically, this data enabled us to build reliable achievement aggregates on all levels in which students were nested, thus allowing us to disentangle the contextual effects of multiple class environments. Note that this is not possible with ordinary representative large-scale data based on cluster sampling procedures[4]—for the German educational context, for instance, data from *PISA* (OECD, 2019), *TIMSS* (Mullis et al., 2016), or *IQB Trend in Student Achievement* (Stanat et al., 2018). More specifically, the PISA study draws random student samples from schools and does not contain classroom identifiers. The TIMSS study draws one intact classroom from a school. As classroom aggregates (e.g., class-average achievement) are then equal to school aggregates, such a design cannot differentiate between school and class effects. The same is true for the IQB Trend in Student Achievement. To date, education-specific applications of cross-classified multilevel models have been, to my knowledge, mainly limited to the scenarios of students

[4] Cluster sampling procedures refer to, for instance, first drawing a random sample of schools, then drawing a random sample of students or intact classrooms.

nested in primary and secondary schools or students nested in schools and neighborhoods. (e.g., Hox et al., 2017). Comprehensive educational monitoring data represents a new field of application for cross-classified multilevel models by decomposing variance components between multiple class levels, and thus more realistically reflecting the reality of systems with course-by-course tracking.

Although the present dissertation successfully tackled the design-based challenge posed by the high correlation between multiple student environments by using comprehensive national educational monitoring data, there was one main obstacle. Often, comprehensive educational monitoring surveys focus on the assessment of one distinct domain. For Study 1, this means that we had no information on students' German and English class membership. Thus, every student was associated with two further class environments that were not included in our analysis.

Study 2 tackled the design-based challenge posed by the confounding of neighborhood and school characteristics by making use of extensive large-scale data in the form of interdisciplinary data from the *German Educational Panel Study* (Blossfeld et al., 2011), which is unique in its provision of information about students' institutional learning environments, such as schools and classrooms, as well as non-institutional environments like neighborhoods. More specifically, this data included information on students' academic self-concept, academic achievement, classroom membership, family background, neighborhood membership, and neighborhood socioeconomic composition, what enabled us to separately and simultaneously estimate the effects of neighborhoods (in terms of neighborhood socioeconomic composition) and educational environments (in terms of classroom-average achievement) on academic self-concept. In the past, researchers with high-quality school data typically studied school context effects, whereas researchers with good neighborhood-level data investigated neighborhood effects. However, because of the overlap between school and neighborhood characteristics, any analysis that omits one of these factors is prone to overstating or misstating the effect of the other (Jargowsky & Komi, 2011). Interdisciplinary large-scale data is an excellent basis for examining interdisciplinary research questions, thus connecting research traditions that have historically developed in parallel. Moreover, it allows for modeling the reality of students' lives in much more detail—for example, by predicting student outcomes using a wide range of individual determinants.

Although the present dissertation successfully tackled the design-based challenge posed by the confounding of school and neighborhood characteristics by using interdisciplinary data,

two major obstacles could not be overcome. First, when using interdisciplinary data, it can be difficult to transfer constructs from one discipline to another. For instance, in Study 2, we operationalized classroom selectivity as average classroom achievement, whereas neighborhood selectivity was operationalized as neighborhood socioeconomic conditions. Due to a low number of students per neighborhood, it was impossible to operationalize neighborhood selectivity as average neighborhood achievement. Second, interdisciplinary data that aims to collect a broad range of information typically has to make compromises concerning the measurement accuracy of these discipline-specific constructs. For instance, in Study 2, the academic self-concept item battery contained only three items, of which one directly referred to school grades. Additionally, some neighborhood characteristics (e.g., income) were calculated using algorithms whose exact functioning was not explained by the data provider *Microm Consumer Marketing*.

7.2.2 Natural Experiments

This dissertation's second subordinate aim was to use natural experiments in the form of a school reform abolishing grades and two detracking school reforms to tackle design-based challenges of previous research on the BFLPE, thus addressing unresolved issues. Because randomized controlled field trials are oftentimes neither ethically nor politically feasible in educational psychology research, there is growing awareness that natural experiments might represent an alternative research design in such cases (Leatherdale, 2019). Thus, there has been rapid growth in the use of quasi-experimental research designs such as natural experiments (Dunning, 2008).

Study 3 tackled the design-based challenge posed by the weak internal validity of traditional mediation models by using data from a natural experiment in the form of a Swedish school reform abolishing grades (Härnqvist, 2000) that is unique in offering as-if random variation in grading practices. More specifically, this data enabled us to compare BFLPEs between non-graded and graded students, thus taking a closer look at the association between grading on a curve and the BFLPE. Therefore, Study 3 substantially expands upon previous work that investigated the association between grading and the BFLPE by examining grades as a mediator with the help of traditional mediation analyses, which are unable to distinguish whether grading on a curve and the BFLPE are two separate processes or are associated with one another, in that grading on a curve contributes to the BFLPE. The school reform abolishing grades provided us with variation in grading practice that typically cannot be induced in the

field for ethical and political reasons, therefore presenting a unique opportunity for investigating the effects of grade provision.

While Study 3 presents an alternative to previous research, there were two main obstacles to overcoming this design-based challenge posed by the weak internal validity of traditional mediation models. First, self-selection natural experiments in education are typically not aimed at answering specific research questions, but represent a transition phase in which certain educational policies are introduced on a voluntary basis. In Study 3, this means that elementary school students in Sweden were provided with written grades only in Grades 3 and 6. Thus, an even better design would be to compare non-graded students with students who receive grades with a much higher frequency. Second, treatment assignment in self-selection experiments is not completely random by definition. In Study 3, municipalities could self-select to abolish grades. Thus, municipality selectivity might have impacted the results. Therefore, ideally, the comparison of non-graded and graded students should be tested in a system with frequent grades in which grading vs. not grading is randomly assigned. An alternative approach to overcome the design-based challenge posed by the weak internal validity of traditional mediation models would be to make use of new methodological approaches for investigating causal mediation (e.g., Imai et al., 2010; Vanderweele & Vansteelandt, 2009).

Study 4 tackled the design-based challenge posed by the non-random variation of tracking practices using two natural experiments in the form of detracking school reforms (Blossfeld et al., 2011; Schreiner et al., 2017), which are unique in offering as-if random variation in tracking practices. More specifically, this data enabled us to compare student cohorts before and after detracking, thus allowing us to take a closer look at the effects of tracking on academic self-concept. To my knowledge, this design provided the strongest test of the BFLPE's predictions concerning tracking, namely differential tracking effects on academic self-concept among high and low achievers, so far. Thus, Study 4 substantially expands upon and advances previous work that tested the BFLPE's predictions regarding tracking either by comparing high-track with low-track students (an indirect test) or by comparing non-tracked with tracked students with non-random variation.

While Study 4 substantially expanded upon and advanced previous research, there was one main obstacle to overcoming the design-based challenge posed by the non-random variation of tracking practices. Cohort-control designs are typically prone to bias resulting from historical events (Shadish et al., 2002). For example, an intervention effect might occur not because of a certain treatment but a historical event that impacts the composition of the student

population (e.g., refugee movements) or youths' living conditions (e.g., wide availability of smartphones). In Study 4, this means that the differential detracking effects on academic self-concept we found might also have been provoked, for example, by youth's increasing exposure to social media, which fosters adolescents' tendency to conduct social comparisons. Alternative approaches for tackling the design-based challenge posed by the non-random variation of tracking practices could be randomized controlled field trials (Duflo et al., 2011) or the use of other natural experiment designs, such as regression discontinuity or instrumental variable designs (Dunning, 2012).

7.3 Strengths and Limitations

When interpreting the present dissertation's results, it is important to keep its general strengths and limitations in mind. All four studies profited from large-scale datasets with appropriate sample sizes that allowed for accurately estimating effect sizes. The four studies focused on academic self-concept in several domains and examined frame-of-reference effects in multiple educational systems as well as a diverse set of age groups. Thereby, Studies 1 and 2 benefited from extensive large-scale data that allowed for the investigation of multiple class environments as frames of reference in systems with course-by-course tracking and for the examination of neighborhood effects on academic self-concept. Studies 3 and 4 profited from unique natural experiments that allowed for an investigation of the association between the BFLPE and grading on a curve as well as the effects of tracking on academic self-concept. Finally, the data was analyzed using state-of-the-art methods, such as complex multilevel models that accounted for missing data (e.g., full information maximum likelihood estimation; FIML; Enders, 2010). Nevertheless, some limitations should be kept in mind.

The first major limitation is the causal interpretation of the BFLPE (negative effect of class-average achievement on academic self-concept) in the cross-sectional models used in Studies 1, 2, and 3. In these studies, variation in the average achievement of educational environments was not experimentally manipulated. This means that third variables that positively impact educational environments' average achievement and negatively affect students' academic self-concept might be potential confounders. Previous work (e.g., Marsh et al., 2004; Marsh, Seaton, et al., 2008) opposed this potential threat of internal validity. These authors stated that controlling for potential confounders (e.g., SES or teacher qualifications) would actually increase the size of the BFLPE, thus arguing that the "uncontrolled" cross-sectional BFLPE represents a conservative estimation of the real causal effect. However, contrary to this assumption, there might indeed be third variables that threaten the internal validity of the cross-sectional BFLPE model. For example, demanding curricula or performance-oriented teachers can be expected to positively affect the average achievement of educational environments while simultaneously decreasing students' academic self-concept. Another argument for the correlational BFLPE design is that it is the best design that is feasible. Randomly assigning students to educational environments that differ in their average achievement is typically practically and ethically impossible (for experiments in university education, see Booij et al., 2016). However, not only randomized controlled trials but also well-designed quasi-experimental studies can create random variation in the average achievement of

educational environments. For example, Loyalka et al. (2018) showed that the internal validity of the BFLPE model could be improved by using a cross-subject student-fixed effects design.

The second major limitation, which has also been mentioned in previous research (e.g., Dai, 2004; Dai & Rinn, 2008), is that in all four studies, the social comparison processes posited to underlie the BFLPE were not directly measured. This means that social comparison processes were hypothesized but not observed. For instance, it is theoretically possible that the negative effect of class-average achievement in the traditional cross-sectional BFLPE model results not from social comparison processes but from teachers adapting their curricula to their students. That means that students in high-achieving learning environments might have a lower feeling of success, resulting in a lower academic self-concept. A more in-depth investigation of the social comparison processes underlying the BFLPE paradigm has indeed been the subject of previous research (e.g., Huguet et al., 2009; Marsh, Trautwein, et al., 2008). However, studies based on large-scale assessments typically have difficulties modeling small-scale processes. On the other hand, laboratory studies are usually not able to model the "real-world" classroom setting. A promising direction for investigating social comparison processes in the BFLPE paradigm in more detail might be the use of new technologies such as virtual reality environments that will are able to manipulate students' classroom experience experimentally.

The third major limitation is that the data used in all four studies were not collected with the aim of answering the specific research questions under evaluation. This limitation typically occurs in studies based on secondary data analyses (Davis-Kean & Jager, 2016). More specifically, the fact that data collection was not specifically tailored to the studies' research questions resulted in measurement problems (e.g., concerning academic self-concept in Study 3) or a lack of information on additional class environments (Study 1). In Study 2, information on students' class composition in elementary school would have helped to investigate whether neighborhood effects were relicts from primary education. Finally, in Study 4, both detracking reforms were "reform packages" that contained additional reform elements. For example, the Austrian detracking reform also included the establishment of team teaching, and the German reform in the state of Thuringia also included changes in the weighting of final examination grades.

7.4 Practical Implications

The present dissertation has diverse implications for educational practice. One of the major practical implications concerns the achievement-related composition of learning environments, also called educational tracking. Tracking is probably one of the most hotly debated issues with regard to the design of educational systems (Oakes, 2005). In recent decades, tracking practices have been criticized for separating students along race and social class lines, thus providing unequal opportunities and increasing social disparities in educational outcomes (Hallinan, 2004; Rubin, 2006). Thus, a trend of detracking educational systems could be observed (Yonezawa et al., 2002). In contrast to the proposed homogenizing effect of comprehensive education, Study 4 suggests that—in line with research on the BFLPE—detracking lowers the academic self-concept of low achievers, while this is not the case for high achievers. As academic self-concept is seen as an important antecedent of students' achievement as well as course choices, academic aspirations, and occupational careers, the social comparison processes that underlie the BFLPE may be seen as an obstacle to exploiting the full potential of detracking reform efforts to close education gaps by race and class. For policy and practice, this means that decision-makers must keep in mind that the desegregation of educational systems may be at the expense of low achievers' motivation.

The emergence of the BFLPE within detracking school reforms also has consequences for gender diversity in the occupational world. For instance, encouraging females to pursue STEM (science, technology, engineering, and mathematics) subjects is often declared a desired aim to (a) support economic growth and (b) improve women's career opportunities as well as close the gender wage gap (OECD, 2007). Generally, females are assumed to be overrepresented in low math tracks (Nagy et al., 2008). Therefore, detracking might decrease young women's academic self-concept, as this group of students would then be exposed to class environments in comprehensive systems with higher average achievement. Empirical evidence for this assumption is presented by Hübner et al. (2017), who showed that detracking was at the expense of girls' math self-concept. As academic self-concept is seen as an important determinant of academic choices, detracking in STEM subjects could potentially discourage females from pursuing a STEM career path. Indeed, there is evidence that detracking in math reduces the share of women graduating from STEM programs (Schwerter & Biewen, 2020).´

Given these proposed negative side effects of mixed-ability classrooms, the question arises how to prevent them. Here, Study 3 comes into play, which found that a Swedish school reform abolishing grades was not able to change the BFLPE, in line with a Darwinian-economic

perspective on social comparison. If one assumes that social comparison processes are unavoidable, one might argue based on the local dominance effect that educational systems in which students attend comprehensive distal learning environments (e.g., schools) but more segregated proximal environments (e.g., classrooms) could potentially counteract the negative side effects described above. For instance, an educational system in which students across the entire achievement spectrum attend the same school but are assigned to different streams within it (e.g., *within-school streaming*) or tracked course-by-course only in subjects of particular relevance, such as mathematics and language (e.g., *setting*), would allow low achievers to maintain a relatively high academic self-concept. Simultaneously, such a system could potentially expose all students to a diverse student body representing the entire spectrum of society and provide all students with equal school-related resources (e.g., teaching quality or technical equipment).

Alongside ability grouping and educational tracking, the present dissertation brings to light new theoretical considerations concerning the provision of school grades. Grading is generally a controversial topic (Kohn, 2011). In relatively ideological battles, grades are often suspected of making students more concerned about pleasing the teacher than thinking critically (Gatto, 1992), inducing unhealthy competition (Purpel, 1989), or undermining students' intrinsic motivation (Kohn, 1998). Another claim concerning school grades is that they impact students' self-concept by demotivating low achievers and additionally motivating high achievers (Romanowski, 2004), thus contributing to educational disparities. The present dissertation's Study 3 supports the notion of social comparison processes as an unalterable aspect of human nature. This suggests that the grading controversy might be overstated, at least in this respect. The Darwinian-economic perspective on social comparison processes indicates that students' academic self-concept is not rooted in their grades but by social comparison information that exists independent of grade provision. Statistically speaking, school grades might be strongly associated with academic self-concept not because they affect academic self-concept but because grades are a proxy measure of students' perceived class rank, which is the result of social comparison processes. Put more simply, school grades may provide students with social comparison information they already know.

The present dissertation also provides an alternative, contradictory perspective on the composition of non-institutional student environments such as neighborhoods. Typically, sociological neighborhood research assumes that "good" neighborhoods positively impact various desirable life outcomes (e.g., Wilson, 1987). Based on this theoretical consideration, it has been argued that neighborhood segregation increases social inequalities because members

of upper social classes profit from their neighborhood's resources, whereas members of lower social classes do not. The assumed benefits of advantageous neighborhood socioeconomic conditions have been used as an argument for neighborhood destratification and seen as a measure for combatting social class inequality. In Study 2, we found advantageous neighborhood socioeconomic conditions to not or even negatively predict academic self-concept. This finding should remind policymakers and urban planners that advantageous neighborhood conditions might not just positively affect educational outcomes via collective socialization mechanisms but might also have null or even negative effects on student outcomes that are susceptible to social comparison processes. This assumption offers an important alternative perspective to policymakers, as neighborhood desegregation might not inevitably lead to a homogenization of all educational outcomes. On the contrary, it is possible that social class gaps in educational outcomes that are susceptible to social comparison processes may actually be exacerbated by neighborhood desegregation.

7.5 Directions for Future Research

Based on the present dissertation, I identified two main areas for future research. First, more high-quality work on academic self-concept's importance for students' academic development is needed. Second, the BFLPE needs to be embedded into a broader peer effects framework integrating contextual effects research on a diverse set of outcomes (e.g., academic self-concept, achievement, choices). This will in turn make it possible to inform educational policy with best practices concerning student composition.

7.5.1 Academic Self-Concept's Importance

One direction for future research concerns the importance of academic self-concept for students' academic development. More specifically, this relates to the causal ordering of academic self-concept and academic achievement, which has been seen as the most important research question in self-concept research (Marsh & Perry, 2005). The assumed high importance of academic self-concept as a motivational variable stems from cross-lagged panel studies that regressed achievement at a later time point (T2) on self-concept at an earlier time point (T1), simultaneously controlling for T1 achievement (e.g., Arens et al., 2017; Marsh & Martin, 2011). Typically, these studies found small to moderate effects of academic self-concept on academic achievement (Huang, 2011; Valentine et al., 2004). The effect of T1 academic self-concept on T2 academic achievement has often been interpreted as causal evidence for the decisive role of students' academic self-concept for their academic development. However, note that a causal interpretation of the effect of academic self-concept on academic achievement depends on the assumption that all potential third variables have been taken into account. Indirect evidence from studies finding high-achieving learning environments to foster academic achievement while simultaneously reducing academic self-concept (e.g., Hübner et al., 2017; Stäbler, 2017) seems to suggest that the causal relationship between the two constructs might be weaker than originally assumed. More rigorous tests of the self-enhancement approach are needed to better understand academic self-concept's importance for students' academic development. This, in turn, would help to further refine the practical implications of the BFLPE. Here, I give several examples of what a rigorous investigation of academic self-concept's importance for students' academic development might look like. First, students' self-concept could be experimentally manipulated in a laboratory or field setting. An example of the latter would be to develop a classroom intervention with a waitlist control group design that aims to foster students' academic self-concept using a peer feedback intervention (Simonsmeier et al., 2020). Alternatively, a more accurate understanding

of the causal relationship between academic self-concept and academic achievement could also be obtained by matching otherwise similar students with high and low academic self-concept and following their educational pathways. Another idea would be to use academic self-concept variation within pairs of twins to conduct a more thorough test of the skill development approach (for an appropriate methodological approach, see, e.g., Turkheimer & Harden, 2013). More research is also needed on the channels through with this effect might operate. Is it mainly that a high self-concept leads students to invest more effort, or is it that high self-concept leads students to select advanced courses, which in turn raise achievement?

Another fundamental question that has to be answered by future research is whether it is desirable to provide all students with the highest possible academic self-concept or with an accurate academic self-concept. To date, research on the importance of an accurate academic self-concept is rather scarce (for an exception, see Eshel & Kurman, 1991). One could argue that an accurate self-concept is a basic requirement for making appropriate educational-related decisions, such as with respect to university entry and major choice. The accuracy of students' academic self-concept could be measured by comparing students' self-evaluations and teacher's external evaluations or by relating students' self-evaluations to their achievement rank in the population. It would be interesting to see whether such self-concept accuracy measures can predict dropout rates from certain university majors.

7.5.2 The BFLPE in a Broader Peer Effects Framework

Another issue for future research is integrating the BFLPE into a broader peer effects framework. Based on research on the BFLPE, there are good reasons to believe that placing students in high-ability learning environments harms their academic self-concept and that comprehensive education harms the academic self-concept of low achievers, who are then exposed to more high-achieving classmates on average. However, what are the effects of selective learning environments on academic achievement? Generally, there are two different views on this issue: First, a relatively small body of research supports the notion that selective educational environments negatively affect academic achievement (Denning et al., 2018; Dicke et al., 2018; Murphy & Weinhardt, 2018). This view is in alignment with research on academic self-concept and the BFLPE. Selective learning environments decrease academic self-concept, which in turn leads to lower academic achievement. However, the majority of studies suggest that selective learning environments positively impact students' academic achievement via *peer spillover effects* (e.g., Ammermueller & Pischke, 2009; Betts & Zau, 2004; Burke & Sass, 2013). This second view seemingly contradicts academic self-concept theory because academic

self-concept is supposed to be a determinant of academic achievement. There are several potential explanations for these conflicting findings. First, positive effects of selective learning environments on students' academic achievement might be artifacts of different forms of bias (e.g., Dicke et al., 2018; Manski, 1993). Secondly, as already mentioned above, it is possible that the causal relationship between academic self-concept and academic achievement is less pronounced than assumed and there is no contradiction in selective learning environments simultaneously fostering academic achievement and dampening academic self-concept. Indirect empirical support for this assumption comes, for instance, from Hübner et al. (2017), who found a detracking school reform to close the gender achievement gap in math while simultaneously increasing math self-concept differences between males and females. As already discussed above, the present dissertation's Study 4 also shows that academic self-concept changes are not necessarily accompanied by respective changes in academic achievement. As one can see from the discussion above, there is still no clear understanding of how selective learning environments, and thus also tracking, impact students' academic development.

Future research needs to investigate the effect of selective learning environments on a broad array of educational outcomes to provide a clear answer to the question: "What happens if one places children in selective learning environments?" The challenge of this endeavor lies in the causal identification of peer effects (Angrist, 2014). Thus, future educational psychology research aiming to simultaneously model peer effects on a broad array of educational outcomes could, for example, apply stronger research designs like fixed-effects models (Betts & Zau, 2004) or natural experiments with instrumental variables (Hoxby, 2000).

7.6 Conclusion

The present dissertation advanced our understanding of the mechanisms, implications, and interdisciplinary integration of frame-of-reference effects on academic self-concept (BFLPE) by applying new research designs. More specifically, it generated new insights into the pivotal frames of reference for academic self-concept formation in school systems with course-by-course tracking, the association between neighborhood and school effects on academic self-concept, the association between grading on a curve and the BFLPE, and tracking effects on academic self-concept. In addition, to these substantive contributions, the present dissertation calls for a higher diversity of research designs to be used—in research on the BFLPE but also in educational psychology research in general. Especially the usage of natural experiments provide an outstanding opportunity for testing educational psychology theory in the field, simultaneously providing implications for educational practice.

8 References

Alexander, K., & Eckland, B. K. (1975). Contextual effects in the high school attainment process. *American Sociological Review*, *40*(3), 402–416. https://doi.org/10.2307/2094466

Ammermueller, A., & Pischke, J.-S. (2009). Peer effects in european primary schools: Evidence from the Progress in International Reading Literacy Study. *Journal of Labor Economics*, *27*(3), 315–348. https://doi.org/10.1086/603650

Angrist, J. D. (2014). The perils of peer effects. *Labour Economics*, *30*, 98–108. https://doi.org/10.1016/j.labeco.2014.05.008

Archambault, I., Eccles, J. S., & Vida, M. N. (2010). Ability self-concepts and subjective value in literacy: Joint trajectories from grades 1 through 12. *Journal of Educational Psychology*, *102*(4), 804–816. https://doi.org/10.1037/a0021075

Arens, A. K., Marsh, H. W., Pekrun, R., Lichtenfeld, S., Murayama, K., & Vom Hofe, R. (2017). Math self-concept, grades, and achievement test scores: Long-term reciprocal effects across five waves and three achievement tracks. *Journal of Educational Psychology*, *109*(5), 621–634. https://doi.org/10.1037/edu0000163

Arum, R. (2000). Schools and communities: Ecological and institutional dimensions. *Annual Review of Sociology*, *26*(1), 395–418. https://doi.org/10.1146/annurev.soc.26.1.395

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Social and Clinical Psychology*, *4*(3), 359–373. https://doi.org/10.1521/jscp.1986.4.3.359

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037//0022-3514.51.6.1173

Baumeister, R. F. (1997). Identity, self-concept, and self-esteem: The self lost and found. In R. Hogan, J. Johnson, & Briggs, S., R. (Eds.), *Handbook of personality psychology* (pp. 681–710). Academic Press. https://doi.org/10.1016/B978-012134645-4/50027-5

Becker, M., & Neumann, M. (2018). Longitudinal big-fish-little-pond effects on academic self-concept development during the transition from elementary to secondary schooling. *Journal of Educational Psychology*, *110*(6), 882–897. https://doi.org/10.1037/edu0000233

Beretvas, S. N. (2011). Cross-classified and multiplemembership models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 313–334). Routledge.

Betts, & Zau (2004). Peer groups and academic achievement: Panel evidence from administrative data. Working paper from University of California.

BIFIE. (2016). *Datensatz zur Standardüberprüfung 2012 Mathematik, 8. Schulstufe. Schülerebene, M812I. Forschungsdatenbibliothek (FDB). Nicht-imputierter Datensatz, v2.0.* BIFIE.

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

Booij, A. S., Leuven, E., & Oosterbeek, H. (2016). Ability peer effects in university: Evidence from a randomized experiment. *Review of Economic Studies*, *84*, 547–578. https://doi.org/10.1093/restud/rdw045

Braun, L., Rieger, S., Spengler, M., Göllner, R., Rose, N., Trautwein, U., & Nagengast, B. (2020). Rethinking the elusive construct of global self-concept: A latent composite score as the apex of the Shavelson model. Advance online publication. https://doi.org/10.31234/osf.io/dbkw6

Brooks-Gunn, J., Duncan, G. J., Klebanov, P. K., & Sealand, N. (1993). Neighborhoods influence child and adolescent development? *American Journal of Sociology*, *99*(2), 353–395. https://doi.org/10.1086/230268

Brown, D. A., Gardener, J., Oswald Andrew, J., & Quian, J. (2008). Does wage rank affect employees' well-being? *Industrial Relations*, *47*(3), 355–389. https://doi.org/10.1111/j.1468-232X.2008.00525.x

Brunner, M., Keller, U., Dierendonck, C., Reichert, M., Ugen, S., Fischbach, A., & Martin, R. (2010). The structure of academic self-concepts revisited: The nested Marsh/Shavelson model. *Journal of Educational Psychology*, *102*(4), 964–981. https://doi.org/10.1037/a0019644

Burke, M. A., & Sass, T. R. (2013). Classroom peer effects and student achievement. *Journal of Labor Economics*, *31*(1), 51–82. https://doi.org/10.1086/666653

Byrne, B. M., & Shavelson, R. J. (1986). On the structure of adolescent self-concept. *Journal of Educational Psychology*, *78*(6), 474–481. https://doi.org/10.1037/0022-0663.78.6.474

Calsyn, R. J., & Kenny, D. A. (1977). Self-concept of ability and perceived evaluation of others: Cause or effect of academic achievement? *Journal of Educational Psychology*, *69*(2), 136–145.

Card, D., Mas, A., Moretti, E., & Saez, E. (2012). Inequality at work: The effect of peer salaries on job satisfaction. *American Economic Review, 102*(6), 2981–3003. https://doi.org/10.1257/aer.102.6.2981

Chanal, J. P., Marsh, H. W., Sarrazin, P. G., & Bois, J. E. (2005). Big-fish-little-pond effects on gymnastics self-concept: Social comparison processes in a physical setting. *Journal of Sport and Exercise Psychology*, *27*(1), 53–70. https://doi.org/10.1123/jsep.27.1.53

Chapman, J. W. (1988). Learning disabled children's self-concepts. *Review of Educational Research, 58*(3), 347–371. https://doi.org/10.3102/00346543058003347

Chmielewski, A. K. (2014). An international comparison of achievement inequality in within- and between-school tracking systems. *American Journal of Education*, *120*(3), 293–324. https://doi.org/10.1086/675529

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type. *American Educational Research Journal*, *50*(5), 925–957. https://doi.org/10.3102/0002831213489843

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

Clark, A. E., & Oswald, A. J. (1996). Satisfaction and comparison income. *Journal of Public Economics, 61*(3), 359–381. https://doi.org/10.1016/0047-2727(95)01564-7

Crabtree, J. (2003). Maintaining positive self-concept: Social comparisons in secondary school students with mild learning disabilities attending mainstream and special schools. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *International advances in self research* (pp. 261–290). Information Age.

Craven, R. G., Marsh, H. W., & Print, M. (2000). Gifted, streamed and mixed-ability programs for gifted students: Impact on self-concept, motivation, and achievement. *Australian Journal of Education*, *44*(1), 51–75. https://doi.org/10.1177/000494410004400106

Dai, D. Y. (2001). A comparison of gender differences in academic self-concept and motivation between high-ability and average chinese adolescents. *Journal of Secondary Gifted Education*, *13*(1), 22–32. https://doi.org/10.4219/jsge-2001-361

Dai, D. Y. (2004). How universal is the big-fish-little-pond effect? *The American Psychologist, 59*(4), 267-268. https://doi.org/10.1037/0003-066X.59.4.267

Dai, D. Y., & Rinn, A. N. (2008). The big-fish-little-pond effect: What do we know and where do we go from here? *Educational Psychology Review*, *20*(3), 283–317. https://doi.org/10.1007/s10648-008-9071-x

Dai, D. Y., Rinn, A. N., & Tan, X. (2013). When the big fish turns small. *Journal of Advanced Academics, 24*(1), 4–26. https://doi.org/10.1177/1932202X12473425

Davis, J. A. (1966). The campus as a fog pod: An application of the theory of relative deprivation to career decisions of college men. *American Journal of Sociology*, *72*(1), 17–31. https://doi.org/10.1086/224257

Davis-Kean, P. E., & Jager, J. (2016). Using secondary data analysis. In D. Wyse, N. Selwyn, E. Smith, & L. E. Suter (Eds.), *The BERA/SAGE handbook of educational research* (pp. 505–522). Sage.

Denning, J. T., Murphy, R., & Weinhardt, F. (2018). Class rank and long-run outcomes. Discussion Paper 118.

Dicke, T., Marsh, H. W., Parker, P. D., Pekrun, R., Guo, J., & Televantou, I. (2018). Effects of school-average achievement on individual self-concept and achievement: Unmasking phantom effects masquerading as true compositional effects. *Journal of Educational Psychology*, *110*(8), 1112–1126. https://doi.org/10.1037/edu0000259

Dietz, R. D. (2002). The estimation of neighborhood effects in the social sciences: An interdisciplinary approach. *Social Science Research, 31*(4), 539–575. https://doi.org/10.1016/S0049-089X(02)00005-4

Dijkstra, P., Kuyper, H., van der Werf, G., Buunk, A. P., & van der Zee, Y. G. (2008). Social comparison in the classroom: A review. *Review of Educational Research*, *78*(4), 828–879. https://doi.org/10.3102/0034654308321210

Domina, T., McEachin, A., Hanselman, P., Agarwal, P., Hwang, N., & Lewis, R. W. (2019). Beyond tracking and detracking: The dimensions of organizational differentiation in schools. *Sociology of Education*, *92*(3), 293–322. https://doi.org/10.1177/0038040719851879

Dompnier, B., Pansu, P., & Bressoux, P. (2006). An integrative model of scholastic judgments: Pupils' characteristics, class context, halo effect and internal attributions. *European Journal of Psychology of Education, 21*(2), 119–133. https://doi.org/10.1007/BF03173572

Donnellan, M. B., Trzesniewski, K. H., Robins, R. W., Moffitt, T. E., & Caspi, A. (2005). Low self-esteem is related to aggression, antisocial behavior, and delinquency. *Psychological Science, 16*(4), 328–335. https://doi.org/10.1111/j.0956-7976.2005.01535.x

Duesenberry, J. S. (1949). *Income, savings, and the theory of consumer behavior*. Harvard University Press.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review, 101*(5), 1739–1774. https://doi.org/10.1257/aer.101.5.1739

Dunn, E. C., Milliren, C. E., Evans, C. R., Subramanian, S. V., & Richmond, T. K. (2015). Disentangling the relative influence of schools and neighborhoods on adolescents' risk for depressive symptoms. *American Journal of Public Health, 105*(4), 732–740. https://doi.org/10.2105/AJPH.2014.302374

Dunn, E. C., Richmond, T. K., Milliren, C. E., & Subramanian, S. V. (2015). Using cross-classified multilevel models to disentangle school and neighborhood effects: An example focusing on smoking behaviors among adolescents in the United States. *Health & Place*, *31*, 224–232. https://doi.org/10.1016/j.healthplace.2014.12.001

Dunning, T. (2008). Improving causal inference. *Political Research Quarterly, 61*(2), 282–293. https://doi.org/10.1177/1065912907306470

Dunning, T. (2012). *Natural experiments in the social sciences*. Cambridge University Press.

Dupriez, V., Dumay, X., & Vause, A. (2008). How do school systems manage pupils' heterogeneity? *Comparative Education Review, 52*(2), 245–273. https://doi.org/10.1086/528764

Easterlin, R. A. (1974). Does economic growth improve the human lot? Some empirical evidence. In P. A. David & M. W. Reder (Eds.), *Nations and households in economic growth* (pp. 89–125). Academic Press.

Eccles, J. S. (2009). Who am i and what am i going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist, 44*(2), 78–89. https://doi.org/10.1080/00461520902832368

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 74–146). Freeman.

Eckert, C., Schilling, D., & Stiensmeier-Pelster, J. (2006). Einfluss des Fähigkeitsselbstkonzepts auf die Intelligenz- und Konzentrationsleistung. *Zeitschrift Für Pädagogische Psychologie*, *20*(1/2), 41–48. https://doi.org/10.1024/1010-0652.20.12.41

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Eshel, Y., & Kurman, J. (1991). Academic self-concept, accurary of preceived ability and academic attainment. *British Journal of Educational Psychology*, *61*(2), 187–196. https://doi.org/10.1111/j.2044-8279.1991.tb00974.x

Felson, R. B. (1984). The effect of self-appraisals of ability on academic performance. *Journal of Personality and Social Psychology*, *47*(5), 944–952. https://doi.org/10.1037/0022-3514.47.5.944

Felson, R. B., & Reed, M. D. (1986). Reference groups and self-appraisals of academic ability and performance. *Social Psychology Quarterly*, *49*(2), 103–109. https://doi.org/10.2307/2786722

Ferryman, K. S., de Souza Briggs, X., Popkin, S. J., & Rendon, M. (2008). *Do better neighborhoods for MTO families mean better schools?* Metropolitan Housing and Communities Center.

Festinger, L. (1957). A theory of social comparison processes. *Human Relations*, *7*, 117–140. https://doi.org/10.1177/001872675400700202

Fleischmann, M. (2017). *Do same-sex comparisons matter? An investigation of sex-specific reference group effects on mathematical self-concept*. Unpublished master thesis.

Frank, R. H. (2011). *The Darwin economy: Liberty, competition, and the common good*. Princeton University Press.

Gaias, L. M., Lindstrom Johnson, S., White, R. M. B., Pettigrew, J., & Dumka, L. (2018). Understanding school–neighborhood mesosystemic effects on adolescent development. *Adolescent Research Review*, *3*(3), 301–319. https://doi.org/10.1007/s40894-017-0077-9

Galster, G. C. (2008). Quantifying the effect of neighbourhood on individuals: Challenges, alternative approaches, and promising directions. *Schmollers Jahrbuch*, *128*(1), 1–42. https://doi.org/10.3790/schm.128.1.7

Galster, G. C. (2012). The mechanism(s) of neighbourhood effects: Theory, evidence, and policy implications. In M. van Ham, N. Manley, L. Bailey, D. Simpson, & D. MacLennan (Eds.), *Neighbourhood effects research: New perspectives* (pp. 23–56). Springer.

Gardener, J., Brown, G., Oswald, A., & Quian, J. (2005). Does wage rank affect employees' well-being? *Industrial Relations*, *47*(3), 355–389. https://doi.org/10.1111/j.1468-232X.2008.00525.x

Gatto, J. T. (1992). *Dumbing us down: The hidden curriculum of compulsory schooling.* New Society Publishers.

Goethals, G., & Darley, J. (1977). Social comparison theory: An attributional approach. In J. M. Suls & R. Miller (Eds.), *Social comparison processes: Theoretical and empirical perspectives* (pp. 259–278). Wiley.

Goldstein, H. (2016). Multilevel cross-classified models. *Sociological Methods & Research, 22*(3), 364–375. https://doi.org/10.1177/0049124194022003005

Göllner, R., Damian, R. I., Nagengast, B., Roberts, B. W., & Trautwein, U. (2018). It's not only who you are but who you are with: High school composition and individuals' attainment over the life course. *Psychological Science*, 1-12. https://doi.org/10.1177/0956797618794454

Greenwald, A. G. (1988). A social-cognitive account of the self's development. In D. K. Lapsley & F. C. Power (Eds.), *Self, ego and identity: Interpretative approaches* (pp. 30–42). Springer.

Guimond, S., Branscombe, N. R., Brunot, S., Buunk, A. P., Chatard, A., Désert, M., Garcia, D. M., Haque, S., Martinot, D., & Yzerbyt, V. (2007). Culture, gender, and the self: Variations and impact of social comparison processes. *Journal of Personality and Social Psychology, 92*(6), 1118–1134. https://doi.org/10.1037/0022-3514.92.6.1118

Guo, J., Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2015). Directionality of the associations of high school expectancy-value, aspirations, and attainment. *American Educational Research Journal*, *52*(2), 371–402. https://doi.org/10.3102/0002831214565786

Guo, J., Marsh, H. W., Parker, P. D., & Dicke, T. (2018). Cross-cultural generalizability of social and dimensional comparison effects on reading, math, and science self-concepts for primary school students using the combined PIRLS and TIMSS data. *Learning and Instruction*, *58*, 210–219.

Hallinan, M. T. (2004). The detracking movement. *Education Next*, *4*(4), 72–76.

Härnqvist, K. (2000). Evaluation through follow-up. A longitudinal program for studying education and career development. In C. G. Janson (Ed.), *Seven Swedish longitudinal studies in behavioral science* (pp. 76–114). Forskningsrådsnämnden.

Harter, S. (1998). The development of self-representations. In W. Damon & N. Eisenberg (Eds.), *Handbook of child psychology: Social, emotional, and personality development* (pp. 553–617). John Wiley.

Hartung, A., & Hillmert, S. (2019). Assessing the spatial scale of context effects: The example of neighbourhoods' educational composition and its relevance for individual aspirations. *Social Science Research*, *83*, 1–13. https://doi.org/10.1016/j.ssresearch.2019.05.001

Hattie, J. (2003). *The status and direction of self-concept research: The importance of importance.* Paper presented at the Human Development Conference, Auckland.

Hattie, J. (2014). *Self-concept*. Psychology Press.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge Taylor & Francis Group.

Hoxby, C. (2000). Peer effects in the clasroom: Learning from gender and race variation, Working paper no. 7867, National Bureau of Economic Research.

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, *49*(5), 505–528. https://doi.org/10.1016/j.jsp.2011.07.001

Hübner, N., Wagner, W., Hochweber, J., Neumann, M., & Nagengast, B. (2020). Comparing apples and oranges: Curricular intensification reforms can change the meaning of students' grades! *Journal of Educational Psychology*, *112*(1), 204–220. https://doi.org/10.1037/edu0000351

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

Huguet, P., Dumas, F., Marsh, H. W., Wheeler, L., Seaton, M., Nezlek, J., Suls, J., & Régner, I. (2009). Clarifying the role of social comparison in the big-fish-little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology*, *97*(1), 156–170. https://doi.org/10.1037/a0015558

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334. https://doi.org/10.1037/a0020761

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development*, *73*(2), 509–527. https://doi.org/10.1111/1467-8624.00421

James, W. (1890). *The principles of psychology*. Holt.

Jargowsky, P., & Komi, M. (2011). Before or after the bell? School context and neighborhood effects on student achievement. In H. Newburger, E. Birch, & S. Wachter (Eds.), *Neighborhood and life chances: How place matters in modern America* (pp. 50–72). University of Pennsylvania Press.

Jencks, C., & Mayer, S. E. (1990). The social consequences of growing up in a poor ceighborhood. In L. E. Lynne & M. G. H. McGreary (Eds.), *Inner-city poverty in the United States* (pp. 111–186). National Academies Press.

Johnson, O. (2012). A sytematic review of neighborhood and institutional relationships related to education. *Education and Urban Society*, *44*(4), 477–511. https://doi.org/10.1177/0013124510392779

Jonkmann, K., Becker, M., Marsh, H. W., Lüdtke, O., & Trautwein, U. (2012). Personality traits moderate the big-fish–little-pond effect of academic self-concept. *Learning and Individual Differences*, *22*(6), 736–746. https://doi.org/10.1016/j.lindif.2012.07.020

Kauppinen, T. M. (2008). Schools as mediators of neighbourhood effects on choice between vocational and academic tracks of secondary education in Helsinki. *European Sociological Review*, *24*(3), 379–391. https://doi.org/10.1093/esr/jcn016

Kohn, A. (1998). *What to look for in a classroom*. Jossey-Bass.

Kohn, A. (2011). The case against grades. *Educational Leadership*, *69*(3), 28–33.

Kulik, C.-L. C. (1985). Effects of inter-class ability grouping on achievement and self-esteem. Paper presented at the 93rd annual convention of the American Psychological Association, Los Angeles.

Kulik, J. A., & Kulik, C.-L. C. (1992). Meta-analytic findings on grouping programs. *Gifted Child Quarterly*, *36*(2), 73–77. https://doi.org/10.1177/001698629203600204

Leatherdale, S. T. (2019). Natural experiment methodology for research: A review of how different methods can support real-world research. *International Journal of Social Research Methodology*, *22*(1), 19–35. https://doi.org/10.1080/13645579.2018.1488449

Leventhal, T., & Brooks-Gunn, J. (2000). The neighborhoods they live in: The effects of neighborhood residence on child and adolescent outcomes. *Psychological Bulletin*, *126*(2), 309–337. https://doi.org/10.1037//0033-2909.126.2.309

Liem, G. A. D., Marsh, H. W., Martin, A. J., McInerney, D. M., & Yeung, A. S. (2013). The big-fish-little-pond effect and a national policy of within-school ability streaming. *American Educational Research Journal*, *50*(2), 326–370. https://doi.org/10.3102/0002831212464511

Lipset, M. (1960). *The social bases of politics*. Johns Hopkins University Press.

Liu, W. C., Wang, C. K. J., & Parkins, E. J. (2005). A longitudinal study of students' academic self-concept in a streamed setting: The Singapore context. *The British Journal of Educational Psychology*, *75*, 567–586. https://doi.org/10.1348/000709905X42239

Loveless, T. (2013). *The 2013 Brown Center report on American education: How well are American students learning?* Brookings Institution.

Loyalka, P., Zakharov, A., & Kuzmina, Y. (2018). Catching the big fish in the little pond effect: Evidence from 33 countries and regions. *Comparative Education Review*, *62*(4), 542–564. https://doi.org/10.1086/699672

Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish-little-pond effect. *Contemporary Educational Psychology*, *30*(3), 263–285. https://doi.org/10.1016/j.cedpsych.2004.10.002

MacKinnon, D. (2012). *Introduction to Statistical Mediation Analysis*. Taylor & Francis.

Makel, M. C., Lee, S.-Y., Olszewki-Kubilius, P., & Putallaz, M. (2012). Changing the pond, not the fish: Following high-ability students across different educational environments. *Journal of Educational Psychology*, *104*(3), 778–792. https://doi.org/10.1037/a0027558

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, *60*(3), 531. https://doi.org/10.2307/2298123

Marsh, H. W., Martin, A. J., & Debus, R. L. (2001). Individual differences in verbal and math self-perceptions: One factor, two factors, or does it depend on the construct? In R. Riding & S. Rayner (Eds.), *Self-perception: International perspectives on individual differences* (pp. 149–170). Able Publishing.

Marsh, H. W. (1984). Self-Concept, Social Comparison, and Ability Grouping: A Reply to Kulik and Kulik. *American Educational Research Journal*, *21*(4), 799. https://doi.org/10.2307/1163002

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. (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. (1989). Age and sex effects in multiple dimensions of self-concept: Preadolescence to early adulthood. *Journal of Educational Psychology*, *81*(3), 417–430. https://doi.org/10.1037/0022-0663.81.3.417

Marsh, H. W. (1990). *Self-description questionnaire manual*. University of Western Australia.

Marsh, H. W. (1991). Failure of high-ability high schools to deliver academic benefits commensurate with their students' ability levels. *American Educational Research Journal*, *28*(2), 445–480. https://doi.org/10.2307/1162948

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

Marsh, H. W. (2016). Cross-cultural generalizability of year in school effects: Negative effects of acceleration and positive effects of retention on academic self-concept. *Journal of Educational Psychology*, *108*(2), 256–273. https://doi.org/10.1037/edu0000059

Marsh, H. W., Abduljabbar, A. S., Morin, A. J., 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. https://doi.org/10.1037/a0037485

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J., Abdelfattah, F., & Nagengast, B. (2014). The Big-Fish-Little-Pond Effect in Mathematics. *Journal of Cross-Cultural Psychology*, *45*(5), 777–804. https://doi.org/10.1177/0022022113519858

Marsh, H. W., Byrne, B. M., & Shavelson, R. J. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology*, *80*(3), 366–380. https://doi.org/10.1037/0022-0663.80.3.366

Marsh, H. W., Chessor, D., Craven, R., & Roche, L. (1995). The effects of gifted and talented programs on academic self-concept: The big fish strikes again. *American Educational Research Journal*, *32*(2), 285–319. https://doi.org/10.3102/00028312032002285

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective. Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science*, *1*(2), 133–163. https://doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., & Hau, K.-T. (2003). Big-fish-little-pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist*, *58*(5), 364–376. https://doi.org/10.1037/0003-066X.58.5.364

Marsh, H. W., Hau, K.-T., & Craven, R. (2004). The big-fish-little-pond effect stands up to scrutiny. *American Psychologist*, *59*(4), 269–271. https://doi.org/10.1037/0003-066X.59.4.269

Marsh, H. W., Köller, O., & Baumert, J. (2001). Reunification of east and west german school systems: Longitudinal multilevel modeling study of the big-fish-little-pond effect on academic self-concept. *American Educational Research Journal*, *38*(2), 321–350. https://doi.org/10.3102/00028312038002321

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology, 78*(2), 337–349. https://doi.org/10.1037//0022-3514.78.2.337

Marsh, H. W., Kuyper, H., Morin, A. J., Parker, P. D., & Seaton, M. (2014). Big-fish-little-pond social comparison and local dominance effects: Integrating new statistical models, methodology, design, theory and substantive implications. *Learning and Instruction, 33,* 50–66. https://doi.org/10.1016/j.learninstruc.2014.04.002

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *The British Journal of Educational Psychology*, *81*, 59–77. https://doi.org/10.1348/000709910X503501

Marsh, H. W., Martin, A. J., Yeung, A. S., & Craven, R. (2016). Competence self-perceptions. In C. Dweck & D. Yaeger (Eds.), *Handbook of competence and motivation.* Guilford Press.

Marsh, H. W., Möller, J., Parker, P., Xu, M. K., Nagengast, B., & Pekrun, R. (2014). Internal/external frame of reference model. In J. D. Wright (Ed.), *International encyclopedia of social and behavioral sciences* (pp. 425–432). Elsevier.

Marsh, H. W., Parada, R. H., & Craven, R. G. In the looking glass: A reciprocal effect model elucidating the complex nature of bullying, psychological determinants, and the central role of self-concept. In C. S. Sanders & G. D. Phye (Eds.), *Bullying: Implications for the classroom* (pp. 63–106). Academic Press.

Marsh, H. W., & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, *47*(1), 213–231. https://doi.org/10.1037/0022-3514.47.1.213

Marsh, H. W., Pekrun, R., Murayama, K., Arens, A. K., Parker, P. D., Guo, J., & Dicke, T. (2018). An integrated model of academic self-concept development: Academic self-concept, grades, test scores, and tracking over 6 years. *Developmental Psychology*, *54*(2), 263–280. https://doi.org/10.1037/dev0000393

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology*, *111*(2), 331–353. https://doi.org/10.1037/edu0000281

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Lichtenfeld, S. (2017). Long-term positive effects of repeating a year in school: Six-year longitudinal study of self-beliefs, anxiety, social relations, school grades, and test scores. *Journal of Educational Psychology*, *109*(3), 425–438. https://doi.org/10.1037/edu0000144

Marsh, H. W., & Perry, C. (2005). Self-concept contributes to winning gold medals: Causal ordering of self-concept and elite swimming performance. *Journal of Sport & Exercise Psychology*, *27*(1), 71–91. https://doi.org/10.1123/jsep.27.1.71

Marsh, H. W., & Rowe, K. J. (1996). The negative effects of school-average ability on academic self-concept: An application of multilevel modelling. *Australian Journal of Education*, *40*(1), 65–87. https://doi.org/10.1177/000494419604000105

Marsh, H. W., & Seaton, M. (2015). The big-fish–little-pond effect, competence self-perceptions, and relativity: Substantive advances and methodological innovation. *Advances in Motivation Science*, *2*, 127–184.

Marsh, H. W., Seaton, M., Kuyper, H., Dumas, F., Huguet, P., Régner, I., Buunk, A. P., Monteil, J.-M., & Gibbons, F. X. (2010). Phantom behavioral assimilation effects: Systematic biases in social comparison choice studies. *Journal of Personality*, *78*(2), 671–710. https://doi.org/10.1111/j.1467-6494.2010.00630.x

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

Marsh, H. W., Tracey, D. K., & Craven, R. G. (2006). Multidimensional self-concept structure for preadolescents with mild intellectual disabilities. *Educational and Psychological Measurement*, *66*(5), 795–818. https://doi.org/10.1177/0013164405285910

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., Trautwein, U., Lüdtke, O., & Köller, O. (2008). Social comparison and big-fish-little-pond effects on self-concept and other self-belief constructs: Role of generalized and

specific others. *Journal of Educational Psychology*, *100*(3), 510–524. https://doi.org/10.1037/0022-0663.100.3.510

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

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

Mayer, S. E., & Jencks, C. (1989). Growing up in poor neighborhoods: How much does it matter? *Science, 243*(4897), 1441–1445. https://doi.org/10.1126/science.243.4897.1441

McCormick, K. (2018). James Duesenberry as a practitioner of behavioral economics. *Journal of Behavioral Economics for Policy*, *2*(1), 13–18.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research*, *95*(4), 203–213. https://doi.org/10.1080/00220670209596593

Meyer, J. W. (1970). High school effects on college intentions. *American Journal of Sociology, 76*(1), 59–70. https://doi.org/10.1086/224906

Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research*, *41*(4), 473–497. https://doi.org/10.1207/s15327906mbr4104_3

Möller, J., & Marsh, H. W. (2013). Dimensional comparison theory. *Psychological Review*, *120*(3), 544–560. https://doi.org/10.1037/a0032459

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79*, 1129–1167. https://doi.org/10.3102/0034654309337522

Möller, J., & Trautwein, U. (2015). Selbstkonzept. In E. Wild & J. Möller (Eds.), *Pädagogische Psychologie* (pp. 178–197). Springer.

Mullis, I.V.S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015. International results in mathematics*. IEA.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Murphy, R., & Weinhardt, F. (2014). Top of the class: The importance of ordinal rank. CESifo Working Paper Series No. 4815.

Murphy, R., & Weinhardt, F. (2018). Top of the class: The importance of ordinal rank. NBER Working Paper No. 24958.

Musu-Gillette, L. E., Wigfield, A., Harring, J. R., & Eccles, J. S. (2015). Trajectories of change in students' self-concepts of ability and values in math and college major choice. *Educational Research and Evaluation*, *21*(4), 343–370. https://doi.org/10.1080/13803611.2015.1057161

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

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

Nagy, G., Trautwein, U., Baumert, J., Köller, O., & Garrett, J. (2007). Gender and course selection in upper secondary education: Effects of academic self-concept and intrinsic value. *Educational Research and Evaluation*, *12*(4), 323–345. https://doi.org/10.1080/13803610600765687

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

Nieuwenhuis, J., & Hooimeijer, P. (2016). The association between neighbourhoods and educational achievement, a systematic review and meta-analysis. *Journal of Housing and the Built Environment*, *31*, 321–347. https://doi.org/10.1007/s10901-015-9460-7

Oakes, J. (2005). *Keeping track: How schools structure inequality*. Yale University Press.

OECD. (2007). *Women in Science, Engineering and Technology (SET): Strategies for a Global Workforce.* OECD.

OECD. (2019). *PISA 2018 Results (Volume I)*. OECD.

Orth, U., Robins, R. W., & Roberts, B. W. (2008). Low self-esteem prospectively predicts depression in adolescence and young adulthood. *Journal of Personality and Social Psychology*, *95*(3), 695–708. https://doi.org/10.1037/0022-3514.95.3.695

Pajares, F., & Schunk, D. H. (2002). Self and self-belief in psychology and education. An historical perspective. In J. Aronson (Ed.), *Improving academic achievement* (pp. 3–21). Academic Press.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, *72*(6), 407–418. https://doi.org/10.1037/h0022602

Parducci, A. (1968). The relativism of absolute judgements. *Scientific American*, *219*(6), 84–90. https://doi.org/10.1038/scientificamerican1268-84

Parker, P. D., Dicke, T., Guo, J., & Marsh, H. W. (2019). A macro context theory of academic self-concept: Ability stratification and the big-fish-little-pond effect. Advance online publication. https://doi.org/10.31234/osf.io/bwy59

Parker, P. D., Marsh, H. W., Ciarrochi, J., Marshall, S., & Abduljabbar, A. S. (2014). Juxtaposing math self-efficacy and self-concept as predictors of long-term achievement outcomes. *Educational Psychology*, *34*(1), 29–48. https://doi.org/10.1080/01443410.2013.797339

Parker, P. D., Marsh, H. W., Thoemmes, F., & Biddle, N. (2019). The negative year in school effect: Extending scope and strengthening causal claims. *Journal of Educational Psychology*, *111*(1), 118–130. https://doi.org/10.1037/edu0000270

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

Pekrun, R., Frenzel, A., Goetz, T., & Perry, R. P. (2007). The control-value theory of achievement emotions: An integrative approach to emotions in education. In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 13–27). Academic Press.

Plieninger, H., & Dickhäuser, O. (2013). The female fish is more responsive: Gender moderates the BFLPE in the domain of science. *Educational Psychology*, *35*(2), 213–227. https://doi.org/10.1080/01443410.2013.814197

Preckel, F., Schmidt, I., Stumpf, E., Motschenbacher, M., Vogl, K., Scherrer, V., & Schneider, W. (2019). High-ability grouping: Benefits for gifted students' achievement development

without costs in academic self-concept. *Child Development*, *90*(4), 1185–1201. https://doi.org/10.1111/cdev.12996

Purpel, D. E. (1989). *The moral and spiritual crisis in education*. Bergin & Garvey.

Repko, A. F. (2012). *Interdisciplinary research: Process and theory*. Sage.

Rheinberg, F., & Enstrup, B. (1978). Selbstkonzept der Begabung bei Normal- und Sonderschülern gleicher Intellligenz: Ein Bezugsgruppeneffekt. *Zeitschrift Für Entwicklungspsychologie and Pädagogische Psychologie*, *9*(3), 171–180.

Richardson, A. C., & Bond, R. M. (2012). Psychological correlates of University students' academic performance: a systematic review and meta analysis.

Richter, S. S., Brown, S. A., & Mott, M. A. (1991). The impact of social support and self-esteem on adolescent substance abuse treatment outcome. *Journal of Substance Abuse*, *3*(4), 371–385. https://doi.org/10.1016/s0899-3289(10)80019-7

Romanowski, M. H. (2004). Student obsession with grades and achievement. *Kappa Delta Pi Record*, *40*(4), 149–151. https://doi.org/10.1080/00228958.2004.10516425

Rubin, B. C. (2006). Tracking and detracking: Debates, evidence, and best practices for a heterogeneous world. *Theory into Practice*, *45*(1), 4–14. https://doi.org/10.1207/s15430421tip4501_2

Rui, N. (2009). Four decades of research on the effects of detracking reform: Where do we stand? - A systematic review of the evidence. *Journal of Evidence-Based Medicine*, *2*(3), 164–183. https://doi.org/10.1111/j.1756-5391.2009.01032.x

Sampson, R. J., Morenoff, J. D., & Gannon-Rowley, T. (2002). Assessing "neighborhood effects": Social processes and new directions in research. *Annual Review of Sociology*, *28*(1), 443–478. https://doi.org/10.1146/annurev.soc.28.110601.141114

Sanbonmatsu, L., Kling, J. R., & Duncan, Greg J., Brooks-Gunn, Jeanne (2006). Neighborhoods and academic achievement: Results from the moving to opportunity experiment. NBER Working Paper No. 11909.

Savickas, M. L. (2002). Reinvigorating the Study of Careers. *Journal of Vocational Behavior*, *61*(3), 381–385. https://doi.org/10.1006/jvbe.2002.1880

Savickas, M. L. (2005). The theory and practice of career construction. In S. D. Brown & R. W. Lent (Eds.), *Career Development and Counseling: Putting Theory and Research to Work* (pp. 42–70). John Wiley.

Schreiner, C., & Breit, S. (2012). *Standardüberprüfung 2012 Mathematik, 8. Schulstufe. Bundesergebnisbericht*. BIFIE.

Schreiner, C., Breit, S., Pointinger, M., Pacher, K., Neubacher, M., & Wiesner Christian. (2017). *Standardüberprüfung 2017 Mathematik, 8. Schulstufe. Bundesergebnisbericht.* BIFIE.

Schunk, D. H., & Pajares, F. (2005). Competence perceptions and academic functioning. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of Competence and Motivation* (pp. 85–104). Guilford Publications.

Schwabe, F., Korthals, R., & Schils, T. (2019). Positive social relationships with peers and teachers as moderators of the Big-Fish-Little-Pond Effect. *Learning and Individual Differences*, *70*, 21–29. https://doi.org/10.1016/j.lindif.2018.12.006

Schwerter, J., & Biewen, M. (2020). Does more math in high school increase the share of female STEM workers? Evidence from a curriculum reform. IZA Discussion Paper No. 12236.

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

Seaton, M., Marsh, H. W., & Craven, R. G. (2010). Big-fish-little-pond effect: Generalizability and moderation - Two sides of the same coin. *American Educational Research Journal*, *47*(2), 390–433. https://doi.org/10.3102/0002831209350493

Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, *55*(1), 5–14. https://doi.org/10.1037/0003-066X.55.1.5

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

Sharkey, P., & Faber, J. W. (2014). Where, when, why, and for whom do residential contexts matter? Moving away from the dichotomous understanding of neighborhood effects. *Annual Review of Sociology*, *40*(1), 559–579. https://doi.org/10.1146/annurev-soc-071913-043350

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, *46*(3), 407–441. https://doi.org/10.2307/1170010

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology*, *27*, 1–60.

Simonsmeier, B. A., Peiffer, H., Flaig, M., & Schneider, M. (2020). Peer feedback improves students' academic self-concept in higher education. *Research in Higher Education, 61*(6), 706–724. https://doi.org/10.1007/s11162-020-09591-y

Simpkins, S. D., Fredricks, J. A., & Eccles, J. S. (2012). Charting the Eccles' expectancy-value model from mothers' beliefs in childhood to youths' activities in adolescence. *Developmental Psychology*, *48*(4), 1019–1032. https://doi.org/10.1037/a0027468

Soares, A. T., & Soares, L. M. (1969). Self-perceptions of culturally disadvantaged children. *American Educational Research Journal*, *6*(1), 31–45. https://doi.org/10.3102/00028312006001031

Spinath, B., & Spinath, F. M. (2005). Longitudinal analysis of the link between learning motivation and competence beliefs among elementary school children. *Learning and Instruction*, *15*(2), 87–102. https://doi.org/10.1016/j.learninstruc.2005.04.008

Stäbler, F. (2017). Die Zusammensetzung der Lerngruppe und ihre Effekte auf psychosoziale Merkmale und Leistung von Schülerinnen und Schülern. Dissertation.

Stäbler, F., Dumont, H., Becker, M., & Baumert, J. (2017). What happens to the fish's achievement in a little pond? A simultaneous analysis of class-average achievement effects on achievement and academic self-concept. *Journal of Educational Psychology*, *109*(2), 191–207. https://doi.org/10.1037/edu0000135

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S., & Henschel, S. (Eds.). (2018). *IQB Trends in Student Achievement 2018*. Waxmann.

Steinmayr, R., & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences*, *19*(1), 80–90. https://doi.org/10.1016/j.lindif.2008.05.004

Stinchcombe, A. L. (1991). The conditions of fruitfulness of theorizing about mechanisms in social science. *Philosophy of the Social Sciences*, *21*(3), 367–388. https://doi.org/10.1177/004839319102100305

Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A., & Williams, R. M. (1949). *The American soldier: Adjustment during army life.* Princeton University Press.

Südkamp, A., & Möller, J. (2009). Referenzgruppeneffekte im simulierten Klassenraum. *Zeitschrift Für Pädagogische Psychologie*, *23*(34), 161–174. https://doi.org/10.1024/1010-0652.23.34.161

Super, D. (1951). Vocational adjustment: Implementing a self-concept. *Occupations*, *30*, 351–357.

Swann, W. B., Chang-Schneider, C., & Larsen McClarty, K. (2007). Do people's self-views matter? Self-concept and self-esteem in everyday life. *The American Psychologist*, *62*(2), 84–94. https://doi.org/10.1037/0003-066X.62.2.84

Sykes, B., & Musterd, S. (2010). Examining neighbourhood and school effects simultaneously. *Urban Studies*, *48*(7), 1307–1331. https://doi.org/10.1177/0042098010371393

Taylor, D. M., & Moghaddan. (1994). *Theories of intergroup relations: International social psychological perspectives*. Praeger.

Thijs, J., Verkuyten, M., & Helmond, P. (2010). A further examination of the big-fish–little-pond effect. *Sociology of Education*, *83*(4), 333–345. https://doi.org/10.1177/0038040710383521

Tracey, D. K., Marsh, H. W., & Craven, R. G. (2003). Selfconcepts of preadolescent students with mild intellectual disabilities: Issues of measurement and educational placement. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *International advances in self research* (pp. 203–230). Information Age.

Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Self-esteem, academic self-concept, and achievement: How the learning environment moderates the dynamics of self-concept. *Journal of Personality and Social Psychology*, *90*(2), 334–349. https://doi.org/10.1037/0022-3514.90.2.334

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., & Möller, J. (2016). Self-concept: Determinants and consequences of academic self-concept in school contexts. In A. A. Lipnevich, F. Preckel, & R. D. Roberts (Eds.), *Psychosocial Skills and School Systems in the 21st Century* (pp. 187–214). Springer International Publishing.

Trowbridge, N. (1972). Self concept and socio-economic status in elementary school children. *American Educational Research Journal*, *9*(4), 525. https://doi.org/10.2307/1162274

Trzesniewski, K. H., Donnellan, M. B., Moffitt, T. E., Robins, R. W., Poulton, R., & Caspi, A. (2006). Low self-esteem during adolescence predicts poor health, criminal behavior, and

limited economic prospects during adulthood. *Developmental Psychology*, *42*(2), 381–390. https://doi.org/10.1037/0012-1649.42.2.381

Turkheimer, E., & Harden, K. P. (2013). Behavior genetic research methods. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 159–187). Cambridge University Press. https://doi.org/10.1017/CBO9780511996481.012

Upshaw, H. S. (1969). The personal reference scale: An approach to social judgment. *Advances in Experimental Social Psychology*, *4*, 315–371. https://doi.org/10.1016/S0065-2601(08)60081-7

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist*, *39*(2), 111–133. https://doi.org/10.1207/s15326985ep3902_3

Vanderweele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, *2*(4), 457–468. https://doi.org/10.4310/SII.2009.v2.n4.a7

Vingilis, E., Wade, T. J., & Adlaf, E. (1998). What factors predict student self-rated physical health? *Journal of Adolescence*, *21*(1), 83–97. https://doi.org/10.1006/jado.1997.0131

Wang, M.-T., Eccles, J. S., & Kenny, S. (2013). Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological Science*, *24*(5), 770–775. https://doi.org/10.1177/0956797612458937

Wang, Z. (2015). Examining big-fish-little-pond-effects across 49 countries: A multilevel latent variable modelling approach. *Educational Psychology*, *35*(2), 228–251. https://doi.org/10.1080/01443410.2013.827155

Wang, Z., & Bergin, D. A. (2017). Perceived relative standing and the big-fish-little-pond effect in 59 countries and regions: Analysis of TIMSS 2011 data. *Learning and Individual Differences*, *57*, 141–156. https://doi.org/10.1016/j.lindif.2017.04.003

Wehrens, M. J.P.W., Buunk, A. P., Lubbers, M. J., Dijkstra, P., Kuyper, H., & van der Werf, G. P.C. (2010). The relationship between affective response to social comparison and academic performance in high school. *Contemporary Educational Psychology*, *35*(3), 203–214. https://doi.org/10.1016/j.cedpsych.2010.01.001

Wicht, A., & Ludwig-Mayerhofer, W. (2014). The impact of neighborhoods and schools on young people's occupational aspirations. *Journal of Vocational Behavior*, *85*(3), 298–308. https://doi.org/10.1016/j.jvb.2014.08.006

Wigfield, A., & Eccles, J. S. (1994). Children's competence beliefs, achievement values, and general self-esteem. *The Journal of Early Adolescence*, *14*(2), 107–138. https://doi.org/10.1177/027243169401400203

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

Wilson, W. J. (1987). *The truly disadvantaged*. University of Chicago Press.

Wodtke, G. T., & Parbst, M. (2017). Neighborhoods, schools, and academic achievement: A formal mediation analysis of contextual effects on reading and mathematics abilities. *Demography*, *54*(5), 1653–1676. https://doi.org/10.1007/s13524-017-0603-1

Wood, J. V. (1989). Theory and research concerning social comparisons of personal attributes. *Psychological Bulletin*, *106*(2), 231–248. https://doi.org/10.1037/0033-2909.106.2.231

Wouters, S., Fraine, B. de, Colpin, H., van Damme, J., & Verschueren, K. (2012). The effect of track changes on the development of academic self-concept in high school: A dynamic test of the big-fish–little-pond effect. *Journal of Educational Psychology*, *104*(3), 793–805. https://doi.org/10.1037/a0027732

Yonezawa, S., Wells, A. S., & Serna, I. (2002). Choosing tracks:"Freedom of choice" in detracking schools. *American Educational Research Journal*, *39*(1), 37–67. https://doi.org/10.3102/00028312039001037

Zeidner, M. (1992). Key facets of classroom grading: A comparison of teacher and student perspectives. *Contemporary Educational Psychology*, *17*(3), 224–243. https://doi.org/10.1016/0361-476X(92)90062-4

Zell, E., & Alicke, M. D. (2010). The local dominance effect in self-evaluation: Evidence and explanations. *Personality and Social Psychology Review*, *14*(4), 368–384. https://doi.org/10.1177/1088868310366144